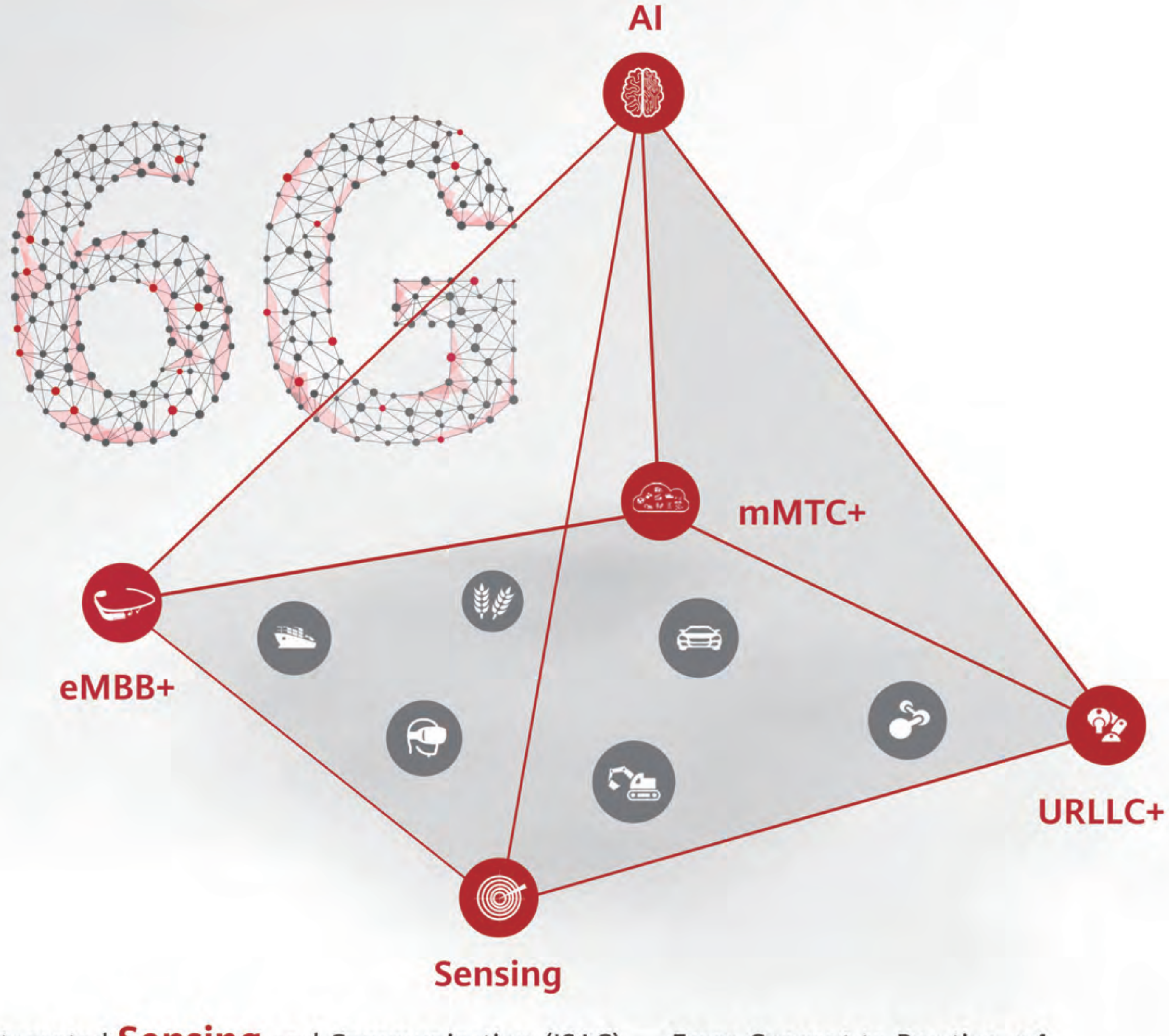
Communications of HUAWEI RESEARCH

September 2022
Issue 2



HUAWEI

6G

From Connected People and Things to Connected Intelligence

HUAWEI

Editorial Note

There is no doubt that the 6G research journey has set sail. All long-term research programs have invested heavily in 6G, and many governments have made 6G research an important agenda in their national strategies — this has become a global phenomenon. We can therefore say with certainty that 6G is on the way.

However, the creation of 6G requires innovative technologies, and the success of 6G requires innovative applications. Revolutionary applications and revolutionary technologies are indispensable prerequisites.

The focus of this special issue is on the innovation of 6G technologies. For engineers, we are exploring what technologies are needed to enable 6G — the direction of this exploration is becoming clear. Because 6G is a market evolution, it cannot be accomplished overnight. We need to bear in mind that the innovation of 6G is in fact that 6G will enable more innovations beyond what we can imagine. 6G is not a technology-only innovation, but a pursuit of innovation to generate greater social value. In this direction, 6G-AI, 6G-ISAC, 6G-Extreme Connectivity, 6G-NTN, 6G-Trustworthiness, and 6G-Green will become the cornerstones of 6G. Therefore, the mobile communication network will truly transform into an intelligent bridge between the physical world and the digital world. As we move toward 6G, the focus of mobile communications is also evolving: from networks to terminals, from low frequency to high frequency, and from B2C to B2B. These changes present both challenges and opportunities.

As a foundational ICT platform — especially a mobile communications platform — for the next decades, 6G will be rooted in deep theories and cutting-edge engineering technologies, and require deep experience and practical expertise. We can see that the future 6G will require more efforts and investment, as it is a huge engineering endeavor.

History tells us that the success of 6G will depend on open cooperation and the commitment to embrace all partners and players. As a result, international cooperation has become a tradition in our industry — a characteristic reflected in this special issue. We would like to thank our friends and colleagues for their contributions.

Finally, let's remind ourselves that this special issue is the beginning, not the end, of the journey to 6G innovation.

Wen Tong

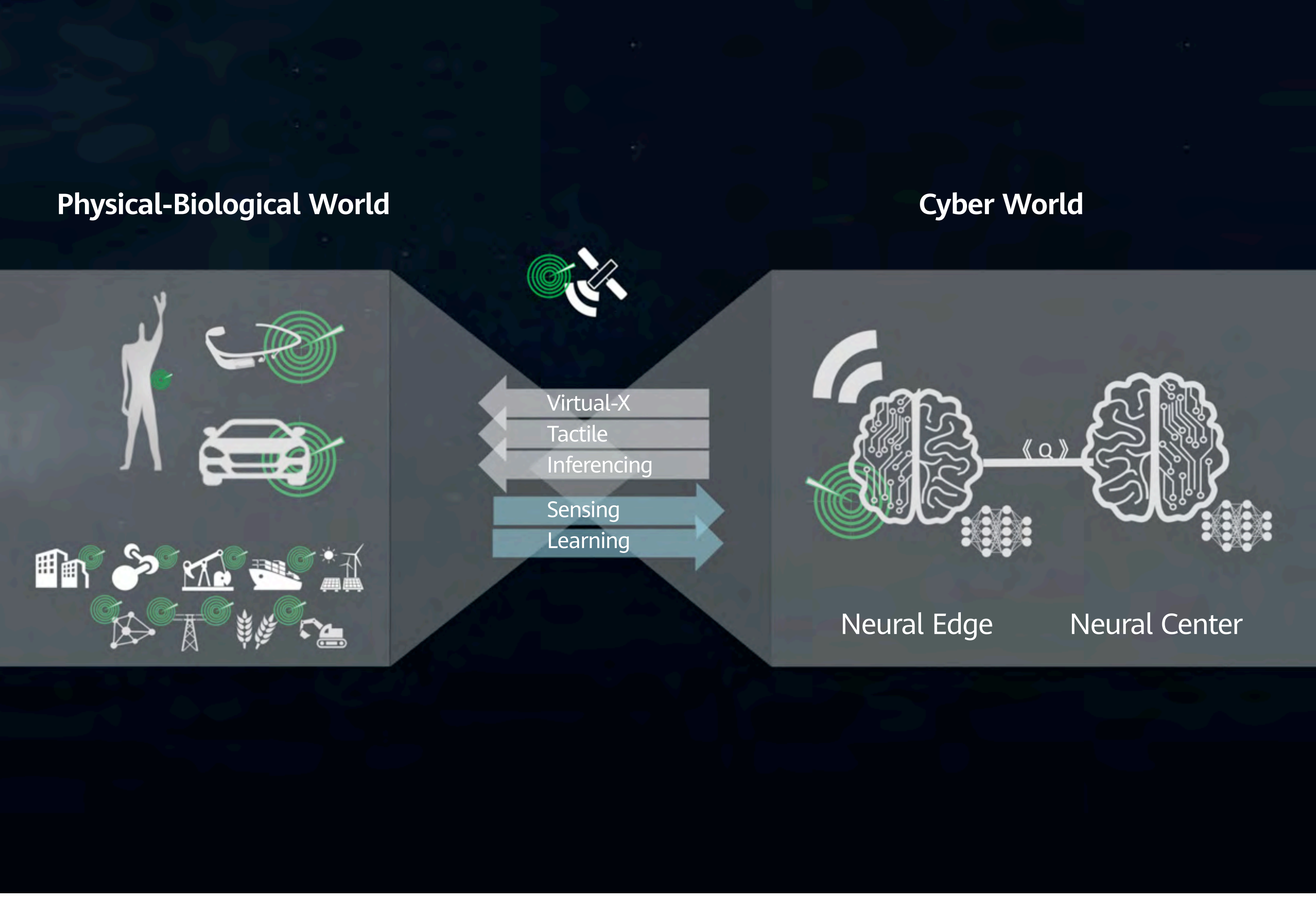
CTO, Huawei Wireless

E-mail: HWResearch@huawei.com

Outlook

Research

Prototype



Integrated Sensing and Communication (ISAC) — From Concept to Practice

Alireza Bayesteh [1], Jia He [2], Yan Chen [1], Peiying Zhu [1], Jianglei Ma [1], Ahmed Wagdy Shaban [1], Ziming Yu [2], Yunhao Zhang [2], Zhi Zhou [2], Guangjian Wang [2]

[1] Ottawa Wireless Advanced System Competency Centre

[2] Wireless Technology Lab

Abstract

6G will serve as a distributed neural network for the future Intelligence of Everything. Network Sensing and Native AI will become two new usage scenarios in the era of connected intelligence. 6G will integrate sensing with communication in a single system. Radio waves can be exploited to "see" the physical world and make a digital twin in the cyber world. This article introduces the concept of integrated sensing and communication (ISAC) and typical use cases, and provides two case studies of how to use 6G ISAC to improve localization accuracy and perform millimeter level imaging using future portable devices. The research challenges to implementing ISAC in practice are discussed.

Keywords

integrated sensing and communication (ISAC), localization, THz imaging, sensing accuracy, sensing resolution, prototype

1 Introduction

In 6G mobile communication systems, the use of higher frequency bands (from mmWave up to THz), wider bandwidth, and massive antenna arrays will enable high-accuracy and high-resolution sensing, which can help implement the integration of wireless signal sensing and communication (ISAC) in a single system for their mutual benefit. On the one hand, the entire communications network can serve as a sensor. The radio signals transmitted and received by network elements and the radio wave transmissions, reflections, and scattering can be used to sense and better understand the physical world. The capabilities to obtain range, velocity, and angle information from the radio signals can provide a broad range of new services, such as high accuracy localization, gesture capturing and activity recognition, passive object detection and tracking, as well as imaging and environment reconstruction [1]. This is called "network as a sensor". On the other hand, the capabilities of high-accuracy localization, imaging, and environment reconstruction obtained from sensing can improve communication performance — for example, more accurate beamforming, faster beam failure recovery, and less overhead when tracking the channel state information (CSI) [2–3]. This is called "sensing-assisted communication". Moreover, sensing is a "new channel" that observes, samples, and links the physical and biological world to the cyber world. Real-time sensing is therefore essential to make the concept of the digital twin — a true and real-time replica of the physical world — a reality in the future.

3GPP has initiated some preliminary study on use cases and potential ISAC requirements using the air interface of 5G advanced [1]. 6G ISAC systems will, however, be further optimized, fully integrated, and will not be constrained by the limitations of the current 5G system. The sensing use cases offered by future 6G ISAC systems will most likely include ultra-high accuracy localization and tracking, simultaneous imaging, mapping, and localization,

Table 1 ISAC use cases as new services in 6G according to different categories

Use Case Category / Application Category	High-Accuracy Localization and Tracking	Simultaneous Imaging, Mapping, and Localization	Augmented Human Sense	Gesture and Activity Recognition
Vertical Industry Intelligent healthcare Intelligent transportation Intelligent factory/ manufacturing Smart agriculture	• Surgery with cooperative robots • Docking drone on a moving vehicle • Device/module placement and installation • Livestock movement and animal migration monitoring	• Sensing glasses with ultra-high resolution imagery • 3D road environment mapping • Warehouse robotics automation system • Crop production and crop physiology	• Tele-surgery • Pollution and air quality detection • Automatic flaw detection on products • Intelligent crop monitoring for nutrients, water stress and disease	• Gesture-controlled smart operation theater • In-cabin monitoring and contactless control • Contactless control for intelligent manufacturing system • Gesture-based robots and machinery control for precision agriculture
Consumer Smart home & entertainment Smart mobile devices	• Collaborative robots for household chores • Precise localization of small objects (tag or active objects) using mobile phones	• Close-in scene and object imaging	• Imaging of water pipes behind walls • Calories count • Contaminated ingredients detection	• Virtual piano • Touchless home appliances • Contactless control on intelligent screen
Public Service Smart city Smart environment Smart security and public safety	• Drone as robotic waiter • Hydrological monitoring {e.g., precipitation, water flow/level} • Crowd management and emergency evacuation for major events	• Wireless SLAM • Drone base stations swarm SAR imaging • In-car sensing for driver and passenger monitoring	• Crack detections in buildings, bridges and man-made structures • Fine particulate matter detection (PM10, PM2.5) • Explosive detection and gas sensing • Security scans on packages	• Gesture-based appliances for enhanced accessibility for seniors and differently abled people • Panic and terrifying emotion recognition

augmented human sensing, gesture and activity recognition, as illustrated in Table 1 [1]. The use cases and performance requirements will be further discussed in Section 2.

The integration of sensing and communication functions can happen at three different levels, from loosely coupled to fully integrated. At the lowest integration level, sensing and communication capabilities can co-exist on hardware by sharing the spectrum, which is more efficient than dedicated spectrum usage. Sensing can benefit from the economies of scale in the mobile communication network, where shared hardware will be cost effective and eases deployment and maintenance issues. The second level of integration calls for the integration of waveform and signal processing where the time, frequency, and spatial domain processing techniques have a common objective and can be combined to serve both sensing and communication functions. A fully integrated system with cross-layer, cross-module, and cross-node information sharing is expected to significantly enhance the mutual performance of both sensing and communication, as well as reduce the overall cost, size and power consumption of the network system.

In addition to the wider spectrum and the larger number of antennas, the sensing functionality and performance will be further enabled by other technology innovations such as the larger scale of cooperation between base stations and user equipment (UE), joint design of communication and sensing waveforms, advanced techniques for interference cancellation, and the native AI capability to better deal with the sensed data.

Next, we will discuss typical ISAC use cases and then elaborate on examples of ISAC application in enhanced localization and millimeter level resolution. Design challenges will be discussed thereafter, followed by the conclusions.

2 ISAC Use Cases

2.1 Overview

Wireless sensing has long been a separate technology developed in parallel with the mobile communication systems. Positioning is the only sensing service that mobile communication systems (until 5G) could offer. General sensing rather than positioning will become a new function integrated into the 6G mobile communication system. This capability will open up brand new services for 6G. These services are currently provided by various dedicated sensing equipment, such as radar, light detection and ranging (LIDAR), and professional CT and MRI equipment.

The ISAC capability will thus enable many new services that mobile communication system operators can offer. These include very high accuracy positioning, localization and tracking, imaging for biomedical and security applications, simultaneous localization and mapping to automatically construct maps of complex indoor or outdoor environments, pollution or natural disaster monitoring, gesture and activity recognition, flaw and material detection and many other services. These services will in turn enable application scenarios in all kinds of business for future consumers and vertical industries. The potential new services that could be supported by future ISAC systems are listed in Table 1. In the table, the use cases are categorized into four functional categories across different applications/industries (vertical industry, consumer and public services):

- High-accuracy localization and tracking
- Simultaneous imaging, mapping and localization
- Augmented human sensing
- Gesture and activity recognition

It is also worth mentioning that, in addition to the preceding services, sensing can also be used to assist communications and positioning, more details of which can be found in Section 5.4.

2.2 High-Accuracy Localization and Tracking

Low-latency high-accuracy localization and tracking enable meaningful association between cyber information and the locations of physical entities in multiple scenarios from factories to warehouses, hospitals to retail shops, and agriculture to mining.

The 6G network will provide services for both device-based and device-free object localization. For 6G device-based localization, the target is a connected device in the network, and the location information is derived from the reference signals or measurement feedback from the device. Localization for 6G device-free objects, on the other hand, does not need the object to be a connected device in the network. The estimation of delay, Doppler, and angle spectrum information (corresponding to the distance, velocity, and angle of the objects) are obtained from the scattered and reflected wireless signals either through monostatic sensing (receiver is the same as transmitter) or bistatic sensing (receiver is another node or device in the network). By processing these wireless signals further, the locations, orientations, velocity, and other geometric information of the objects in a physical 3D space can be extracted. With higher bandwidth and increased antenna aperture, the 6G ISAC system can have strong capabilities to separate multipaths, through which better localization and tracking performance can be achieved, and the localization accuracy for outdoor use cases can be up to the centimeter level.

Having high-accuracy relative localization is important when two or more entities exist and they are approaching one another, or the entities have coordinated moving direction and speed. In automatic warehousing applications, centimeter-level accuracy enables device-level placement, and the near-millimeter-level accuracy can further enable module-level installation and placement in tight spaces, allowing for efficient storage of components that have a small form factor. Relative localization is necessary as a viable alternative for close-in maneuvering owing to the fact that complexity, physical limitations, and external infrastructure are mission critical for each robot to accurately determine its location in relation to a common datum. An example will be a drone docking onto a moving vehicle with an extremely small margin for landing, due to the limited area of the moving vehicle's cargo platform.

Future ISAC systems that are empowered by native AI can provide semantic localization capability with context awareness. To support future smart home/shopping mall/restaurant/hotel, and automatic factory applications, objects and parts need to have dispatchable localization information such as shelf level, seat number, table number, etc. In a restaurant, robotic waiters, which have semantic localization capability, can accurately deliver food to guests and even go a step further and set different level of protections according to different task specifics, e.g., fragile and rigid objects can be treated with different levels of location and velocity accuracy during transportation.

2.3 Imaging, Mapping, and Environment Reconstruction

In simultaneous imaging, mapping, and localization, the sensing capabilities from these three perspectives are mutually enhanced. Particularly, the imaging function is used to capture the images of the surrounding environment, and the localization function is used to obtain the locations of surrounding objects. These images and/or locations are then used by the mapping function to construct a map. The mapping function helps the localization function improve the inference of locations. ISAC will leverage on advanced algorithms, edge computing, and AI to produce super-resolution and highly recognizable images and maps in which the vast network of objects, including vehicles and base stations, act as sensors to provide a remarkably extended imaging area. Moreover, performance will significantly improve due to the ease of fusing results that are shared with cloud-based services across the entire network.

6G-based super-resolution and high accuracy sensing applications open up a range of possibilities in 3D indoor imaging and mapping, which in turn enable various applications, such as scene reconstruction, spatial localization, and navigation for indoor scenarios, and help provide the most up-to-date knowledge of an environment for networks and devices. The accurate mapping information can then be applied to determine the multipath reflection points. Owing to the fact that scattered signals bounce multiple times where the LOS surfaces act as mirrors, compensated images of NLOS objects can be reconstructed by applying mirroring techniques. Once the environment is reconstructed, the next step will be the localization and imaging of the NLOS targets. Target locations can then be detected with good accuracy when prior information regarding the scene's geometry is known.

In an outdoor imaging and mapping scenario involving a mobile vehicle, its sensors usually have a restricted view and limited coverage due to weather, obstacles, and the sensors' power control. That said, nearby stationary base stations may have a greater field of view, longer sensing distance, and higher resolution because they collect and use their own sensing data or sensing data of UE. Therefore, mobile vehicles can achieve higher levels of autonomy by utilizing the maps reconstructed by the base stations to determine their next move. Moreover, the sensing resolution and accuracy performance will significantly improve due to the fusion of imaging results across the network. The densely distributed base stations in an urban area and ISAC make environmental reconstruction and 3D localization possible, which in turn form the virtual city. The reconstructed map used for smart traffic control scenarios, such as traffic flow monitoring, queue detection, and accident detection, are an important use case in the dynamic virtual city.

2.4 Augmented Human Senses

Technology progress makes augmented human sensing a reality. Augmented human sensing aims to provide a safe, high-precision, low-power, sensing and imaging capability that exceeds human abilities, by means of a portable terminal (e.g., 6G-enabled mobile phones, wearables, or medical equipment implanted beneath human skin), to sense the surrounding environments. With the help of scientific and technological advancement, augmented human sensing can be achieved to facilitate information collection and integrate the maximum number of environmental messages into the 6G network.

In the 6G network, high-resolution imaging and detection sensing techniques will open the door for numerous applications, such as remote surgery, cancer diagnostics, detection of slits on products, and sink water-leakage detection. A surgeon may be able to conduct surgery at a different location through the help of an ultra-high-resolution imaging monitor system and remote operation platform system. In addition, intelligent factories will leverage these superior sensing solutions to implement contactless ultra-high-precision detection, tracking, and quality control, where millimeter-level radial-range resolution and ultra-high cross-range resolution based on higher bandwidth and increased antenna array aperture, respectively, are required. 6G communication technologies, with high THz frequency and corresponding short wavelength that is less than 1 mm can increase the bandwidth and decrease the array size, so that these augmented human sensing functions can be integrated or installed in portable devices.

While ultra-high-resolution scenarios require higher bandwidth and increased antenna aperture, another application of "seeing beyond the eye" that can sense the changes beneath the skin, behind occlusion, or in darkness, poses different requirements. 6G radio wave (up to THz) based sensing can achieve the NLOS imaging ability, where technology for detecting hidden objects can be equipped on portable devices that have a powerful imaging capability. As such, mobile phones can be used to detect pipelines behind walls or perform security scans on packages by utilizing the penetration characteristics of electromagnetic waves. Moreover, 6G ISAC can enable atraumatic medical detection which plays an important role in eHealth procedures such as diagnosis, monitoring, and treatments. It provides ultra-high reliability and accuracy and does not harm human bodies.

Spectrogram recognition is another interesting application that could be supported by a 6G ISAC system. It can identify targets through spectrogram sensing of their electromagnetic or photonic characteristics. This includes the analysis of absorption, reflectivity, and permittivity parameters, which helps distinguish the type and quality of materials. Pollution and product quality management are some of the prospective applications of this technology. Spectrogram recognition can also be used in food sensing applications to detect the food type and ingredients through the transmission and reflection of THz signals. This technology will help identify different types of food, calorie content, presence of contaminated ingredients, etc.

2.5 Posture and Gesture Recognition

Device-free gesture and posture recognition using machine learning is the key to promoting human–computer interfaces that allow users to convey commands and conveniently interact with devices through body postures, hand gestures, etc. In 6G system, the higher-frequency band will enable higher resolution and accuracy to capture finer postures and gestures, and the detection of motion activities (resulting in Doppler shifts) will be more sensitive in the higher-frequency band. Furthermore, the massive antenna arrays allow for recognition with significantly improved spatial resolution and accuracy. Another important benefit of gesture and posture sensing by 6G is the fact that there is no risk of personal privacy information being compromised, as is the case with cameras now, which makes it ideal for many scenarios, especially smart home scenarios. In a future gesture and posture recognition system that utilizes the densely distributed 6G network, devices will be collectively used to sense the surrounding environments, and sensing data association and fusion at an extended range will significantly improve the overall recognition performance.

There will be advanced gesture and posture recognition features in smart hospitals in the foreseeable future. The medical rehabilitation system in future smart hospitals will enable the automatic supervision of patients. This ensures that their gestures and movements during physiotherapy conform to the standard requirements of rehabilitation exercises. There will be prompt alerts on incorrect movements or gestures, significantly improving patients' rehabilitation. In addition, an alarm alerting the hospital's control center will be generated if a patient falls during an exercise, or if a suspicious person is detected intruding into a restricted area.

The future smart home will be equipped with an advanced hand gesture capturing and recognition system where it allows a hand's 3D position, rotation, and gesture to be tracked. Thus, by simply waving our hands and other gestures, many household appliances such as smart light, smart TV, etc., can be remotely controlled. Looking ahead, more complicated functionalities can be realized by the advanced hand gesture capturing and recognition function in the 6G network, such as playing a virtual piano in the air, in order to provide a completely immersive experience anywhere, anytime. Without doubt, this futuristic concept would open up a range of possibilities for many more innovative applications related to high-accuracy finger motion detection and tracking.

2.6 Key Performance Indicators

Within the ISAC context, several new key performance indicators (KPIs) are introduced for sensing capability and they are listed in Table 2.

Table 3 presents the relevant key performance indicators along with the requirements that must be met in order to realize the important use cases discussed in the earlier sections.

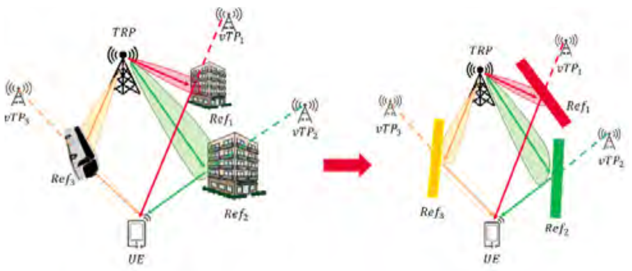
3 ISAC for Centimeter-Level Positioning

3.1 Background, Motivation and High-Level Scheme

6G requires solutions for sub-centimeter level positioning techniques for various future applications and use cases. This level of accuracy for positioning requires much more detailed knowledge of the radio signal propagation environment where sensing comes into play. By learning the environment RF map and the way the transmitted waveform is manipulated by it, the UE position can be obtained as a function of the measurement parameters. This way, the multipath nature of the propagation channel will be helpful [4]. Moving to higher frequencies can further facilitate such sensing-assisted positioning because the channel becomes sparser, and hence, characterizing the mapping between UE position and its propagation channel takes less effort. In a reflection-dominant environment (which is the case in higher frequencies), one such mapping can be obtained by decomposing the multipath channel as multiple LOS channels coming from multiple anchors. Those anchors are obtained by mirroring the transmission point (TP) over the surface of the corresponding reflector for each path. Those virtual anchors are referred to as virtual TPs or vTPs.

Table 2 Sensing key performance indicators and their descriptions

Key Performance Indicator	Description
Coverage	Range and field of view limits within which objects can be detected by the system.
Accuracy	Difference between the sensed and real values in range, angle, velocity, etc.
Resolution	Separation between multiple objects in range, angle, velocity, etc.
Detection/False alarm probabilities	Probabilities that an object will be detected when one is present/not present.
Availability	Percentage of time for which a system is able to provide the sensing service according to requirements.
Refresh rate	Rate at which positioning/localization data is refreshed.



(a) the real reflectors map

(b) the extracted geometrical representation to the visible reflectors

Figure 1 Mapping the objects/reflectors of the environment to virtual anchors, i.e., mapping multipath components to vTPs

The advantage of such characterization is two-fold:

- The vTPs are totally synchronous with the actual TP. This solves one of the prominent problems of current positioning technologies, which rely on multiple TPs that are not synchronous.
- The channels between vTPs and the UE are LOS, which means that there is no NLOS bias.

Hence, it can solve the two limiting problems of NR positioning (i.e., synchronization error and NLOS error).

However, implementing such technology in a real cellular system is fraught with various challenges and the goal of this section is to provide solutions for these challenges and pave the way for utilizing sensing-assisted positioning in future 6G networks.

- **Potentially large number of vTPs**: In the initial stage of environment sensing, the reflections of the TP location with respect to all objects in the map are obtained. The issue is that number of vTPs grows linearly with the number of reflection planes and grows exponentially

Table 3 ISAC use cases along with key performance indicators and requirements

Use Case Category	Coverage	Resolution	Accuracy	Probability	Availability	Refresh Rate
High-accuracy localization and tracking						
Module installation and placement	10 m	-	1 mm	-	99.99%	< 100 ms
Docking drone on a moving platform	50 m	-	1 cm	-	99.99%	< 10 ms
Robot/Drone as waiter	50 m	-	1 cm	-	99.9%	< 100 ms
Simultaneous imaging, mapping, and localization						
SLAM	50 m	5 cm	1 cm	-	99.9%	< 10 ms
Indoor NLOS localization	100 m	5 cm	1 cm	-	99.9%	< 10 ms
Urban environment reconstruction (virtual city)	100-200 m	0.5 m	0.1 m	-	99%	< 1s
Augmented human sensing						
Remote surgery and medical diagnostics	2 m	1 mm	< 0.5 mm	-	99.9999%	< 1 ms
Security scans on packages via mobile devices	0.5 m	1-2 mm	0.5 mm	-	99%	< 100 ms
Spectrogram recognition for calories	0.5 m	1 mm	0.5 mm	-	99%	< 100 ms
Posture and gesture recognition						
Medical rehabilitation activity recognition	10 m	1 cm	0.5 cm	-	99.9%	< 1s
Virtual piano anywhere, anytime	10 m	0.5 mm	0.1 cm	-	99%	< 1 ms

with the number of allowed bounces. This is, in particular, problematic in outdoor scenarios.

- **Association of the multipath measurements to vTPs**: Another major challenge in implementing the multipath assisted positioning techniques is that a UE has no idea how to match each measurement parameter vector (consisting of angles, delay and Doppler) to a vTP and this can potentially produce a large positioning error. In general, the matching between the observations and the visible vTPs is a combinatorial problem with exponential complexity.

To solve the above issues, we introduce our proposed *sensing-assisted position estimation* (SAPE) scheme. The basic concept of SAPE is to utilize the high resolution capabilities of the massive MIMO and mmWave technologies in space, angular, and time domains in order to increase the resolvability of the multipath components and exploit the environment RF map to identify the potential reflectors of such multipath components, thereby sensing the environment while localizing UEs with high resolution and accuracy. This allows for exploiting the multipath components (including NLOS) to enhance the accuracy of the position, velocity, and orientation information by providing the association between the observations reported from the UEs and the prior information corresponding to the main environment reflectors. Efficient association and accurate mapping need careful design of specific sensing signals, novel transmission and reception signal processing techniques, and their corresponding measurement and signaling mechanisms.

In particular, the proposed SAPE scheme comprises two main steps:

1. **First step sensing or environment sensing**, in which the network (TP) tries to find/update the location of the main reflectors of the environment and obtains the subspace for the next step sensing;

2. **Second step sensing**, in which the TP sends tailored, specific sensing signals in the subspaces obtained in the first step sensing in order to enhance the multipath resolvability and association. The UE performs measurements over the received sensing signals, and by proper mapping of the measurements to the vTPs, the UE position, as well as velocity vector and orientation (which is also referred to as *pose estimation*) can be obtained.

The proposed SAPE scheme is in contrast to most SLAM techniques where all the localization burden/processing is at the UE side.

3.2 Detailed Proposed SAPE Scheme

3.2.1 First and Second Step Sensing

In the initial environment sensing stage (first step), the TP senses the entire communication space by using a relatively wide beam or small bandwidth in order to generate a coarse RF map to the main reflectors/objects of the communication space. The main goal of this stage is to identify the potential reflectors and map them to vTPs. A static RF map is then available at the TP through this first stage sensing, based on which the location and orientation of the static objects or reflectors can be pre-calculated.

In the second step of sensing, which is the stage of environment sensing update or dedicated sensing, the TP starts targeted sensing based on the obtained RF map and coarse UE location. Particularly, the TP senses certain subspaces, based on the coarse UE location and location of the main reflectors, and processes the reflected signals to obtain finer sensing information of those reflectors. Simultaneously, the UE also performs measurements on the sensing signal to obtain information including multipath identification, range, Doppler, angular and orientation measurements in order to obtain the UE position. Therefore, the second step sensing refines the pre-calculated information obtained in the first phase and thus supports quasi-static environment. In addition, this step can correct the potentially large location errors of the vTP locations obtained from the first step sensing. The impact of vTP location errors will be studied in Section 4.

3.2.2 Multipath Parameter Estimation

The problem involves estimating the parameters of the dominant J multipath components of the received signal at the UE per transmitted beam. The parameters to be estimated are the delay τ_j, Doppler ν_j, channel path coefficients β_j and angle of arrivals ϑ_j^r, ϕ_j^r, i.e., elevation and azimuth angles of the j-th path. All these parameters are collected into one vector denoted by θ_j, for all j. Given the transmitted signal $\boldsymbol{s}_m(t)$ over the m^{th} beam, the received signal is given by:

$$Y_{(m)}(t)=\sum_{j=1}^{J_{(m)}} X_{j,(m)}(t;\,\theta_j) \qquad (1)$$

where $X_j(t;\ \theta_j)$ is the received signal of the j-th path. We note here that $X_j(t;\theta_j)$ subsumes the effect of the beamformer at the transmitter and the additive white Gaussian noise at the receiver. We note also that the TX beamforming, during the second step sensing stage, makes $Y_{(m)}(t)$ sparse, i.e., $J_{(m)}$ is small. The joint estimation of these space-time-frequency parameters results in complex noncovex optimization problems. Moreover, the entanglement of the paths' parameters limits the accuracy and reduces the resolution of the estimated parameters, thereby impeding their resolvability. In addition, the high dimensionality in space, time, and frequency, and real-time processing requirements necessitate taking the computational complexity of the parameter estimation algorithm into consideration. Thus, we are looking for a low-complexity super-resolution channel parameters estimator. The literature on the multipath parameters estimation can be classified into four categories, namely, spectra-based [5–6], subspace-based [7–8], compressive sensing-based (sparse signal recovery/reconstruction) and maximum likelihood-based (ML) approaches [9–11]. A high-level comparison between the four categories is provided in Table 4.

Among these algorithms, space alternating generalized expectation (SAGE) maximization is known to be a reasonable approach for reducing the computational complexity, and the slow convergence rate of the maximization step in the EM algorithm is improved by employing the alternating optimization concept over the estimated parameters for each path. Similar to EM, the SAGE consists of two consecutive and iterative steps, i.e., expectation and maximization. In the expectation step, the unobservable data (in our case they are the multipath components θ_j) is estimated based on the observation of the incomplete data and a previous estimate $\hat{\theta}^{(i)}$ of the parameters vector θ. In the maximization step, the parameters vector of j-th path θ_j is re-estimated iteratively by alternatingly optimizing the components of θ_j, i.e., delay, Doppler, channel coefficients and angle of arrivals. In this way, the multi-parameter optimization problem is reduced to multiple single-parameter optimization problems.

3.2.3 Multipath Parameter Association

The multipath parameter association problem requires finding a way to associate the estimated parameters, i.e., delays and angles of arrival, of the different channel

Table 4 Comprehensive comparison between channel parameters estimation

Category	Description	Pros	Cons
ML-based	Maximum likelihood estimator	The optimal solution	Prohibitive complexity
ML-based	Importance sampling ML	Superior performance	High complexity Slow convergence
ML-based	Expectation maximization	Superior performance	High complexity Slow convergence
ML-based	SAGE	Super resolution	Medium complexity Fast convergence
Sparse signal reconstruction	OMP and its variant	Competitive performance	Medium complexity
Sparse signal reconstruction	Based on convex relaxations such as l_1 norm, nuclear norm, and atomic norm	Super resolution	High complexity
Subspace-based	MUSIC, ESPRIT, and their variants	Medium resolution	Medium complexity
Subspace-based	FFT	Low resolution	Low complexity

multipath components, $Z = \{\hat{\theta}_1,..,\hat{\theta}_{N_{obs}}\}$, where N_{obs} denotes the number of observation and Z is obtained by the UE for each measurement-beam pair, to their relevant visible vTPs represented by the set of the ground truth values $G = \{\boldsymbol{g}_1,..,\boldsymbol{g}_{N_{vtp}}\}$, where N_{vtp} denotes the number of visible vTPs and G is obtained by the TP through the first and second step sensing, as shown in Figure 2.

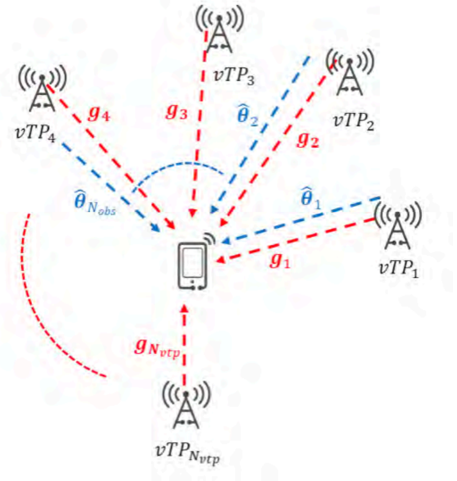


Figure 2 Illustration of measurements — vTP association problem

Extensive research has been conducted in order to alleviate the association error and reduce the computational complexity of the association algorithms. These works can be categorized into two main lines of thoughts, namely, soft-decision/probabilistic data association and hard decision data association [12]. In the probabilistic approach [13–16], all the vTPs are assigned to a certain measurement/observation with different probabilities, with the probabilities indicating how likely a given measurement is due to a particular vTP. This requires a proper selection of the statistical model for assigning these probabilities. In the hard decision data association approach [12, 17], each measurement/observation is associated only to one vTP. The techniques within this approach can be divided into two categories, namely, probabilistic-based hard decision algorithms and distance metric-based selection algorithms. In the former, the measurement is associated to the most likely association event according to a certain probabilistic measure such as maximum likelihood or a posteriori; in the latter, the measurement is associated to the nearest association event according to a certain distance metric such as the Mahalanobis distance [12, 17]. The main drawback of this approach is that it depends heavily on the accurate knowledge of the UE position as a prior.

Prior art adopts measuring the distance between the set of the measurements $Z = \{\hat{\theta}_1,.., \hat{\theta}_{N_{obs}}\}$ and the set of the expected ground truth values $G = \{\boldsymbol{g}_1,..,\boldsymbol{g}_{N_{vtp}}\}$, where $\hat{\theta}_i$ is the ith vector that contains range and angles of arrival of the measurement and $\boldsymbol{g}_k$ is the distance vector between the expected UE position $\hat{\boldsymbol{p}}$ and the k-th vTP obtained from previous estimations. However, this technique has two main shortcomings. First, it requires prior information about the UE's position which might not be available in many scenarios. Second, it requires using long training periods and measurements to iteratively update the prior knowledge of the UE's position to make the association algorithm converge. To cover these shortcomings, we propose a new technique that mainly exploits the differential/mutual distances between the members of the two measurement sets. The key idea of the proposed algorithm is to match the relative/differential distances between the members of the measurements set to the relative distances between a subset of visible/expected vTPs as shown in Figure 3.

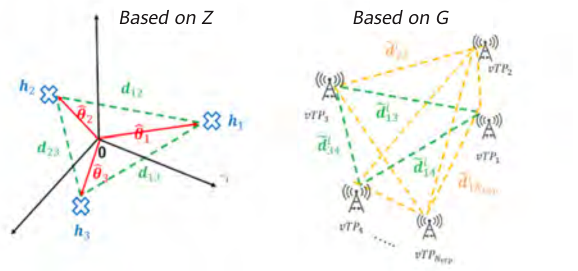


Figure 3 Illustration of relative/tdifferential distance (a) based on Z (b) based on G

This is mainly aimed at avoiding the dependency of G on $\hat{\boldsymbol{p}}$. This requires two different modifications on the two sets, Z and G. First, instead of directly using the N_{obs} measurements, we convert them into N_{obs} hypothetical vTP locations relative to the origin point. We denote this set of hypothetical vTP locations by $H_z = \{\boldsymbol{h}_1,..,\boldsymbol{h}_{N_{obs}}\}$, where h_i is the relative location vector of the i-th hypothetical vTP. Based on these relative location vectors, we calculate the differential/mutual distances between these relative locations, i.e., $\boldsymbol{d}_{ij} = \boldsymbol{h}_i - \boldsymbol{h}_j \forall_i, j, i \neq j$. The set of the differential/mutual Euclidean distances between the hypothetical vTPs' locations is defined as D and has a size of $\binom{N_{obs}}{2}$ elements where its n-th element is denoted by $\boldsymbol{d}_{ij}(n)$. The second modification is to build the set of the differential/mutual distances between the real vTPs' locations, i.e., $\tilde{D}$. Because the set of hypothetical vTPs' locations, i.e., H_z and the set of the actual vTPs locations, i.e., R usually have a different cardinality, we divide the later set into $\binom{N_{vtp}}{N_{obs}}$ subsets of size

N_{obs}, with each containing a different combination of vTPs' locations. We denote these subsets by $r_i = \{\boldsymbol{r}_{i1,\ldots,}\boldsymbol{r}_{iN_{obs}}\}$, $i \in \{1,\ldots,\binom{N_{vtp}}{N_{obs}}\}$, where $\boldsymbol{r}_{i1}$ is the location vector of first vTP in the ith subset. For each r_i, we define the differential/mutual distances between its members as $\tilde{\boldsymbol{d}}^{i}_{mn} = \boldsymbol{r}_{im} - \boldsymbol{r}_{in}, \forall m,n,m \neq n$. The set of the mutual/differential between the members of r_i is denoted by $\tilde{D}^i$. We measure the distance between the sets D and $\tilde{D}^i$ by:

$$d^{i}_{DD} = \min_{\pi_j} \sum_{n=1}^{\binom{N_{obs}}{2}} \left| d(n) - \tilde{d}^{i}_{\tilde{D}^{i}_{\pi_i}}(n) \right| \tag{2}$$

where, with a slight abuse of notation, d(n) and $\tilde{d}^{i}_{\tilde{D}^{i}_{\pi_j}}(n)$ are n-th elements of D and $\tilde{D}^{i}_{\pi_j}$, respectively, and π_j is the j-th permutation of the elements of $\tilde{D}^{i}$. Using this new metric, the association is given by:

$$\pi_{opt} = \arg_{i} \min d^{i}_{DD} \tag{3}$$

π_{opt} contains the indices of those entries of R that are optimally assigned to D.

3.3 Performance Evaluation of the Proposed SAPE Scheme

3.3.1 Link-Level Evaluation

In this subsection, we mainly evaluate the average behaviour of the proposed association algorithm using statistically generated measurements. In the simulation, we generate 8 uniformly distributed vTP positions. We further assume each received observation θ (can be range or angle) has Gaussian noise with standard deviations of $\sqrt{c * \text{CRLB}(\theta)}$, where CRLB($\theta$) denotes the Cramer-Rao Lower Bound for mean square error estimate of θ and the constant factor c accounts for the non-ideal factors in the detection which results in the gap between the real estimation and the lower-bound (in the evaluation results, c = 6). We calculate the association error by counting the number of different indices of the associated vTPs from the observed ones.

As shown in Figure 4, the proposed association algorithm provides a very good performance in the synthetic scenario. In addition, better performance is observed with more measurements (more vTPs) because more mutual/differential distances are used for association. We note that the association error in Figure 4 represents the ratio of the number of vTPs wrongly associated to the total number of vTPs on average. For instance, in Figure 4, it is shown that with 3 observations, one gets an average association error of 0.07 at SNR of 5 dB, i.e., just 7 out of 100 vTPs on average will be associated wrongly. It is also noteworthy to mention that the average association error is different from the average position error. However, the former affects the later. In other words, wrongly associating one out of 10 vTPs might not produce significant position error if the measurements for this vTP have a lower weight in calculating the position error.

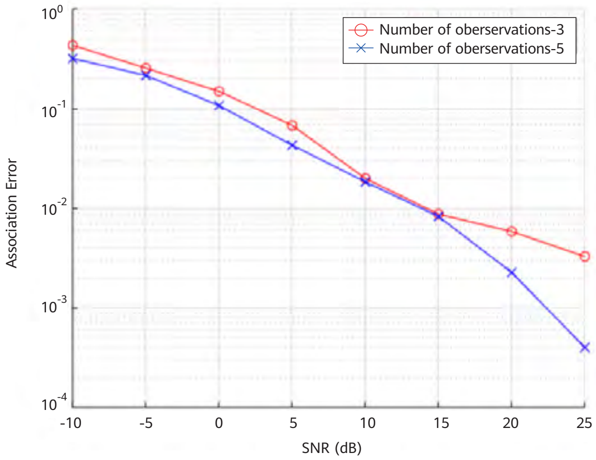


Figure 4 Average association error of the proposed association algorithm when the number of visible vTPs is 3 and 5

We further evaluate the performance of the proposed association algorithm and the impact on the positioning error in a real multipath environment. Without generality, we assume that the estimation error due to the channel parameters estimation stage follows the CRLB. Figure 5 presents the CDF of the UE positioning error bound (PEB) due to the association scheme and compares it with the case of ideal association, which shows the promising performance of the proposed SAPE technology and its potential for 6G positioning.

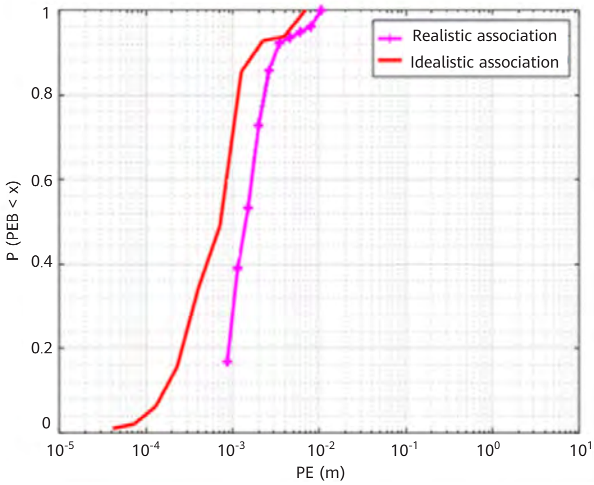


Figure 5 Overall positioning performance of the proposed SAPE technology at the link level

3.3.2 System-Level Evaluation

In this section, we provide the performance of the proposed SAPE scheme at the system level. Similar to the system-level communication performance evaluation, the key step in such an evaluation is abstracting the PHY-level performance at the network level, which is the so-called PHY abstraction. The rationale behind sensing PHY abstraction is to map the system parameters in terms of SINR, bandwidth, time duration, and antenna configuration to a sensing performance (i.e., range, Doppler or angle mean square errors). The proposed SAPE scheme is evaluated and compared with baseline NR in terms of PEB, based on the proposed PHY abstraction methodology in two scenarios:

Idealistic scenario: where the sensing is assumed to be perfect. In this case, the evaluation is based on applying the proposed PHY abstraction methodology in SLS and evaluating the candidate schemes in two scenarios: indoor hotspot (InH) and outdoor urban micro (UMI). Both scenarios are evaluated over mmWave bands and the simulation parameters are given in Table 5.

Table 5 Parameters for SLS evaluation

Parameter	Value
Bandwidth	80 MHz
Sensing time	14 symbols
Sub-carrier spacing	60 kHz
Number of subcarriers	1024
Deployment	Indoor hotspot, 256 × 32 UMI 32 × 16 (outdoor only), 20 RRUs and 200 UEs
Channel model	SCM (stochastic)
Carrier frequency	60 GHz
Simulation methodology	Based on SLS using the proposed sensing PHY abstraction
Non-idealities modeled	Sensing error
Synch. error between TRPs	0 (perfect synch.) or 1 ns

Based on these parameters, the simulation results are given in Figure 6.

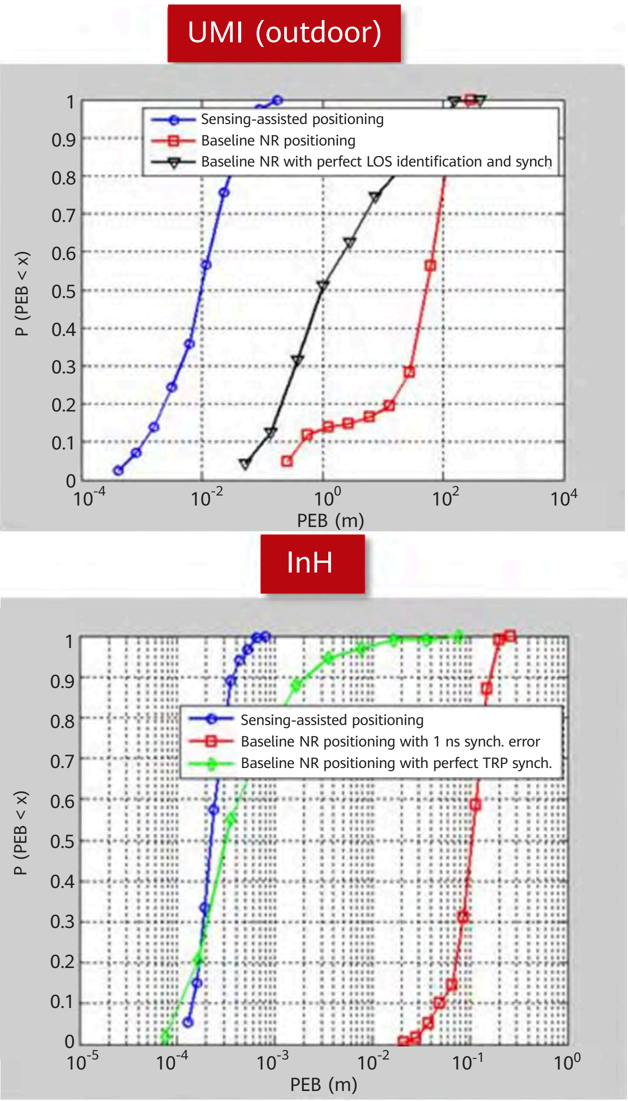


Figure 6 SLS results of the proposed SAPE vs. baseline NR in idealistic scenario

Based on the results, we can observe that under ideal conditions (no RF impairments, no sensing error, no diffraction), SAPE can achieve an order of magnitude better accuracy compared to NR. In addition, the NR baseline cannot achieve good performance in any scenario, even under ideal conditions, due to NLOS bias and synchronization error between the TPs.

Realistic scenario: assuming sensing error, the candidate schemes are evaluated in outdoor UMI. The simulation parameters are the ones given in Table 5. For modeling the sensing error, we assume the vTPs corresponding to each path/cluster are Gaussian-distributed with some variances which are also modeled as random variables. In addition, the vTP location variance for the LOS link is set to 0 as it corresponds to the actual TP. Based on these parameters, the simulation results are given in Figure 7.

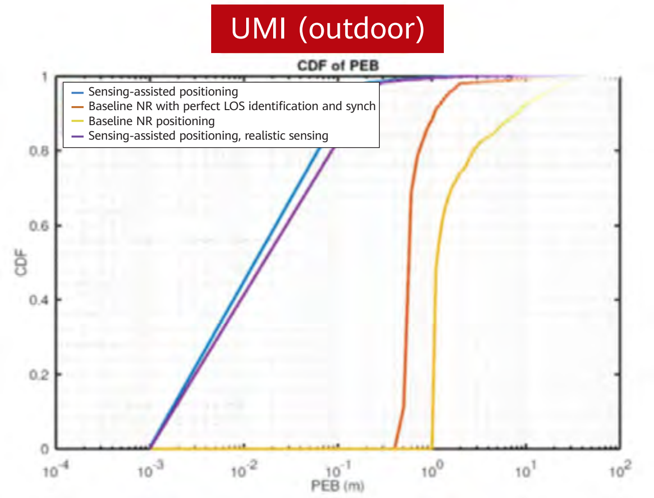


Figure 7 SLS results of the proposed SAPE vs. baseline NR in the realistic scenario

Based on the results, we can observe that SAPE can achieve an order of magnitude better accuracy (cm-level accuracy) when compared with NR, even with the sensing error.

4 ISAC for Millimeter-Level Imaging at the THz Band

THz lies between the mmWave and infrared frequencies, and thus has millimeter-level and even sub-millimeter-level wavelength, making the ISAC system at the THz band (ISAC-THz) particularly suitable for high resolution sensing applications such as millimeter-level resolution 3D imaging. Like the other lower frequency radio waves, THz can penetrate some obstacles, achieving high-precision sensing in all weather and lighting conditions.

Recent developments in semiconductor technology have bridged the "THz band gap" and made the hardware feasible at the terminal side. ISAC-THz based portable devices will thus open the door for numerous new sensing applications such as augmented human sensing with very high resolution. Table 6 shows the allocated mobile frequency bands with a contiguous bandwidth greater than 5 GHz. The ultra-wide bandwidth in THz will also enable Terabits/second data rate transmission, especially in short-range communications. The corresponding range resolution (from equation 4) based on Heisenberg's Uncertainty Principle is also provided in this table. Under the assumption of synthesized aperture, the cross-range resolution is provided in Table 7 based on equation 5 where λ is the wavelength, D is the aperture size, and r is the distance between transceiver and target. Because THz can provide high sensing resolution in addition to high communication throughput, the integration of THz sensing and communication has become an attractive and active research area.

$$\triangle t = \frac{C}{2B} \tag{4}$$

$$\triangle d = \frac{\lambda r}{2D} \tag{5}$$

The application of ISAC-THz design is expected to provide many opportunities for brand new services especially on future mobile devices or even wearables as illustrated in Figure 8. In addition to the localization and imaging applications, molecular spectrogram analysis is another interesting application area that could be enabled by ISAC-THz, as discussed in Section 2.

Table 6 Maximum contiguous bandwidth in the range of 100-450 GHz and the corresponding range resolution

Freq. (GHz)	Contiguous Bandwidth (GHz)	Range Resolution (mm)
102–109.5	7.5	20
141–148.5	7.5	20
151.5–164	12.5	12
167–174.8	7.8	19
191.8–200	8.2	18
209–226	17	8.8
252–275	23	6.5
275–296	21	7.1
306–313	7	21
318–333	15	10
356–450	94	1.6

Table 7 Aperture size and corresponding cross-range resolution at 140 GHz

λ = 2.1 mm, r = 30 cm	
Aperture Size (cm)	Cross-Range Resolution (mm)
1	32
5	6.4
10	3.2
20	1.6

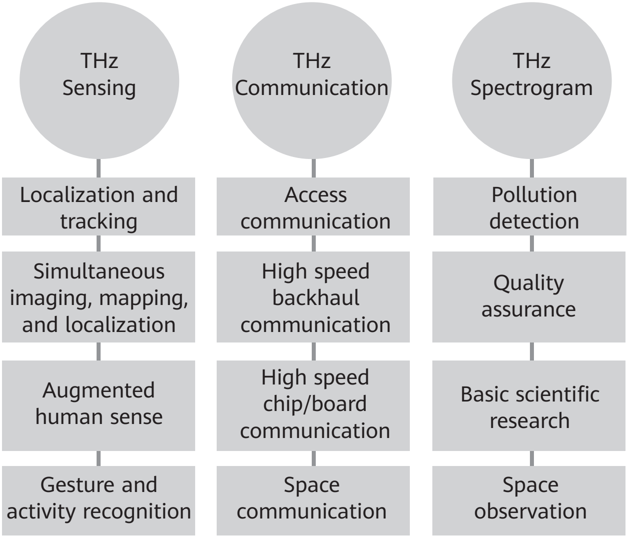

THz Sensing
THz Communication
THz Spectrogram
Localization and tracking
Access communication
Pollution detection
Simultaneous imaging, mapping, and localization
High speed backhaul communication
Quality assurance
Augmented human sense
High speed chip/board communication
Basic scientific research
Gesture and activity recognition
Space communication
Space observation

Figure 8 THz application in sensing and communication

In this section, we elaborate on our ISAC prototype of THz imaging on portable devices that achieves millimeter-level resolution. A robot arm equipped with ISAC-THz module is used to represent a human arm holding a THz imaging camera. The prototype is built to operate at 140 GHz carrier frequency with a bandwidth of 8 GHz.

4.1 Hardware Architecture of the ISAC-THz Module

From the THz imaging aspect, thousands of antenna elements are required to create a large aperture for high cross-range resolution. However, it is clear that physically packing thousands of antenna elements into the portable device is infeasible due to the size and power constraint requirements of the device [18–19]. To solve this problem, virtual aperture techniques are applied in the prototype system [3]. In particular, the virtual MIMO antenna array design in the hardware transceiver architecture using the sparse sampling design in the scanning process is proposed [3, 20–21].

First, a virtual MIMO antenna array structure is constructed to form a virtual aperture that can achieve the same performance with respect to its equivalent physical aperture array as illustrated in Figure 9. Next, a sparse scanning approach is applied, transforming the degree-of-freedom in time and space into a larger virtual aperture, as shown in Figure 9. The scanning performed by the robot arm thus mimics a user holding a smartphone and imaging an object with a zigzag scanning trajectory.

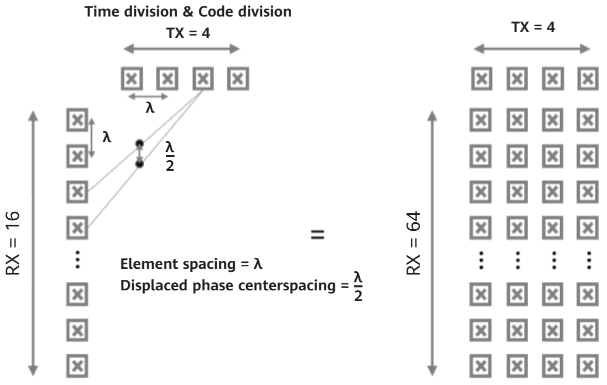

Time division & Code division
TX = 4
λ
λ
λ/2
RX = 16
=
Element spacing = λ
Displaced phase centerspacing = λ/2
TX = 4
RX = 64

Figure 9 MIMO virtual aperture

To implement the overall solution of a virtual aperture, the following three requirements need to be satisfied in the hardware design:

- Multiple transceiver (TRX) chains to support the MIMO antenna array structure as the first step for the overall virtual aperture.
- Wide antenna pattern to cover the target scanning area in order to maintain the correlation among the reflected samplings.
- Real-time position information of the device to perform coherent processing of the received signals.

The schematic of the prototype architecture is shown in Figure 10. The transmitter antenna array has 4 RF ports and the receiver antenna array has 16 RF ports, forming a 4T16R MIMO antenna array structure [3]. The per unit antenna radiation pattern is a wide beam design with a 3 dB beam width of 50° and gain of 7 dBi.

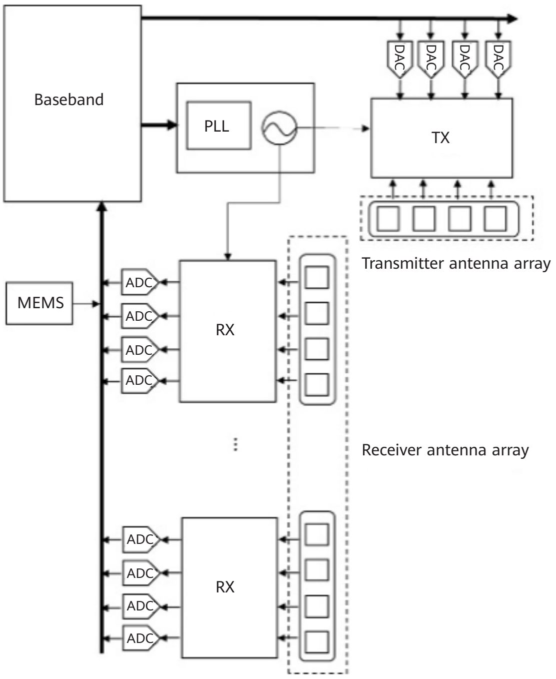


Figure 10 Illustration of the architecture of ISAC prototype

4.2 Compressed Sensing-based Tomography Imaging

A major challenge for the virtual aperture imaging technique is the irregular scanning trajectory caused by the user moving the ISAC imaging module to perform THz scanning on an object. Assume a zigzag scanning routine is used to image an object, as shown in Figure 11. The echo samplings in the horizontal direction are continuous, i.e., the spatial spacing between sampling points is comparable to the wavelength of the echo signal. However, continuous sampling cannot be maintained in the vertical direction. As a result, the echo samplings in the vertical direction are sparse, which will cause high and non-uniform sidelobe effects, giving rise to false artifacts, which may lead to imaging failure.

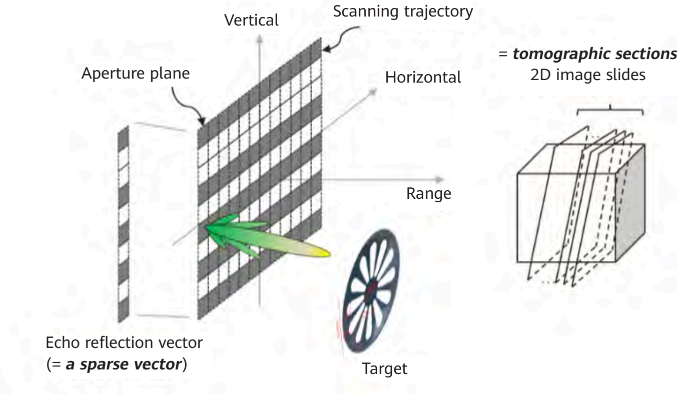
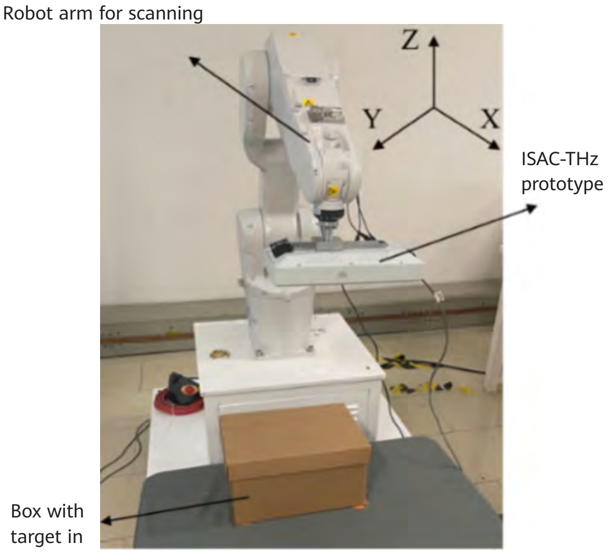


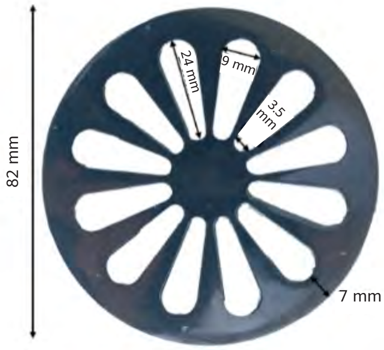
Figure 11 Illustration of the sparse scanning approach and the tomographic imaging techniques

To solve this challenge, we consider decomposing the scanning trajectory on a two-dimensional (2D) plane into several sets of linear scanning tracks along the horizontal direction, where the sparseness of the sampling signals in the vertical domain is then equivalent to the sparseness between horizontal tracks, as illustrated in Figure 11. In this case, the reflected/echo information from the object can be retrieved from these vertically sparse samplings via compressed sensing techniques [3].

As depicted in Figure 12a, the robotic arm scans at a speed of 1 m/s with the scanning area set as 10 cm by 12 cm in the prototype. The longitudinal spacings of the scan trajectories are controlled to simulate the sparsity in the trajectories of the user's hand-held scanning behavior. The target object to be imaged, as shown in Figure 12b, is put in a box with a cap on top of it. As we can see from Figure 12b, the smallest distance in the hallowed pattern is 3.5 mm, so the highest resolution of the imaging results can be 3.5 mm.



(a) Prototype setup for THz sensing where the ISAC-THz module is held by a robot arm representing a human arm

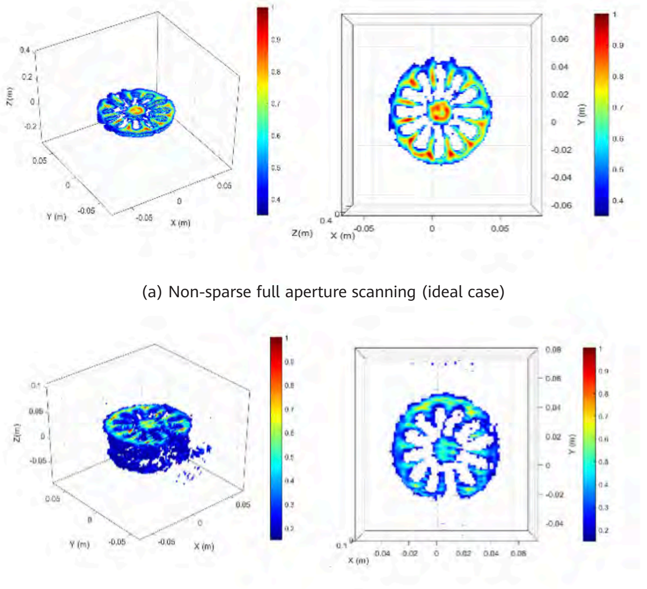


(b) Target object in the box

Figure 12 Setup of the ISAC-THz prototype

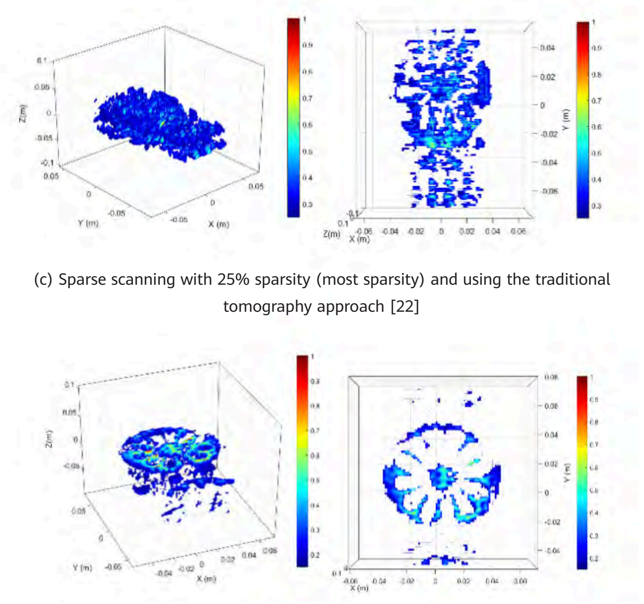
The proof-of-concept THz imaging performances with different sparsity configurations in the scanning patterns are presented and compared in Figure 13. In each of the figures, the 3D imaging results are shown on the left and the cross-range profile perceived from top down is shown on the right.

The non-sparse full aperture scanning in Figure 13a is an ideal case, in which the vertical sampling is half wavelength adjacent. This achieves the best PSLR and ISLR performance, which is set as an upper bound performance reference. Then, in order to simulate the sparsity in real free hand scanning, we assume different sparsity configurations in tests, from 50% (medium sparsity) to 25% (most sparsity), where *X*% sparsity means that there are *X*% of the full samplings remaining in the vertical direction. With the collection of fewer samplings, stronger side-lobe interference occurs at the resulted aperture, resulting in worse imaging performance. From the comparison of Figure 13c and Figure 13d, we see that when the sparsity is too high, the traditional tomography algorithm is not enough to recover the images. In this case, the compressed sensing based tomography approach showed its superior performance.



(a) Non-sparse full aperture scanning (ideal case)

(b) Sparse scanning with 50% sparsity (medium sparsity)



(c) Sparse scanning with 25% sparsity (most sparsity) and using the traditional tomography approach [22]

(d) Sparse scanning with 25% sparsity (most sparsity) and using the compressed sensing based tomography approach

Figure 13 Imaging results at different sparsity configurations

4.3 Multi-Channel Imaging

The multi-channel imaging process can be treated as a time-domain coherent combination of electromagnetic signals from multiple receiving channels. Theoretically, n receivers can reduce the sampling time to 1/n compared with one receiver with the same imaging quality. Less sampling time will reduce the difficulty of motion error compensation, which in turn will improve the imaging quality.

However, in multi-channel imaging, one major challenge arises from the imbalance in gain and time delay of different receiver channels due to hardware imperfection. The antenna mounting positional imperfection will introduce the displaced phase center error, as shown in Figure 14. Multi-channel amplitude and phase imbalance will lead to azimuth ghosting, which will significantly degrade the imaging quality. The amplitude imbalance can be easily compensated by multichannel amplitude equalization methods [23], while phase imbalance compensation needs auto-focusing algorithm such as gradient descent.

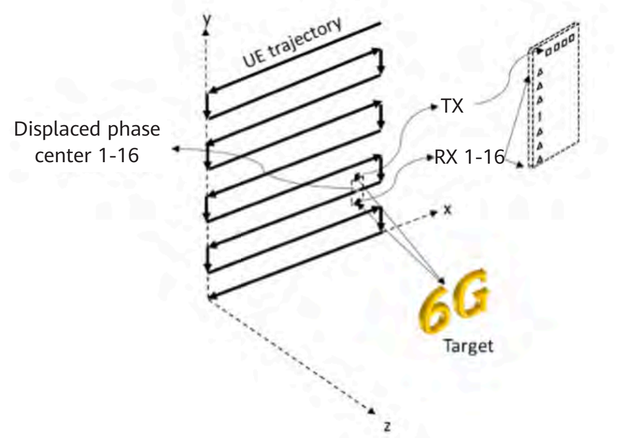
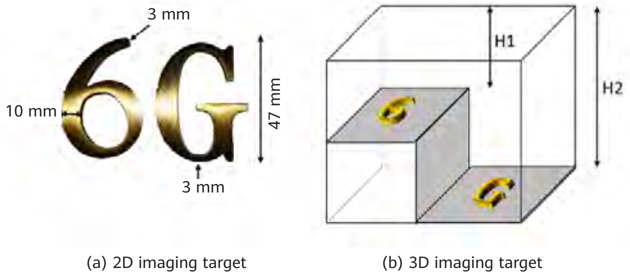


Figure 14 Illustration of the displaced phase center caused by multi-channel imaging

To validate the performance of multi-channel imaging, we use the same testbed described in the last subsection but a different target (resolution is similar, i.e., 3 mm) as shown in Figure 15. In this case, we tried both the 2D target shown in Figure 15a and the 3D target shown in Figure 15b, where the 3D imaging target was formed by placing the two characters at different heights inside the box. With the aforementioned benefit of multi-channel imaging, very sparse sampling is needed for a good imaging quality. In the prototype, only 12% sparsity is configured in the scanning trajectory.



(a) 2D imaging target

(b) 3D imaging target

Figure 15 Imaging targets used in multi-channel imaging

Figure 16 shows the imaging results of the 2D target. The imaging results without multi-channel phase error compensation are illustrated in Figure 16a where severe ghosting on the final image that significantly degrades the imaging quality can be prominently seen. Using the geometric interpretation algorithm, the sidelobe due to the multi-channel imbalance has been duly suppressed as can be clearly seen in Figure 16b.

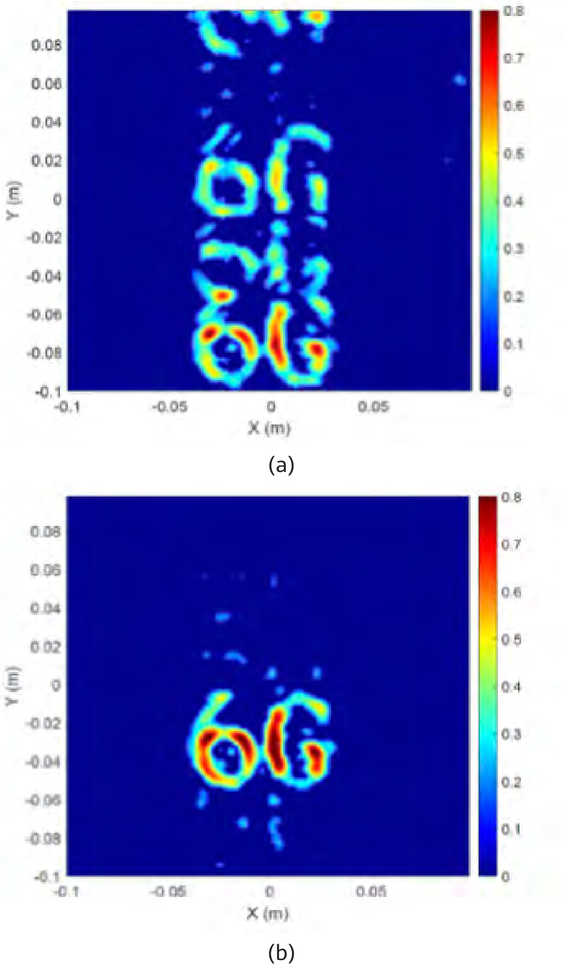


Figure 16 Multi-channel imaging result of the 2D target

Subsequently, the imaging results of the 3D target are shown in Figure 17. The imaging result clearly depicts the shape of the two characters and their relative distance in the 3D space.

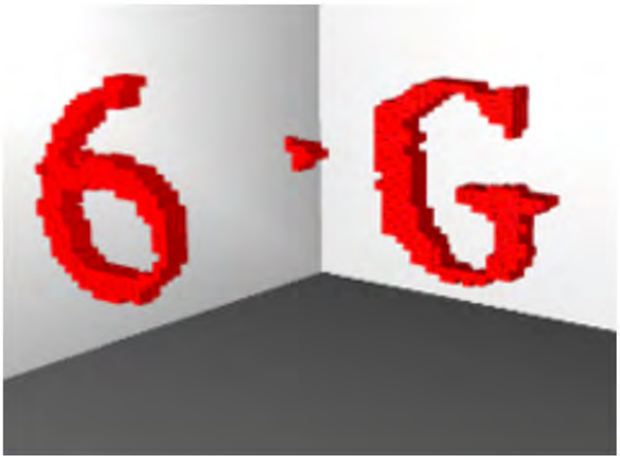


Figure 17 3D imaging result

5 Major Challenges for Making ISAC a Reality

5.1 Channel Modeling and Evaluation Methodology

In 6G, the channel model needs to be considered for both communication and sensing services. This brings significant challenges to the channel modeling methodology. Until 5G, because it has low computational complexity and is easily standardizable, stochastic channel modeling methodology dominated the evaluation of wireless communications, and is used in many projects and standards such as 3GPP-SCM, WINNER-I/II, COST2100, and MESTIS. It is adequate in evaluating the communication performance. However, there is a doubt as to whether it still can meet the more diverse requirements from different sensing applications.

One typical sensing channel is the echo channel, which consists of the backscattering RCS characteristics from the object and its surroundings. This type of propagation channel brings new requirements for the physical electromagnetic (EM) characteristic which are not supported in the current communication channel models. One typical use case is the high resolution imaging application. This type of application requires the deterministic channel coherence of the antenna array aperture with the geometry information. This requirement is contradictory to the typical stochastic channel modeling approach. Therefore, the traditional channel modeling methodologies deserve some rethinking and innovation.

Another major challenge is the evaluation performance metrics based on the new sensing requirements. Conventionally, throughput, latency, and reliability are the main evaluation performance metrics for communication systems. However, due to the different sensing applications, there are new dimensions of evaluation metrics that need to be considered, such as sensing resolution, accuracy, detection probability, and update rate. So far, no KPIs have been proposed for the joint performance characterization and evaluation of both the communication and sensing services. This implies that a new scenario-dependent evaluation methodology may need to be investigated.

To address the challenges mentioned above, the following research directions are proposed:

- Typical scenarios and evaluation methodology

The traditional indoor hotspot, urban micro, urban macro are defined in the 3GPP 38.901 communication channel. The environment and the purpose of the application will deeply affect the channel model parameters and even the channel model generation approach. Therefore, the ISAC typical scenario should be categorized and the typical use cases of each category should be highlighted for further evaluation.

For a given evaluation use case, metrics to characterize the joint performance between communication and sensing are needed in order to optimize the performance trade-off for both services simultaneously. To characterize the performance of both functions as well as mutual enhancement, scenarios and metrics need to be implemented into the system level simulations and the ISAC performance must be evaluated in a fully integrated network.

- Channel measurement and modeling methodology

Regarding ISAC channel modeling, a single channel modeling scheme may not meet the need to evaluate all ISAC applications. Instead, stochastic, deterministic, and even hybrid channel models must be considered. For instance, in the sensing-assisted beamforming use case, the stochastic channel modeling could be adopted, whereas for localization and tracking application, ray tracing could be considered as a strong candidate for channel modeling because the detailed contours for the object reflection/scattering are not strictly required. On the other hand, for imaging and recognition applications, there is a need to consider EM algorithm when the size of the scatterers is close to the signal wavelength and therefore the interaction of the signal to the scatterers are strongly correlated with the EM characteristics.

5.2 Joint Waveform and Signal Processing Design

Most of the works on the joint design of sensing and communications mainly focus on the joint waveform design. The main challenge for the joint waveform design is the contradicting KPIs for communications and sensing. In particular, the main target for communications is maximizing the spectral efficiency, whereas the optimum waveform design for sensing is focused on estimation resolution and accuracy. Because CP-OFDM has been proven to be a favorable option for communication, many researchers have considered this waveform for sensing as well. Although the introduction of cyclic prefix (CP) has been shown to degrade auto-correlation in the time domain [24], a novel approach of frequency domain processing [25] allows for efficient parameter estimation of CP-OFDM, achieving the maximum processing gain. Furthermore, CP-OFDM has been shown to be free of the range-Doppler coupling problem, which means that the range and Doppler estimation can be performed independently [25]. However, these favorable properties of CP-OFDM depend on perfect synchronization (in both time and frequency domains) between the transmitter and the receiver, which may not be present, especially for bistatic sensing. In addition, the large peak to average power ratio (PAPR) of CP-OFDM is another major issue for radar applications where power efficiency is very important.

Alternatively, frequency modulated continuous wave (FMCW) waveform, which has traditionally been used for radar, is not capable of carrying data at transmission rates desirable for communication services. Some researchers proposed to modify the FMCW waveform to make it more communication-friendly. Among the many contributions in this line of research, we can mention [26], in which the authors propose to use up-chirp for communication and down-chirp for radar, and [27], introducing trapezoidal frequency modulation continuous-wave (TFMCW) modulation in which the radar cycle and communication cycles are multiplexed in the time domain. Although these techniques enable efficient multiplexing of communication data in the sensing signal, they still suffer from low spectral efficiency due to the existence of the chirp-like sensing signal. Another line of research is devoted to using single-carrier waveform based on the code domain spreading of joint radar and communication signals. For this class of waveforms, the radar performance has been shown to be affected by the auto-correlation of the sequences and the long spreading codes result in good auto-correlation at the expense of communication spectral efficiency [27]. In addition, Doppler estimation requires more complicated algorithms [27]. The current state of the art suggests that there is still room for waveform design to strike a balance between good communication and sensing performance to meet 6G ISAC requirements.

5.3 Hardware Co-design

In the design of the ISAC system, the solution that integrates the baseband and RF hardware reduces the overall power consumption, system size, and information exchange latency between the two systems. The hardware converging strategy facilitates the mutually beneficial functions of sensing and communication in distortion calibration and compensation. The common impairments in the ISAC system due to hardware imperfections are demonstrated in Figure 18.

It should be noted that in light of the differences in evaluation metrics and algorithms between communication and sensing, hardware requirements are quite different. Considering the cost and size of the historical communication and radar systems, the ISAC system hardware design will closely resemble the traditional communication architecture. As a tradeoff, we need to consider the impact of distortion parameters on sensing performance. For instance, a communications system depends on full duplex isolation to achieve high capacity.

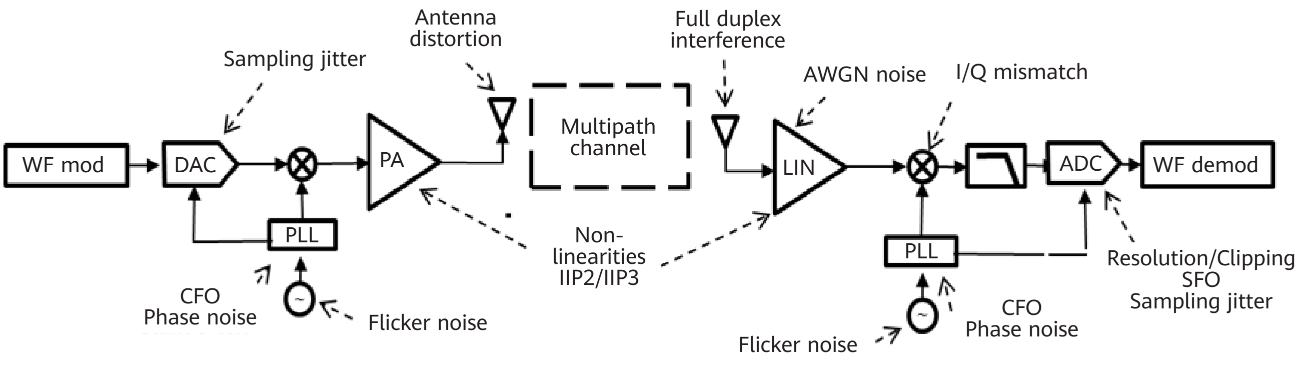


Figure 18 Impairments of ISAC transmission system

In contrast, from the OFDM ISAC perspective, limited transmitter-receiver isolation is a primary concern in the detection of static targets [28]. Proper design of integrated RF architecture and self-interference cancellation in the ISAC system are key technical problems that need to be solved. Another issue is that sensing requires the accumulation of coherent signals to ensure performance, which makes the system more sensitive to sampling jitter, frequency offset, and phase noise [29]. This in turn leads to higher requirements on synchronization and stability of the system. In short, we need to consider these hardware challenges in the selection of ISAC waveforms, sensing algorithms, and non-ideal distortion compensation schemes.

5.4 Sensing-assisted Communication

Although sensing will be introduced as a separate service in the future, it might still be beneficial to look at how the information obtained through sensing can be used in communication. The most trivial benefit of sensing will be environment characterization, which enables sensing-assisted communication due to more deterministic and predictable propagation channels. It has been shown that the environment knowledge provided by sensing not only improves the accuracy of channel estimation in mmWave, but also significantly reduces the overhead because the environment is shared by potentially many UEs and sensing-based channel acquisition does not repeat the channel estimation process for each individual link [30]. Another example would be sensing-assisted beam alignment, especially in mmWave vehicular communication where the main challenge is the frequent link reconfiguration resulting in significant overhead. In [31], it has been proposed to use the information obtained from a radar mounted on an infrastructure operating in a given mmWave band to configure the beams of the vehicular communication system operating in another mmWave band. Moreover, the users' location information and the environment map obtained by sensing helps identify the link blockage caused by large objects, especially in dense urban networks, so that the power and beams can be adjusted accordingly to improve the communication throughput [32]. Other examples of sensing-assisted communication can also be considered and studied to reduce the latency and overhead of communication systems in future 6G networks with the help of information provided by the sensing system. Another benefit of sensing for communication would be improving the users' positioning accuracy by combining the advantages of active localization and passive localization and thus overcoming their shortcomings to satisfy 6G localization requirements.

5.5 Communication-assisted Collaborative Sensing

6G ISAC takes advantage of the mobile communication network to support synchronized, collaborative multi-node sensing. Sensing through cooperation refers to the sensing nodes that share their observations with each other and attempt to reach a common consensus on the surrounding environment. This will significantly improve localization performance. Integration of sensing capabilities into the existing communication network will be the most viable and cost effective option where the multiple network nodes (base stations, UEs, etc.) can function as a complete sensing system to enable network sensing operations for the use cases highlighted in Section 2. The process involves the collaborating nodes forming a dynamic reference grid through distributed sensing and processing. The collaboration reduces measurement uncertainty and provides greater coverage as well as higher sensing accuracy and resolution through sensing data fusion. In addition, this offers interesting possibilities for being able to carry out sensing under non-line-of-sight (NLOS) conditions. The major research challenges here would lie in the synchronization, joint processing, and network resource allocation in order to achieve the optimum sensing fusion results.

6 Conclusion

With the concept of ISAC being commonly accepted as one of the key technology trends for 6G, this paper takes a step forward and elaborates two case studies on how 6G ISAC technologies can be applied to improve localization and to perform high resolution imaging. In particular, the proposed SAPE scheme utilizes the joint benefit of device-free and device-based sensing and greatly improves the positioning accuracy compared with the current NR scheme. The prototype of the THz camera justifies the feasibility of mm-level imaging resolution on portable devices for both 2D and 3D objects placed in a box. Joint efforts from both academia and industry are needed to address further challenges in the system level evaluation of ISAC, new channel modeling methodology, new waveform design, low complexity algorithm design, and low cost hardware design.

References

[1] W. Tong, P. Zhu, *et al.*, "6G: the next horizon: from connected people and things to connected intelligence," Cambridge: Cambridge University Press, 2021.

[2] D. K. P. Tan, J. He, Y. Li, A. Bayesteh, Y. Chen, P. Zhu, and W. Tong, "Integrated sensing and communication in 6g: motivations, use cases, requirements, challenges and future directions," in *1st IEEE International Online Symposium on JC&S*, 23-24 February, 2021.

[3] O. Li et al., "Integrated sensing and communication in 6G: a prototype of high resolution THz sensing on portable device," in *European Conference on Networks and Communications (EuCNC)*, 8-11 June, 2021.

[4] K. Witrisal et al., "High-accuracy localization for assisted living: 5G systems will turn multipath channels from foe to friend," in *IEEE Signal Processing Magazine*, vol. 33, no. 2, pp. 59-70, March 2016, doi: 10.1109/MSP.2015.2504328.

[5] F. Talebi and T. Pratt, "Channel sounding and parameter estimation for a wideband correlation-based MIMO model," in *IEEE Transactions on Vehicular Technology*, vol. 65, no. 2, pp. 499-508, Feb. 2016, doi: 10.1109/TVT.2015.2404571.

[6] N. Ben Rejeb, I. Bousnina, M. B. Ben Salah, and A. Samet, "Channel parameters estimation using cross-correlation matrix of a wireless SIMO system," *Fourth International Conference on Communications and Networking, ComNet-2014*, 2014, pp. 1-5, doi: 10.1109/ComNet.2014.6840915.

[7] L. Wei, Q. Li, and G. Wu, "Direction of arrival estimation with uniform planar array," *2017 IEEE 86th Vehicular Technology Conference (VTC-Fall)*, 2017, pp. 1-5, doi: 10.1109/VTCFall.2017.8287882.

[8] F. Wen, N. Garcia, J. Kulmer, K. Witrisal, and H. Wymeersch, "Tensor decomposition based beamspace ESPRIT for millimeter wave MIMO channel estimation," *2018 IEEE Global Communications Conference (GLOBECOM)*, 2018, pp. 1-7, doi: 10.1109/GLOCOM.2018.8647176.

[9] Pei Chen and H. Kobayashi, "Maximum likelihood channel estimation and signal detection for OFDM systems," *2002 IEEE International Conference on Communications. Conference Proceedings*. ICC 2002 (Cat. No.02CH37333), 2002, pp. 1640-1645 vol.3, doi: 10.1109/ICC.2002.997127.

[10] R. Carvajal, J. C. Aguero, B. I. Godoy, and G. C. Goodwin, "EM-based maximum-likelihood channel estimation in multicarrier systems with phase distortion," in *IEEE Transactions on Vehicular Technology*, vol. 62, no. 1, pp. 152-160, Jan. 2013, doi: 10.1109/TVT.2012.2217361.

[11] M. Feder and E. Weinstein, "Parameter estimation of superimposed signals using the EM algorithm," in *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 36, no. 4, pp. 477-489, April 1988, doi: 10.1109/29.1552.

[12] Y. Bar-Shalom, F. Daum, and J. Huang, "The probabilistic data association filter," in *IEEE Control Systems Magazine*, vol. 29, no. 6, pp. 82-100, Dec. 2009, doi: 10.1109/MCS.2009.934469.

[13] Paul Meissner, Christoph Steiner, and Klaus Witrisal, "UWB positioning with virtual anchors and floor plan information," *2010 7th Workshop on Positioning, Navigation and Communication*, IEEE, 2010.

[14] Paul Meissner, Thomas Gigl, and Klaus Witrisal, "UWB sequential Monte Carlo positioning using virtual anchors," *2010 International Conference on Indoor Positioning and Indoor Navigation*, IEEE, 2010.

[15] Paul Meissner *et al.*, "Analysis of an indoor UWB channel for multipath-aided localization," *2011 IEEE International Conference on Ultra-Wideband (ICUWB)*, IEEE, 2011.

[16] Deissler, Tobias, and Jörn Thielecke., "UWB SLAM with rao-blackwellized Monte Carlo data association," *2010 International Conference on Indoor Positioning and Indoor Navigation*, IEEE, 2010.

[17] Paul Meissner and Klaus Witrisal, "Multipath-assisted single-anchor indoor localization in an office environment," *2012 19th International Conference on Systems, Signals and Image Processing (IWSSIP)*, IEEE, 2012.

[18] C. B. Barneto, T. Riihonen, M. Turunen, L. Anttila, M. Fleischer, K. Stadius, J. Ryynänen, and M. Valkama, "Full-duplex OFDM radar with LTE and 5G NR waveforms: challenges, solutions, and measurements," in *IEEE Transactions on Microwave Theory and Techniques*, vol. 67, no. 10, pp. 4042-4054, Oct. 2019.

[19] K. Siddiq, R. J. Watson, S. R. Pennock, P. Avery, R. Poulton, and B. Dakin-Norris, "Phase noise analysis in FMCW radar systems," *2015 European Radar Conference (EuRAD)*, Paris, pp. 501-504, 2015.

[20] C. Wang, Y. Li, Z. Li, K. Zeng, J. He, and G. Wang, "A 3D imaging method for future communication and imaging integrated terminal," *The 2021 CIE International Conference on Radar*.

[21] X. Li, J. He, Z. Yu, G. Wang, and P. Zhu, "Integrated sensing and communication in 6G: the deterministic channel models for THz imaging," *2021 IEEE 32nd annual international symposium on personal, indoor and mobile radio communications (PIMRC)*, 2021, pp. 1-6, doi: 10.1109/PIMRC50174.2021.9569384.

[22] T. Jin, X. Qiu, D. Hu, and C. Ding, "Unambiguous imaging of static scenes and moving targets with the first Chinese dual-channel spaceborne sar sensor," Sensors 17(8), 1709 (2017).

[23] A. Vertiy and S. Gavrilov, "Near-field millimeter wave and microwave tomography imaging," *Proc. Int. Kharkov Symp. Phys. Engrg. Millim. Sub-Millim. Waves (MSMW)*, Jun. 2007, pp. 104-108.

[24] B. Paul, A. R. Chiriyath, and D. W. Bliss, "Survey of RF communications and sensing convergence research," IEEE Access, pp. 252 - 270, 2016.

[25] M. Braun, C. Strum, and F. K. Jondral, "Maximum likelihood speed and distance estimation for OFDM radar," in Proc. *2010 IEEE Radar Conf.*, Washington, DC, May 2010.

[26] G. N. Saddik, R. S. Singh, and E. R. Brown, "Ultra-wideband multifunctional communications/radar system," *IEEE Trans. Microw. Theory Tech.*, pp. 1431-1437, July 2007.

[27] K. Wu and L. Han, "Joint wireless communication and radar sensing systems - state of the art and future prospect," *IET Microwaves Antennas & Propagation*, vol. 7, no. 11, pp. 876-885, 2013.

[28] C. B. Barneto, T. Riihonen, M. Turunen, L. Anttila, M. Fleischer, K. Stadius, J. Ryynänen, and M. Valkama, "Full-duplex OFDM radar with LTE and 5G NR waveforms: challenges, solutions, and measurements," in *IEEE Transactions on Microwave Theory and Techniques*, vol. 67, no. 10, pp. 4042-4054, Oct. 2019.

[29] K. Siddiq, R. J. Watson, S. R. Pennock, P. Avery, R. Poulton, and B. Dakin-Norris, "Phase noise analysis in FMCW radar systems," *2015 European Radar Conference (EuRAD)*, Paris, pp. 501-504, 2015.

[30] C. Jiao, Z. Zhang, C. Zhong, and Z. Feng, "An indoor mmWave joint radar and communication system with active channel perception," in *2018 IEEE International Conference on Communications (ICC)*, Kansas City, MO, 2018.

[31] N. González-Prelcic, R. Méndez-Rial, and R. W. Heath, "Radar aided beam alignment in mmWave V2I communications supporting antenna diversity," in *Information Theory and Applications Workshop (ITA)*, La Jolla, CA, 2016.

[32] Z. Li, S. Yang, and T. Clessienne, "Exploiting location information to enhance throughput in downlink V2I systems," in *2018 IEEE Global Communications Conference (GLOBECOM)*, Abu Dhabi, United Arab Emirates, 2018.

NET4AI: Supporting AI as a Service in 6G

Xu Li [1], Hang Zhang [1], Chenghui Peng [2], Zhe Liu [2], Fei Wang [2]

[1] Ottawa Wireless Advanced System Competency Centre

[2] Wireless Technology Lab

Abstract

In this paper, we address the emerging privacy-preserving deep learning problem through a systemic approach. We generalize split learning (SL) and combine it with federated learning (FL) to obtain a two-level learning framework. This framework inherits the advantages of both SL and FL, yet avoids their drawbacks and allows for learning approach customization through cut layer selection. We further propose a 6G system architecture, named NET4AI, to support the two-level learning framework (for example, perform learning approach customization) and to provide *k*-anonymity and data confidentiality protection. It is worth noting that the NET4AI is not limited to privacy-preserving deep learning, but designed to be a generic architecture supporting any computing-oriented services and artificial intelligence (AI) applications in 6G. The NET4AI leverages pervasive edge computing capabilities in 6G and offers an end-to-end solution to network-based AI, from deployment to operations.

Keywords

privacy, AI, 6G, edge computing

1 Introduction

Artificial intelligence (AI) refers to intelligence as exhibited by machines. It perceives external data and takes appropriate actions to achieve goals. AI techniques have become an essential tool for solving challenging problems in business analytics and decision-making. It is anticipated that AI applications will reach every possible sector of the global economy and affect all aspects of society [1].

AI techniques can be broadly classified as rule-based AI and learning-based AI. In rule-based AI (for example, an expert system), human knowledge is encoded into rules that apply to input data for problem solving. Rule-based AI is limited by its knowledge base and cannot solve unknown problems. In contrast, learning-based AI aims to learn rules from historical data and uses these rules to achieve specific goals. Learning-based AI is at the center of the current resurgence of AI research and development, with machine learning approaches becoming mainstream.

At the core of a machine learning algorithm is a learning model that describes the relationship between input data and output rules. Typical learning models include artificial neural networks (ANNs), genetic algorithms, and regression analysis, to name just a few. This paper draws attention to ANN-based machine learning, in particular deep learning, which is able to extract features from training data and identify which are relevant to a target problem. It is suitable for correlated data and prevails in a variety of applications [2]. In the sequel, we use "learning model" and "AI model" interchangeably.

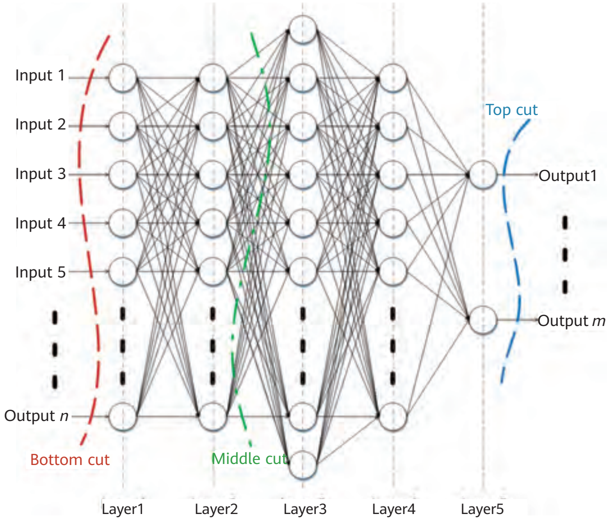


Figure 1 A DNN with three hidden layers (layers 2–4)

ANN in deep learning includes multiple hidden layers between the input and output layers, and is often referred to as deep neural network (DNN). Figure 1 shows a DNN with three hidden layers. Deep learning relies on frequent data access and intensive computation to train the DNN. To reduce training time, distributed systems have been exploited for parallelizing time-consuming computation and slow I/O access in deep learning [3].

In parallel to the proliferation of machine learning, mobile computing has entered an exciting new era, where personal devices such as smart phones and tablets are becoming the primary computing platform for many people and applications. These devices have access to an unparalleled amount of data that is often not only personal but also private in nature.

When deep learning meets mobile computing, a new paradigm of privacy-preserving deep learning with decentralized data is revealed [4]. As collecting and storing such sensitive data comes with associated privacy risks, as well as the responsibility to protect the privacy embedded in the data, a large amount of research efforts have recently been devoted to differential privacy [5] in deep learning, which aims to protect the exact training data of individual devices to the point that they are indistinguishable.

A learning model is used for inference after it has been trained. Model inference can be performed at different places, depending on how the model is distributed. If the learning model is distributed to the client, inference happens locally. Local inference may place a large computational workload on the client. If the model is held on a server, the client will need to upload inference data to that server, and this can cause information leakage. Recent research [6] suggests splitting the model between the client and the server, so that the client sends intermediate results rather than raw data to the server. This split inference can reduce communication overhead and latency, and protects data privacy as the server cannot derive information about the raw inference data from the intermediate results. This is similar to split learning (SL), which is described later in Section 1.1.

Model inference is technically similar to a forward propagation step in model training. It is triggered by an inference data item, instead of a training data item, and it returns a classification result from the output layer rather than triggers loss function evaluation and backpropagation. Due to this technical similarity, we will focus on model training here in this paper. However, our solution can be readily applied to model inference. Below we will briefly review existing work related to privacy-preserving deep learning.

1.1 Related Work

In terms of deep learning, differential privacy approaches include adding noise to training data without jeopardizing its statistical properties so that the trained model still captures features in the original dataset [7], and applying cryptographic techniques so that learning is based on encrypted data without decryption [8].

In this paper we draw attention to an alternative, where instead of sending raw training data, clients forward information that appears random. Federated learning (FL) [9] and SL [10] are two typical examples of this approach, and both train a deep learning model (such as a DNN) without requiring raw training data to leave the clients (for example, uploaded to a training server).

In FL [9], individual clients each train a local model using their own datasets only, and update the model parameters to a training server where a global model (specifically, global model parameters) is maintained. The training server aggregates updates received from the clients to adjust the global model, the parameters of which are then returned to the clients. Based on this information, the clients update the local model and continue the training. The procedure then repeats until the global model converges. FL can be viewed as a generalized implementation of stochastic gradient descent [11] with flexible batch size and participating clients.

In SL [10], the DNN is split into two disjoint components by a cut layer. The lower component includes the input layer and is run on the client side, while the remaining upper component runs on the server side. A cut normally occurs between two layers of the DNN (for example, between layers 2 and 3 in Figure 1), although in theory it can be freely defined as long as it produces two disjoint partitions of the DNN. Consequently, the two components can be viewed as two concatenated learning models — a client-side model and a server-side model — with the client-side model feeding its output to the server-side model as input. Clients (such as devices) interact with the training server sequentially to train the DNN using their local data, by iteratively sending intermediate results (such as the output of the client-side model) to the server and receiving the corresponding gradients from the server. When a client finishes the training with the server using its local data, it provides the latest model parameters to the next client, which continues the training using its own dataset. Training then proceeds sequentially among clients until all are finished. A new round of training may be initiated as needed.

FL combines simultaneously and individually trained local models to generate a global model. As the local models are based on pure local data that is usually non-IID (independent and identically distributed), FL converges slowly. SL essentially trains the global model directly using all local datasets and can therefore converge fast (in terms of duration, but not necessarily in terms of training rounds). However, it requires synchronization among clients due to its sequential learning nature. A comparative study of FL and SL can be found in [4]. As clients do not send raw training data to the training server, FL and SL both offer differential privacy.

Study [12] has shown that an insider adversary with complete knowledge of the learning model can construct information that is very similar to the training data by taking advantage of the gradual course of model convergence. In FL, this can lead to information leakage to malicious clients without violating differential privacy. SL, in contrast, does not suffer from this problem, as none of its clients have complete knowledge of the deep learning model. That being said, if the learning process involves only a small number of clients, severe information leakage is possible as information similarity is narrowed down to the local data of the small set of clients. Consequently, it is desirable to ensure a minimum number k of participating clients, where k is a system parameter. This provides further privacy protection and is known as k-anonymity [13].

A secure aggregation protocol is proposed in [14] to achieve k-anonymity in FL. The protocol is built on the concept of secret sharing and runs between devices and the server. As k-anonymity relies on the server's involvement, the proposal may not ease devices privacy concerns especially if the devices do not trust the server. This anxiety is largely due to fear of concentrated power: the server is not only the primary entity of the learning (it owns the learning model and oversees the entire learning process) but also knows which devices contributed to the learning. The work presented in this paper takes a systemic approach to address this anxiety and resolve the problem.

The work presented in this paper is also related to the shared machine learning system described in [15]. This system protects data security and privacy using specialized hardware that provides trusted execution environments. However, it is not suitable for large-scale public systems such as the telecommunication network.

1.2 Our Contribution

Privacy-preserving deep learning with decentralized data is tightly coupled with the telecommunication network when the training data sources are mobile terminal devices. It is envisioned that the 6G wireless system will go beyond connectivity provisioning to enable connected intelligence — in other words, distributed learning and inference. The 6G wireless system should therefore provide native support to this new AI computing paradigm.

In this paper, we first generalize SL [10] by extending cut layer definition so that it covers FL [9] and centralized learning (CL) as special cases, and combine it with FL to obtain a two-level learning framework. The framework inherits the advantages of FL and SL, but not their drawbacks. We propose customizing the learning approach at the bottom level of the two-level framework by selecting proper cuts for the AI model. The cut layer selection takes into account a number of factors, and the goal is to balance the learning overhead on devices, on servers, and on the network. Such an optimization can result in a mix of local learning (at client side), CL (at server side) and SL (at both sides) to appear at the bottom level concurrently.

We then propose a 6G wireless system architecture, referred to as NET4AI, which bears a service-oriented design. It supports the two-level learning framework and offers learning approach customization as a value-added service. With the NET4AI architecture, the system can provide end-to-end support to network-based AI applications, from deployment to operations. During the operation phase, the AI computing modules of AI applications and user devices (such as UEs) communicate anonymously to finish AI computing (for example, model training and model inference), as coordinated by the system. The system can further enforce k-anonymity for user privacy protection and apply proxy re-encryption to ensure data confidentiality.

The remainder of the paper is organized as follows: we describe important concepts and assumptions in Section 2; we present the two-level learning framework and the NET4AI architecture in Section 3; we identify a number of challenges associated with NET4AI in Section 4; and we offer our conclusions in Section 5.

2 Terminologies and Assumptions

In this section, we will describe our assumptions about the network infrastructure and introduce radio computing node (RCN), a radio access network (RAN) node equipped with edge computing capabilities. We will also introduce some AI computing related concepts, including job, task, and routine. The networking logic of an AI computing service is expressed using these concepts.

2.1 Pervasive Edge Clouds

Algorithm, data, and computing power are the main driving factors of AI innovation. Conventionally, AI computing power is provided by a centralized cloud platform, and data is pushed to the central cloud for processing. In the era of big data and AI, this introduces a number of issues relating to data privacy, latency, and efficiency, which further trigger a paradigm shift in the other direction, that is, bring computing to data.

Edge computing (EC) is an approach to computing that reduces transmission delay and bandwidth consumption by moving computation and storage close to the network edge, and therefore to the end user and data. As EC techniques are developing into maturity, it is anticipated that EC capabilities, in the form of edge clouds, will be pervasively deployed in the network, for example, collocated with RAN nodes or base stations. When a RAN node or base station is equipped with EC capabilities, it is referred to as an RCN in this paper.

As the wireless network infrastructure offers both computing resources (such as CPU cycles, memory, storage, and I/O access) and communication resources (including radio resources and transport resources), it becomes possible to jointly optimize utilization of the distributed and diversified infrastructure resources, and to enable ultimate, optimal end-to-end performance to end users. It is assumed that there are one or more entities in the network (for example, resource managers) performing infrastructure resource management.

2.2 Computing-related Concepts

An AI computing service can be an application-layer service. It can also be a network service that aims to optimize network management or operations.

The AI computing service includes one or more computing modules, called *service functions*. The AI computing service is associated with well-defined computing logic and delivers computational results according to input data. This computing logic includes algorithm implementations for each of these service functions and interactions between them. The latter can be specified in terms of job, task, and routine, as defined below.

The AI computing service includes one or more independent jobs, each of which is focused on a computational goal and exposed to the service consumer. An execution can be triggered at the job level, upon request, on some events, or under certain conditions. When a job is being executed, the service consumer can join in or contribute to the execution by providing input data and/or by receiving computational results, for example.

A job comprises one or more tasks that may have interdependencies. During execution of a job, its tasks are executed in accordance with their interdependencies. A task that depends on another must be executed after the other task has been executed. Each task is associated with a service function chain (SFC), which comprises one or more routines, each of which corresponds to two adjacent service functions (routine server function and routine client function) in the SFC. Within a routine, the routine client function can represent devices, implying that the computing logic of the function runs on the devices. However, the routine server function cannot represent devices.

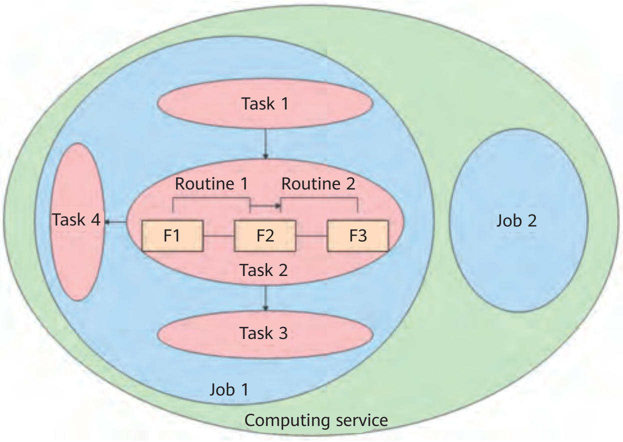


Figure 2 Illustration of a service, jobs, tasks, routines, and service functions (F)

During execution of a task, its routines are executed in sequence along the SFC. When a routine is being executed, its client and server functions are triggered to communicate with each other. Figure 2 illustrates the relationship between a service, jobs, tasks, routines, and service functions, where arrowed lines indicate dependency. Let us take an AI computing service or application as an example. Model training and model inference can be two separate jobs of the AI computing service. The (model) training job can include a model training task and its dependent model validation task. The model training task can be associated with a three-function SFC, as illustrated in Figure 3, and it includes two routines, which respectively correspond to the bottom- and top-level learning in the two-level learning framework introduced later in Section 3.1.

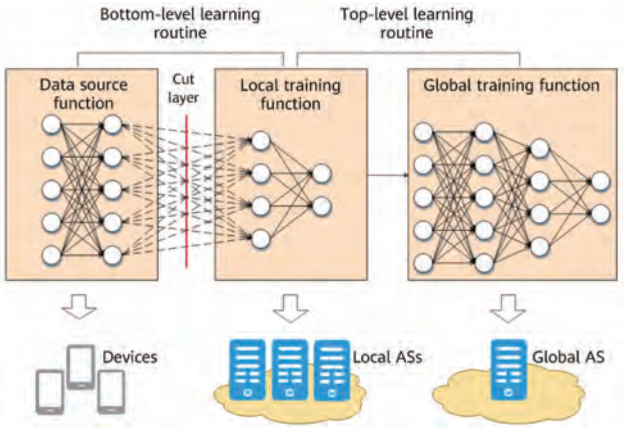


Figure 3 Two-level learning framework

A service function (whether a routine client function or a routine server function) is regarded as a concrete service function if it does not represent devices. Alternatively, a service function representing devices is considered to be an abstract service function. When an AI computing service is deployed in the network, concrete service functions from the AI computing service are instantiated at network locations, such as edge clouds. After a concrete service function is instantiated at a network location, an instance of the service function runs on an application server (AS) at the network location. Given a routine, we refer to an instance of the routine client function as a routine client, and an instance of the routine server function as a routine server. When the routine client function represents devices, these devices are considered routine clients.

3 Native Support of AI Applications

In this section, we will discuss how to offer native support for AI applications in 6G, particularly in regard to privacy-preserving deep learning. This includes a two-level learning framework, which allows learning approach customization, and a service-oriented system architecture, which provides an end-to-end solution to AI computing services. We will also discuss protocol design in RCNs — RAN nodes that provide radio-and-computing integrated resources — and show how the proposed solution works through a case study.

3.1 Learning Approach Customization

We generalize SL [10] by extending the definition of cut layer to the point where FL [9] and CL become two special cases of SL. In FL, each device has complete knowledge of the AI model and trains the model using its local dataset. FL can be viewed as SL applying a top cut, where the cut layer is above the output layer. On the other hand, CL requires devices to send raw training data to a server and learning happens purely on the server side. As such, CL can be viewed as SL with a bottom cut applied, where the cut layer is below the input layer. Traditional SL [9] corresponds to cases where the AI model is partitioned by a middle cut. Bottom cut, middle cut, and top cut are illustrated in Figure 1.

We combine FL and the generalized SL to obtain a generic two-level learning framework. As shown in Figure 3, the framework can be expressed by a task composed of two routines, a bottom-level learning routine and a top-level learning routine. The generalized SL may be applied at the bottom level (the bottom-level learning routine) of the framework, while FL at the top level (the top-level learning routine). The bottom-level learning routine runs between two service functions: a data source function and a local training function. The data source function represents devices, while the local training function can be instantiated at local ASs to train local AI models. These local ASs can be located on RCNs, for example. The top-level learning routine runs between the local training function and a global training function, the latter of which can be instantiated at a global AS. The global AS can be located at an edge cloud relatively far from the RAN so that it can efficiently serve multiple local ASs (RCNs). Within this two-level learning framework, learning approach customization is realized through cut layer selection at the bottom level.

When a middle cut is selected for the bottom-level learning (routine), the two-level learning framework offers the advantages of both FL and SL. As the bottom-level learning is based on the combined datasets of multiple devices, which are less non-IID than a single device's dataset, the trained local AI models at the local ASs are more accurate than those trained by individual devices using their own dataset in FL. Generally, improved local model accuracy leads to accelerated convergence of the global model. As devices do not have complete knowledge of the AI model, information in the training data is not leaked to adversary devices as described in [13].

When the top cut is selected for the bottom level, local AI models are trained at individual devices, and the local training function receives local AI model parameters and sends them to the global training function in an aggregate form. In this case, the framework reduces to FL and suffers from the information leakage problem. When the bottom-level learning applies the bottom cut, the framework does not offer differential privacy. The top cut and the bottom cut are not recommended as far as privacy is concerned. However, they may be used due to other factors as described below.

Devices associated with the same instance of the local training function for bottom-level learning should be assigned with the same cut layer, so that they and the local training function instance exhibit consistent behavior during the learning. Under this constraint, and considering model structure, device status (such as computing power and energy levels), server conditions (for example, AS loading), device locations, deployment locations of the local training function, and network conditions (for example, bandwidth or congestions), the bottom-level learning can apply a mixed cut to optimize device, server, and network performance all at the same time. When a mixed cut is applied, a different cut layer can be selected for different groups of devices, with each group associated with a different local training function instance, as illustrated in Figure 4.

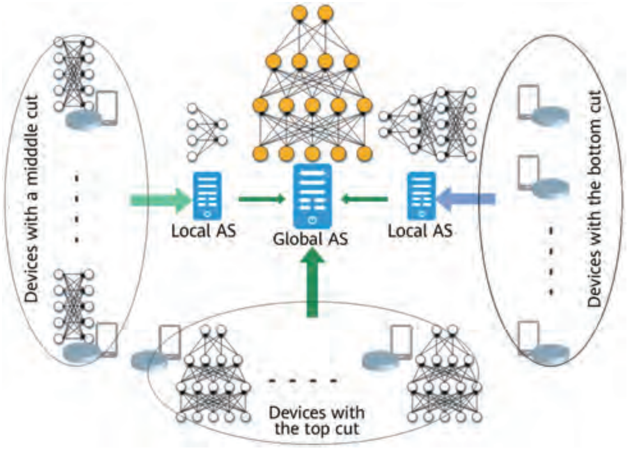


Figure 4 A mixture of different cuts applied to model training

3.2 NET4AI System Architecture

Adhering to the concepts of job, task, and routine, we propose a 6G wireless system architecture to support the two-level learning framework. This architecture is known as NET4AI. A 6G wireless system using this architecture is referred to as a NET4AI system, and such a system can offer NET4AI services to its customers, providing end-to-end support of AI computing services. The NET4AI architecture operates under the assumption that service functions can be located at or run on devices or edge clouds. Figure 5 illustrates the NET4AI architecture.

The NET4AI architecture includes a control plane (CP) and a compute plane (CmP). The control plane manages AI computing services and controls executions of jobs and tasks for those services, in addition to providing traditional device-related management functionalities. It includes a number of CP entities, such as a service manager, orchestrator, resource manager, access manager, job manager, and task manager. Depending on implementation, some of these control plane entities can be merged — for example, the job manager can be merged into the service manager. Furthermore, and also depending on implementation, the control plane can span the RAN and the core network segments of the system, or dedicate itself to just one of them. The compute plane controls executions of routines for an AI computing service and supports data communication between service functions. It comprises one or more routine managers as well as a forwarding sub-plane, which can be simply called a forwarding plane (FP). The forwarding plane can correspond to the user plane (UP) in 3GPP 5G system architecture [17], and includes one or more forwarding plane functions (FPFs) and RAN nodes (specifically, the user plane part of RAN nodes). Note that each FPF can be integrated with a RAN node.

The links between entities in Figure 5 indicate the interfaces they use to communicate with each other and can be defined or created at a per-service granularity. Special attention should be paid to the T2 interface. When the FPF is separate from the RAN node, as illustrated by scenario 1 in Figure 5, the T2 interface corresponds to the RAN node and maps to a radio bearer. When the FPF is integrated with the RAN node, as illustrated by scenario 2 in Figure 5, the RAN node implements the FPF's functionality. In this case, the interface T2 corresponds to the device and can be supported by a radio bearer. In either scenario, the radio bearer can be shared among multiple devices, such as a computing radio bearer (CRB) as described in Section 3.3. Note that in scenario 2, the T4 interface becomes integral to the RAN node in cases where the RAN node integrates with the edge cloud (such as when the RAN node is an RCN).

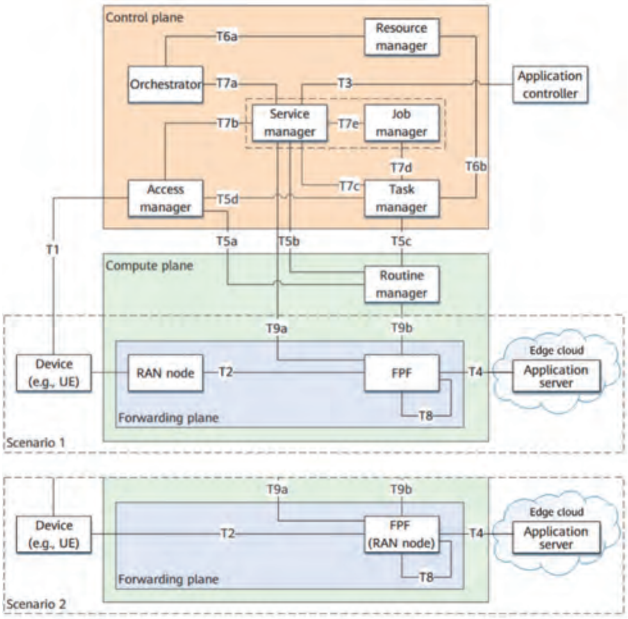


Figure 5 NET4AI architecture

An authorized application controller (AC) can register an AI computing service with the NET4AI system. The AC belongs to the service provider, and is responsible for managing the AI computing service. During registration, the computing service is instantiated in the system. Every concrete service function of the AI computing service is instantiated at one or more network locations (such as edge clouds). After registration, the NET4AI system supports the AI computing service's operations by coordinating executions of routines, tasks, and jobs of the AI computing service. We describe service instantiation and service operations in detail below.

3.2.1 Service Instantiation

The AC registers the AI computing service with the NET4AI system by sending information describing service functions, routines, tasks, and jobs of the AI computing service to a service manager, which interacts with the orchestrator to instantiate the AI computing service accordingly. During this process, the orchestrator determines the deployment of the AI computing service and selects an appropriate compute plane.

The deployment decision includes locations of instances of concrete service functions and the resources needed for each of the service function instances. These service function instances include routine servers and possibly routine clients of individual routines of the AI computing service. The deployment decision also includes logical links between routine clients and routine servers for each of the routines. In cases where there is a logical link between a routine client and a routine server, they can communicate with each other via the compute plane during execution of the routine.

The routine client is connected to an FPF in the compute plane via a T2 or T4 interface, while the routine server is also connected to an FPF in the compute plane, via a T4 interface. The two FPFs are either the same entity, or different entities that are interconnected via a T8 interface. The routine client and the routine server are assigned to the same routine manager in the compute plane. The routine manager manages execution of the routine with respect to the routine client and the routine server, by triggering data communication between them.

The orchestrator implements the deployment decision using the resource manager. The orchestrator informs the service manager about the compute plane selection result, and the service manager configures the compute plane accordingly.

3.2.2 Service Operations

The AI computing service comprises a job that includes a task for training an AI model using the two-level learning framework described in Section 3.1. We refer to this task as a model training task, and it involves a bottom-level learning routine and a top-level learning routine, as illustrated in Figure 3. The service functions involved in the model training task include a data source function, local training function, and global training function. The data source function is the routine client function of the bottom-level learning routine. It represents devices and acts as a source of training data. The local training function is the routine server function of bottom-level learning routine. We will now elaborate on how the NET4AI system supports service operations related to the job.

- Control plane behavior

The AC requests execution of the job by sending a job order to the service manager, which then informs a selected job manager to execute the job. As described earlier, the service manager and the job manager can be combined in some implementations.

When executing the job, the job manager selects a task manager for each of the tasks (including the model training task) within the job, and triggers the task manager(s) to execute the tasks in accordance with their interdependencies. When executing a task, a task manager identifies related routine manager(s) in the compute plane, which correspond to the routines within the task. The task manager requests the routine manager(s) to execute the routines in accordance with their interdependencies.

If the routine client function (for example, the data source function) of a routine represents devices, the task manager selects these devices as routine clients and assigns each of them to a routine manager. The devices are selected among those that are allowed to access the AI computing service, have registered, and have given their consensus (on contributing to the job). A device can register to the NET4AI system and provide its consensus via the access manager, which may have functionalities similar to those of the access and mobility management function (AMF) in the 3GPP 5G system architecture [17].

When executing the model training task, the corresponding task manager performs cut layer selection for the bottom-level learning routine, as described in Section 3.1, in order to customize the learning approach for the AI model. The cut layer selection can be performed jointly with routine client selection described above. The task manager informs the routine manager about the cut layer selection result.

- Compute plane behavior

The NET4AI compute plane is deeply involving in the execution of each routine of the AI computing service. It is responsible for coordinating the participation of routine clients and routine servers in the routine execution, enforcing *k*-anonymity, and enabling anonymous data communication between the routine clients and the routine servers, as described below.

In the compute plane, a routine manager is assigned with routine clients and routine servers for the corresponding routine, and each routine server is associated with one or more routine clients. The server-client association is determined by the orchestrator during service instantiation if the routine client function is a concrete service function (as described in Section 3.2.1), and dynamically by the task manager and/or the routine manager in all other cases. The server-client association is not known at the application layer, which includes the routine clients, routine server, and AC.

When executing the routine, the routine manager triggers or invites the routine clients associated with a routine server into a data communication with the routine server. This communication is mutually anonymous as the communicating parties do not know about each other. The routine manager first invites the routine server via the forwarding plane. During this step, the routine manager can provide cut layer information to the routine server, if applicable. It then invites the routine clients, and these invitations can be sent via the access manager (for example, when the routine client is a device) or via the forwarding plane.

During data communication, data traffic is routed between a routine client and the routine server by the forwarding plane. Either the routine client or the routine server can notify the routine manager when it finishes communication, so that the routine manager can proceed to the next step, for example, inviting the next routine client to communicate, or notifying the task manager that the routine execution has completed.

Figure 6 illustrates a routine execution procedure in accordance with the above description. In step 6, the routine server can request to restart the data communication or submit a notification regarding its completion. In the case of a restart request, the routine manager repeats steps 3–6 in step 7. If all routine servers have sent a completion notification, the routine manager notifies the task manager in step 8 that the execution of the routine has completed.

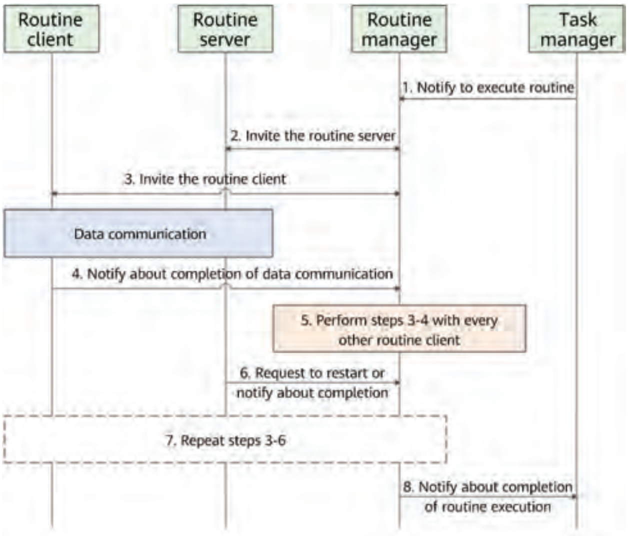

Routine client
Routine server
Routine manager
Task manager
1. Notify to execute routine
2. Invite the routine server
3. Invite the routine client
Data communication
4. Notify about completion of data communication
5. Perform steps 3-4 with every other routine client
6. Request to restart or notify about completion
7. Repeat steps 3-6
8. Notify about completion of routine execution

Figure 6 Routine execution procedure, with respect to a routine server

Note that step 5 can be performed in parallel to steps 3 and 4, unless the routine requires sequential communication. For example, the bottom-level learning routine may require sequential communication if a middle cut is selected for it. In this case, when the routine server is associated with multiple routine clients, the routine manager invites one such client to communicate with the routine server at a time. When inviting a routine client, the routine manager can provide it with cut layer information.

During data communication, there may be a strong requirement for user privacy in cases where the routine client is a device. To address such privacy concerns, *k*-anonymity can be enforced in the compute plane, where the value of *k* is a system parameter or a requirement from the AC. Assume that the routine manager is assigned with $m - 1$ other routine servers during service instantiation. The task manager assigns at least $m*k$ routine clients to the routine manager, which then associates at least *k* routine clients with the routine server.

There may also be a data confidentiality requirement for data communication. As communicating parties do not know about each other, end-to-end encryption is not applicable in this setting. Instead, data confidentiality can be achieved through proxy re-encryption in the forwarding plane (i.e., FPFs). That is, data originating from a sender is in ciphertext and is forwarded by the forwarding plane without decryption. Before the data is forwarded, it is re-encrypted using a re-encryption key, enabling the receiver to recover the original data (plaintext) using their own private key. The re-encryption key can be obtained by the routine manager in step 2 or 3 and configured into the forwarding plane before data communication begins.

3.3 Compute Plane Protocols in RCNs

The radio interface at a RAN node includes three protocol layers: layer 1 — physical layer, layer 2 — data link layer, and layer 3 — network layer. The NET4AI architecture focuses on protocols at layers 2 and 3.

A radio bearer is a layer 2 logical channel. It bridges layer 1 and layer 3 to support the transfer of user or control data. Legacy radio bearer management assumes that the RAN node is a network access point, and that computing happens on the other side of the network. It may not be efficient or suitable for RCNs, where communication and computing are integrated to allow computing within the network. Consequently, the NET4AI compute plane introduces a CRB at layer 2 to enable RCNs to distinguish between data at the computing and user planes for self-loop computing services within RCNs. A CRB connects a UE served by an RCN and one or more service functions deployed on the RCN; in this sense, it provides the T2 interface functionality shown in Figure 5. A deep edge protocol (DEP) is a new simplified protocol, which is proposed at layer 3 to enable efficient exchange of data between the UE and service functions. The core reason for introducing DEP is to enable service functions to be deployed in a wireless network like the RAN. As a result, data transmission protocols can be simplified to a greater extent than through the use of traditional cloud deployments.

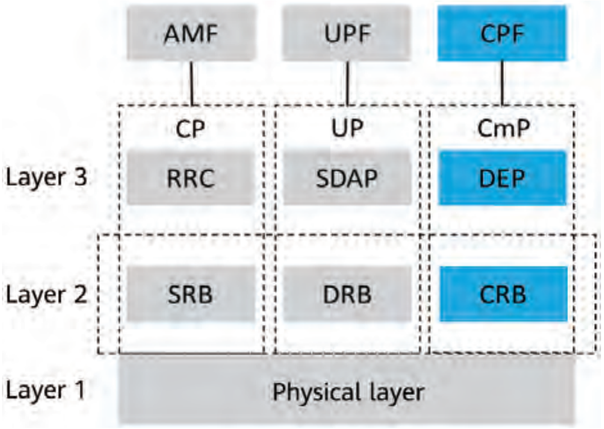


Figure 7 Illustration of CRB and DEP

Figure 7 illustrates CRB and DEP, along with 5G layer 2 radio bearers and layer 3 protocols. In the figure, the AMF and the user plane function (UPF) are 5G network functions, which respectively provide some of access manager and FPF functionalities in the NET4AI architecture. The compute plane function (CPF) is a module within the RCN that implements the FPF's functionality. These radio bearers and protocols can all be present on an RCN to support different scenarios or needs. For instance, data radio bearers (DRBs) and Service Data Adaptation Protocol (SDAP) can be used to support UE connection to a service function that is not deployed on the RCN.

Assume that a UE is being served by an RCN. When a UE accesses a service function deployed on the RCN for computing, the RCN allocates a CRB to the UE, and the CRB connects the UE to the service function. An example basic procedure is as follows: The UE sends a computing request to the control plane of the NET4AI system via the RCN. When the NET4AI control plane notifies the UE that the request has been accepted, the NET4AI control plane also notifies the RCN to establish a DEP session for the UE. The RCN (the control plane part) will then allocate a DEP session to the UE accordingly. The DEP session maps to a CRB, which can be either newly created or an existing one. The RCN then transmits the CRB parameters to the UE over radio resource control (RRC) signaling, enabling the UE and service function to exchange data.

As illustrated in Figure 8, a DEP session may map to one or more CRBs, and a CRB can support one or more DEP sessions. Every CRB is configured with specific QoS capabilities. When a DEP session is mapped to multiple CRBs, these CRBs are configured with different QoS

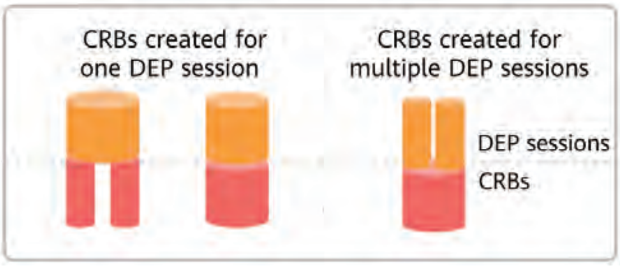

CRBs created for one DEP session
CRBs created for multiple DEP sessions
DEP sessions
CRBs

Figure 8 Mapping between CRBs and DEP sessions

capabilities. Consequently, the DEP session supports multiple QoS flows via the corresponding CRBs. If a DEP session is mapped to a single CRB, the DEP session supports only one QoS flow.

3.4 Case Study

A healthcare institute is building a blood pressure model. The model is a DNN-based AI model designed to capture blood pressure ranges during various times of the day for different age groups in connection to certain geographic regions. The healthcare institute wants to bring the model online for a large population of users to contribute to model training. The healthcare institute registers an AI computing service with the NET4AI system to achieve this goal.

When registering the AI computing service, the healthcare institute provides information about the AI model, including the number of layers, number of neurons per layer, and number of links between every two adjacent layers. The AI computing service includes two concrete service functions: a local training function and a global training function. The local training function supports CL and SL for the AI model, while the global training function implements FL model aggregation logic.

The AI computing service includes a model buildup job, which in turn includes a model training task. The model training task includes a bottom-level learning routine and a top-level learning routine, as shown in Figure 3. The bottom-level learning routine is associated with the data source function, which represents devices, and the local training function. Meanwhile, the top-level learning routine is associated with the local training function and the global training function. The model buildup job can further include a model verification task intended to be executed after the model training task. For simplicity, we ignore the model verification task in this case study.

According to information about the AI computing service received from the healthcare institute, the NET4AI system deploys the AI computing service in the network (for example, at edge clouds). The deployment includes an instance of the global training function and multiple instances of the local training function. Each of these service function instances is attached to the NET4AI system via an attachment point, which is either an FPF or an RCN (when the service function instance is deployed on the RCN).

After the AI computing service is deployed, the healthcare institute requests the NET4AI system to execute the model buildup job. According to the request, the NET4AI system executes the model buildup job by notifying the instances of the local training function to execute the bottom-level learning routine. As this routine requires devices to participate, execution does not start until such participation begins.

Once a device connects to the network, the NET4AI system informs the device about the model buildup job of the AI computing service. The device may then volunteer to contribute to the model buildup job and provide its consensus to the NET4AI system. At the same time, the device can provide information about its status (for example, computing power and energy levels) and privacy requirements, the latter of which indicates if the device is willing to provide raw data for the model buildup job.

Based on the information received from devices as well as other information (such as network conditions), the NET4AI system divides consenting devices into multiple groups. The NET4AI system selects a cut layer for each of the groups for the bottom-level learning routine, and associates each group of devices with a service function instance, which is either an instance of the local training function or the instance of the global training function. For example, as illustrated in Figure 4, where the three ellipses represent three groups of devices, the NET4AI system selects the bottom cut for a group, a middle cut for another, and the top cut for the remainder; it associates the first two groups respectively with two instances of the local training function, and the third group with the instance of the global training function.

The NET4AI system informs the devices within each of the groups about the cut layer selection result, and connects the devices to the attachment point of the corresponding service function instance. Devices that are assigned with a top cut or a middle cut further obtain the local training function from the NET4AI system (configured with the cut layer) and run it locally.

When sufficient devices are available (at least k devices within a group), the NET4AI system invites these devices into the execution of the bottom-level routine. These devices can then send data to the corresponding instance of the local training function via the NET4AI system. During data communication, the devices and the local training function instance do not know about each other. The data sent from the devices includes blood pressure readings and user age information in cases where the devices are assigned with the bottom cut, corresponds to intermediate results as described in SL in cases where the devices are assigned with a middle cut, or comprises local model parameters as in FL in cases where the devices are assigned with the top cut.

When the instances of the local training function that are associated with devices finish computing (when local AI models are established), the NET4AI system knows that the execution of bottom-level learning routine has finished and notifies the local training function instances and global training function instance to execute the top-level learning routine. The model buildup job can be executed this way in rounds, until the global training function instance notifies the NET4AI system to stop (for example, when the global model converges).

The global training function instance can provide the model parameters to the healthcare institute via the NET4AI system. The healthcare institute then considers the blood pressure model to be successfully trained.

4 Technical Challenges

In order to support a vision of inclusive intelligence, 6G networks need to consider native AI design instead of overlay design at an architectural level, which introduces new technical challenges to 6G networks beyond traditional connectivity issues, involving issues such as data privacy, heterogeneous resources, and energy saving. Of particular concern is the complexity of the wireless edge environment, which occurs due to unstable wireless connections, large-scale distribution, and heterogeneity of edge resources. The following issues require further study:

- Energy efficiency

NET4AI may need to support large-scale distributed training within 6G networks. Data communication will be increased significantly for model and parameter synchronization, which, when combined with rising computing costs, results in severe energy consumption challenges.

Considering the training process, there are generally two ways to reduce communication overheads. The first is to reduce the amount of exchanged data per round, and includes model compression methods such as quantization, sparsification, and knowledge distillation [18–21]. The second is to reduce the number of communication rounds. For example, FedAvg performs multiple rounds of local updates before aggregation [9]. Further investigation is required in order to reach a compromise between communication overhead and AI training performance.

Topological structure design for communication networks is another effective method for improving communication efficiency [22]. For example, the Ring Allreduce solution in high-performance computing (HPC) may reduce communication bandwidth. However, the topology of wireless networks is not as flexible as IoT servers, and actually applying such mature HPC technologies to wireless networks still requires a lot of research.

- Adapting to a dynamic environment

Distributed learning systems may consider certain fault tolerant mechanisms, such as the classical Byzantine [23]. However, this becomes more challenging with wireless networks. One of the reasons for this is because resources can dynamically change in wireless network environments that include connections. For example, an AI training task on a base station may be blocked by extreme burst traffic, or the sudden drop-out of terminal devices due to unstable wireless connections. As a result, to ignore failed nodes [24–25] or perform redundant backups [26], further research may be needed for natively adapting to wireless dynamic environments.

- Heterogeneous resource scheduling

To achieve NET4AI, we need to manage and schedule large-scale heterogeneous resources within a 6G network. However, designing an efficient distributed scheduling framework and algorithm is a challenging prospect.

To properly and efficiently deploy AI tasks to a wireless network, we must first perform general modeling for various heterogeneous resources, including computing power, memory, storage, and communication bandwidth. After that, the state and action spaces of resource scheduling can be defined, and then the distributed scheduling framework and algorithm can finally be designed. The scheduling algorithm needs to consider how to efficiently allocate tasks among large-scale heterogeneous resources, which is an NP problem. The computational complexity increases greatly with the scaling.

- Data service

Huge amounts of data (such as sensing data) may be generated, processed, and consumed in 6G networks to drive network intelligence and data sharing, and to improve network operation efficiency. However, this introduces challenges for 6G data services, including how to fully exploit data value while also ensuring data security and natively complying with data regulations such as the General Data Protection Regulation (GDPR) in the European Union (EU) and the European Economic Area (EEA) from architectural perspectives.

NET4AI needs to ensure that users have full control of their personal data and can decide whether to share, monetize, or offer the data for training. Certain standalone data protection solutions, such as *k*-anonymity, *l*-diversity, *t*-closeness [31], and differential privacy, may not be enough. Constructing a complete architecture-level data service framework with a transparent multi-party mechanism is a key challenge for 6G.

5 Conclusion

In this paper, we proposed NET4AI — a 6G system architecture intended to support future computing services, AI computing services in particular. The NET4AI offers end-to-end support to AI computing services, from the deployment phase to the operation phase. It addresses the emerging problem of privacy-aware deep learning and allows for learning approach customization. It can enforce *k*-anonymity and ensure data confidentiality in the compute plane, offering strong privacy protection. We discussed protocol design at layers 2 and 3 of RCNs, i.e., RAN nodes equipped with edge computing capabilities, and we introduced CRB and DEP to support efficient combined communication and computation on RCNs. We also identified a number of technical challenges associated with NET4AI. However, development of NET4AI architecture is still in its initial stages. Details of actual implementations (for example, mobility and connection management), with respect to the execution of routines, tasks, jobs, and other system procedures, are yet to be developed.

References

[1] R. Sagramsingh, "AI: A global survey," IEEE-USA, 2019.

[2] S. Dargan, M. Kumar, M.R. Ayyagari, and G. Kumar, "A survey of deep learning and its applications: A new paradigm," Archives of Computational Methods in Engineering, 2019.

[3] J. Verbraeken, M. Wolting, J. Katzy, J. Kloppenburg, T. Verbelen, and J.S. Rellermeyer, "A survey on distributed machine learning," arXiv:1912.09789, 2019.

[4] P. Vepakomma, T. Swedish, R. Raskar, O. Gupta, and A. Dubey, "No peek: A survey of private distributed deep learning," CoRR abs/1812.03288, 2018.

[5] C. Dwork, "Differential privacy," Proc. ICALP, pp. 1-12, 2006.

[6] Y. Kang, J. Hauswald, C. Gao, A. Rovinski, T. Mudge, and J. Mars L. Tang, "Neurosurgeon: Collaborative intelligence between the cloud and mobile edge," Proc. ASPLOS, 2017.

[7] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," Proc. TCC, pp. 265 - 284, 2006.

[8] M. Minelli, "Fully homomorphic encryption for machine learning," PhD thesis, PSL Research University, 2018.

[9] H.B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. Aguera y Arcas, "Communication-efficient learning of deep networks from decentralized data," Proc. AISTATS, 2017.

[10] O. Gupta and R. Raskar, "Distributed learning of deep neural network over multiple agents," Journal of Network and Computer Applications, vol. 116, pp. 1 - 8, 2018.

[11] L. Bottou, F.E. Curtis, and J. Nocedal, "Optimization methods for large-scale machine learning," SIAM Review, vol. 60, no. 2, pp. 223 - 311, 2018.

[12] B. Hitaj, G. Ateniese, and F. Perez-Cruz, "Deep

models under the GAN: Information leakage from collaborative deep learning," Proc. ACM CCS, pp. 603 - 618, 2017.

[13] V. Ayala-Rivera, P. McDonagh, T. Cerqueus, and L. Murphy, "A systematic comparison and evaluation of k-anonymization algorithms for practitioners," Trans. Data Privacy, vol 7, no. 3, pp. 337 - 370, 2014.

[14] K. Bonawitz, V. Ivanov, B. Kreuter, A. Marcedone, H. Brendan McMahan, S. Patel, D. Ramage, A. Segal, and K. Seth, "Practical secure aggregation for privacy-preserving machine learning," Proc. ACM CCS, pp. 1175 - 1191, 2017.

[15] Technical framework for a shared machine learning system, Recommendation ITU-T F.748.13.

[16] Z. Zheng, S. Xie, H.-N. Dai, X. Chen, and H. Wang, "Blockchain challenges and opportunities: A survey," Int. J. Web and Grid Services, vol. 14, no. 4, pp.352-375, 2018.

[17] System architecture for the 5G System (5GS), 3GPP TS 23.501.

[18] Jakub Konečný, H Brendan McMahan, Felix X Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon, "Federated learning: Strategies for improving communication efficiency," arXiv preprint arXiv:1610.05492, 2016b

[19] L. Liu, J. Zhang, S. Song, and K. B. Letaief, "Hierarchical quantized federated learning: Convergence analysis and system design," arXiv preprint arXiv:2103.14272, 2021.

[20] F. Haddadpour, M. M. Kamani, A. Mokhtari, and M. Mahdavi, "Federated learning with compression: Unified analysis and sharp guarantees," in International Conference on Artificial Intelligence and Statistics, pp. 2350-2358, PMLR, 2021.

[21] Cheng Y, Wang D, Zhou P, *et al.*, "A survey of model compression and acceleration for deep neural networks," arXiv preprint arXiv:1710.09282, 2017.

[22] G. Neglia, C. Xu, D. Towsley, and G. Calbi, "Decentralized gradient methods: Does topology matter?" in Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics(S. Chiappa and R. Calandra, eds.), vol. 108 of Proceedings of Machine Learning Research, pp. 2348-2358, PMLR, 26-28 Aug 2020.

[23] D. Yin, Y. Chen, K. Ramchandran, and P. Bartlett, "Byzantine-robust distributed learning: Towards optimal statistical rates," in Proc. Int. Conf. Machine Learning, 2018, pp. 5650-5659.

[24] K. Bonawitz, H. Eichner, W. Grieskamp, D. Huba, A. Ingerman, V. Ivanov, C. Kiddon, J. Konecny, *et al.*, "Towards federated learning at scale: System design," in Proc. Conf. Machine Learning and Systems, 2019

[25] T. Li, A. K. Sahu, M. Sanjabi, M. Zaheer, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," in Proc. Conf. Machine Learning and Systems, 2020.

[26] Z. Charles and D. Papailiopoulos, "Gradient coding using the stochastic block model," in Proc. Int. Symp. Information Theory, 2018, pp. 1998-2002. doi:10.1109/ISIT.2018.8437887

[27] Liu Y and Passino K M, "Swarm intelligence: Literature overview[J]," Department of electrical engineering, the Ohio State University, 2000.

[28] Xing Xu, Rongpeng Li, Zhifeng Zhao, and Honggang Zhang, "Stigmergic Independent Reinforcement Learning for Multi-Agent Collaboration," IEEE Trans. Neural Networks & Learning Systems (TNNLS), February 2021.

[29] W. Niu *et al.*, "GRIM: A general, real-time deep learning inference framework for mobile devices based on fine-grained structured weight sparsity," in IEEE Transactions on Pattern Analysis and Machine Intelligence, doi: 10.1109/TPAMI.2021.3089687.

[30] Y. E. Sagduyu, Y. Shi, A. Fanous, and J. H. Li, "Wireless network inference and optimization: Algorithm design and implementation," in IEEE Transactions on Mobile Computing, vol. 16, no. 1, pp. 257-267, 1 Jan. 2017, doi: 10.1109/TMC.2016.2538233.

[31] Rajendran, Keerthana, Manoj Jayabalan, and Muhammad Ehsan Rana, "A study on k-anonymity, l-diversity, and t-closeness techniques," IJCSNS 17.12 (2017): 172.

Very-Low-Earth-Orbit Satellite Networks for 6G

Hejia Luo [1], Xueliang Shi [1], Ying Chen [1], Xian Meng [1], Feiran Zhao [1], Michael Mayer [2], Peter Ashwood Smith [2], Bill McCormick [2], Arashmid Akhavain [2], Daqing Liu [1], Huailin Wen [1], Yu Wang [1], Xiaolu Wang [1], Ruonan Yang [1], Rong Li [1], Bin Wang [1], Jun Wang [1], Wen Tong [2]

[1] Wireless Technology Lab

[2] Ottawa Wireless Advanced System Competency Centre

Abstract

With the breakthrough of advanced satellite launching and manufacturing technologies in recent years, both the academia and industrial communities are making considerable efforts to study mega constellations for Very-Low-Earth-Orbit (VLEO) satellites. The non-terrestrial network (NTN) is widely believed to be a part of the 6G network. In this paper, the vision for the evolution of VLEO satellites-based NTN towards 6G is proposed, as well as the technical challenges and potential solutions.

Keywords

VLEO, mega constellation, 6G, NTN

1 Introduction

The idea of Very-Low-Earth-Orbit (VLEO), which is at an altitude of around 350 km, has the potential to change the paradigm for the Internet because it is much lower than the traditional low Earth orbit (LEO) of 600 km to 1200 km or geostationary Earth orbit (GEO) of 35768 km. Accordingly, communications based on mega VLEO constellations are envisioned owing to attractive features such as low transmission delay, smaller propagation loss, high area capacity, and lower manufacturing and launching cost when compared with traditional LEO or GEO satellites. All these features will contribute to wider global utilization.

Satellite-based communications are believed to be an important part of 5G-Advanced and 6G by the global communication ecosystem. The 3rd Generation Partnership Project (3GPP) [1] has officially started researching on integrating satellite communications with 5G New Radio (NR) techniques titled "non-terrestrial network (NTN)." The study item (SI) of NTN (Release 14 to Release 16) identifies NTN scenarios, architectures, basic NTN issues and related solutions, and 12 potential use cases by considering the integration of satellite access in the 5G network including roaming, broadcast/multicast, and Internet of Things (IoT) [2-4]. In Release 17, the first work item (WI) of New Radio Non-terrestrial Network (NR-NTN) and Internet of Things Non-terrestrial Network (IoT-NTN) were approved at the end of 2019. NR basic features will be supported by both regenerative and transparent satellite systems in Release 17 to Release 19. 6G NTN will begin from Release 20 and more enhancements and new features will be discussed, including but not limited to support for integrating terrestrial networks (TN) and NTN and improved spectral efficiency compared to 5G and 5G-Advanced NTN. NTN with ultra-dense VLEO constellation will be a part of the 6G network and play an essential role in ensuring extremely flexible communication access services.

To achieve successful commercialization of VLEO-based NTN, new usage scenarios and applications need to be explored and a few technical challenges need to be addressed. A comprehensive discussion of the vision and challenges of VLEO-based NTN for 6G will be presented in this paper. The remaining parts of this paper are organized as follows. Section II introduces the driving factors and motivations of VLEO-based NTN networks according to the latest progress from the industrial community. Section III summarizes the usage scenarios and applications that mostly have gained the consensus of both the academia and the industry. Section IV identifies the challenges, and the potential solutions that might require a long-term effort for developing a competitive VLEO-based NTN. Section V draws the conclusion.

2 Driving Factors and Motivations

2.1 Requirements

Non-terrestrial communications such as satellite communications will be leveraged to build an inclusive world and enable new applications in a cost-effective way. The wireless coverage is expected to expand coverage from 2D "population coverage" on the ground surface to the 3D "global and space coverage." Integrating non-terrestrial and terrestrial communications systems will achieve 3D coverage of the Earth. They will not only provide communications with broadband and wide-range IoT services around the world, but also provide new functions such as precision-enhanced positioning and navigation and real-time earth observation.

With the development of new High-Throughput Satellite (HTS) as well as Non-Geostationary-Satellite Orbit (NGSO) systems such as the Medium-Earth-Orbit (MEO) system O3b and many proposed LEO and VLEO systems, such as Oneweb [5], Starlink [6], and TeleSat [7], it is expected that the cost will become much lower, the access capabilities will increase, and time delay of satellite connections will be reduced by VLEO constellations. SpaceX's Starlink project has launched over 1900 satellites by the end of December 2021, which made this company the world's largest satellite communications operator [8]. The decreasing cost of the satellite manufacturing and launching service is making the advent of huge fleets of small satellites in low earth orbits a reality. In addition to bridging the "digital divide", the role of satellite communications in 2030 and beyond is perceived to be pivotal in ensuring data connectivity to both fixed and mobile users.

2.2 Benefits of Integrating TN and NTN

Compared with the cellular network, the satellite communications service still calls for dedicated and expensive user terminals, which are out of reach of common users. A fundamental integration of TN and NTN will change the status quo and significantly improve user experience. With this integration, the satellite communications industry can fully utilize the fast development and economies of scale of the cellular industry, thus reducing the cost of terminals and the service price to more attractive levels. By achieving unified design of TN and NTN, the barrier among different satellite systems will also be eliminated, allowing users to freely roam among terrestrial networks and non-terrestrial networks of different operators.

3 Usage Scenarios and Applications

The VLEO-based NTN is expected to provide couples of usage scenarios and applications, as shown in Figure 1.

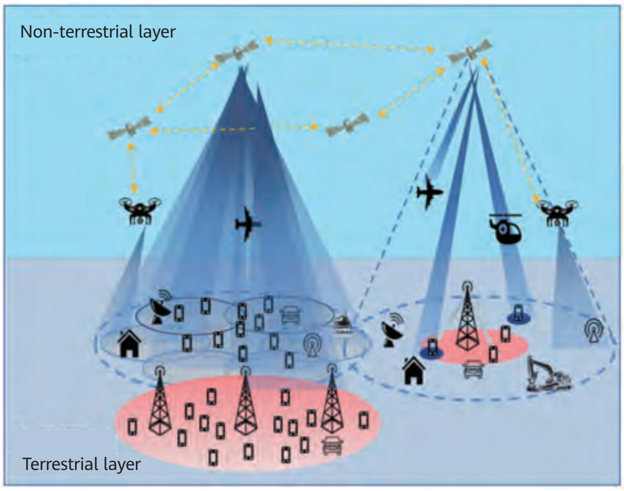


Figure 1 Usage scenarios and applications

3.1 Extreme Coverage

Today, almost half of the world's population lives in rural and remote areas that do not have basic Internet services. Non-terrestrial networks can provide affordable and reliable connectivity and broadband services for areas where telecommunications operators cannot afford to build terrestrial networks. By using non-terrestrial network nodes, such as satellites, unmanned aerial vehicles, and high-altitude platforms, non-terrestrial networks can be flexibly deployed, connecting people through various devices such as smartphones, laptops, fixed-line phones, and televisions.

3.2 Mobile Broadband for the Unconnected

Current commercial satellite communications systems have low transmission rates and high costs. In addition, satellite mobile phones are not integrated with the terminal equipment of the traditional terrestrial cellular network, and people will have to use two different mobile phones to access the satellite network and the cellular network respectively. In the future, we believe satellites can directly connect to mobile phones, providing broadband connectivity, with data rates similar to those of cellular networks in remote areas. For example, the user data rate should be able to reach 5 Mbit/s for download and 500 kbit/s for upload.

3.3 Broadband Connection on the Move

People should be able to access the Internet anytime, anywhere, no matter what kind of transportation they take. Take the air traffic scenario for example. In 2019, over 4 billion people traveled by aircrafts, which means almost 12 million people fly somewhere every day [9]. Most of them have no Internet connection during the flight or experience Internet access at very low speed. Future communications systems should provide MBB experience connections for all aircraft passengers.

3.4 Wide-Range IoT Services Extended to Unconnected Locations

Currently, IoT communications are implemented based on cellular network coverage. However, cellular-based IoT communications cannot guarantee connection continuity in many scenarios. In the future, IoT devices should be able to connect and report information anytime, anywhere. As a result, it will become easier to use NTN to collect information, such as the status of Antarctic penguins, the living conditions of polar bears, and animal and crop monitoring, from remote and uninhabited areas.

3.5 High-Precision Positioning and Navigation

In the future, most cars will have the capability to connect to the terrestrial network. However, the terrestrial network may not be able to provide high-quality vehicle-to-everything (V2X) services for users in remote areas. The integrated network can implement high-precision positioning and navigation and improve the positioning accuracy from meters to centimeters. On this basis, automatic driving navigation, precision agriculture navigation, mechanical construction navigation, and high-precision user positioning services can be provided.

3.6 Real-Time Earth Observation and Protection

With the development of remote sensing technology and the fast deployment of mega constellations, the future remote sensing technology will ultimately be in real time and feature high resolution. With these two significant features, earth observation can be introduced to more scenarios, such as real-time traffic dispatch, real-time remote sensing maps available to individual users, high-precision navigation combined with high-resolution remote sensing and positioning, and quick response to disasters.

4 Challenges and Solutions

To realize the usage scenarios and applications listed earlier, critical challenges and possible solutions are identified in this section, as enablers to a fully integrated network with TN and NTN for the 6G era.

4.1 Integrated Network Architectures

To provide unified services with a single device, new integrated network architectures composed of both TN and NTN need to be proposed. However, several challenges need to be overcome to realize a truly integrated network.

- The integrated network is in general a 3D heterogeneous network, with each layer having different coverage ranges and link quality. It is critical to coordinate each layer in the network to achieve unified network access. In addition, user equipment (UE) should have the flexibility to communicate with the most appropriate layer based on its own capability and context.
- A wide range of services with different quality of service (QoS) requirements will be supported by the integrated 6G network. However, the resources in the integrated network must exhibit a high level of heterogeneity. The resource availability for multiple services will vary over time, as each segment may dynamically allocate resources with high priority to support legacy services [10].
- The global span of an integrated network calls for reliable control anywhere, anytime. This typically requires a large number of ground stations to be deployed all over the world, which induces high complexity and increases expenditure. The end-to-end delay can be quite large since core network functions are implemented at very few sites on the ground and inter-plane inter-satellite link (ISL) communications are rather limited due to visibility and velocity constraints.

The following are potential technical solutions to overcome the preceding challenges.

- 3D UE-centric cell-free communication [11] is a promising solution for the integrated network. With UE-centric methodology, the cell boundaries can be eliminated efficiently. This leads to interference-free and reduced handover communications in scenarios with many heterogeneous access points. On the other hand, spatial multiplexing for cell-free communications means that the beams from multiple nodes, e.g., satellites and terrestrial base stations, can be resolved by using different phase gradients on the receiving array. It is thus possible to serve any location on the ground from multiple distinct sites and directions by fully exploiting the space-air-ground dimensions of the entire network.
- Network slicing enables multiple logical networks to run as independent tasks on a common shared physical infrastructure. Each network slice represents an independent virtualized end-to-end network and allows operators to perform multiple functions based on different architectures. As a consequence, a set of customized services with distinct QoS levels can be

provisioned by deploying multiple isolated and dedicated network slices on top of the integrated network.

- A hierarchical control framework with very few ground stations and GEO satellites is used to achieve global network control, while MEO satellites and LEO/VLEO satellites with ISL capabilities are used for regional and local control. The concept of a space-based core network can be exploited to facilitate global control and reduce propagation delay. For example, some core network functionalities, e.g., user plane function (UPF) and access and mobility management function (AMF), can be placed onboard satellites, so that both control messages and UE traffic are not required to traverse to ground stations with many hops.

4.2 Air Interface Technologies

LEO/VLEO constellation will be an important component of 6G networks. The capacity density at each location on Earth can be used to understand the service capability of a constellation. Take Starlink "Gen2" constellation (including about 30000 satellites) [12] for instance. The peak average capacity density after full deployment is in the middle-latitude area, which is about 3.6 Mbit/s/km^2, as shown in Figure 2.

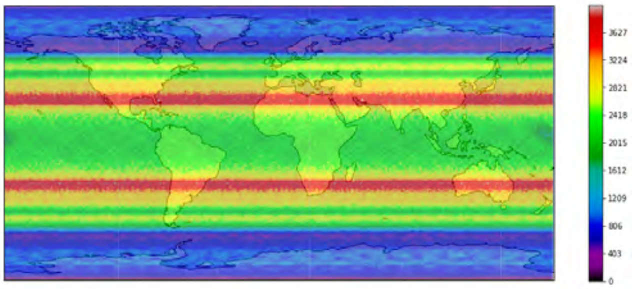


Figure 2 Starlink "Gen2" capacity density

The peak average capacity density is still very low compared with that of cellular services although the constellation has been optimized to maximize the service capability in the middle latitudes. This is partly because the metric of average capacity density implicitly assumes that the service capability is averaged over the ground surface whereas populated ground areas, shown in Figure 3, occupy a small proportion of the Earth's total area, resulting in a large percentage of the capability being wasted on oceans and unpopulated ground.

Another concern is the limited link budget. The single-user throughput provided by a single satellite is very limited, leading to less utilization of the spectrum assigned to satellite communications when compared with the terrestrial scenario.

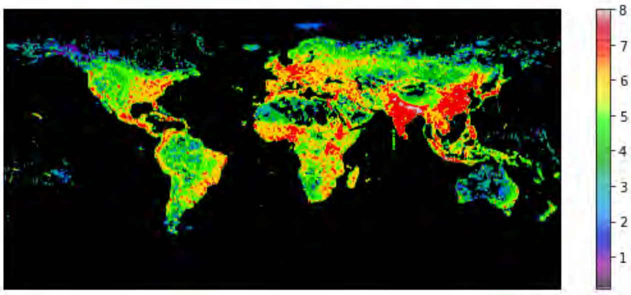


Figure 3 Global population density (generated from [13])

To fully unleash the service capabilities and address these fundamental challenges, two potential solutions are provided.

- On-demand coverage for imbalanced requirements

The beam-hopping concept is introduced to adapt the imbalanced requirements over the satellite coverage area [14–15]. Satellites can scan through a set of predefined beam hopping patterns, during which beams are active for a period of time for different areas to fulfill the service requests.

Beam-hopping technology can use all the available satellite resources to provide services to specific locations or users. By adjusting the beams' illumination duration and period, different offered capacity values can be achieved, i.e., the imbalanced requirements in different beams can be satisfied.

Additionally, beam hopping can reduce co–channel interference by placing inactive beams as barriers between co–channel beams [15]. However, beam hopping brings new challenges to LEO/VLEO satellite communications, e.g., designing beam-hopping illumination patterns to completely satisfy location-based service requirements, and considering restrictions of on-board capabilities.

Figure 4 demonstrates a snapshot of beam-hopping scheduling during a period of satellite movement. The target area where UEs are located is covered by 4 satellites (cells) at the time of observation, whose coverage topologies are shown in red, green, blue, and black, respectively. Each satellite uses at most eight beams (i.e., the highlighted beams out of all the candidate beam locations in the figure) to provide services to the connected UEs. In the LEO/VLEO system, due to the high mobility of satellites and the fact that traffic demand and buffer status of UEs will vary over time, both the candidate beams and the highlighted (illuminated) beams will be different between snapshots.

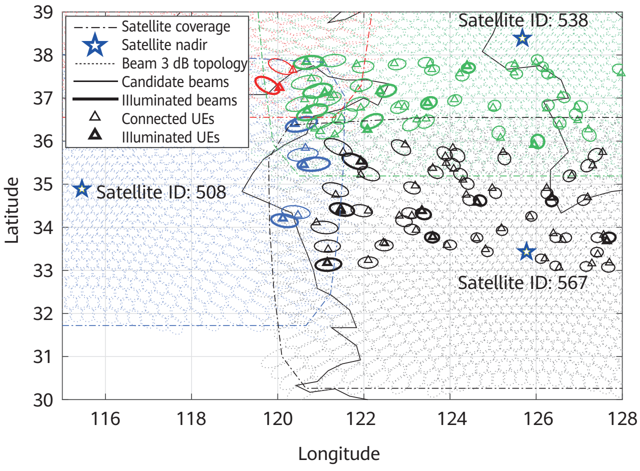


Figure 4 A snapshot of beam-hopping scheduling

Figure 5 illustrates the average throughput for different scheduling algorithms based on the simulation of a time period. As can be observed, the throughput of the beam-hopping based algorithm better matches the UEs' required capacity than the baseline Round Robin-scheduling, especially for UEs with higher traffic demands.

Multi-satellite cooperative transmission is another enabler to achieve on-demand coverage. This technology enables one user to receive multi-satellite signals simultaneously. Future LEO/VLEO mega constellations will include tens of thousands of satellites, which is the basis of multi-satellite cooperative transmission.

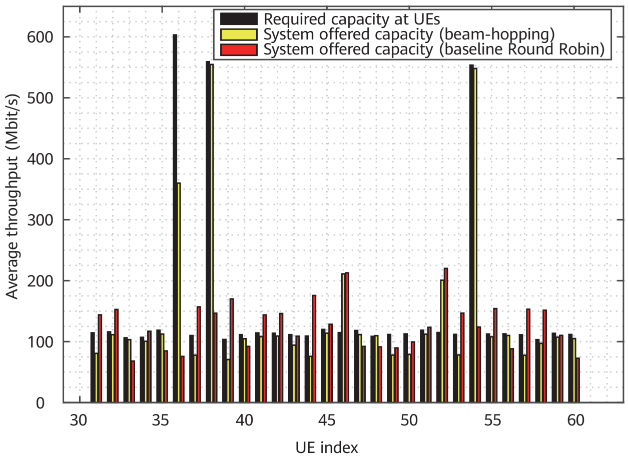


Figure 5 Throughput with and without beam hopping

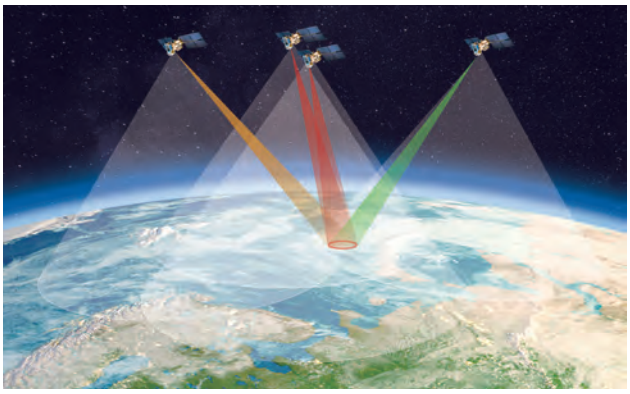


Figure 6 Multi-satellite cooperative transmission

Accordingly, the transmission rate can be increased when a user receives signals from multiple satellites at the same time, or when multiple satellites receive signals from the user, as shown in Figure 6. Based on cooperative transmission, the peak capacity density can be significantly increased as shown in Table 1. Such a scheme makes sense considering the fact that only a very small percentage of the covered area is in service and multiple satellites are usually visible with a mega constellation. The multi-satellite cooperative transmission technique can also resolve the insufficient link budget problem that arises due to the limited transmit power of one user or satellite.

Table 1 Performance of multi-satellite cooperative transmission

Satellite Coverage Area (km^2)	About 2,000,000
Beam Coverage Area (km^2)	About 1,000
Average Capacity Density (Mbit/s/km^2)	About 3.6
Cumulated Capacity Density (Mbit/s/km^2)	About 7,200

- Multi-beam precoding for high spectral efficiency.

The spectral efficiency of existing satellite communications is much lower than that of terrestrial networks due to the insufficient link budget and co-channel interference among beams. Multi-color frequency reuse is usually adopted to mitigate the co-channel interference in satellite communications, which leads to very low system spectrum efficiency. Precoding technique, which is widely used in terrestrial communications, can be employed to mitigate the co-channel interference [16]. As shown in Figure 7, multi-beam precoding can provide full-frequency reuse and improve the spectrum efficiency in VLEO/LEO satellite communications scenarios.

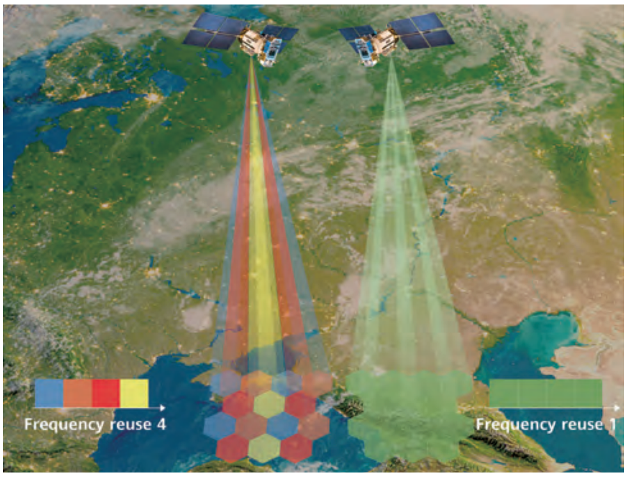


Figure 7 Multi-beam precoding

Multi-beam precoding for satellites based on full channel feedback is not preferred as there will be a large feedback delay due to the long transmission delay. As the main characteristic of the satellite channel is Line of Sight, the multi-beam precoding matrix can be calculated based on the large-scale channel which is approximately decided by the relative location between the UE and the satellite. The performance of location-based multi-beam precoding is shown in Figure 8. It can be observed that, compared with no precoding (blue bar), the introduction of multi-beam precoding (green bar) can result in a huge gain in terms of total throughput during the time the satellite provides services.

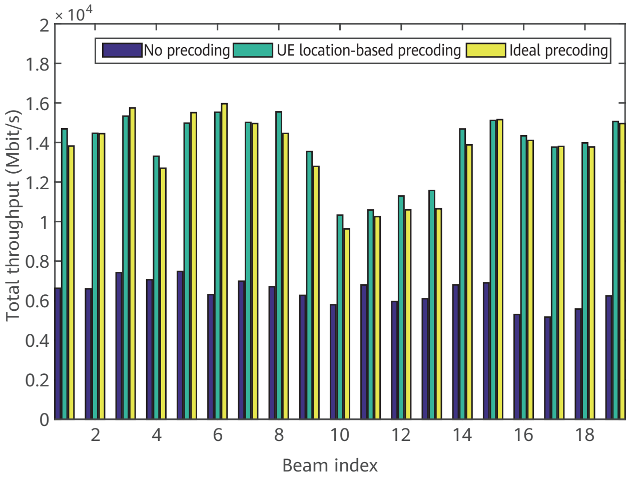


Figure 8 Throughput with and without multi-beam precoding

4.3 Dynamic Topology and Routing Algorithm

The end-to-end delay based on the VLEO constellation is expected to be lower than that based on the terrestrial Internet. Figure 9 shows the ISL-based route between Beijing and New York with the shortest distance as well as the ping Round-Trip-Time (RTT) comparison between ISL-based route and typical Internet-based route. The ping RTT of the typical Internet-based route is about 250 ms while that of the ISL-based route can be as low as 100 ms along the satellite-based route.

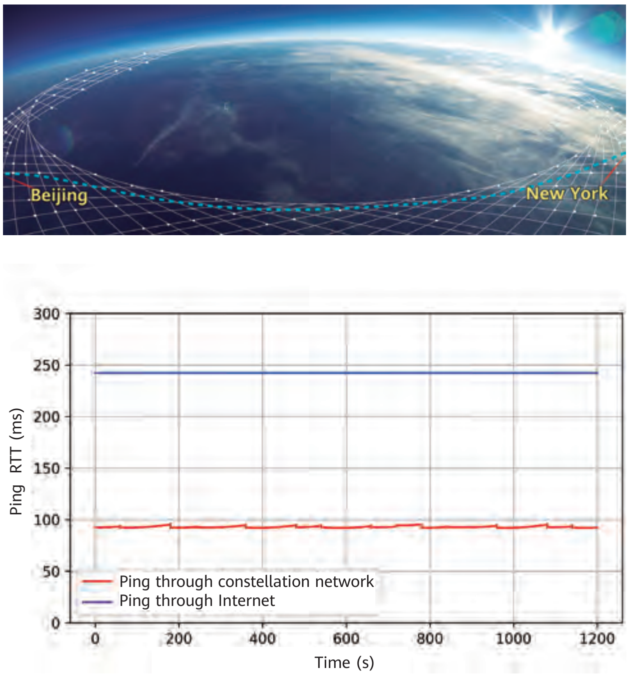


Figure 9 ISL-based route (upper) and ping RTT comparison between ISL-based and terrestrial-based route (bottom)

The potential size of mega constellations is a concern when considering routing and forwarding. Specifically, routing table sizes can grow dramatically as the satellite network grows in size. In terrestrial networks, large networks are generally partitioned into smaller networks, either by creating subnets, or by utilizing some functionality, for example Open Shortest Path First (OSPF) areas or Intermediate System to Intermediate System (IS-IS) link levels. In a satellite network, the network is in continual motion and therefore will require continual network segmentation. Highly dynamic subnetting will have detrimental consequences for the data plane.

However, each network node in a satellite network follows a predefined orbit around the Earth. Predictive routing is a specific class of routing and forwarding mechanism that takes advantage of the predictable nature of network topology changes. Unlike traditional routing and forwarding, where network nodes use flooding to signal topology changes, predictive routing allows the nodes to periodically switch routing tables that reflect the network topology graphs at different points of time. Each node contains an almanac that includes information such as the topology and time validity period. Provided that all nodes coordinate and have an accurate notion of time, the resulting network topology will appear stable. The periodicity of these changes will depend on factors such as the LEO/VLEO altitude and can be calculated by the satellite, or by a ground-based network control center.

Although the predictive routing mechanism works well in small networks as long as there are no unexpected events, an unpredicted link failure may result in a routing failure, the duration of which depends on the almanac update period. Typically, almanac updates are usually scheduled at a much slower rate than those of the traditional link state protocols, leaving nodes with outdated topology for a longer period of time. Furthermore, it requires precise timing synchronization among the satellite nodes to get all nodes updated, resulting in the data plane becoming unreliable during this time period.

Orthodromic Routing (OR) is a promising solution to address the above problems by trading some packet losses (especially when there are sufficiently large holes in the ISL mesh) against massive scalability. Since a sub-arc of the great circle between two points A and B is referred to as the Orthodrome, OR is defined as the shortest path routing on the surface of a unit sphere. Figure 10 shows the Orthodrome.

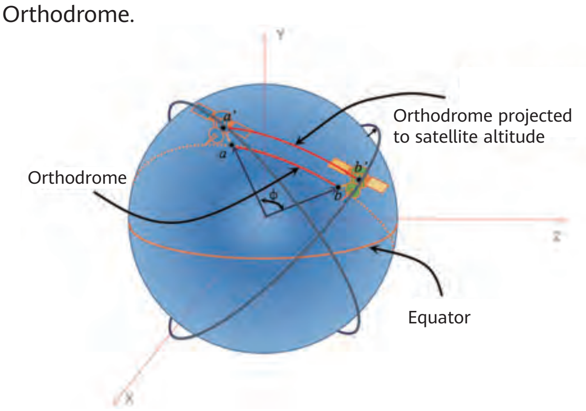


Figure 10 Orthodrome relative to the great circle

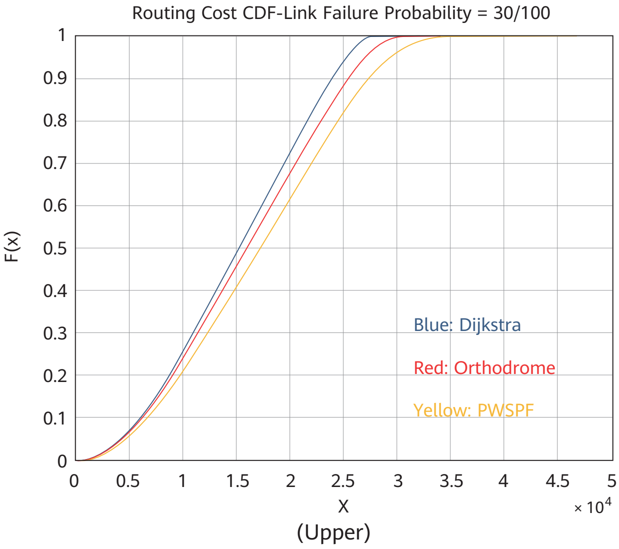
OR consists of an addressing and forwarding plane, a path computation algorithm, and a limited flooding algorithm. The addressing plane of OR embeds the <X,Y,Z> coordinates of a point on the unit sphere for both the source and destination into the IP header thus obviating the need for constant translation of identification and location. The data plane then forwards packets to the closest satellite within a relatively small flooding vicinity along the shortest path (following the ISLs) to that satellite. All satellites also have coordinate-based addresses which are a strict function of time. Therefore, all satellites can calculate their own addresses and the addresses of the satellites in their flooding region as a function of time. Flooding over a limited radius of hops and then performing path computations on those limited radius graphs is well known and Dijkstra produces the first hop needed by the forwarding plane.

Based on the above concepts a class of OR algorithms are defined as OR(r) where r is the radius in hops of the floods. OR(∞) functions as link state protocols while OR(1) performs simple geographic routing (forwarding data to the closest neighbor). Of interest is to determine which OR(r) is to be used for a given constellation size and expected link failure probability. We have conducted simulation tests on some of these algorithms and the simulations show that OR(r) can produce robust distributed routing for a relatively small r value and 10-20% link failure probabilities. This means that OR(r) can be tailored to a given constellation size and worst case failure probabilities to provide fully distributed forwarding at low loss rates.

Additionally since OR(r) may require a forwarding table with $O(r^2)$ entries, we explore several hardware solutions to pick the appropriate entry with maximum parallelization and thus appropriate for minimum clock cycle hardware forwarding at line rates.

The OR(r) algorithm as described above is executed at each hop. Therefore, the choice of gateway/intermediate nodes can change at every step towards the destination. We also have a slight variation on OR(r), where once a gateway/intermediate node is chosen, the packet is encapsulated with a source route such that the gateway can be used prior to extending its path further towards the destination. We refer to this as Piece-Wise Shortest Path OR(r) algorithm OR(r)-PWSPF.

Simulations are set up to compare the OR(r)-PWSPF with the basic OR(r) algorithm against a theoretical but non-existent full knowledge Dijkstra algorithm. The Dijkstra algorithm based on full knowledge represents an upper bound on what is possible for a given constellation. Figure 11 shows the CDF for path lengths (costs) for the different algorithms, with full knowledge Dijkstra in blue, OR(20) in red and OR(20)-PWSPF in yellow. A comparison of failed routing pairs i.e. source-destination pairs that cannot be reached under 30% link failure probabilities shows that both OR(20) and OR(20)-PWSPF come within 0.25% of full knowledge Dijkstra.



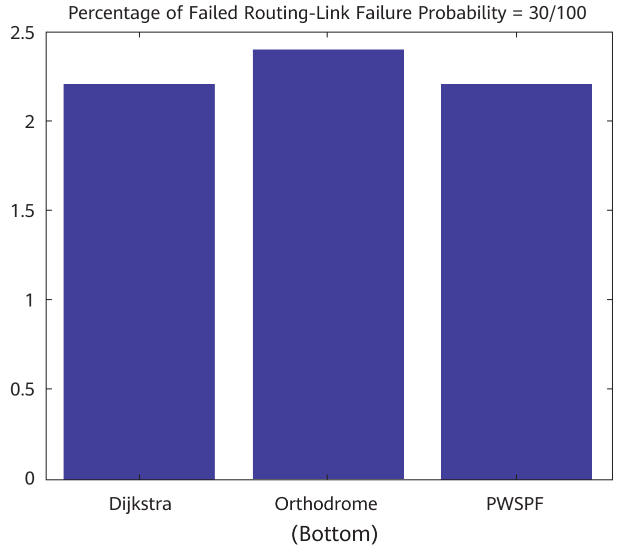


Figure 11 OR(20) routing cost comparisons and failed routing pairs for 30% link failure probabilities

It can be concluded that both OR and OR-PWSPF are capable of delivering performance that is very close to performance in an ideal scenario, but require much less control (flooding) traffic and are thus more favorable to use in a highly dynamic network.

The orthodromic family of routing algorithms employs precise local topology views at each node for global routing. Nodes in these methods only react to network events that happen in their own region, and they are unaware of events that happen elsewhere in the network. These techniques as discussed earlier, provide good performance in comparison with the traditional link state protocols. But the lack of convergence with respect to the global topology in these approaches might result in prolonged sub-optimal paths during network failures.

To solve the above mentioned issue, the routing can be via multiple-precision regions. Each node's link state database and topology graph consists of multiple zones/levels/regions/radii. Each zone has a degree of precision with respect to the network event refresh time. The following illustrates an example of a multiple-precision region network graph in a node.

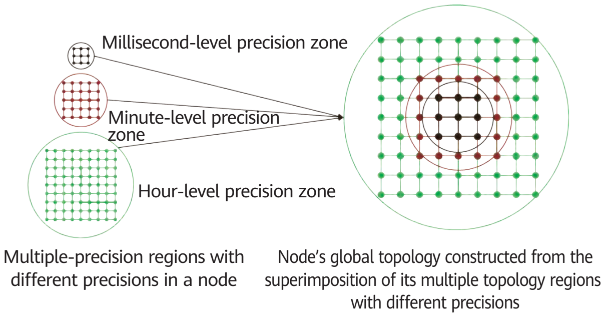

Millisecond-level precision zone
Minute-level precision zone
Hour-level precision zone
Multiple-precision regions with different precisions in a node
Node's global topology constructed from the superimposition of its multiple topology regions with different precisions

Figure 12 Multi-precision region graph: regions

Different techniques and strategies can be employed to deliver updates to a node for each of its topology zones based on the zone's precision requirement. While one zone can use an almanac, for example, the other can use traditional or limited flooding.

The nodes use the shortest path to the destination based on their global view of the network which now consists of multiple precision levels. This method can be applied to networks that employ traditional routing or networks that employ OR or OR-PWSPF algorithms to deal with node mobility in satellite networks and employ geographical addressing.

This technique shares the advantages provided by OR algorithms and allows the use of large flat topologies in network operations.

Finally, in order to limit the dynamic change of the satellite constellation topology, ISLs are usually assumed within the same constellation layer, and each satellite can have only two intra-plane and two inter-plane ISLs. This greatly compromises the communication capability of the entire network, and the optimal bandwidth and minimum delay cannot be achieved. Therefore, new routing algorithms are expected to accommodate constellations with more free connections among satellites e.g. across layer connections, thus extending the capability boundaries of the LEO/VLEO constellation.

4.4 Powerful On-board Capabilities

The on-board capabilities call for thorough enhancements to accommodate communication requirements of NTN for 6G, mainly in on-board processors, radio frequency subsystem, antennas, and data transmission algorithms. Massive-beam satellites with on-board data processing capabilities and advanced algorithms will play a key role in future low-orbit satellite communications, providing more linking capabilities for users over the coverage area through frequency and beam traffic reconfiguration.

In future NTN, massive-beam high-gain phased array antennas will be equipped to prevent the extremely high path loss from space to ground. Assuming the altitude of the satellite is 300 km, the free space path loss is around 170 dB at Ka band with an extra loss of 6 dB due to rain. When the diameter of the satellite payload antenna is 1.0 m, the maximum antenna gain can be assumed as 45 dBi and the equivalent isotropically radiated power (EIRP) may reach 50 dBW, which is subject to the power restrictions on satellites. The typical diameter of a ground UE antenna for the Ka band is 0.5 m, which leads to a maximum gain of 34 dBi and a G/T value of 8.5 dB. Approximate calculation shows that the downlink signal-to-noise ratio (SNR) may reach up to 27 dB with a bandwidth of 400 MHz. The signal quality is sufficient to support higher-order modulation of 64QAM. The data rate achieved by a single beam is 1200 Mbit/s and the spectrum efficiency is 4.8 bit/s/Hz, considering interference.

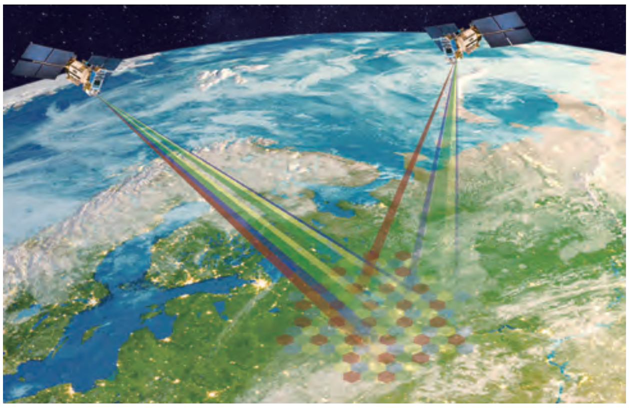


Figure 13 Massive-beam satellites

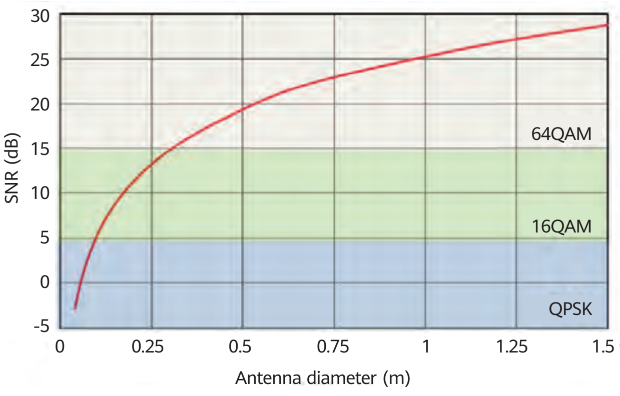


Figure 14 Available SNR for different satellite antenna diameters

The challenge is to find a way to generate these beams by utilizing the limited physical space on the satellite. The digital beam forming (DBF) method is considered a promising solution for future phased antenna arrays, in which multiple beams are generated in the digital domain. The digitization of the Tx/Rx data can also provide maximum flexibility and dynamic range in large systems [17]. The practical challenges to implementing DBF are the large amounts of data that needs to be processed and the use of sophisticated transceivers that consume high amounts of power, which cannot be provided by satellites. The development of digital integrated circuits and mixed-signal integrated circuits makes the DBF implementation realistic. In [18], a full DBF transceiver is designed for millimeter wave (mmWave) application. A maximum of 20 digital beams are generated from 64 RF channels. In the future, the number of beams will extend to over 1000 and RF channels to over 4000. The progress in RF components and materials also helps reduce the power consumption and improve the on-board capabilities.

4.5 Low-Cost Manufacturing & Service

Reducing the satellite components' manufacturing cost and the service price is the prerequisite for making satellite communications a part of daily life.

Regarding the manufacturing, a full integration of satellite communications into the cellular system is expected to be the most effective way to reduce the cost of communications components in ground segment devices like UEs, gateways, as well as the on-board processing system. With a unified air interface design capable of satellite communications and terrestrial communications, the baseband chips and components of satellite communications can make full use of the economies of scale of the cellular industry, leading to much lower chip and device costs.

It is a challenge to reduce the cost of the space segment to achieve low-cost manufacturing. The space-class components are radiation-hardened and screened to make sure they are reliable enough in the space environment. Because this process is not industrialized, the cost is extremely high. In addition, because the quantity of radiation-hardened devices is very small, manufacturers have no incentive to perform radiation-hardening for the latest products, leading to a delay in the delivery of space products by several years or more when compared with their latest commercial counterparts. Low cost, high performance and low lead time are the requirements for commercial satellite parts. In recent years, there have been some explorations on using commercial-class devices i.e. the Commercial-Off-The-Shelf (COTS) parts in spacecrafts. Optimized processes, such as a better balance of the cost and reliability in the screening, new shield designs, and a fault detection and recovery mechanism, are needed to ensure the stability and commercial efficiency of spacecrafts.

The service cost will also benefit from the full integration between satellite communications and cellular communications since it is also related to the economies of scale. Currently, the ecosystems of different constellations are isolated from each other and the number of users of each constellation is insufficient to make full use of constellation capacity, resulting in the cost per bit in existing satellite communications being much higher than that of terrestrial networks. In 6G, the wireless standards should be unified around the world, and with a single device, people should be able to freely roam between TN and NTN and between different NTNs. In this way, the network capacity of a satellite system can be much better utilized to reduce the overall service cost.

4.6 Interference Reduction and Co-existence

It is critical to find a way to prevent the interference between TN and NTN in order to ensure communication service quality. Frequency sharing between cellular and satellite communications is a hot topic that has been discussed in both the academia and the industry. However, the current frequencies allocated to cellular and satellite communications are usually isolated from each other. In actual practice, a gap is introduced to ensure that the out-of-band leakage of the waveform signal due to non-linear devices can be sufficiently low. Owing to the fast development of cellular communication, the spectral efficiency of terrestrial networks has dramatically increased, and is much higher than that of satellite communications. The frequency resources allocated to cellular networks contribute more to human communication requirements. This motivates cellular operators to obtain more frequency resources from satellite operators to provide users with a better cellular experience.

Considering the fact that very limited frequency resources are available, it is more important than ever to design a frequency sharing mechanism that not only considers the comprehensive utilization of the spectrum, but also meets the needs of different types of communications scenarios from a technical and neutral perspective. In general there are several hierarchical frequency sharing technologies that can be considered to reduce the interference among the different types of satellite communications and cellular communications.

- Space isolation

The most straightforward method for interference reduction is space isolation. The same frequency resource can be allocated to both cellular and satellite networks that are geographically far away from each other to prevent any possible interference. For example, the frequency assigned for cellular operators in terrestrial networks can also be used for satellite communications on an ocean provided that the two deployment areas are geographically far away so that the maximum transmitted signals from the cellular base station would be much lower than the background thermal noise of the satellite communications terminal receiver after long propagation, and vice versa.

An application of space isolation in scenarios where the cellular and satellite networks are striving for sharing the frequency resource is shown in Figure 15. For a cellular base station, the space area around this base station will be noted as an "electronically fenced area" where satellite beams are not allowed.

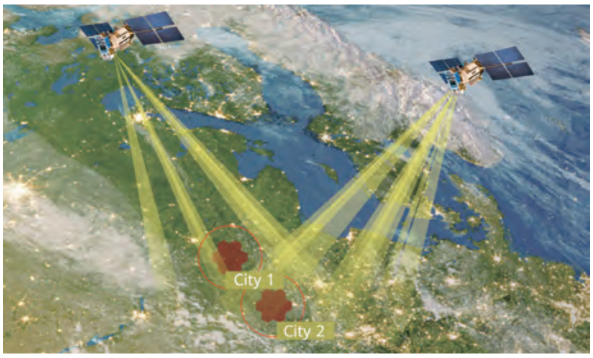


Figure 15 A demo of "electronically fenced area"

The size of the electronically fenced area will significantly affect the possible interference level from the LEO/VLEO satellites. Figure 16 shows one snapshot of the interference level in terms of interference noise ratio (INR). The satellite beams causing INR above -10 dB and -5 dB are marked in yellow and red, respectively, with difference in the size of the electronically fenced area. A 54 km-wide electronically fenced area is sufficient to eliminate all interference above -5 dB for the considered case.

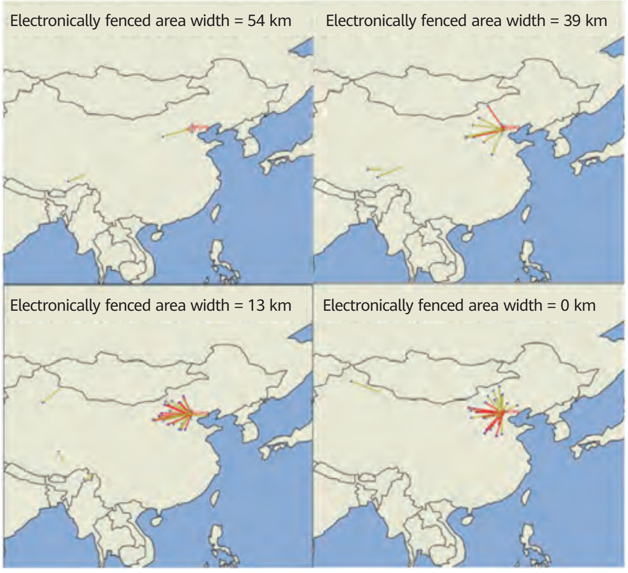

Electronically fenced area width = 54 km
Electronically fenced area width = 39 km
Electronically fenced area width = 13 km
Electronically fenced area width = 0 km

Figure 16 Application of space isolation in interference avoidance between satellites and cellular base stations

Figure 17 shows the interference along a time interval with two electronically fenced areas of width 0 km and 54 km. A larger isolation distance can effectively reduce the probability of receiving high INR.

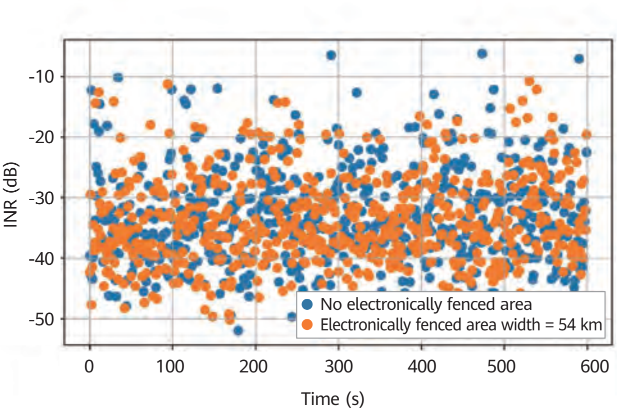


Figure 17 INR levels within a time interval of 600s with different electronically fenced area widths

- Angle isolation

For scenarios targeting mmWave bands where only UEs with directional antenna are deployed, angle isolation can be considered to prevent the interference caused from different systems. Considering a serving area illuminated by signals of the same frequency band from different systems, the arrival angle of the signals may be far different from each other. At the receiver side, the huge side-lobe reduction of directional UE provides good spatial filtering and can eliminate the interference. The potential interference to other systems can also be eliminated because the transmitted signals will experience huge attenuation due to the directional antenna.

- Scheduling-based interference coordination

Scheduling-based interference coordination has been deployed in cellular communications systems to alleviate the interference in cell edge areas. With close interaction among neighboring base stations, joint decisions can be made among those stations to send signals to UEs at the cell edge with staggered frequency resources in order to prevent interference. Compared with the traditional sensing-and-decision procedure, coordination-based scheduling attempts to solve the interference issue in a proactive way, and thus provide better user experience.

However, coordination-based scheduling is rarely used between the cellular and non-terrestrial networks for the time being since they are isolated from each other. By taking the advantage of integrating cellular and satellite communications, scheduling-based interference coordination is expected to become possible.

5 Conclusion

The successful realization of LEO/VLEO-based NTN communications calls for joint efforts from the academia and industrial communities. The ongoing development of new technologies and the growing interest and investments in space applications is extending the boundaries of potential LEO/VLEO-based communications to new heights. In addition to the technical aspects of satellite communications itself, a fundamental integration of cellular- and satellite-based communications at the physical layer from day one is also the key to the commercial success of LEO/VLEO-based satellite communications in 6G. The NR-based NTN discussion in 3GPP provides an excellent platform that traditional cellular and satellite communities can use to work together to build a fully integrated network. As the advanced frequency sharing schemes between the cellular network and NTN mature, regulatory authorities may have more room to assign frequency resources in an efficient way.

References

[1] http://www.3gpp.org

[2] 3GPP TR 38.811, "Study on New Radio (NR) to support non-terrestrial networks (Release 15)."

[3] 3GPP TR 38.821, "Solutions for NR to support non-terrestrial networks (NTN) (Release 16)."

[4] 3GPP TR 22.822, "Study on using Satellite Access in 5G; Stage 1 (Release 16)."

[5] https://oneweb.net

[6] https://www.spacex.com

[7] https://www.telesat.com

[8] "Starlink Statistics." planet4589.org - Jonathan's Space Report. Archived from the original on 5 May 2021.

[9] E.Mazareanu. "https://www.statista.com/topics/1707/air-transportation/.", 2020.

[10] M. Sheng, Y. Wang, J. Li, R. Liu, D. Zhou, and L. He, "Toward a flexible and reconfigurable broadband satellite network: Resource management architecture and strategies," IEEE Wireless Communication, vol. 24, no. 4, pp. 127-133, Aug. 2017.

[11] M. Y. Abdelsadek, H. Yanikomeroglu, and G. K. Kurt, "Future ultra-dense LEO satellite networks: A cell-free massive MIMO approach," IEEE International Conference on Communication. Workshops (ICC Workshops), pp. 1-6, 2021.

[12] FCC Application File Number: SAT-LOA-20200526-00055;

[13] CIESIN S. Gridded population of the world, version 4 (GPWV4): population density. Center for International Earth Science Information Network-CIESIN-Columbia University. NASA Socioeconomic Data and Applications Center (SEDAC), 2015.

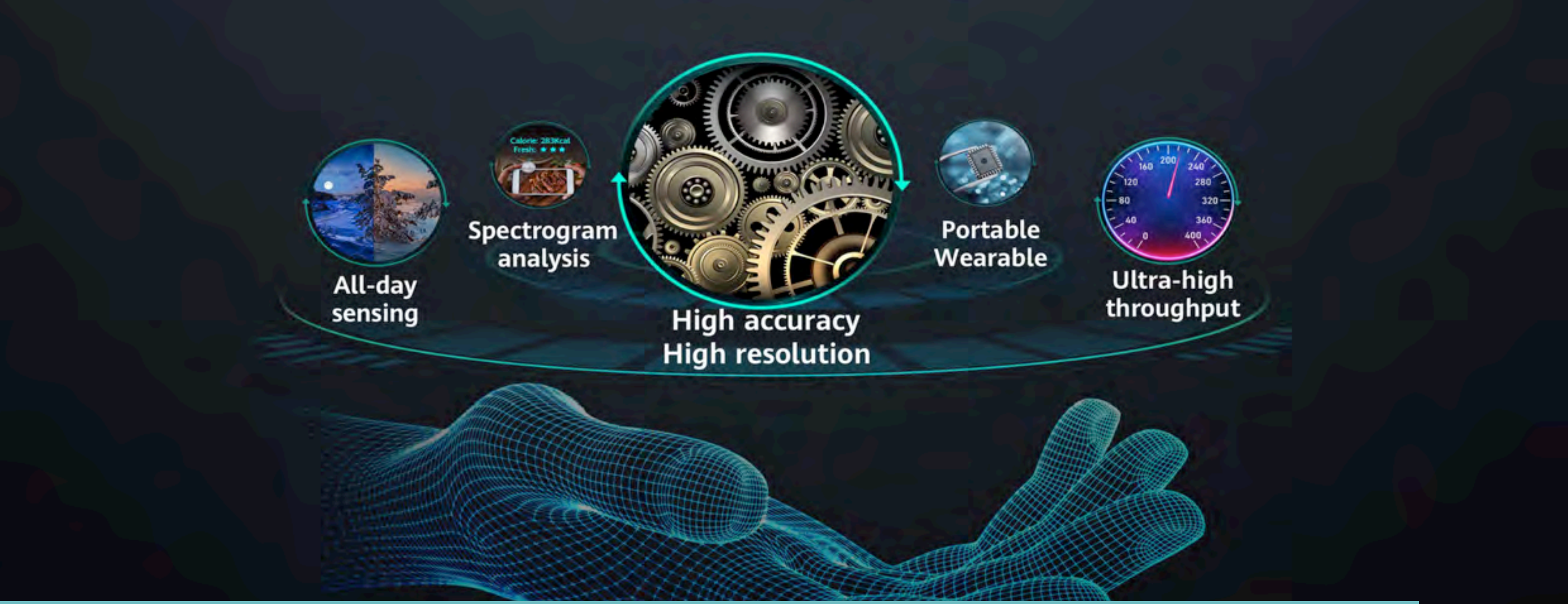
[14] J. Anzalchi, A. Couchman, P. Gabellini, et al. "Beam hopping in multi-beam broadband satellite systems: System simulation and performance comparison with non-hopped systems," IEEE 5th Advanced Satellite Multimedia Systems Conference and the 11th Signal Processing for Space Communications Workshop, pp. 248-255, 2010.

[15] J. Tang, D. Bian, G. Li, J. Hu and J. Cheng, "Resource allocation for LEO beam-bopping satellites in a spectrum sharing scenario," IEEE Access, vol. 9, pp. 56468-56478, 2021.

[16] G. Zheng, S. Chatzinotas and B. Ottersten, "Generic optimization of linear precoding in multibeam satellite systems," IEEE Transactions on Wireless Communications, vol. 11, no. 6, pp. 2308-2320, June 2012.

[17] C. Fulton, M. Yeary, D. Thompson, J. Lake, and A. Mitchell, "Digital phased arrays: Challenges and opportunities," Proceedings of the IEEE, vol. 104, no. 3, pp. 487-503, 2016.

[18] B. Yang, Z. Yu, J. Lan, R. Zhang, J. Zhou, and W. Hong, "Digital beamforming-based massive MIMO transceiver for 5G millimeter-wave communications," IEEE Transactions on Microwave Theory and Techniques, vol. 66, no. 7, pp. 3403-3418, July 2018.



Terahertz Sensing and Communication Towards Future Intelligence Connected Networks

Guangjian Wang [1], Huanhuan Gu [2], Xianjin Li [1], Ziming Yu [1], Oupeng Li [1], Qiao Liu [1], Kun Zeng [1], Jia He [1], Yan Chen [2], Jianmin Lu [1], Wen Tong [2], David Wessel [2]

[1] Wireless Technology Lab

[2] Ottawa Wireless Advanced System Competency Centre

Abstract

The terahertz (THz) technologies are extremely promising for future 6G wireless communication and sensing systems because of the advantages of ultra-wide available bandwidth. However, the increased operating frequency and bandwidth pose higher requirements in terms of propagation channel modeling, new transmission technologies, high performance components, and signal processing complexity. Therefore, the challenge lies in how to achieve system demands as well as improve the system data rate and sensing resolution with limited hardware complexity and power consumption.

In this paper, the latest progress on spectrum and potential application scenarios for the THz band are examined. A hybrid channel modeling framework for the THz band is proposed to improve the accuracy and efficiency of the modeling. The potential intermediate radio frequency (IRF) architecture, key components, and antenna and integration technologies have also been investigated. In particular, the THz subsystem with silicon and III-V compound semiconductor material heterogeneous integration is proposed to promote the performance by utilizing the advantages of different processes and materials. Finally, the prototype and measurement campaign are conducted, which illustrate the advantages of the THz band for high throughput communication and high resolution sensing scenarios. A variety of measurement campaign examples show 210 Gbit/s data transmission rate at 330 m distance, and up to 3 mm invisible imaging, which is the highest performance in this field.

Keywords

channel model, phased array, reconfigurable intelligent surface, THz, THz antenna, THz integration technologies, integrated sensing and communication, wireless communication

1 Introduction

With the rapid development of wireless cellular communication from 1G to 5G, it is not just humans that are being better connected, but also an increasing number of intelligent things such as industrial equipment, cars, sensors, and home devices. This trend will continue beyond 2030, leading to intelligent connection of everything, anywhere, all-time [1]. If these intelligent things also have the capability of sensing their surroundings and sharing this information with other intelligent things, this will make connections even more intelligent. Joint radar and communication technology have been considered in this regard. Co-located radar and communication systems have been emphasized for the goal of minimizing interference to each other in previous research [2–3]. However, it has stringent requirements on the information exchange between these two systems, hence leading to limitations in practice. Effective integrated sensing and communication (ISAC) systems, including those that are loosely coupled to those fully integrated, are expected to reduce the system size and information exchange latency between the co-located radar and communication systems. With each new generation, higher spectrum with larger bandwidth is utilized. This is also beneficial for sensing. If we can make THz work for 6G, many new opportunities will be within our grasp. For the reasons previously stated, we believe the sensing will be one of key new services for 6G in addition to the continued expansion of 5G services.

To this end, the terahertz (THz) band (0.1–10 THz) is one of many promising pillar technologies that meets the requirements of 6G for 2030 and beyond, accommodating a massive number of connected devices and featuring ultra-high user data rates in the order of terabit per second (Tbit/s) [1]. This is because the THz band has ultra-large available bandwidth resources and ultra-high communication rates. Therefore, THz communication is considered as an important alternative air interface technology for achieving Tbit/s communication rate. It is also expected to be applied to scenarios such as holographic communication, small-scale communication, ultra-large-capacity data backhaul, and short-distance ultra-high-speed transmission. In addition, high-precision positioning and high-resolution sensing imaging of a network and/or a terminal device are performed by using a feature of an extremely large bandwidth, which is also an expansion direction of a THz communication application [4].

As mentioned previously, THz can provide high-quality imaging resolution equivalent to optics (about 100 microns). THz waves can penetrate many infrared opaque materials such as paper, plastics, ceramics and semiconductors. They can interact with molecular hydrogen bonds or van der Waals forces without any ionizing radiation and can be used for spectroscopic identification of organic materials. Terahertz photons have low energy (1 THz is equivalent to 4 meV) and are not harmful to human beings, unlike high energy X-rays. The vibrational and rotational energy levels of molecules, as well as the phonon vibrational energy levels of semiconductors and superconducting materials are all within the THz band, so THz waves have great advantages in spectral analysis and material identification. Because THz can be used for both communication and sensing, it is a strong candidate for ISAC [5].

Compared with millimeter waves and microwaves in low frequency bands and visible light in high frequency bands, the THz channel characteristics are quite different. Compared with millimeter waves, THz waves have stronger frequency selectivity, more obvious scattering effect, and larger transmission loss. Compared with a light wave, a THz wave has less path loss, stronger volatility, stronger reflected energy, and is less likely to be blocked. Therefore, the existing channel models and measurement methods of millimeter waves, microwaves, and visible light systems cannot be directly applied to the THz band, which highlights the necessity of developing THz channel measurement instruments. In the field of wireless channel modeling, there are two modeling methods. One is statistical channel modeling methodology based on measured data, and the other is deterministic channel modeling methodology based on ray tracing or electromagnetic (EM) field boundary solving theory [6]. Statistical channel modeling theory is widely used in mobile communication standard channel modeling scenarios [7], such as the 3GPP standard channel model. However, with the continuous enrichment of next-generation mobile communication scenarios, the demand for a new spectrum increases, and a statistical channel modeling method cannot completely meet the new channel requirements. Therefore, deterministic channel modeling methods are gradually being studied, and high-precision channel modeling in specific scenarios is carried out by using computational electromagnetics (CEM) methods [8].

The device based on the THz semiconductor technology mainly refers to a transistor in the THz frequency band. The solid-state circuit based on the solid-state device can implement the THz source and perform frequency mixing, frequency multiplication, and amplification on the THz signal to generate and detect the THz wave at a specific frequency [9].

The Schottky barrier diode (SBD) can work at normal temperature, and has a low turn-on voltage and a very short reverse recovery time. At present, the SBD in the THz band is mainly based on GaAs material because of its high saturation electron rate and electron mobility. GaAs-based SBD is used in THz solid-state active circuit and represented by American VDI Company since 1960s, which has been very mature and industrialized. Currently, the cutoff frequency of the component is higher than 30 THz, and the frequency mixer and frequency multiplier basically cover the THz band.

Chip integration has become the most important research direction of THz semiconductor technology. Based on semiconductor materials, semiconductor devices used in THz band amplifiers can be divided into two types: Si-based devices and III-V compound-based devices [10–11]. Si-based devices are mainly complementary metal–oxide–semiconductor (CMOS) devices and SiGe bipolar complementary metal-oxide-semiconductor (BiCMOS) devices. Group III-V compound devices include GaAs pseudomorphic high-electron-mobility transistor (PHEMT), GaAs metamorphic high-electron-mobility transistor (MHEMT), InP high-electron-mobility transistor (HEMT), InP heterojunction bipolar transistor (HBT), and GaN HEMT.

The selection of the various process and material properties for THz devices is primarily based on the characteristic frequency and cutoff frequency. To select the right technology for THz applications, many parameters must be considered, such as cost, output power, efficiency, maturity of interconnection and packaging technologies, and integration capabilities. In the THz band, a large-scale antenna array is usually required to ensure the transmit power. As the operating frequency increases, high integration becomes increasingly important. Obviously, a small wavelength of THz is very beneficial for implementing a large-scale antenna array with a small size. But the small wavelength also poses corresponding challenges.

Up to now, much progress has been made in key technological breakthroughs and prototype system development for THz communication and sensing systems. For example, Zhejiang University developed a multi-channel THz wireless communication system based on photoelectric combination [12]. The system uses an eight-channel THz carrier for modulation to achieve ultra-high-speed wireless communication, which is with a working frequency of 0.4 THz, a modulation method of 16QAM, and a transmission rate of 160 Gbit/s. The advantage of the system lies in achieving ultra-high transmission rate and improving bandwidth utilization. In 2020, the University of Electronic Science and Technology of China achieved THz high-speed wireless communication with a working frequency of 0.22 THz, a communication distance of more than 1000 m, a bit error rate of less than 1E-6, and a transmission rate of more than 20 Gbit/s [13]. For the sensing technologies, terahertz time-domain spectroscopy (THz-TDS) has been used for material's characterization and process control [14].

This article is organized as follows. After stating the latest progress on THz spectrum and potential application scenarios in Section 2, we discuss THz channel propagation and our latest measurement and modeling results on the THz band in Section 3. In Section 4, we focus on the key components and intermediate radio frequency (IRF) architecture, including THz components and chips, THz antennas and THz integration technologies. The prototypes and measurement results of THz high throughput communication and high precision sensing system will be depicted in Section 5. It includes the prototype description, measurement environment and configurations, and the measurement results. Finally, we offer some conclusions and suggestions for future research in Section 6.

2 THz Spectrum and Application Scenarios

THz spectrum usually refers to the frequency bands between 0.1 THz to 10 THz with a corresponding wavelength of 0.03 mm to 3 mm, and lies somewhere between microwaves and optical waves, as illustrated in Figure 1. Due to its unique position in the EM spectrum, THz has the characteristics of microwaves, such as penetration and absorption, as well as the spectral resolution of optical waves.

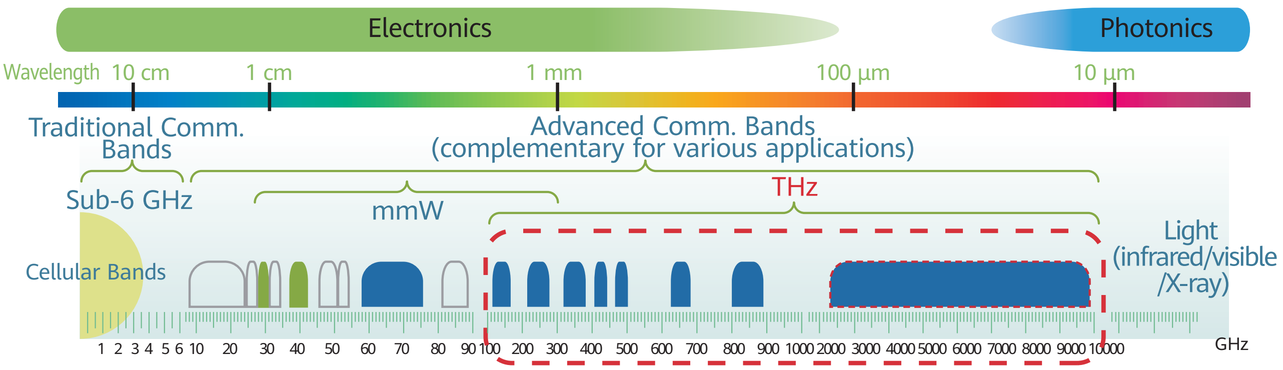


Figure 1 Position of THz waves in the radio spectrum

For a long time, THz spectrum is described as the last virgin land of the radio spectrum. Only a few scientific and astronomical services are deployed in these frequency bands, especially in bands above 275 GHz. In spite of having abundant spectrum and supporting a high transmission rate and strong anti-interference, there are still many practical technical limitations.

However, this has partially changed with the development of integrated components and circuits, and the emergence of various services that require ultra-high data rate transmission. At the World Radio Communication Conference 2019 (WRC-19) [15], RR No. 5.564A was approved, and four globally harmonized frequency bands with a total bandwidth of 137 GHz (i.e., 275–296 GHz, 306–313 GHz, 318–333 GHz, and 356–450 GHz) were allocated for the implementation of land mobile and fixed service application in the frequency range of 275 GHz to 450 GHz, on the basis of study outcomes of Agenda Item 1.15 (WRC-19). Therefore, with the addition of the spectrum allocated at the previous WRCs, there are more than 230 GHz Mobile Service spectrums. Table 1 describes the allocated mobile frequency bands with a contiguous bandwidth greater than 5 GHz.

Table 1 Allocated frequency bands of mobile service in the frequency range of 100–450 GHz

Frequency (GHz)	Contiguous Bandwidth (GHz)
102–109.5	7.5
141–148.5	7.5
151.5–164	12.5
167–174.8	7.8
191.8–200	8.2
209–226	17
252–275	23
275–296*	21
306–313*	7
318–333*	15
356–450*	94

* Allocated for Fixed Service/Mobile Service at WRC-19, 2019.

Such abundant spectrum reserves will drive the rapid development of the THz communication technologies. First of all, it can facilitate extremely high-data-rate connection of the existing wireless transmission applications, such as fixed wireless access (FWA), wireless cellular front-hauling and backhauling, and some short-range link communications [16]. It can increase their connection rates from tens of Mbit/s or several Gbit/s to the unpredictable hundreds of Mbit/s or even several Tbit/s, which is truly comparable to the connection experiences of optical fibers.

Furthermore, the ultra-fine beam generated by the ultra-large-scale antenna array can be implemented in the THz frequency band, which makes high-precision positioning and high-resolution sensing possible. This will support the emergence of new services that are beyond just communication. Shorter wavelengths imply smaller antennas, so small devices can be packed with tens or hundreds of antennas, which are beneficial for angle estimation. The gesture recognition on smartphones is a good example. From the perspective of the base station side, enabling the sensing/imaging feature in future International Mobile Telecommunications (IMT) systems are also important application scenarios for supporting external environment recognition and map reconstruction in THz communications [6].

3 THz Channel Propagation and Modeling

Propagation channel modeling is a fundamental part of wireless communications. Historically, the stochastic channel modeling methodology has dominated the wireless communication channel model. The stochastic channel model can describe the propagation channel using simple statistical parameters with a low computational complexity of implementation. Many projects and standards, such as 3GPP-SCM, WINNER-I/II, COST2100, and MESTIS, belong to this family. In 2015, the 3GPP 38.901 has released the spatial channel model (SCM) from 0.5 GHz to 100 GHz, which become the 5G standard channel model. However, in 6G communication, the spectrum requirement has extended from millimeter-wave bands to THz bands. In the THz band channel modeling, we will face some new challenges and propagation features which are quite different with millimeter waves.

For the propagation attenuation aspect, THz waves will experience higher path loss than millimeter waves and in some situations the molecular absorption should be considered. In the THz band, the ultra-large bandwidth will result in frequency response inconsistency and higher delay resolution. With frequency increase, the wavelength will decrease to the millimeter level. This implies the wavelength will be comparable with the surface roughness of most of the furniture in the environment, which means the new scattering feature should be modeled. Furthermore, the new small-scale parameters including the delay spread, angular spread, and clusters should be restudied under the stochastic channel model.

Apart from communication, the THz band can be a candidate for sensing applications. In contrast to the communication channel, the sensing channel focuses on different parameters and methodology. For example, the imaging channel requires the deterministic channel coherence of the aperture antenna and the geometry information, and this feature is contradictory to the traditional stochastic channel modeling approach. This means that the stochastic channel models are not suitable for sensing applications, whereas deterministic modeling approaches are favored. However, a single channel modeling scheme may not meet the evaluation requirements of all ISAC applications. For sensing applications such as sensing assisted beamforming, the stochastic modeling can be adopted. However, localization and tracking cases, since description of EM information is not required, the ray tracing can be considered as a strong candidate. On the other hand, imaging and recognition need to take the EM algorithm into account when the sizes of scatterers are approximate to wavelength [17].

Based on the new challenges and requirements, we propose a hybrid channel modeling methodology to support the THz band communication and sensing. Depending on different applications, different approaches are used for channel generation and system-level or link-level evaluation. A few suggested channel modeling methods are listed in Table 2.

Table 2 Terahertz channel modeling methodology

Application	Channel Modeling Method
Communication	Stochastic
Positioning	Stochastic (GBSCM)
Localization and mapping	Ray-based
Imaging and recognition	EM-based

In the next section, we will introduce our current THz band channel measurement campaign progress, and relevant stochastic channel modeling results.

3.1 Channel Measurement System and Measurement Campaign

The THz channel measurement platform consists of a radio frequency (RF) front-end with horn antennas at both transmitter and receiver sides, and a vector network analyzer (VNA). The intermediate frequency (IF) signal is generated by the VNA and then mixed with the multiplied oscillator signal to the RF band, and finally, emitted/received by a horn antenna. With the wide bandwidth, a higher delay resolution can be achieved. To ensure wide-angle coverage, we use a wide beam width antenna at the transmitter sides. On the receiver side, a high gain antenna is mounted on a mechanical rotator to achieve the angular channel response and complement the high path attenuation at the THz band. The measurement campaigns have been achieved at 140 GHz, 220 GHz, and 280 GHz. The detailed parameters are listed in Table 3.

Table 3 Parameters of the measurement system

Parameter	Value	
Frequency band [GHz]	140	220
Local oscillator [GHz]	10.667	18
Start frequency [GHz]	130	201
End frequency [GHz]	143	209
Bandwidth [GHz]	13	8
Sweeping points	1301	801
Transmitter antenna gain [dBi]	15	10
Receiver antenna gain [dBi]	25	25
Azimuth rotation range [degree]	[0:10:360]	[0:10:360]
Elevation rotation range [degree]	[-20:10:20]	[-20:10:20]
Delay resolution [ps]	76.9	125
Maximum excess delay [ns]	100	100

We carry out the channel measurement in a typical meeting room and open office area. The realistic environment can be seen in Figure 2. In the meeting room with an area of 10.15 m x 7.9 m and a ceiling height of 4 m. A 4.8 m x 1.9 m desk with a height of 0.77 m is placed in the center, and some chairs are around the desk. The dimension of the office room in our channel measurement campaign is 30 m x 20 m, including a hallway and an office area. The furniture in the environment includes desks, chairs, tiny plants, screens, etc.

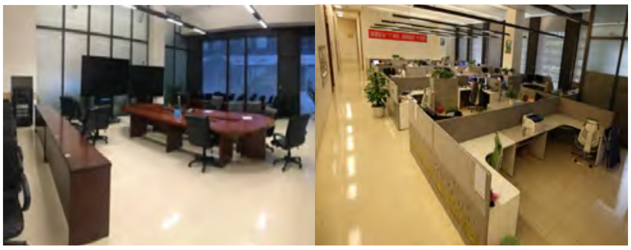


Figure 2 Meeting room (left) and open office (right)

3.2 THz Channel Characterization and Analysis

Path loss is a large-scale fading which reveals the signal power level of the receiver at different places. We evaluate the multi-frequency alpha-beta-gamma (ABG) path loss model for all the measurement sets. As we know, the multi-frequency path loss models cover the relationship between path loss and both distance and frequencies. As a widely used multi-frequency path loss model, ABG model is obtained by adding a frequency-dependent optimization parameter to the alpha-beta (AB) model used in 3GPP. The ABG model can be expressed as

$$PL^{ABG}[dB] = 10\alpha log_{10}\left(\frac{d}{d_0}\right) + \beta + 10\gamma log_{10}\left(\frac{f}{f_0}\right) + X_{\sigma}^{ABG} \qquad (1)$$

Where f and f_0 denote the carrier frequency and the reference frequency in gigahertz, respectively. d and d_0 represent the distance between the transmitter and receiver and reference distance. X_{σ}^{ABG} is a zero-mean Gaussian random variable with standard deviation σ_{SF}^{ABG} in dB, which represents the fluctuation caused by shadow fading. In addition, we can see from formula (1) that α and γ represent the dependence of path loss on distance d and frequency f, respectively, while β is an offset parameter.

Based on the measurement campaign, in the meeting room environment, the ABG path loss results on 140 GHz, 220 GHz, and 280 GHz bands are depicted in Figure 3. The proposed ABG path loss model is as follows:

$$PL^{ABG}[dB] = 20.7log_{10}\left(\frac{d}{d_0}\right) + 26.72 + 22.2log_{10}\left(\frac{f}{f_0}\right) + 2.53 \qquad (2)$$

More measurement campaign and modeling results can be referred to in literature [18–20].

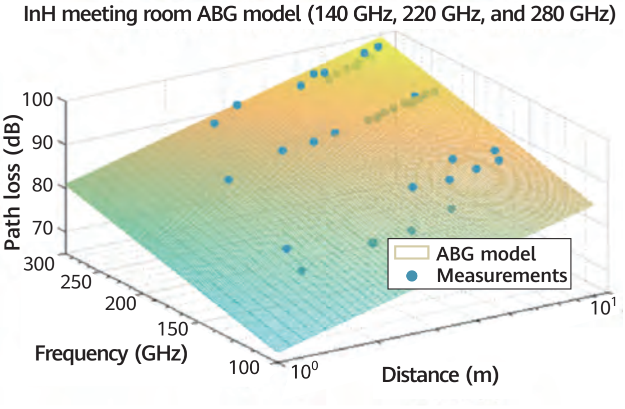
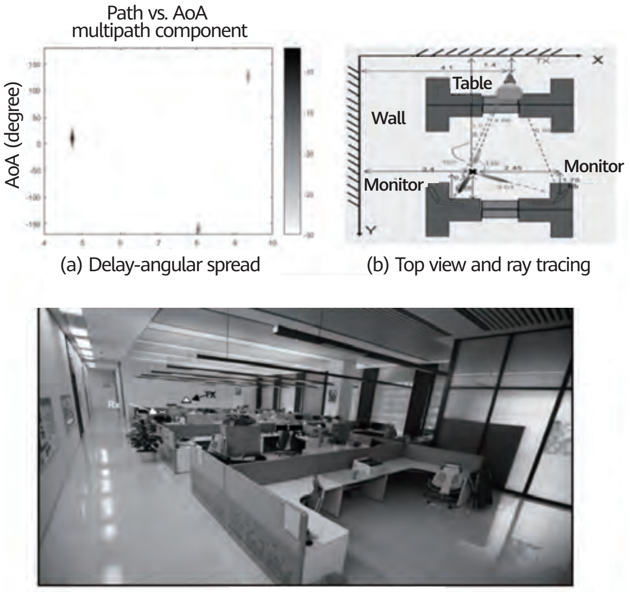


Figure 3 ABG path loss model results at the meeting room

On the small-scale aspect, the THz channel also exhibits different propagation characteristics from the millimeter wave channel. The rough surfaces at the THz band need to be considered carefully as the THz wavelength is comparable to the roughness of object surfaces. The measurements indicate that in the 140 GHz band, multipath components are still rich in the open office due to the abundant furniture. For further study, we choose one receiver position (shown in Figure 4) as an example to analyze the spatial angle of arrival (AoA). We can observe that it is an obviously sparse propagation channel, in which 3 clusters are extracted with 30 dB cutoff threshold. Furthermore, we use the ray tracing mechanism to map the clusters into the geometry space environment in Figure 4b. The propagation paths perfectly match with the geometry map and measurement results. The transmitter antenna beam width is about 30° to cover the entire area. The north direction (upwards direction on the 2D map) is defined as the zero degrees with clock-wise rotation. Based on the geometry reconstruction, from the receiver point of view, the direct path is traced between the TX and RX lines with 8.7° AoA. The second path is coming from the rear left reflected by a monitor screen, and the AoA is –160°. The third path is also a reflection path provided by the rear right monitor screen, and the AoA is 135°. There is at least 90° separation among the three paths. This implies that there are three orthogonal spatial channel streams we can utilize for beamforming design, which is good for beam management and single-user multiple-input multiple-output (SU-MIMO) performance [1].



(a) Delay-angular spread

(b) Top view and ray tracing

(c) Picture of scenario

Figure 4 Indoor hotspot cell (InH) meeting room 140 GHz measurement scenario

To investigate the penetration characteristic of the THz band, we measure 14 different typical materials to study the penetration loss. Three penetration loss types classified based on the analysis results are presented in Table 4. The THz wave can penetrate the carton and cotton coat easily with only several dB, even transparently. This tells us that the THz wave can be used for safety detection application, such as detecting a knife hidden in the pocket. The second category is the typical outdoor-to-indoor (O2I) material, like the single layer glass, and wooden door. The THz wave shows above 10 dB penetration loss which significantly affects the coverage and capacity. The third category we presented is the ultra-high loss for THz waves. This kind of material contains conductive molecules that severely obstruct THz wave propagation.

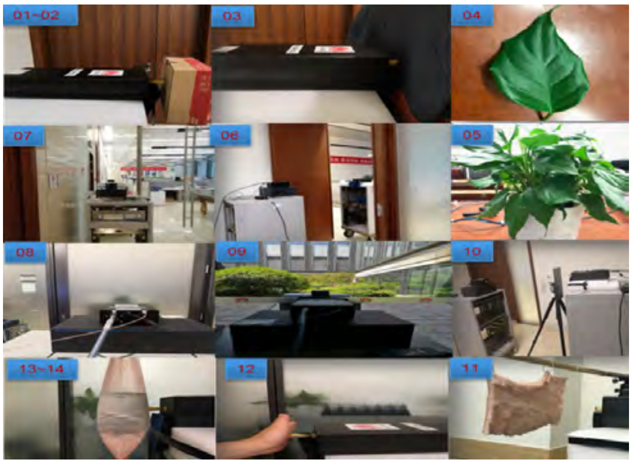


Figure 5 Penetration loss measurement for typical materials

Table 4 Penetration loss of different materials at 140 GHz

Index	Material	Blockage Thickness	Object-Antenna Distance	Penetration Loss (dB)	Loss Model
01	Single -layer carton	2.4 cm	0 cm	2.8	Low loss
02	Double-layer carton	5 cm	0 cm	3.6	
03	Coat	5 mm	0 cm	0	
04	Single leaf	0.5 mm	0 cm	10.6	Middle loss
05	Small indoor vegetation	25 cm	0 cm	29.6	
06	Wooden door	7 cm	50 cm	19.4	
07	Double -layer glass	1.3 cm	50 cm	15.7	
08	Frosted glass	0.6 cm	50 cm	15.2	
09	Double silver low-E tempered glass	1.8 cm	50 cm	55	High loss
10	Metal	3 mm	50 cm	49.6	
11	Pigskin	1.5 cm	10 cm	59	
12	Hand	2 cm	0 cm	59.1	
13	Water	4 cm	0 cm	53.2	
14	Water	7 cm	0 cm	58.4	

A hybrid channel modeling methodology is proposed for the THz communication and sensing. We achieve the THz band sounding system and typical indoor measurement. The SCM-based path loss and multipath components are presented. Based on the analysis results, the THz band shows sparse spatial clustering propagation channel characteristics and sensing capability. In future work, we will investigate the outdoor propagation features for THz band communication and sensing.

4 THz Hardware and Components

In order to meet the requirements of diversified application scenarios in future 6G, it is also necessary to gradually realize the industrialization of THz components and key technologies, so as to realize the large-scale commercialization of the THz communication and sensing systems. The key THz components, THz antenna, intelligent surfaces, and integration technologies are investigated here.

4.1 THz Components

The "THz gap" is due to the lack of compact source and detector technology. So highly pure THz sources, high gain and high power amplifier that operate at the THz band and the highly sensitive THz receivers are key technologies that enable THz applications. What's exciting is that the silicon-based THz components and system have shown a continuing growth in sensing, imaging and communication applications beyond 100 GHz. In addition, by integrating III-V material and device on silicon, the system performance can further be leveraged beyond 500 GHz. Both silicon microelectronic and photonic devices [21] can benefit from this integration approach. The different technologies available to build THz and sub-THz sources are shown in Figure 6.

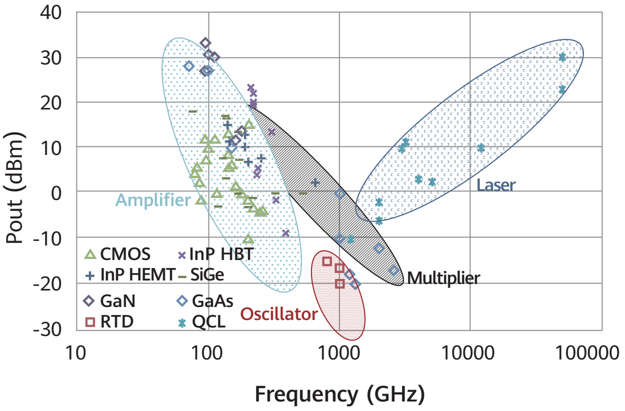


Figure 6 Terahertz gap with respect to source technology

The maximum oscillation frequency $f_{\max}$ of transistors determines the speed of the system. The demand of higher density electronics and faster system performance drives higher CMOS scaling, i.e., higher $f_{\max}$. In conventional CMOS and BiCMOS technology, the $f_{\max}$ of a transistor is between 200 GHz and 350 GHz, i.e., 45–65 nm nodes [1]. With SiGe BiCMOS technology, the transistor can reach $f_{\max}$ of 0.5 THz, i.e., 130 nm node [1]. Beyond $f_{\max}$, the CMOS performance degrades with device scaling. Although nonlinear effects of the device may be exploited to generate harmonic power and detect signals, the efficiency is low. III-V compound semiconductors can drive the $f_{\max}$ far beyond 0.5 THz. InP-based HEMTs can reach $f_{\max}$ of 1.5 THz [22] and double heterojunction bipolar transistors (DHBTs) can bring the $f_{\max}$ up to 1.15 THz [23]. Another example is GaN HEMTs ($f_{max} \approx$ 0.58 THz [24]).

Frequency multiplication and higher harmonic extraction from on-chip oscillators are two common ways to generate THz signals. At the THz band, the planar Schottky diode technology plays a crucial role, and at room temperature, demonstrates powers of 100 µW at 1.2 THz, 15–20 µW at 1.5–1.6 THz, and 3 µW at 1.9 THz. [25–26] provided a comparison of state-of-the-art THz sources in CMOS and SiGe technologies. Sources with both conducted and radiated power were discussed and compared. Due to the parasitic effects at the THz frequency, it is preferred that antennas are integrated on chip to simplify the packing process and prevent unnecessary signal losses. Equivalent isotropically radiated power (EIRP), defined as the product of the radiated power and directivity of the antenna radiation pattern, is used to characterize this type of THz sources. It was shown that by utilizing the power combining technique, the radiating antenna array can significantly increase the output power [27], which is essential for THz beamforming and beam steering application. In a resonant tunneling diode (RTD) oscillator, the on-chip antenna is also directly decoupled for the signal-coupled output, and fundamental oscillation up to 1.98 THz and output power of 0.7 mW at 1 THz by a large-scale array have been reported [28].

Amplification of weak THz signals is a very important function in the system. An effective THz amplifier operates approximately 1/2 of the transistor's $f_{\max}$ and can reach 2/3 of $f_{\max}$ with proper design [29]. Currently, amplifiers using the advanced 35 nm InP HEMT process have achieved 1.1 THz signal amplification. The power amplifiers monolithic microwave integrated circuit (MMIC) designed with the InP DHBT process can output 220 mW power at 220 GHz [30]. Using the three-dimensional (3D) additive fabrication process, a 16-way solid state power amplifier module reaches 820 mW output at 210 GHz, making low-end THz applications possible. Figure 7 demonstrates a GaN HEMT power amplifier and its package operation at 220 GHz. The saturation output power reaches 18 dBm. Compared to the III-V counterpart, the output power and operation frequency of CMOS amplifiers are much lower. The best amplification application for CMOS is at 140 GHz, and the gain can be achieved at 200–300 GHz using positive feedback technology [31]. BiCMOS currently operates at a maximum amplification frequency of 310 GHz, achieving 4 dBm of output power [29].

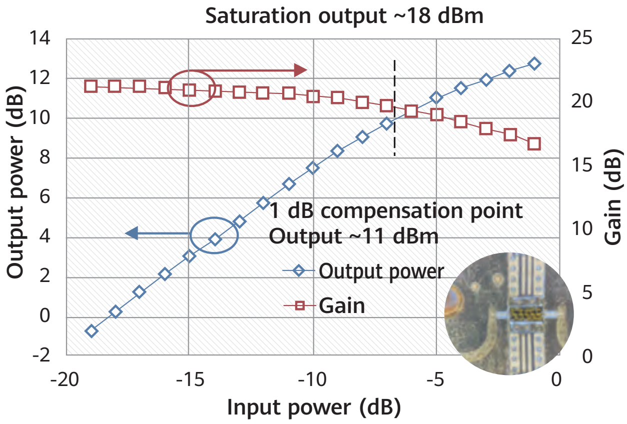


Figure 7 Output performance of 220 GHz GaN HEMT power amplifier

THz receivers can be classified as heterodyne and direct detector receivers. A heterodyne receiver down-converts the THz signal to an IF frequency driven by a local oscillator and it can acquire both phase and amplitude information from the THz radiation, i.e., coherent detection. [1, 26] discussed a variety of THz receivers based on both silicon and III-V technologies, ranging from 200 GHz up to near 1 THz. It showed that InP-based receiver can offer similar (below 300 GHz) or better (beyond 500 GHz) performance in terms of gain and noise figure in general. Coherent receiver has been demonstrated in both THz imaging and communication applications [32].

However, practical implementation of high density 2D on-chip antenna array remains challenging due to the system complexity and high power consumption. Moreover, multiport receiver technology is a desired feature to enable multifunction and multimode THz communication and imaging and sensing application. In [33], multiport receiver technology was reviewed and different multiport architectures were discussed. A 6-port receiver system based on multiport interferometer technique was detailed which is capable of handling AoA detection as well as data communication. This architecture can potentially find its application in future THz joint radar-communication, simultaneous localization and mapping, and imaging/sensing systems.

On the other hand, a direct detector can convert the illuminated THz radiation power to a measurable DC current. The receiver system usually consists of a CMOS integrated field-effect transistor (FET) or SBD with simple antennas such as loop or patches coupled to it and a readout circuitry that rectifies the impinging THz radiation power to a readable DC current. In [1], comparisons of state-of-the-art direct receivers were provided. It showed that direct receivers in general have higher noise power compared to their heterodyne counterparts and are therefore mostly used in THz imaging and sensing application. Direct receivers can better integrate with silicon in large numbers/pixels due to their simple architecture and low power consumption. They have been widely used in THz camera devices when packaged into focal-plane array configuration [34]. The different technologies available to build THz and sub-THz receivers are shown in Figure 8.

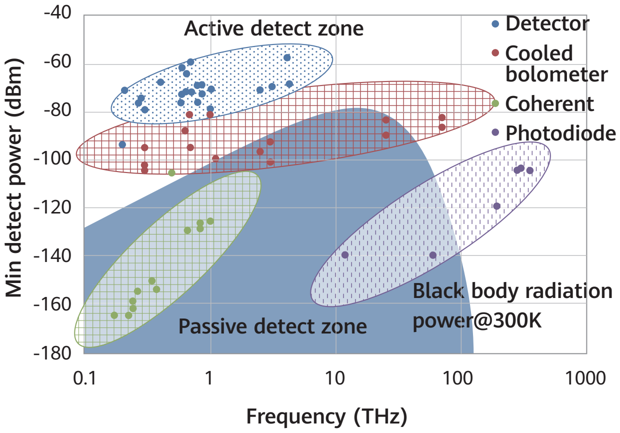


Figure 8 Terahertz receivers with respect to source technology, with 1 Hz resolution bandwidth (RBW)

4.2 THz Antenna

THz "tile-able" array is attractive as it allows for high radiated power by beam manipulations from large-scale THz antenna arrays. In a THz "tile-able" array, large number of antennas are integrated on chip to ensure the structural compactness and high power efficiency to prevent unwanted parasitic loss. In this case, the EIRP is used to indicate the effective output power of the THz source, which combines both the radiated power and directivity of array. High-level integration and scalability are two important considerations in tile design, and smartly using existing on-chip structures has become a promising approach. In [35], a fully "tile-able" array was proposed that uses an existing on-chip slot mesh structure in multiple functional ways, so a high radiated power of 80 μW was achieved. Such structural multi-functionality was further leveraged in [36], in which a de-centralized architecture was proposed and the scale is pushed to be comparable with that of direct detector arrays, however with an approximately 4300x sensitivity improvement. In [37], a densely integrated and unified paralleled amplifier and antenna architecture

was demonstrated. A patch antenna with co-existing topologically paralleled transistors was designed, and it can perform power radiation and amplification simultaneously. The concept was then validated using a standard 65-nm CMOS process. A set of chips were fabricated at 146 GHz and the compact unified prototypes showed an amplified radiation with 3.4 dB gain enhancement through a single element and 6 dB gain enhancement through 2×2 layout. It is also noted that frequency tuning can be achieved by varying bias.

Programmability is another desired feature of THz "tile-able" array. Limited configurability has been demonstrated mainly through electrically, mechanically, or thermally controlled reconfigurable materials [38–39]. The ultimate programmability is one that can configure the transmitted THz fields digitally and receive the THz fields with arbitrary specification. This not only includes beams synthetization with desired characteristics through beam steering or beamforming to enhance the radio performance, but also includes beam with "pixelated" or "voxelated" configuration to form a desired image or video at the user end. In addition, CMOS integration is an important consideration to allow for low-cost fabrication. [40] proposed a THz sensing surface with a log-periodic antenna loaded with 16 distributed detectors. By changing the detector capacitance bank configurations, the antenna is reconfigured to different working states. The system was fabricated with a standard 65 nm CMOS process and tested from a wide frequency range from 0.1–1 THz with responsivity to different angles of direction and polarization. [41] demonstrated a dynamically programmable array made of split-ring resonators loaded with 8 switches fabricated using a 65 nm CMOS process. 256 states (8 bits) were reported combining both amplitude and phase control. The coded surface was shown to project simple letter holography using measured near fields. Though the image resolution is low, this field projection provides a powerful approach for many applications such as sensing, qualitative imaging, and beamforming/beam steering with a silicon compatible approach.

4.3 Intelligent Surface

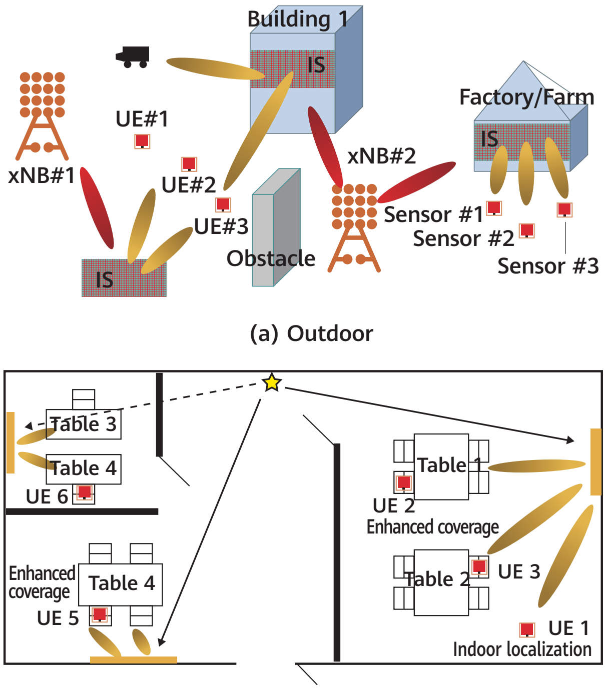
Signal deterioration is one of the major issues in THz communication. The high propagation loss at the THz band results in a very short communication distance. The signal blockage and misalignment impacts are more severe at the THz band. This can affect the THz network coverage and number of user accesses. Moreover, the multipath environment can cause the signal to be "null" at some locations. Therefore, an intelligent wireless system capable of adapting to a time-varying wireless environment is needed to meet these challenges.

Programmable surface is one promising candidate to provide an intelligent and controllable wireless communication system. When applied on the surface of various objects, such as buildings, it can realize various functions such as beamforming and polarization control and provide seamless connections.

In [42], intelligent surface is classified into passive surface (also termed as reconfigurable intelligent surface, RIS) and active surface (also termed as large intelligent surface, LIS). The passive RIS performs some basic functions such as beam reflection, collimation, and polarization. It operates in an energy-efficient way since it is usually composed of low-cost passive components that are self-power-sustaining. The active LIS, on the other hand, performs the RF role partially or fully. It is therefore usually equipped with some RF circuitry and signal processing unit which can be power-consuming. In addition to the basic beam manipulation, active LIS can further amplify the impinging wave, synthesize the desired beam pattern and perform simple signal processing function. Intrinsically, both passive and active intelligent surfaces are made of reconfigurable radiators or scatterers. These radiators can be made of reconfigurable material, such as phase changing material or liquid crystal, or they can be controlled through a programmable interface. Either way, they enable the surface to perform in an "intelligent" way in response to the time-varying wireless communication environment.

From the microscopic perspective, the radiator element can be of simple antenna geometry, such as loop, patch, and wire. By loading the antenna with different passive components, such as varactor diodes, through a digital controller, the working states of antenna can be changed, and therefore its beam pattern can be steerable. In [43], a loop antenna loaded with 8 small extra loops was designed as the radiator element. Each small loop can be digitally controlled with two states ON and OFF, and therefore 8-bit control can be realized through a programmable interface. A chip tiled with 576 such elements was fabricated which provided both amplitude and phase control, dynamic beamforming and multi-beam formation at 0.3 THz. It was also demonstrated that the surface can project simple holographic letter images qualitatively.

Figure 9a shows a wireless communication environment enabled by intelligent surfaces. When installed on the exterior of buildings, intelligent surfaces can be used to create connections between buildings, vehicles, or automated guided vehicles (AGVs) in cases where there is no direct link between them or the link is blocked by obstacles. Intelligent surfaces can further extend the radio coverage from outdoor base stations to indoor users. By programming their working states, intelligent surfaces can perform beamforming and direct its beams to the target end users dynamically and relay information to the desired locations with attenuation compensation. Beamforming from intelligent surfaces can also help transfer power to Internet of Things (IoT) devices and sensors.



(a) Outdoor

(b) Indoor

Figure 9 Typical RIS use cases for THz

When deployed in an indoor environment as shown in Figure 9b, e.g., attached to a wall, intelligent surface can help direct the signal to the target user locations where the signal experiences multipath fading and path loss due to wall blocking and scattering from furniture, plants, etc. Intelligent surfaces can also be used for high precision indoor localization as its large surface aperture can help increase location precision.

In another embodiment, the scattering elements can be of high-dielectric-index nanopillars or nanofins. Each element is modified geometrically to manipulate the amplitude, phase, and polarization of the incident EM wave. In [44], the working mechanism underlying this type of metasurface was discussed. The phase modulation of the transmitted wave is due to different propagation constant of the nanopillars. For example, the induced geometric phase can occur in crossed polarized light linearly with respect to the orientation angle of the nano-element [44]. This provides a possibility for optical holography application. [45–46] demonstrated that this all-dielectric metasurface can be used to encode the hologram by using unit element with varying orientations. Both amplitude and phase information can be recorded and controlled independently. Then, by collecting the transmitted light at the image plane, the original object can be faithfully reconstructed pixel by pixel using a standard computer-generated hologram algorithm.

Though the optical metasurface shows its ability in holography application with high fidelity image reconstruction, these surfaces are static after fabrication. A dynamic control of the holograms is desired to achieve a true holographic display. Although the concept of programmable metasurface and reconfigurable material such as phase changing material can be applied here similarly, it is still very challenging to configure pixel level wavefront representation that can reflect both phase and amplitude information of the original object dynamically in the visible spectrum. Multiplexed metasurface is another way to address this problem. There are various multiplexing methods. Wavelength division multiplexing [47–48] uses nanostructures that are multiplexed in a subwavelength scale and capable of manipulating wavefronts of multiple frequencies. Angle and polarization multiplexing can respond to light of different angles of incidence and polarization. In [49], an orbital angular momentum (OAM) holographic metasurface capable of reconstructing a range of OAM-dependent holographic images was demonstrated using a single meta-hologram with high spatial-resolution. It showed that incident OAM beams of 4 different modes can independently reconstruct distinctive holographic images of alphabet letters from the same multiplexed OAM meta-hologram. Recently, [50] demonstrated a space-multiplexed metasurface that can achieve 2^{28} different holographic frame/image at a maximum rate of 9523 frames per second. In [50], the entire metasurface is divided into many different sub-regions which are combined at different times in a specified configuration modulated by a high-speed dynamic structured laser beam modulation module to project images like an electronic meter. Strings containing digits (0 to 9) and letters can be fully reconstructed and displayed in a meaningful way using this approach.

4.4 Integration Technologies

A crucial element in THz system is the packaging and integration technology. The most important parameters are losses and reflections in the chip-to-substrate transition. At present, the commonly used metal module package has lower integration level, and higher cost. It will be replaced by high-density integrated technologies in the future. Multichip module (MCM), system-in-package (SiP), and heterogeneous integration are promising candidates. MCMs built on high temperature co-fired ceramic (HTCC) or low temperature co-fired ceramic (LTCC) substrates have been used in THz packages. Antenna and silicon lenses are integrated in the package to reduce connection loss and enhance system EIRP [51]. The through-silicon via (TSV) process has better integration and process precision and can be used at higher frequencies [52]. The embedded wafer level ball grid array (eWLB) technology usually used at low frequency (with interposer or distribution layers) can also be used in lower end of THz [53] as SiP technology.

Silicon-based integrated circuits (ICs) prevail due to their low cost and high level of on-chip integration; III–V compound semiconductors represented by GaAs and GaN can provide a higher transmission power. A heterogeneous integration platform can provide better performance, e.g., higher output power, while still retaining the silicon's advantage. Wafer-level integration using Benzocyclobutene (BCB) provides a 2D integration method [54]. The alternative approach bonds the wafer or die of III-V materials onto a patterned silicon wafer, for instance, a 3D BCB-based wafer bonding integration scheme or wafer-scale low-temperature oxide-to-oxide bonding [55]. These methods seem promising as they retain the silicon's advantages while leveraging the high power and high frequency operation ability of III-V semiconductor.

5 THz System and Testing

In this section, our contributions on THz communication and sensing systems are described in detail, which includes link simulations, testing and result analysis of THz communication and sensing systems and prototypes.

5.1 THz Communication System

Recent technology progress in electronic, photonic and material technologies are closing the gap in THz transceiver design. Consequently, THz signal generation, modulation, and radiation methods are converging, and corresponding channel model, noise cancellation, and hardware-impairment compensation and ultra-wideband signal processing techniques for wireless communications are also emerging.

As shown in Figure 10, there are so many significant technical differences between normal frequencies and the THz band as result of channel propagation characteristics, e.g., large atmospheric propagation loss, strong directivity, and ultra-narrow beams. They limit signal coverage and mobile access. The impairment characteristics of broadband RF device, e.g., strong phase noise, frequency selective memory in-phase and quadrature-phase imbalance (IQI) and in-band flatness, require ingenious algorithm design, and ultra-wide bandwidth requires an ultra-high speed analog-to-digital converter (ADC)/digital-to-analog converter (DAC) conversion rate. So from baseband to RF, the design of a complete THz communication system is faced with great technical challenges.

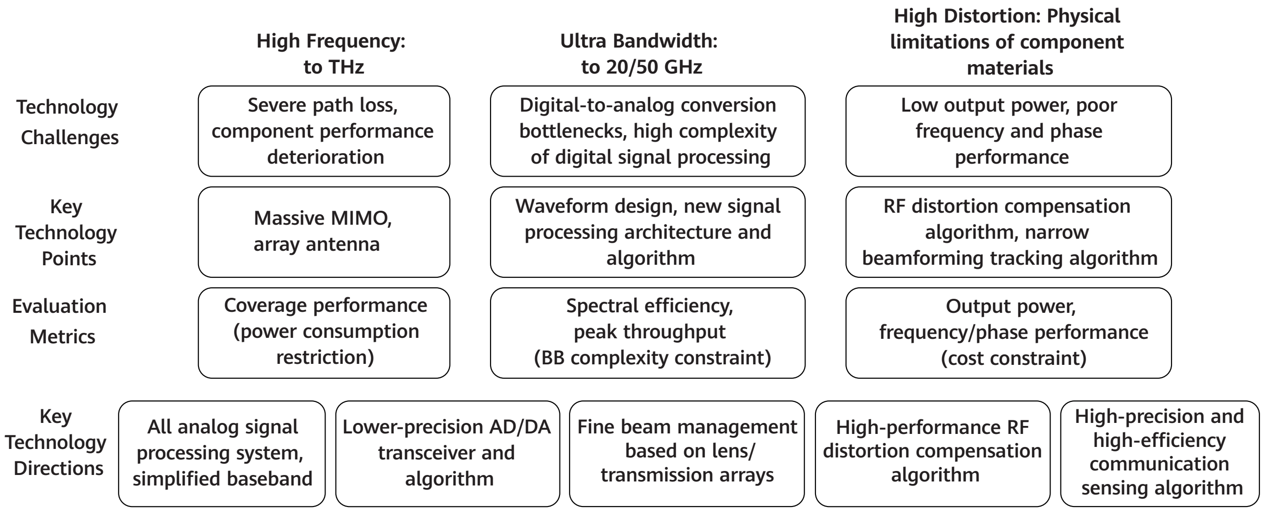
As mentioned previously, it is valuable to explore novel signal processing architectures, waveform design, and corresponding compensation algorithms to solve the challenges of Ultra-High Frequency and Ultra-Wideband.

Up to now, because of the absence of experimental platforms for true THz communications, the majority of THz-band communication works are mainly theoretical and limited experimental validation. In this paper, our THz communication platform, i.e., the integrated testbed for ultra-broadband wireless communications at 220 GHz frequency, is presented, and the block diagram of the testbed is illustrated in Figure 11. As shown, advanced spatial and polarization multiplexing technologies are used to improve spectral efficiency.

The THz communication system consists of an RF transmitter and an RF receiver. At the transmitter, the data bits are organized in frames, modulated into symbols, undergo pre-equalization, pulse shaping, fractional delay pre-compensation, and digital IF modulation, and are fed to the high-speed DAC board. Then analog signal with a center frequency of 12.5 GHz is output. Analog IF signal is connected with the THz analog Front End.

At the receiver, the well-designed digital baseband physical algorithm is employed, the received THz signals from analog front end are digitized by the high-speed ADC boards which are synchronized, and undergo channel estimation and equalization, phase noise estimation and cancellation, interference cancellation, nonlinearity compensation, demodulation and decoding.

Field trial experiment was conducted with 2 × 2 polarization-MIMO. The field trial experiment was conducted at Chengdu, China. The TX/RX link distance is 330 m from the rooftop to the ground, and the channel is almost line-of-sight as shown in Figure 12 [56].



	High Frequency: to THz	Ultra Bandwidth: to 20/50 GHz	High Distortion: Physical limitations of component materials
Technology Challenges	Severe path loss, component performance deterioration	Digital-to-analog conversion bottlenecks, high complexity of digital signal processing	Low output power, poor frequency and phase performance
Key Technology Points	Massive MIMO, array antenna	Waveform design, new signal processing architecture and algorithm	RF distortion compensation algorithm, narrow beamforming tracking algorithm
Evaluation Metrics	Coverage performance (power consumption restriction)	Spectral efficiency, peak throughput (BB complexity constraint)	Output power, frequency/phase performance (cost constraint)

Key Technology Directions	All analog signal processing system, simplified baseband	Lower-precision AD/DA transceiver and algorithm	Fine beam management based on lens/ transmission arrays	High-performance RF distortion compensation algorithm	High-precision and high-efficiency communication sensing algorithm

Figure 10 THz communication technical challenges

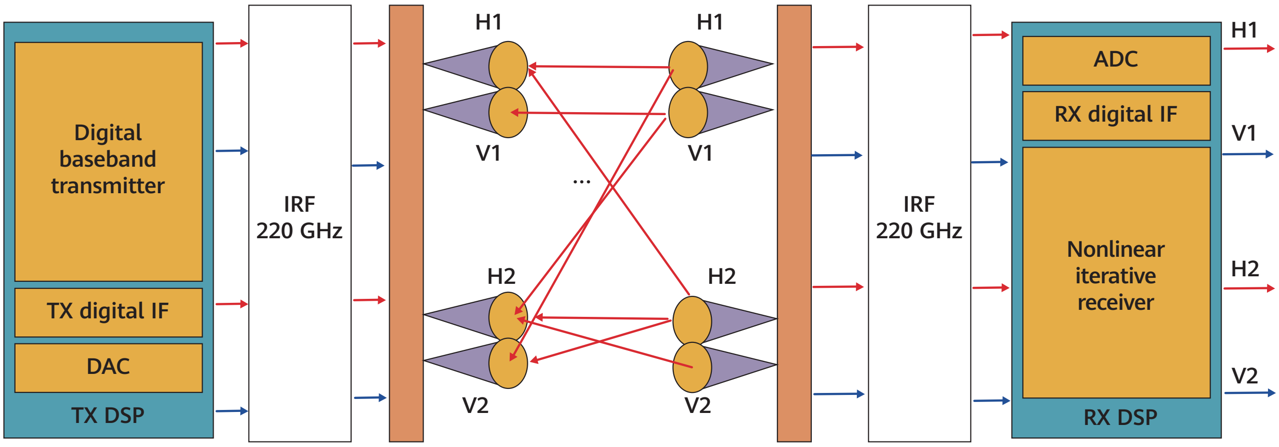


Figure 11 THz communication testbed system architecture

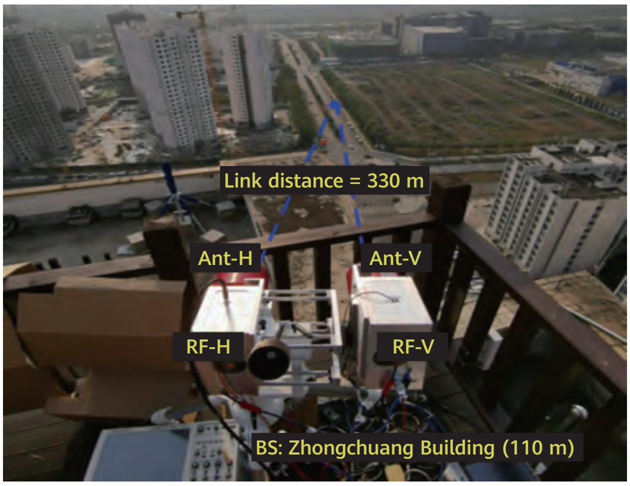


Figure 12 THz field trial, 2 × 2 polarization-MIMO system

In order to compensate the constant frequency selective response of the THz RF components, a digital pre-equalization (DPEQ) filter at the transmitter is implemented. The frequency response H(k) is calculated by comparing the IF signals transmitted and received within the bandwidth of the system components:

$$H(k) = \left(\frac{P_r(k) - P_n}{P_s(k)}\right)^{1/2} \qquad (3)$$

Where $P_r(K)$ is the received signal power with noise at the kth frequency, $P_s(K)$ is the transmitted signal power, and P_n is the noise power for the whole observation bandwidth.

We theoretically and experimentally tested a time-domain DPEQ scheme for wide-bandwidth THz communication systems, which is based on the feedback of channel characteristics from the receiver-side blind and adaptive equalizers. Based on the proposed DPEQ scheme, we theoretically and experimentally studied its performance in terms of various channel conditions as well as resolutions for channel estimation. Besides, the significant improvement in channel flatness and mean squared error (MSE) performances were also demonstrated.

The channel frequency response curve is shown in Figure 13, indicating that the flatness in the signal band is irregularly fluctuating with about 16 dB maximum fluctuation in 12 GHz bandwidth. When pre-equalization is enabled, the flatness is compensated and performance improvement is noticeable.

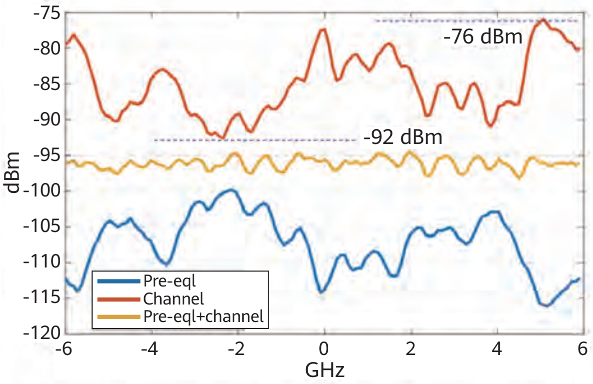
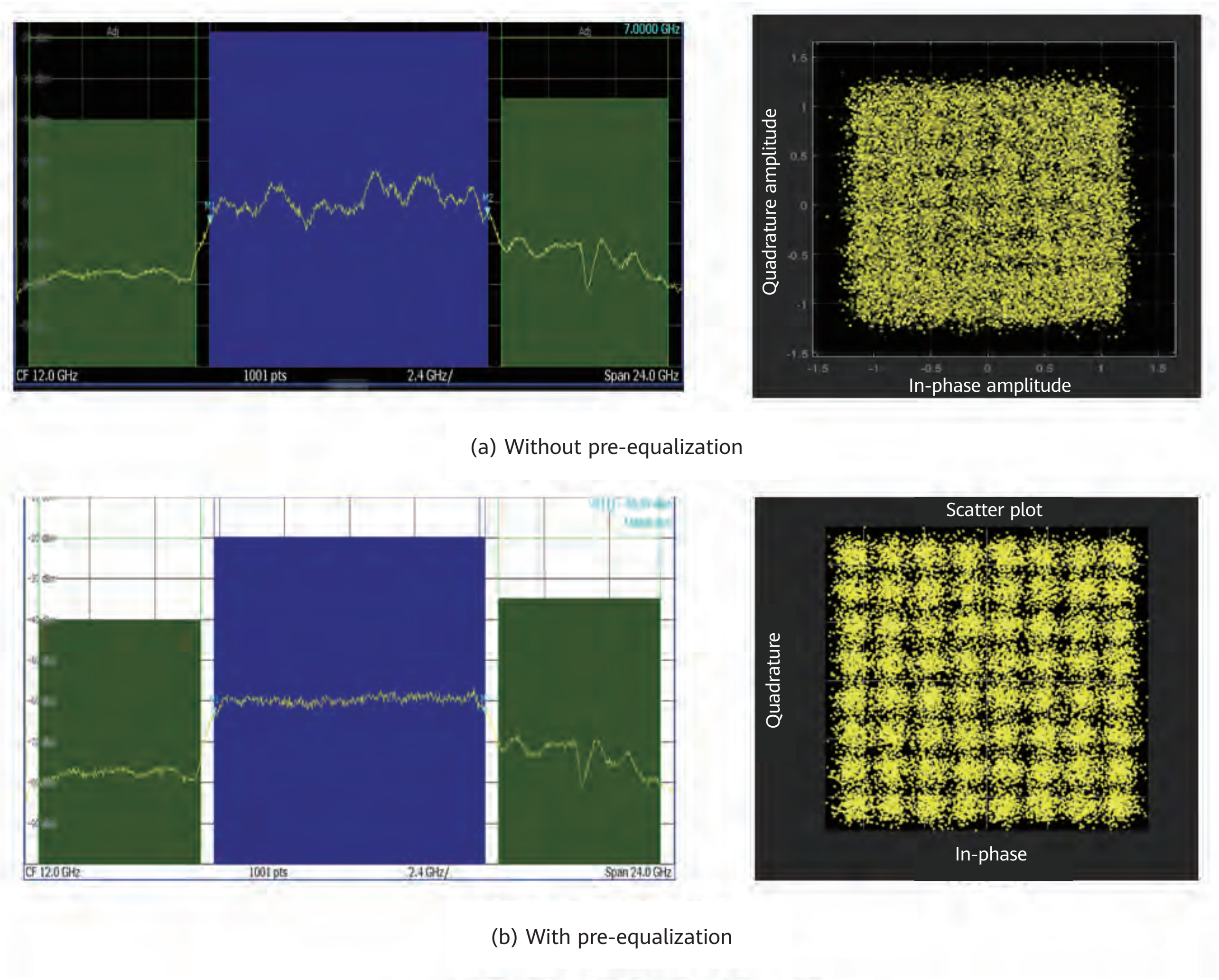


Figure 13 Channel and pre-equalization power spectral density (PSD)

The corresponding demodulation constellation is shown in Figure 14. Pre-equalization at the transmitter does not amplify channel noise, and equivalent signal-to-noise ratio (SNR) is improved. Therefore, resolution of constellation map is increased even for higher order modulation, e.g., 64QAM.

To maximize spectral efficiency and increase THz link capacity, polarization multiplexing is considered in our prototype. The vertically or horizontally polarized waves mean that the electrical field is oscillating in the vertical or horizontal direction respectively. Two ideal polarized antennas results in two independent channels doubling the capacity of the system.



(a) Without pre-equalization

(b) With pre-equalization

Figure 14 PSD and demodulation constellation

However, a real system always experiences imperfections, such as crosstalk between polarizations. The interference between the signals inevitably occurs because of cross-polarization discrimination (XPD) of the antenna and channel degradation. This is because antenna polarization is not ideally isolated. Different polarizations may have different propagation characteristics in different channel scenarios (e.g., under raining environment), resulting in polarization leakage between channels. This leakage can be quantified using the channel XPD factor. It describes how much power from one polarization leaks into another polarization, thus reducing the system's ability to separate between the two polarizations. It is defined for the vertical and horizontal components respectively as [57]

$$XPD_V = \frac{E\{|h_{V,V}|^2\}}{E\{|h_{H,V}|^2\}},\ XPD_H = \frac{E\{|h_{H,H}|^2\}}{E\{|h_{V,H}|^2\}} \qquad (4)$$

where $h_{V,H}$ is the flat channel impulse response between the vertically polarized transmitter and horizontally polarized receiver with the subscripts V and H representing the vertical and horizontal antennas.

The XPD measurement is shown in Figure 15. The recording operation was performed 10 times during one day to measure the receive powers of two polarizations. The results show that mean XPD is about 19 dB with ±2 dB fluctuation influenced by mechanical deformation and beam misalignment.

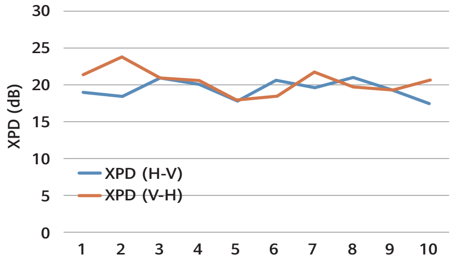


Figure 15 XPD measurement

To eliminate this interference, cross-polarization interference cancellation (XPIC) technology is used to receive signals horizontally and vertically. The signals in the two directions are then processed, and the original signals are recovered from the interfered signals. The assignment of the same frequency to both the vertical and horizontal polarization on a link is allowed.

The consistency of RF components is difficult to achieve since the signal bandwidth is very wide, in-band channel characteristics difference between H polarization and V polarization is significant, and the impairment between the effect of XPD and the frequency selectivity of the channel is coupled. It is necessary to design an ultra-large bandwidth polarization interference cancellation algorithm.

Polarization interference cancellation performance is shown in Figure 16. It can be observed that the contrast convergence curve in time domain is stable, and it has about a 2 dB performance gain.

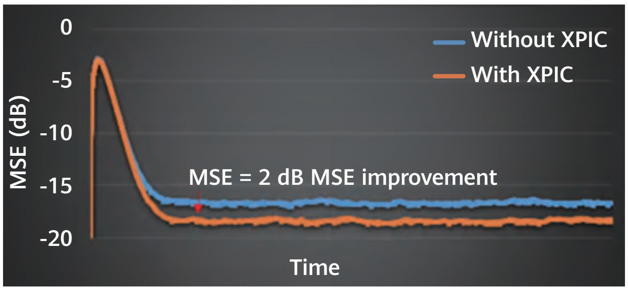


Figure 16 Polarization interference cancellation performance

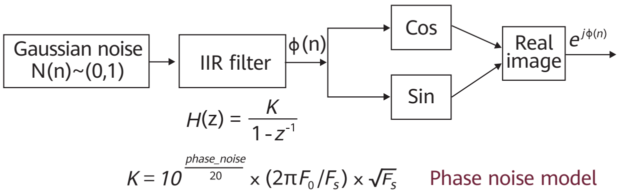
Classical coherent architectures are combined with high spectral efficiency schemes. This entails numerous constraints on the design of RF components especially at the oscillator level. Indeed, high frequency oscillators severely impair THz systems with phase noise [58].

There are several different methods to model the random process of phase noise, such as the well-known Wiener random process and Gaussian random process. In our design, to model the influence of phase noise, a zero-mean White Gaussian Noise is first generated and then is passed through an infinite impulse response (IIR) filter. After that, the filtered noise is added to the angle component of the input signal. This generation process is shown in Figure 17, in which F_0 is the frequency offset, *phase_noise* is the prescribed phase noise level at the frequency offset F_0 , F_s is the sampling frequency, and K is the gain factor controlling the phase noise level at the frequency offset F_0.

$$\phi(n) = \phi(n-1) + Kw(n) \tag{5}$$

The power spectrum of the phase noise $e^{j\phi(n)}$ is equal to

$$\mathrm{P}(f) = \frac{1}{F_s} \cdot \frac{1 - e^{-K^2}}{1 + e^{-K^2} - 2e^{-K^2/2}\cos(2\pi \frac{f}{F_s})} \cdot \mathrm{rect}(\frac{f}{F_s}) \tag{6}$$



Gaussian noise
N(n)~(0,1)
IIR filter
ϕ(n)
Cos
Sin
Real image
$e^{j\phi(n)}$
$H(z) = \frac{K}{1 - z^{-1}}$
$K = 10^{\frac{phase\_noise}{20}} \times (2\pi F_0/F_s) \times \sqrt{F_s}$
Phase noise model

Figure 17 Phase noise model

Normally, in a multi-channel system, the phase noise at each channel will be independent with each other due to the distributed local oscillator at each channel. It will degrade the system performance, and the new phase estimation and compensation schemes will be involved to solve this problem. A specially constructed pilot code is used to project the mixed phase noise onto the space-time orthogonal code space. The distributed Master and Slave phase-locked loop (MS-PLL) and quasi-linear interpolation phase noise suppression (PNS) algorithms are used to track and compensate the phase noise in the multi-channel signal space dimension, which is a low overhead (< 5% pilot proportion) solution, and is able to effectively suppress typical distributed independent phase noise.

Although THz frequency oscillators have an absolute strong phase noise, a higher symbol rate means a shorter symbol switching time, and a stronger phase noise correlation between consecutive symbols.

As is shown in Figure 18, the powerful MS-PLL architecture can track and compensate for the impact of phase noise greatly. When MS-PLL is enabled, the pilot interval is changed, and the impact is not obvious. Compare pilot interval 16 with 256, the performance gap is not more than 1 dB.

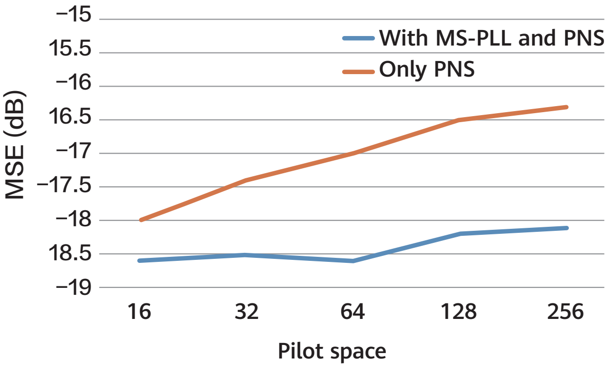


Figure 18 PNS performance

Outlook

Due to the narrow beam (3 dB beam-width is only 1°), mechanical installation and antennas alignment are particularly important to support sufficient receive power. In Figure 19, the transmit power is 16 dBm, and the antenna gain is 43 dBi. According to the link budget, for a 330 m link distance, the RX antenna ports should receive power of −43.6 dBm. Considering the line loss and atmospheric absorption, the measured receive power is −46.7 dBm.

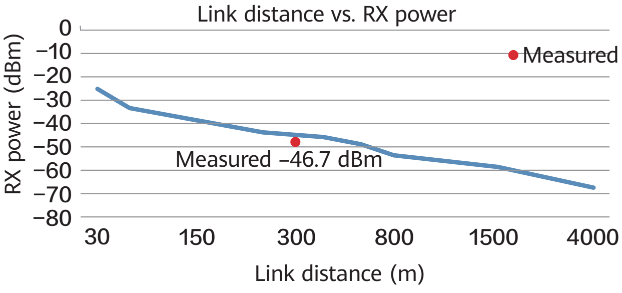


Figure 19 Receive power vs. Link distance

For medium distance outdoor transmission experiments over a distance of 330 m, we connected small 43 dBi lens antennas to the THz-wireless front-ends. At the receiver, the modem implements digital signal processing (DSP) to mitigate the effects of transmission impairments, considering both single-carrier and orthogonal frequency-division multiplexing (OFDM). The measurement results show different Baud rate from 4 GBd to 17.5 GBd. The reference curves correspond to a bit error rate (BER) at the experimentally used soft-decision forward error correction (SD-FEC) threshold of 2.1E-2, which can be achieved assuming error-free decoding with 20 percent overhead [59]. From the performance comparison results, single-carrier is better than OFDM since the former has a small peak-to-average ratio and is less sensitive to phase noise.

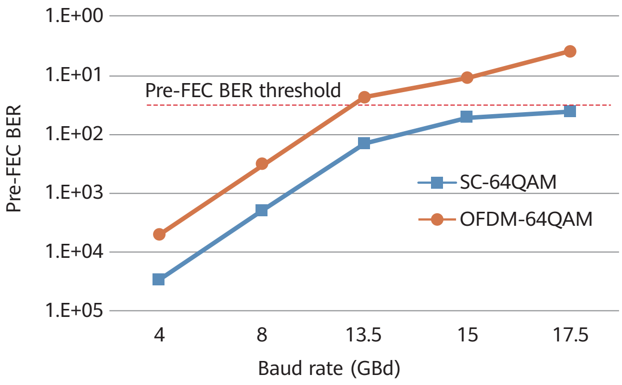


Figure 20 Pre-BER performance

Maximum throughput is 210 Gbit/s (17.5 x 4 x 2), and net data rate is 168 Gbit/s (210 x (1 – 0.2)). After removal of the forward error correction (FEC) overhead, the corresponding demodulated constellation is shown in Figure 21.

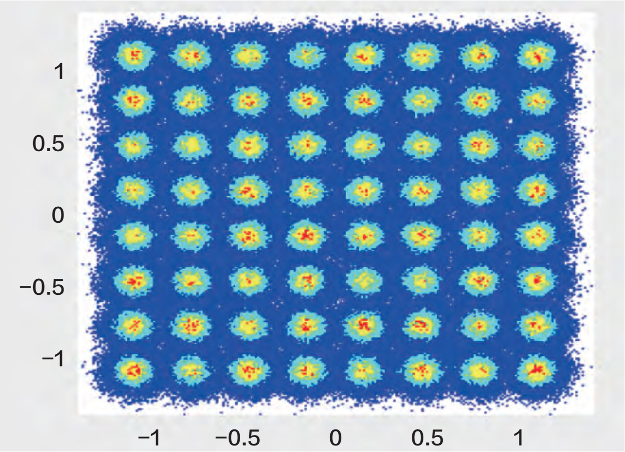


Figure 21 64QAM demodulated constellation

In addition, due to sufficient transmit power, high-gain antenna, and high-sensitivity DSP algorithm, the long distance field trial for single-input single-output (SISO) is also considered in our verification. As shown in Figure 22, the link distance is 3.6 km with high-humidity weather.

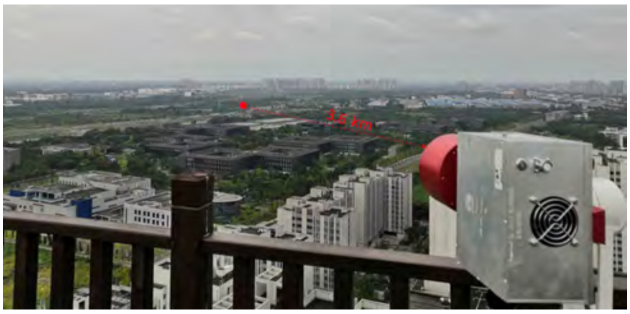


Figure 22 3.6 km long distance test

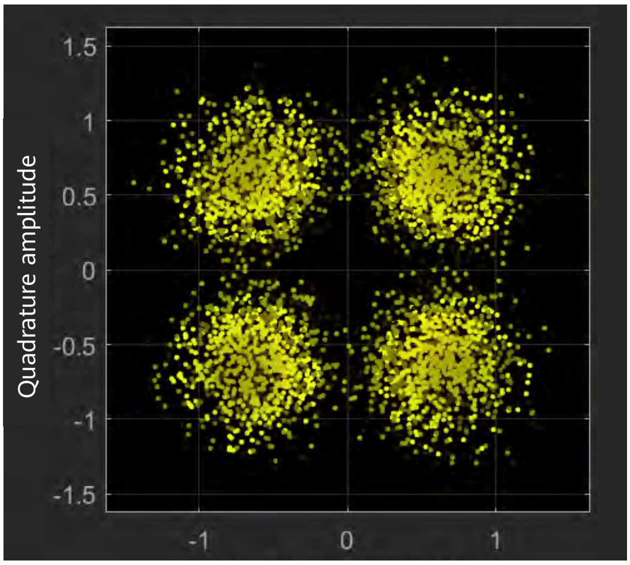


Figure 23 3.6 km, demodulation constellation

The corresponding link budget can be found in Table 5.

Table 5 Corresponding link budget

Parameter	Value
Frequency (GHz)	220
Symbol rate (Gbaud)	175
Transmitter EIRP [dBm]	50.5
Antenna gain [dBi]	42.0
Noise [dBm]	-71.7
RX power@3.6 km, dBm	-64.8

Maximum throughput is 35 Gbit/s, and net data rate is 28 Gbit/s after removal of the FEC overhead.

In order to explore the goal of realizing the ISAC in THz, we use a similar system architecture and device to carry out the sensing experiment. By using the concept of virtual MIMO and compression sensing algorithms, the EM imaging of metal objects hidden in a paper box is rebuilt successfully, and mm-level imaging resolution is achieved.

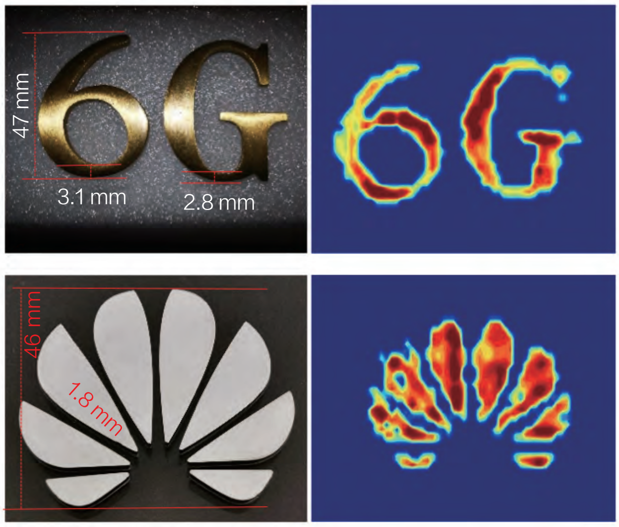


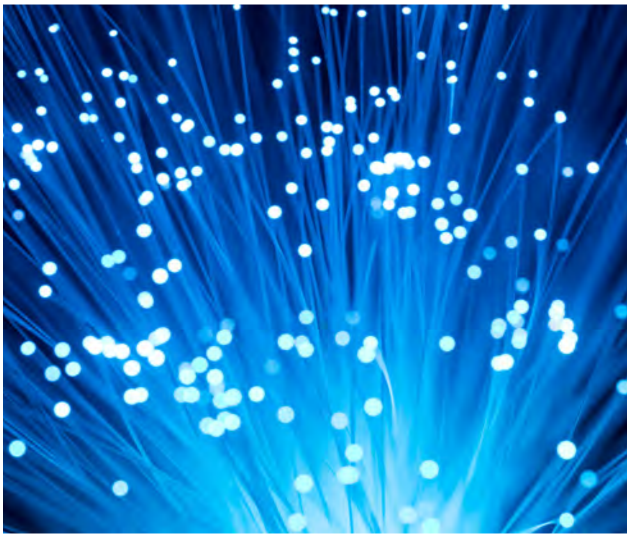
Figure 24 Imaging results of non-sparse full aperture scanning

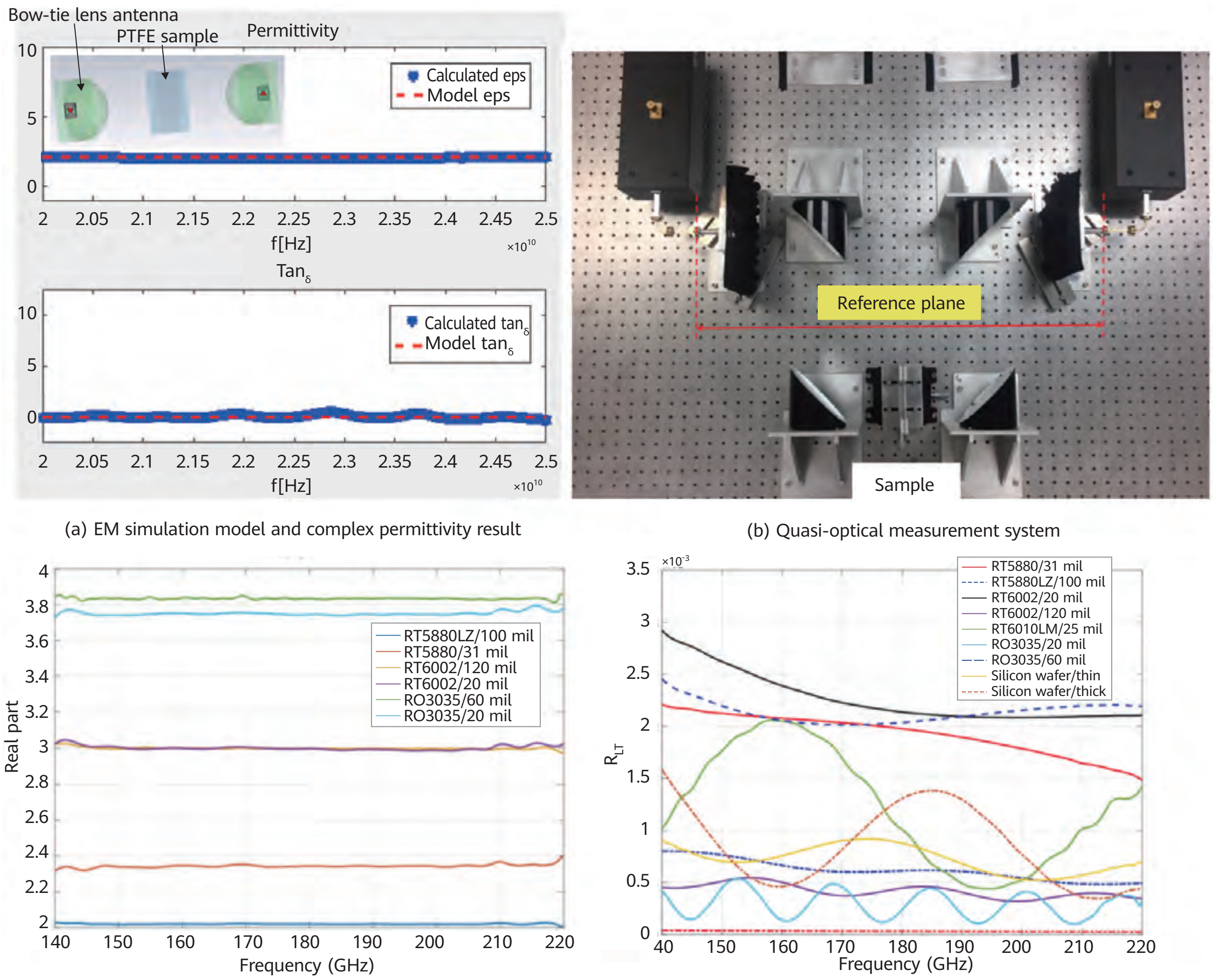
5.2 THz Sensing System

Material characterization is a potential THz application that can be used to study properties of dielectric materials. The THz-TDS allows non-invasive measurement of various material parameters through some mathematical operations, which is by sending a broadband pulse signal to the material sample, and measuring the output signal either in transmission or reflection mode. We developed both simulation model and measurement setup for THz-TDS material characterization application.

CST Microwave Studio was used as simulation tool for THz-TDS setup. In the EM model as shown in Figure 25, two lens bow-tie antennas were designed and placed at each side of a dielectric Polytetrafluoroethylene (PTFE) sample. Then a time-domain simulator was set up and both simulations of with and without the sample were performed. After the simulation, both reference signal (without sample) and output signal (with sample), together with the input pulse, were collected and sent to an optimization algorithm to solve for the material properties, i.e., permittivity and loss tangent. Here, the Nelder-Mead algorithm was applied at each frequency to solve for the material property. It is shown that both permittivity and loss tangent extracted from the signal agree well with the theoretical values.

[60] demonstrated a quasi-optical system that performs complex material property measurement at sub-THz. As shown in Figure 25, a set of two-parabolic-mirror system and four parabolic-mirror system was developed. Two 80 mm-length corrugated horn antennas were designed to achieve a wide plane wave zone. After obtaining the S parameters at both ports, a closed mathematical form expression based on multiple reflection models was applied to calculate the complex material property. It is shown that the complex permittivity of various Rogers RT/duroid series printed circuit board (PCB) substrates had agreed well with the literature.





(a) EM simulation model and complex permittivity result

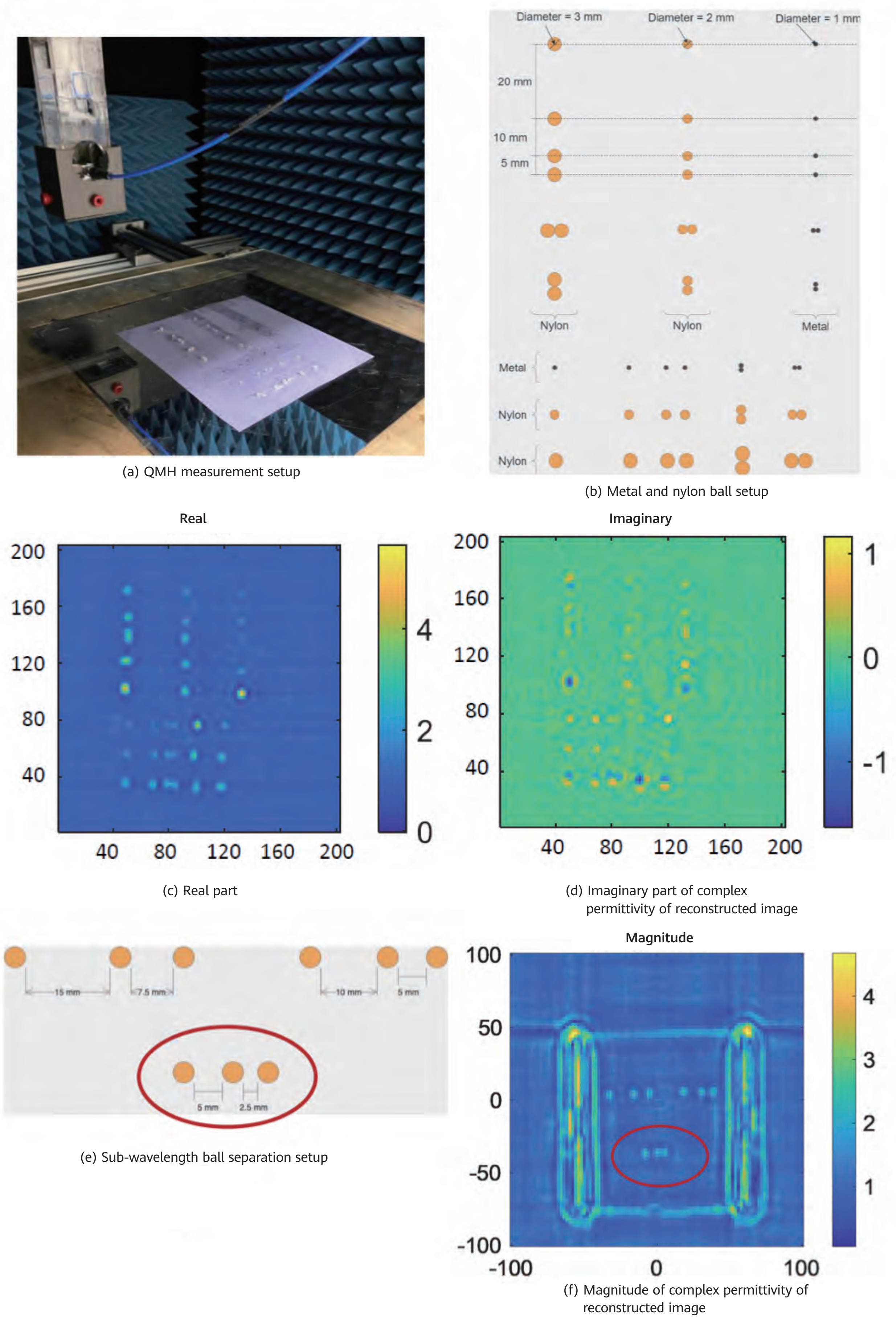
(b) Quasi-optical measurement system

(c) (d) Measurement results of complex permittivity

Figure 25 Simulated and measurement results of material characterization application; measurement from [60]

THz imaging is promising for biomedical applications, due to the non-ionizing THz radiation. Fast and efficient image reconstruction algorithms can help accelerate the imaging acquisition speed. In this paper, a qualitative microwave holography (QMH) imaging method was demonstrated to perform the imaging and material mapping application [61–62]. QMH is a real-time direct inversion algorithm that can reconstruct the object image from all the S parameter measurements on the image plane. The S parameters are then used in two linearization models, Born's and Rytov's approximations, to reconstruct the object image and map its complex material property.

The following image test-bed was then set up to validate the QMH method, as shown in Figure 26. Nylon and metal balls of various sizes were set up, with diameters ranging from 1 mm to 3 mm and separation distance from 5 mm to 20 mm. Four S parameters were collected in 2-port measurements with frequency sweep from 26 GHz to 40 GHz. It is shown that QMH can achieve spatial resolution close to λ/4, even under a far field measurement setup. A comparison of Born and Rytov approximation in image qualities can be found in [62–63].



(a) QMH measurement setup

(b) Metal and nylon ball setup

(c) Real part

(d) Imaginary part of complex permittivity of reconstructed image

(e) Sub-wavelength ball separation setup

(f) Magnitude of complex permittivity of reconstructed image

Figure 26 QMH reconstruction of metal and nylon balls of various sizes. Results from [62]

6 Conclusion

In this paper, we discussed the advantages and typical scenarios of THz communication and sensing application. We also proposed a hybrid channel modeling framework to improve the modeling accuracy and efficiency at the THz frequency. In particular, the THz subsystem with silicon and III-V compound semiconductor materials heterogeneous integration is analyzed and proposed to improve the performance by using the advantage of different processes and materials. Finally, the prototype and measurement campaigns were conducted to illustrate the advantages of THz frequency for high throughput communication and high resolution sensing scenarios. A variety of measurement campaign examples show 210 Gbit/s data transmission rate at 330 m distance, and up to 3 mm invisible imaging, which achieves the highest performance in this field.

Future work will concentrate on following topics:

- Unified channel modeling frameworks and parameters compatible for all frequency bands and all application scenarios
- Unified air-interface and signal processing frameworks
- Technologies to improve the operating frequency and output power of components, and the low cost large array solutions at the THz frequency
- Real-time prototype and field trial with multiple points to further study the performance advantages of THz frequency for future 6G

References

[1] W. Tong and P. Zhu, "6G: the next horizon, from connected people and things to connected intelligence," Cambridge University Press, 2021.

[2] Zheng, Le, *et al.*, "Radar and communication coexistence: an overview: a review of recent methods," *IEEE Signal Processing Magazine*, vol. 36, no. 5, pp. 85-99, 2019.

[3] Mazahir, Sana, Sajid Ahmed, and Mohamed-Slim Alouini, "A survey on joint communication-radar systems," 2020. https://doi.org/10.3389/frcmn.2020.619483.

[4] Li O, He J, Zeng K, *et al.*, "Integrated sensing and communication in 6G A prototype of high resolution THz sensing on portable device," *2021 Joint European Conference on Networks and Communications & 6G Summit (EuCNC/6G Summit)*, 2021, pp. 544-549.

[5] Kürner T, "Turning THz communications into reality: status on technology, standardization and regulation," *2018 43rd International Conference on Infrared, Millimeter, and Terahertz Waves (IRMMW-THz)*. Nagoya, 2018, pp. 1-3.

[6] Tan Danny Kai Pin, *et al.*, "Integrated sensing and communication in 6G: motivations, use cases, requirements, challenges and future directions," *2021 1st IEEE International Online Symposium on Joint Communications & Sensing (JC&S)*. IEEE, 2021.

[7] P. Almers, *et al.*, "Survey of channel and radio propagation models for wireless MIMO systems," J Wireless Com Network, 2007.

[8] W. C. Chew, J. M. Jin, E. Michielssen, and J. Song, "Fast and efficient algorithms in computational electromagnetics," Boston, MA: Artech House, 2001.

[9] Sengupta K, Nagatsuma T, and Mittleman D M, "Terahertz integrated electronic and hybrid electronic-photonic systems," *Nature Electronics*, vol. 1, no. 12, pp: 622-635, 2018.

[10] Xu Y, Unseld F K, Corna A, *et al.*, "On-chip integration of Si/SiGe-based quantum dots and switched-capacitor circuits," *Applied Physics Letters*, vol. 117, no. 14, 2020.

[11] Mokkapati S and Jagadish C, "III-V compound SC for optoelectronic devices," *Materials Today*, vol. 12, no. 4, pp: 22-32, 2009.

[12] Jia S, Zhang L, Wang S, *et al.*, "2 × 300 Gbit/s line rate PS-64QAM-OFDM THz photonic-wireless transmission," *Journal of Lightwave Technology*, vol. 38, no. 17, pp: 4715-4721, 2020.

[13] Niu Z, Zhang B, Wang J, *et al.*, "The research on 220GHz multicarrier high-speed communication system," *China Communications*, vol. 17, no. 3, pp: 131-139, 2020.

[14] Neu J and Schmuttenmaer C A, "Tutorial: an introduction to terahertz time domain spectroscopy (THz-TDS)," *Journal of Applied Physics*, vol. 124, no. 23, 2018.

[15] Final Acts WRC-19, https://www.itu.int/pub/R-ACT-WRC.14-2019.

[16] Technology trends of active services in the frequency range 275-3000 GHz, https://www.itu.int/pub/R-REP-SM.2352.

[17] Xianjin Li, Jia HE, Ziming Yu, *et al.*, "Integrated sensing and communication in 6G: the deterministic channel models for THz imaging," pimrc 2021.

[18] He J, Chen Y, Wang Y, *et al.*, "Channel measurement and path-loss characterization for low-terahertz indoor scenarios," arXiv preprint arXiv:2104.00347, 2021.

[19] Y. Chen, C. Han, Z. Yu, and G. Wang, "140 GHz channel measurement and characterization in an office room," ICC 2021 - IEEE International Conference on Communications, 2021, pp. 1-6, doi: 10.1109/ICC42927.2021.9500596.

[20] Y. Chen, Y. Li, C. Han, Z. Yu, and G. Wang, "Channel measurement and ray-tracing-statistical hybrid modeling for low-terahertz indoor communications," arXiv:2101.12436

[21] J. M. Ramirez *et al.*, "III-V-on-silicon integration: from hybrid devices to heterogeneous photonic integrated circuits," *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 26, no. 2, pp. 1-13, March-April 2020, Art no. 6100213, doi: 10.1109/JSTQE.2019.2939503.

[22] Mei, X. *et al.*, "First demonstration of amplification at 1 THz using 25-nm InP high electron mobility transistor process," *IEEE Electron Dev. Lett*. 36, 327-329 (2015).

[23] Urteaga, M., Grifth, Z., Seo, M., Hacker, J. & Rodwell, and M. J. W, "InP HBT technologies for THz integrated circuits," *Proc. IEEE 105*, 1051-1067 (2017).

[24] Shinohara, K. *et al.*, "Scaling of GaN HEMTs and Schottky diodes for submillimeter-wave MMIC applications," IEEE Trans. Electron Dev. 60, 2982-2996 (2013).

[25] A. Maestrini, J. Ward, J. Gill, H. Javadi, E. Schlecht, G. Chattopadhyay, F. Maiwald, N.R. Erickson, and I. Mehdi, "A 1.7 to 1.9 THz local oscillator source," *IEEE Microwave Wireless Compon. Lett*. 14 (6) (June 2004) 253-255.

[26] U. R. Pfeiffer, R. Jain, J. Grzyb, S. Malz, P. Hillger, and P. Rodríguez-Vízquez, "Current status of terahertz integrated circuits - from components to systems," *2018 IEEE BiCMOS and Compound Semiconductor Integrated Circuits and Technology Symposium (BCICTS)*, 2018, pp. 1-7, doi: 10.1109/BCICTS.2018.8551068.

[27] R. Han *et al.*, "A SiGe terahertz heterodyne imaging transmitter with 3.3 mW radiated power and fully-integrated phase-locked loop," *IEEE Journal of Solid-State Circuits*, vol. 50, no. 12, pp. 2935-2947, 2015.

[28] Asada, M. and Suzuki, S., "Terahertz emitter using resonant-tunneling diode and applications," *Sensors* 2021, 21, 1384. https://doi.org/10.3390/s21041384.

[29] X. Li *et al.*, "A 250-310 GHz power amplifier with 15-dB peak gain in 130-nm SiGe BiCMOS process for terahertz wireless system," in *IEEE Transactions on Terahertz Science and Technology*, doi: 10.1109/TTHZ.2021.3099057.

[30] Z. Griffith, M. Urteaga, P. Rowell, and R.Pierson, "A 23.2dBm at 210GHz to 21.0dBm at 235GHz 16-way PA-cell combined InP HBT SSPA MMIC," *in Proc. IEEE Compound Semiconductor IC Symp*., La Jolla, CA, USA, Oct. 2014, pp. 1-4.

[31] D. Parveg, D. Karaca, M. Varonen, A. Vahdati, and K. A. I. Halonen, "Demonstration of a 0.325-THz CMOS amplifier," *2016 Global Symposium on Millimeter Waves (GSMM) & ESA Workshop on Millimetre-Wave Technology and Applications*, 2016, pp. 1-3.

[32] Jiang, C. *et al*., "A fully integrated 320 GHz coherent imaging transceiver in 130 nm SiGe BiCMOS," *IEEE J. Solid State Circuits*, 51, 2596-2609 (2016).

[33] J. Moghaddasi and K. Wu, "Multifunction, multiband, and multimode wireless receivers: a path toward the future," *in IEEE Microwave Magazine*, vol. 21, no. 12, pp. 104-125, Dec. 2020, doi: 10.1109/MMM.2020.3023223.

[34] Hichem Guerboukha, Kathirvel Nallappan, and Maksim Skorobogatiy, "Toward real-time terahertz imaging," *Adv. Opt. Photon*. 10, 843-938 (2018).

[35] Z. Hu, M. Kaynak, and R. Han, "High-power radiation at 1 THz in silicon: a fully scalable array using a multi-functional radiating mesh structure," *in IEEE Journal of Solid-State Circuits*, vol. 53, no. 5, pp. 1313-1327, May 2018, doi: 10.1109/JSSC.2017.2786682.

[36] Z. Hu, C. Wang, and R. Han, "A 32-unit 240-GHz heterodyne receiver array in 65-nm CMOS with array-wide phase locking," *in IEEE Journal of Solid-State Circuits*, vol. 54, no. 5, pp. 1216-1227, May 2019, doi: 10.1109/JSSC.2019.2893231.

[37] S. N. Nallandhigal, P. Burasa, and K. Wu, "Deep integration and topological cohabitation of active circuits and antennas for power amplification and radiation in standard CMOS," *in IEEE Transactions on Microwave Theory and Techniques*, vol. 68, no. 10, pp. 4405-4423, Oct. 2020, doi: 10.1109/TMTT.2020.2997049.

[38] Liu, C., Ye, J., and Zhang, Y., "Thermally tunable THz filter made of semiconductors," *Opt. Comm*. 283, 865-868 (2010).

[39] Sanphuang, V., Ghalichechian, N., Nahar, N. K., and Volakis, J. L., "Reconfigurable THz filters using phase-change material and integrated heater," *IEEE Trans. THz. Sci*. Technol. 6, 583-591 (2016).

[40] Wu, X., Lu, H., and Sengupta, K., "Programmable terahertz chip-scale sensing interface with direct digital reconfiguration at sub-wavelength scales," *Nat Commun 10*, 2722 (2019). https://doi.org/10.1038/s41467-019-09868-6

[41] Venkatesh, S., Lu, X., Saeidi, H. *et al*., "A high-speed programmable and scalable terahertz holographic metasurface based on tiled CMOS chips," Nat Electron 3, 785-793 (2020). https://doi.org/10.1038/s41928-020-00497-2

[42] C. Huang *et al*., "Holographic MIMO surfaces for 6G wireless networks: opportunities, challenges, and trends," *in IEEE Wireless Communications*, vol. 27, no. 5, pp. 118-125, October 2020, doi: 10.1109/MWC.001.1900534.

[43] Venkatesh, S., Lu, X., Saeidi, H. *et al*., "A high-speed programmable and scalable terahertz holographic metasurface based on tiled CMOS chips," *Nat Electron 3*, 785-793 (2020). https://doi.org/10.1038/s41928-020-00497-2

[44] Qiang Jiang, Guofan Jin, and Liangcai Cao, "When metasurface meets hologram: principle and advances," *Adv. Opt. Photon*. 11, 518-576 (2019)

[45] Overvig, A.C., Shrestha, S., Malek, S.C. *et al*., "Dielectric metasurfaces for complete and independent control of the optical amplitude and phase," *Light Sci Appl* 8, 92 (2019). https://doi.org/10.1038/s41377-019-0201-7

[46] Pengfei Qiao, Li Zhu, and Connie J. Chang-Hasnain, "High-efficiency aperiodic two-dimensional high-contrast-grating hologram," *Proc. SPIE 9757, High Contrast Metastructures V*, 975708 (15 March 2016)

[47] X. Li, L. Chen, Y. Li, X. Zhang, M. Pu, Z. Zhao, X. Ma, Y. Wang, M. Hong, and X. Luo, "Multicolor 3D meta-holography by broadband plasmonic modulation," *Sci. Adv. 2*, e1601102 (2016).

[48] B. Wang, F. Dong, Q.-T. Li, D. Yang, C. Sun, J. Chen, Z. Song, L. Xu, W. Chu, Y.-F. Xiao, Q. Gong, and Y. Li, "Visible-frequency dielectric metasurfaces for multiwavelength achromatic and highly dispersive holograms," *Nano Lett*. 16, 5235-5240 (2016)

[49] H. Ren, G. Briere, X. Fang, P. Ni, R. Sawant, S. Héron, S. Chenot, S. Vézian, B. Damilano, V. Brändli, S. A. Maier, and P. Genevet, "Metasurface orbital angular momentum holography," *Nat. Commun*. 10, 2986 (2019)

[50] Gao Hui, Wang Yuxi, Fan Xuhao, Jiao Binzhang, Li Tingan, Shang Chenglin, Zeng Cheng, Deng Leimin, Xiong Wei, Xia Jinsong, and Hong Minghui, "Dynamic 3D meta-holography in visible range with large frame number and high frame rate," *Science Advances*. 6. eaba8595. 10.1126/sciadv.aba8595.

[51] T. Tajima, H. Song, and M. Yaita, "Compact THz LTCC receiver module for 300 GHz wireless communications," *in IEEE Microwave and Wireless Components Letters*, vol. 26, no. 4, pp. 291-293, April 2016, doi: 10.1109/LMWC.2016.2537044.

[52] S. Hu *et al.*, "TSV technology for millimeter-wave and terahertz design and applications," *in IEEE Transactions on Components, Packaging and Manufacturing Technology*, vol. 1, no. 2, pp. 260-267, Feb. 2011, doi: 10.1109/TCPMT.2010.2099731.

[53] A. Hassona *et al.*, "Demonstration of +100-GHz interconnects in eWLB packaging technology," IEEE Trans. Compon., Packag., Manuf. Technol.,vol. 9, no. 7, pp. 1406-1414, Jul. 2019.

[54] X. Yang *et al.*, "Low-loss heterogeneous integrations with high output power radar applications at W-band," *in IEEE Journal of Solid-State Circuits*, doi: 10.1109/JSSC.2021.3106444.

[55] M. Urteaga *et al.*, "THz bandwidth InP HBT technologies and heterogeneous integration with Si CMOS," *2016 IEEE Bipolar/BiCMOS Circuits and Technology Meeting (BCTM)*, 2016, pp. 35-41, doi: 10.1109/BCTM.2016.7738973.

[56] Peiying Zhu, "6GWFF 2021 - 6G: connected intelligence (keynote 1)," https://www.youtube.com/watch?v=PU0wKfwssk0.

[57] Coldrey M, "Modeling and capacity of polarized MIMO channels," VTC Spring 2008-IEEE Vehicular Technology Conference. IEEE, 2008: 440-4

[58] Kasdin N J, "Discrete simulation of colored noise and stochastic processes and 1/f/sup/spl alpha//power law noise generation," Proceedings of the IEEE, 1995, 83(5): 802-827.

[59] Koenig S, Lopez-Diaz D, Antes J, *et al.*, "Wireless sub-THz communication system with high data rate," Nature Photonics 7, 2013, pp. 977-81.

[60] H. T. Zhu and K. Wu, "Complex permittivity measurement of dielectric substrate in sub-THz range," *in IEEE Transactions on Terahertz Science and Technology*, vol. 11, no. 1, pp. 2-15, Jan. 2021, doi: 10.1109/TTHZ.2020.3036181.

[61] Daniel Tajik, Aaron D. Pitcher, and Natalia K. Nikolova, "Comparative study of the Rytov and Born approximations in quantitative microwave holography," *Progress In Electromagnetics Research B*, Vol. 79, 1-19, 2017.

[62] D. Tajik and N.K. Nikolova, "Real-time imaging with simultaneous use of Born and Rytov approximations in quantitative microwave holography," *IEEE Trans. Microwave Theory Tech*. (submitted Aug. 9, 2021)

[63] D. Tajik, R. Kazemivala, and N.K. Nikolova, "Combining the Born and Rytov approximations in

quantitative microwave holography," *The IEEE 19th Int. Symp. on Antenna Technology and Applied Electromagnetics (ANTEM 2021)*, Winnipeg, Canada, Aug. 8-11, 2021.

Internet of Things (IoT) Connectivity in 6G: An Interplay of Time, Space, Intelligence, and Value

Petar Popovski, Federico Chiariotti, Victor Croisfelt, Anders E. Kalør, Israel Leyva-Mayorga, Letizia Marchegiani, Shashi Raj Pandey, Beatriz Soret

Department of Electronic Systems, Aalborg University, Danmark

Abstract

Internet of Things (IoT) connectivity has a prominent presence in the 5G wireless communication systems. As these systems are being deployed, there is a surge of research efforts and visions towards 6G wireless systems. In order to position the evolution of IoT within the 6G systems, this paper first takes a critical view on the way IoT connectivity is supported within 5G. Following that, the wireless IoT evolution is discussed through multiple dimensions: time, space, intelligence, and value. We also conjecture that the focus will broaden from IoT devices and their connections towards the emergence of complex IoT environments, seen as building blocks of the overall IoT ecosystem.

1 Introduction

The term Internet of Things (IoT), although present for several decades, started to gain a significant traction with the emergence of the 5G cellular systems and standards [1–2]. An IoT device is a physical object equipped with sensors and/or actuators, embedded computer and connectivity. As such, it can be seen as a two-way micro-tunnel between the physical and the digital world: physical information gets a digital representation and, vice versa, digitally encoded actions get materialized in the physical world. From a different perspective, related to service and product design, IoT capabilities have significantly transformed many products by expanding the functionality and transcending the traditional product boundaries [3].

The ambition of 5G has been to push the boundaries of connectivity beyond the offering of high wireless data rates and expand towards interconnecting humans, machines, robots, and things. This leads to an enormously complex connected ecosystem: a large number of connections that pose a vast diversity of heterogeneous Quality of Service (QoS) requirements in terms of data rate, latency, reliability, etc. To deal with this complexity, the approach of the 5G system design has been to define three generic services: eMBB (enhanced mobile broadband), mMTC (massive machine-type communication), and URLLC (ultra-reliable low-latency communication) [4–5]. The latter two represent the approach of 5G to natively support the requirements of IoT connectivity. This is in contrast to the 4G and other prior generations, where IoT connections were supported in an ad hoc manner, as an afterthought in system deployment.

In some sense, 5G is a step in the direction of obtaining an ultimate connectivity system that is capable of flexibly supporting all conceivable wireless connectivity requirements in the future. One can think of the three generic connectivity types as three dimensions of a certain "service space" and any single connectivity service can be realized as a suitable combination of eMBB, mMTC, and URLLC. For example, in an advanced agricultural scenario, a remotely-controlled machine needs to support real-time reliable actuation of commands (URLLC), while occasionally sending a video feed (eMBB) as well as gathering data from various sensors and IoT devices in the agricultural environment (mMTC).

But is 5G indeed defining the ultimate connectivity framework? This is an important question, as its affirmative answer would obviate the need to redefine and conceptually upgrade the connectivity types towards 6G. On the contrary, a negative answer entails a critical view on 5G and identification of connectivity scenarios that are not well represented by eMBB, mMTC and URLLC or a combination thereof.

As an attempt to answer the question above, this paper takes the, rather general, perspective depicted in Figure 1 to position the role of IoT connectivity and assess its requirements. The general framework from Figure 1 will be first used to take a critical view on IoT as defined in 5G and identify cases that are not well represented by the two categories mMTC and URLLC. Next, the framework will be used as a blueprint to formulate the features of IoT connectivity in beyond-5G/6G systems. The evolution of future wireless IoT technology will be discussed through multiple dimensions: time, space, intelligence, and value. We also conjecture that the focus will broaden from IoT devices and their connections towards the emergence of complex IoT environments, seen as building blocks of the overall IoT ecosystem.

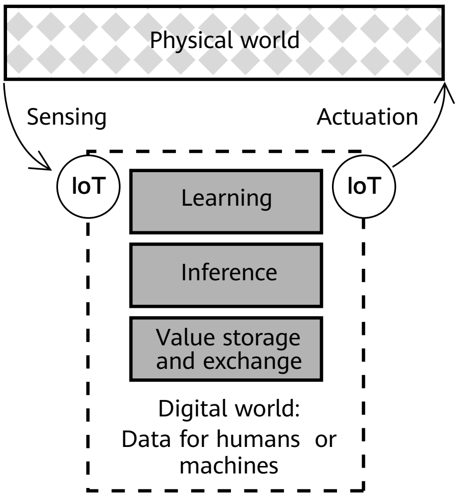

Physical world
Sensing
Actuation
IoT
IoT
Learning
Inference
Value storage and exchange
Digital world: Data for humans or machines

Figure 1 Physical world versus digital world

2 A General IoT Framework

IoT devices reside at the interface between the physical and the digital world and facilitate the two types of information transfer: (1) sensing, creating a digital representation of the physical reality and (2) actuation, converting digital data into commands that exhibit an impact on the physical world. After the information gets converted into a digital data, it can be used in three principal ways:

- Learning: The data is used in a process of training a module that relies on machine learning (ML) or another form of gathering knowledge and building up artificial intelligence (AI).
- Inference: The data is used by an algorithm, AI module or similar to infer conclusions or devise a command that needs to be actuated in the physical world.
- Value storage or exchange: The data is stored for potential use at a future point, such that it possesses a latent value. The data can also be exchanged through the connectivity infrastructure and thus get an actual valuation/monetization.

The "digital world" box includes anything that can store, process or transfer digital data, including the global Internet. There are two general modes that involve IoT communication: machine-to-machine (M2M), that includes interaction and communication only among machines as well as machine-to-human (H2M) (or vice versa), where the overall IoT communication also includes a human in the loop. The principal difference between these two modes is that, when there is a human in the loop, the timing and processing constraints should conform to the ones of the human, while in the case of M2M they are subject to design and specification. In the diagram in Figure 1, a human belong to the physical world. In that sense, the actuation can be understood in a more general way, such as displaying a multimedia content to for the human.

3 A Critical View on IoT in 5G

This section discusses the typical ways in which IoT requirements are articulated within 5G. The objective is to take a critical view by pointing out important scenarios and requirements that are not well covered by the two categories, mMTC and URLLC.

3.1 mMTC

We start by considering mMTC and a view on its typical requirements is depicted in Figure 2. It aims to support a large number of devices, dominantly with an uplink traffic, which is also indicated on the figure. A possible rationale behind this can be formulated as follows. Consider a large set of nodes (sensors) that are generating data locally. The data of different nodes is not correlated, such that each new data packet sent by a different node contributes with a new information about the physical world. Furthermore, each node is only sporadically active, such that the time instant at which it is active and has a data to transmit is unpredictable. Equivalently, this implies that the subset of active nodes at a given time instant is unpredictable. Hence, there are two sources of randomness: the data content and the node activity. This is the basis for the major challenge in mMTC: How to maximize the uplink throughput from a large set of connected nodes, where the subset of active nodes at a given time instant is unknown? This has led to a surge on research in the area of massive random access [6–7]. The challenge requires maximization of the throughput in the uplink, which is more difficult than the downlink, as the devices are uncoordinated and compete for the same shared wireless spectrum. An additional challenge for mMTC devices is the power consumption, which should be optimized to ensure long battery lifetime and unattended operation; this is the case, for example, for sensors embedded in buildings, production plants, or agricultural facilities. More generally, the challenge for mMTC (and we will see that it is similar for URLLC) is made in a maximalist way: it is tacitly assumed that if the most difficult mode of communication is supported, then the easier modes (such as downlink communication towards a subset of nodes) are implied. Finally, following the architectural practice of layered design and modularization, the part of connectivity on Figure 2 is decoupled from and oblivious towards the goals/usage of the mMTC data in the digital world, that is, learning, inference, or value storage/exchange.

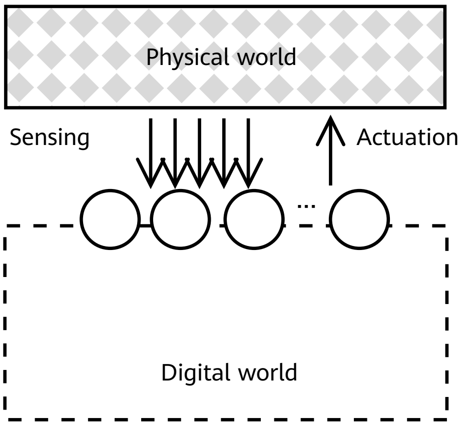


Figure 2 A view on the typical mMTC requirements

Let us now look at a scenario of massive access in which we change some of the assumptions behind the canonical mMTC use case, described above. Consider the case in which the nodes are sensing a physical phenomenon in order to sense an anomalous state and report it to an edge server, which embodies an inference module that is capable of detecting reliably if an anomalous state has occurred. In the simplest case, each sensing node can make a local binary decision whether an anomalous state has occurred (1) or not (0) and send it to the edge server. This violates the assumption that the data across nodes is not correlated, as all of them will try to report about the same observed phenomenon. Furthermore, if the anomalous state occurs within a short time interval, it will trigger response from all sensing nodes in a correlated way, which impacts the statistical properties of the subset of active mMTC nodes. In the ideal case, when a node detects the anomalous state perfectly and the wireless link to the edge server is error-free, then only a single node needs to transmit. Hence, the technical problem is not anymore "throughput maximization from an unknown random subset" but rather a "leader election from an unknown subset with a certain correlation structure in the node activation". If the ideal assumptions on sensing/communication are relaxed, then a sufficient number of nodes should report the detected alarm, such that the edge server can infer a reliable decision about the state of the physical world. The problem can be further relaxed by considering that the sensing node is not directly detecting the anomalous state, but rather a data related to it; then the edge server needs to fuse this data to make inference. The technical problem is now "collect a sufficient number of data points to reliably detect anomaly".

These examples show that the consideration of the data purpose/usage has a significant impact on the technical challenges posed to the wireless connectivity part. For all these new technical problems, a system optimized for mMTC with typical requirements will lead either to inefficient operation (collecting much more data points than needed to make inference) or failing to meet the timing requirement (the detection of the alarm will be delayed due to channel congestion). The case of transmission of identical alarm messages from a massive set of devices that should enable timely and reliable detection at an edge node illustrates that massiveness, reliability and latency may not be separable (as treated in 5G) when the data content/purpose is taken into account. Clearly, following the (overused) approach of cross-layer optimization, one may immediately jump to the conclusion that the access should be designed jointly with the high-level objective of the transmitted data. This is not feasible, as it does not contribute to a scalable architectural design. However, the described problems indicate that, rather than sticking to the problem of "throughput maximization from an unknown random subset", we need to identify a small set of connectivity-related challenges that provide a better span of the IoT requirements with a massive number of devices and design systems that can solve them efficiently.

3.2 URLLC

We now provide a critical view on URLLC, the second generic service related to IoT. In order to illustrate the URLLC requirements, we consider the sense-compute-actuate cycle depicted in Figure 3. In this example we observe the timing of the following loop. An IoT gathers information from the physical world, digitalizes it and transmits it wirelessly to a server that performs computation and inference. Based on that, the server sends a command wirelessly to an actuating device; in the special case, this device is the same one as the sensing IoT device. Figure 3 illustrates the total timing budget for these operations. The specification of URLLC has been done with the motivation to use a small part of this timing budget on the wireless radio link and ensure that transmission is done within this short allocated time with a very high reliability. This would leave a sufficient timing budget to perform the other operations, regardless of whether the total timing budget is 10 ms or 50 ms.

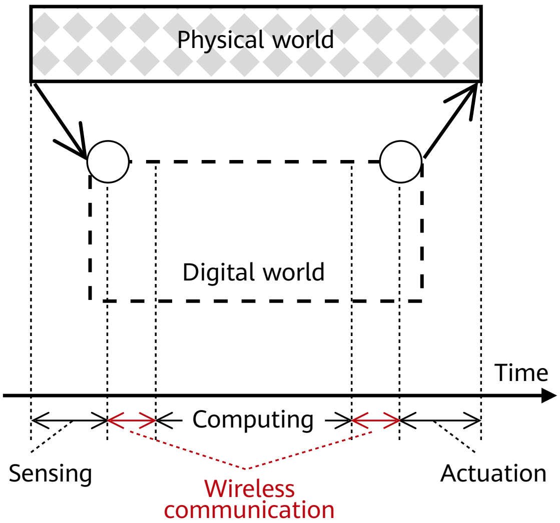
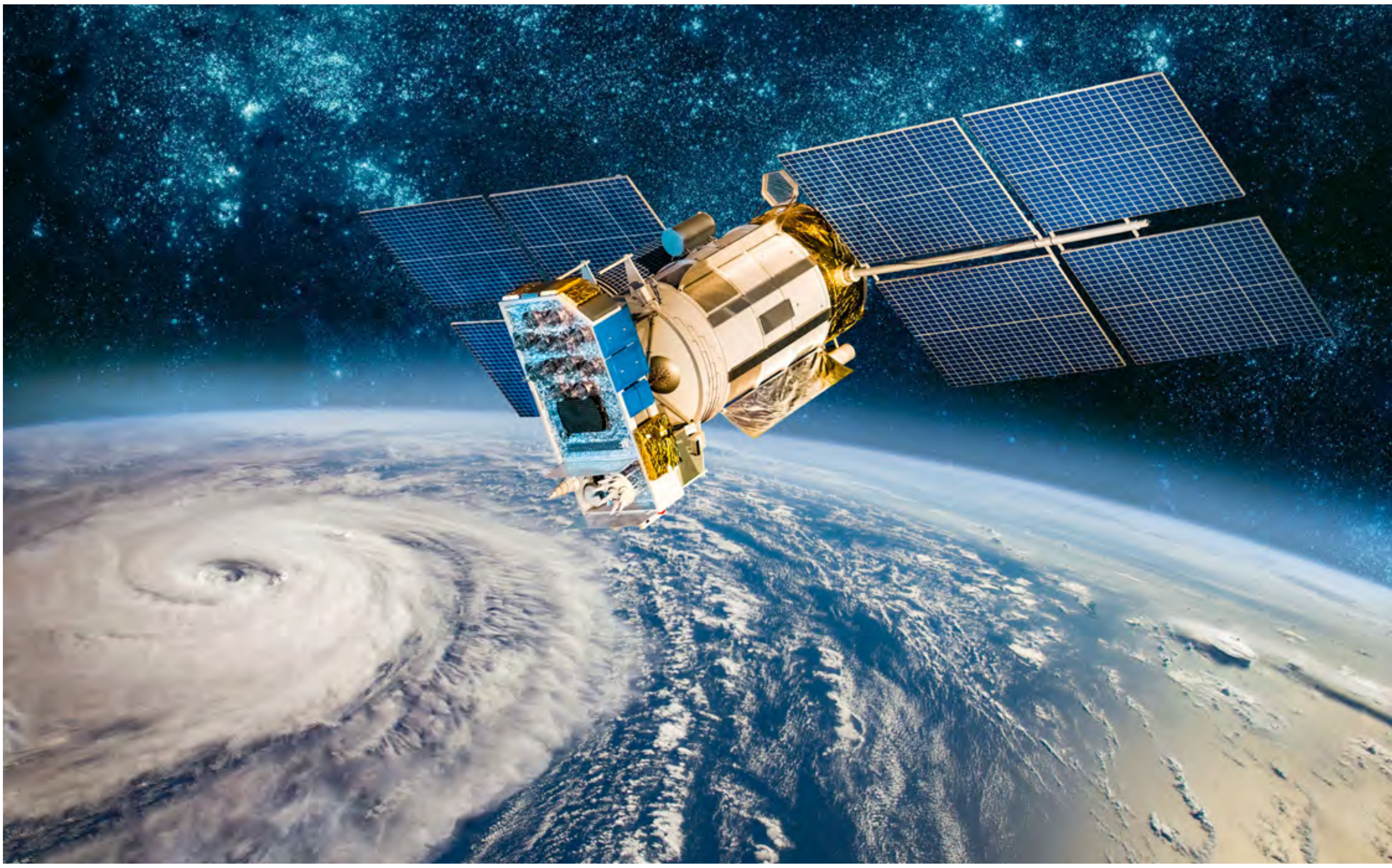


Figure 3 The context for defining of URLLC requirements through a timing budget of a sense-compute-actuate cycle



This is again a rather maximalist approach towards the radio link in the quest to satisfy the end-to-end requirements on timing and reliability. In the early specification of URLLC, the allocated time was 1 ms and the required reliability was 99.999%. Achieving high reliability is associated with the use of high level of diversity (e.g., bandwidth) and power. Relaxing the requirements on the wireless transmission could lead to more efficient operation, while still meeting the overall goal of communication. Specifically, in the example in Figure 3 the computation part may be capable of compensating for the data loss on the sensing wireless link and make a predictive decision that can be passed on to the actuator. Or, as indicated in the early paper on ultra-reliable communication on 5G [8], one can take a holistic perspective on URLLC and ensure that the overall system degrades gracefully if the data is not delivered within a given deadline.

Expressing it in a similar way as we have done in the previous section, the basic problem of URLLC has been formulated as "deliver the data of size X within Y milliseconds with reliability Z". Instead, timing in a communication system can be put in a more general framework and define a set of basic problems that are capable of capturing various timing requirements. For instance, instead of looking at the latency of the packet, one can jointly consider the data generation process and the state of the computation process. In that sense, a more relevant measure than latency can be information freshness or age of information. This is further discussed in Section IV, while for a more general discussion on the timing concepts towards 6G the reader, refer to [9].

4 Time

The definition of real time is highly dependent on the application and its final user. Specifically, real time is dependent on whether the overall system is intended for one of the following three communication setups: (1) human-to-human (H2H), (2) H2M, including setups of communication among machines with a human in the loop (HITL); and (3) M2M. Even fully interactive human-type communication such as augmented reality and virtual reality (AR/VR) does not require millisecond timescales, as human perception becomes the limiting factor [10]: For example, the human eye cannot perceive images that are shown for less than about 13 ms, setting a hard ceiling on the network timing requirements for this type of communication. The same is true in HITL scenarios, where machines can operate faster than human perception limits, but the system operating faster than the perceivable latency threshold will be experienced by the human as instantaneous and seamless [11].

There is no universally defined timing threshold for M2M communication, as the timing requirements depend on the type of applications and on the capabilities of the specific cyber-physical system (CPS). This is in a way reflected in the split between mMTC and URLLC in 5G, which represent two extreme cases. As also discussed in Section III, these two extremes do not cover the full range of use cases. A more accurate view of CPS timing requirements should go beyond the isolated characterization of the wireless communication latency and consider all the contributors in Figure 3. Furthermore, the use of age of information (AoI) [12] instead of latency as a metric can have significant advantages, as AoI can better represent the discrepancy between the model that the system can construct from the sensor transmissions and the actual physical reality. The limits on the allowed AoI depend on the tolerance of the control algorithm and of the application: advanced control algorithms in highly predictable scenarios will be able to work even with very old information, while complex and unpredictable scenarios which require fast reaction times will necessarily have stricter requirements [13]. One step further is to use the content of the packets themselves to define latency and reliability requirements: if the controller employs some form of predictive algorithm, new information that fits the expected model will be relatively unimportant, while unexpected deviations from it will need to be delivered quickly and reliably. This approach can be measured with the value of information (VoI) [14], a metric that combines the age and content of the packet to directly measure the usefulness of communication. The difference between AoI and VoI is clearly shown in Figure 4: While AoI increases linearly and then drops to 0 after a transmission (assuming the latency is negligible), the increase in the VoI depends on the behavior of the system, and might be nonlinear. In the figure, the first period between 0s and 25s has a relatively slow increase, while the period between 40s and 60s has a steeper one, and indeed gets to a higher VoI in a shorter time: This is due to the different behavior of the system, which strays farther from the estimated value at the receiver.

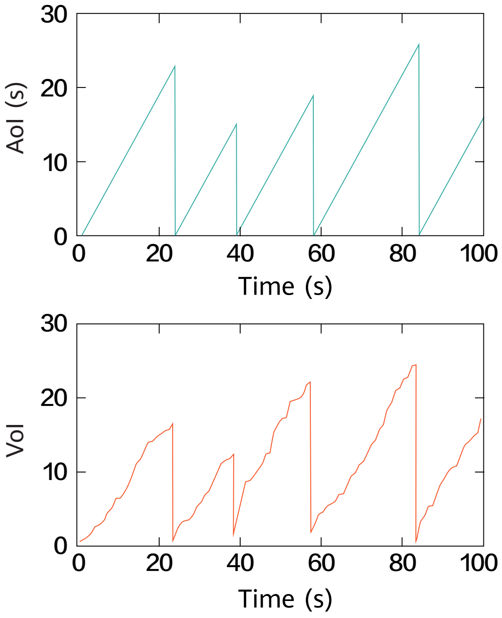


Figure 4 Example of the difference between AoI and VoI in a system with cumulative estimation errors

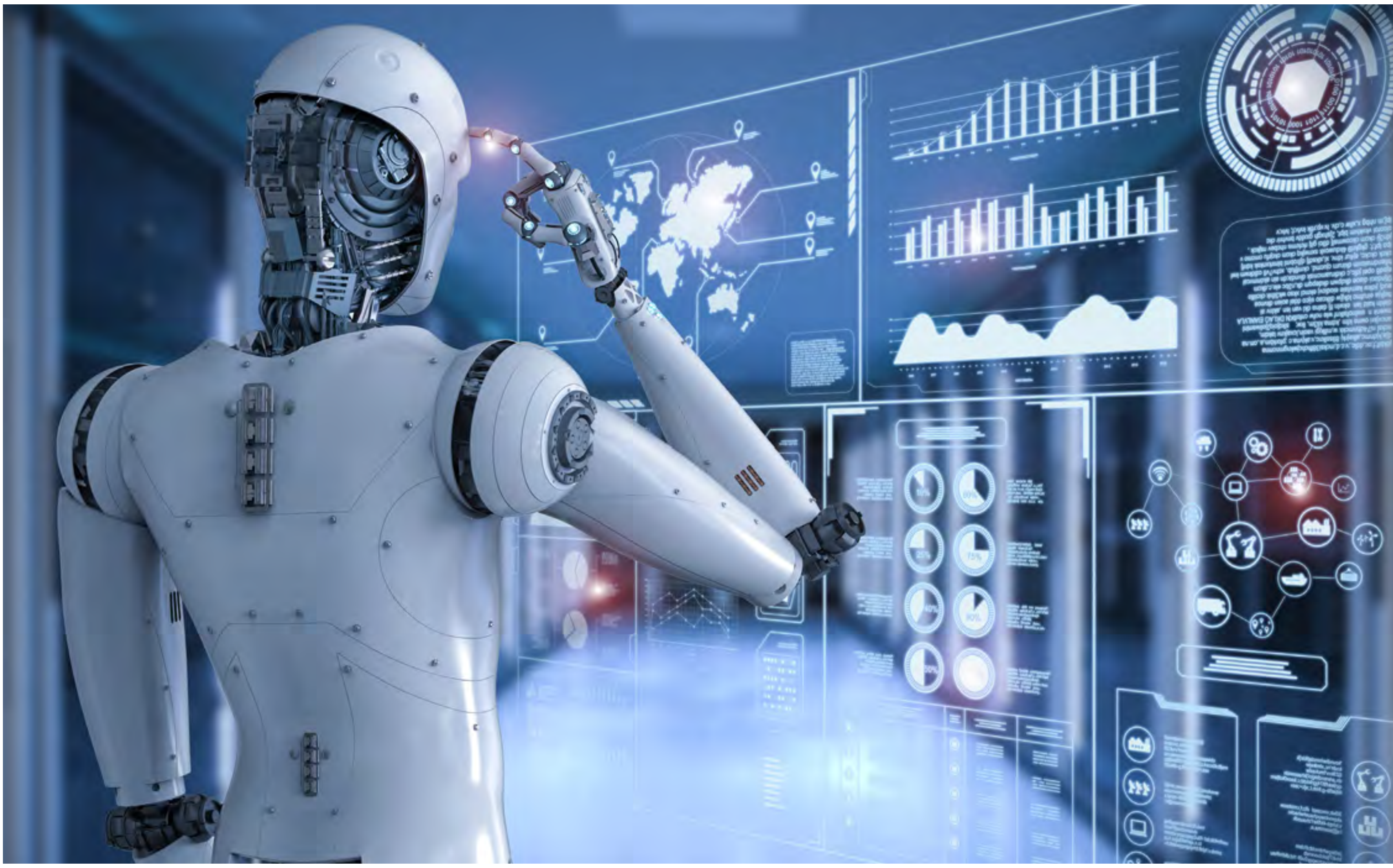
Using VoI as a metric is a step towards semantics-oriented communication. The classical design of a CPS assumes a total independence between the content of the data and their transmission, i.e., uncontrolled arrival of exogenous traffic to the communication system. This sets design boundaries to the communication protocols and relaxing this rigid separation allows us to tackle the system design process holistically and improve the performance. In control and HITL applications, the urgency of information (UoI) approach [15] defines VoI in such a way that the packets with the highest value are the ones that affect the control performance the most. This definition of value is also closely tied to the market value of data, which we will describe below: In both cases, samples from the sensors are more valuable if they are surprising, i.e., if they contain information that is not currently represented in the model of the system. The difference between the two is in the way the data is used: in the data market case, this new information is used to improve that model, while in VoI applications, it is used to track and control a system.

5 Space

Evolving towards 6G, we need to look in the changes that occur in the space in which IoT devices are deployed to operate. In this context, we use the term space to refer to: (1) the environment where sensing and actuation take place and (2) the propagation medium where the electromagnetic waves travel to transfer information between two or more points. Hence, by delimiting the space in which the interface between the physical and digital worlds resides, the definition of time (space-time) and frequency as resources for communication is inherent. Therefore, space sets the basis for resource sharing and competition among devices.

As the optimization of frequency and time resources becomes insufficient, the next frontier towards increased network capacity is to optimize the use of space. Thence, increasing the network capacity per unit area has been one the major objectives of every subsequent generation of mobile networks. However, this objective has encountered a major challenge: the optimization has been limited to the placement and capabilities of the networking devices — the infrastructure — whereas there has been little to no control on the user side. That is, IoT and other mobile user devices possess limited capabilities and, as a consequence, their wireless channels are mostly determined by nature. Because of this, the traditional approach towards a greater network capacity is pre-planned network densification in combination with frequency reuse to minimize inter-cell interference. Only in recent years, precoding, beamforming, and beam steering techniques have enabled a much more flexible and agile exploitation of the space resources through massive multiple input-multiple-output (mMIMO) [16] and the development of cell-free networks [17]. In mMIMO, the channel state information, based on spatial reciprocity, is exploited to achieve communication with multiple devices in the same block of time-frequency resources. Furthermore, distributed or cell-free mMIMO allows us to exploit the macro-diversity of the environment by allowing the IoT devices to communicate to antenna elements at different locations to combat blockages and eliminate coverage holes.

Despite these advances, the capabilities of the IoT devices will remain limited in order to keep their cost down. Nevertheless, new developments on distributed infrastructures, AI/ML, and signal processing techniques will enable the network infrastructure and the environment itself to become intelligent. This will enable the real-time self-optimization of heterogeneous architectures that can relax the hardware requirements of IoT devices while exploiting the spatial resources. In the recent developments, the propagation environment may act as an ally to the simple IoT devices rather than only a major challenge that needs to be overcome. For instance, reconfigurable intelligent



surfaces (RISs) [18] consist of elements that can alter the properties of the incident signals adaptively and, hence, allow for a much greater control over space than that of the IoT devices. Specifically, RISs can be used to take advantage of the location of the devices to create highly directive and interference-free beams towards the base station in real time. This allows for a new interpretation and exploitation of overlapping signals and also alleviates the hardware requirements of the devices, since part of the hardware on the device can be outsourced to the environment.

The structure of the physical space plays a major role on the feasibility of deploying network infrastructure and, hence, on the availability of Internet connectivity. Historically, we have seen the infrastructure as deployed in the 2D space; usually encompassing ground-level infrastructure while considering the air and (outer) space infrastructure to be, oftentimes, alien to it and, in the best case, complementary (i.e., global positioning and navigation services). It is only in recent years, that the New Space era and the advances in unmanned aerial vehicles (UAVs) have expanded our view of the network infrastructure to the 3D space [19]. Satellites deployed in the low earth orbit (LEO) can serve as a global network, capable of achieving low end-to-end latency while providing coverage in remote regions (e.g., Arctic and maritime) where deploying terrestrial infrastructure is infeasible [20]. Besides, while LEO satellite constellations are moving infrastructure, they present deterministic space-time dynamics that can be exploited for resource optimization [21]. Due to this combination of characteristics, one of the major objectives for 6G is to achieve a full integration of the terrestrial infrastructure with satellites, drones, and other aerial devices to fully exploit the 3D nature of space [22–24].

In the digital world, space influences a series of characteristics of the data beyond quantity, such as its content and, hence, relevance. This calls for a characterization of how the optimization of wireless resources for a given delimited space affects the overall learning, inference, and value of data. In this sense, the network-level optimization objectives must be redefined to consider the role that space plays in the use that the data will have. Consequently, there is the need for a new interpretation of resource efficiency depending on the context: What is the data content and what will it be used for? For example, the concept of over-the-air computation exploits the superposition property of the medium to effectively merge data from multiple sensors or model updates in the case of distributed learning [25]. This indicates that the 5G interpretation of space is far from being a definitive vision, as the capitalization of space, now dynamically, depends mainly on the utility of the data.

6 Intelligence: Learning and Inference

A general and rather certain trend in the coming years is that the intelligence in networks, network nodes, but also connected devices, will continuously increase. As the number of applications relying on IoT technology has grown, we have seen the capabilities of those things evolving accordingly. Devices that used to be exclusively employed as sense-and-transmit entities are now equipped with different levels of embedded intelligence directly operating on the information collected. This need for increasingly smarter communicating parties opens the way to the definition of a "smarter" content to exchange and novel ways of making sure this is done efficiently and correctly. There have been already some efforts in this perspective [26–27], where the communication effort is optimized so that only the most useful data for the actual data consumer is transmitted. In a machine learning perspective this could, for instance, get translated into "communicate only the most significant features". But what is actually determining the significance of the information exchanged in this context? And how do we make sure this is transmitted efficiently? One natural option would be to consider relevant whatever maximizes the performance of the receiver at executing a specific task, while relying on a compression strategy able to extract exactly this relevant information from the data. Communication frameworks based on neural network autoencoders to encode and decode messages would fulfill the requirements described above, as they are inherently able to find compressed input representations which are the most useful for the task at hand (e.g., [28]). Yet, this is not enough. Systems would end up being ad-hoc systems, able to operate correctly only on a small set of tasks (by leveraging multi-task learning schemes [29]; otherwise only on a single task), where all the parties involved in the communication share the same model (i.e., same network with same structure, parameters and weight values), and interpretability would remain a crucial challenge. Preliminary studies have shown the potential of data-driven techniques (such as autoencoders) in this context [30–31]; though, it is evident how such paradigms create strong constraints against generalization. This is why we will eventually need to move away from those systems and start associating semantics and meaning to the data, as also suggested by a series of recent papers on semantic communication [32–34] as well as advances in graph and semantic networks [35–36]. Indeed, by integrating semantics-based systems (i.e., systems with knowledge representations, such as knowledge graphs, and reasoning capabilities), well studied and long exploited in more traditional AI, with more recent data-driven frameworks, we would eventually enable efficient communication of relevant and meaningful information among entities sharing the same view of the world in the form of a knowledge base, rather than a specific task-dependent model [37].

7 Value

The inter-networked CPSs in the IoT networks are readily accumulating and processing data at a large scale. When operated with ML tools, these massively distributed data stimulate real-time and non-real time inference and decision-making services that create an economic value of data. For example, services defining prediction, localization, automation and control heavily consume large data samples for training learning models and improve its performance, i.e., model accuracy. These are certainly a few of the several promising outlooks with data in general. In particular, data is a valued commodity for trade that has gathered significant economic value and a multitude of social impacts. However, it is an overstatement if the narratives on the economic value of data leave the fundamental inter-dependencies between the data properties themselves, the contextual, time-space information it encodes, and its value.

At the other end of the story, the utility of distributed data in the physical space is constantly challenged by a conventional perspective at an IoT device. The hindrance resides in the fundamental attributes of IoT networks and their components: data are not readily usable and highly localized, the data sources are resource-constrained, and the system is under stress due to unreliable connectivity during data transfer. Whereas, data in the digital space permits flexibility in its storage, mobility, customization and inter-operation to extract meaningful information, behaving likewise digital goods. Therein, data can be monetized and exchanged for added value, as shown in Figure 5. The platform is a primary interface of interaction between buyers/sellers, leveraging connectivity for value storage or exchange in the data market, which quantifies pricing schemes and facilitates the overall data trading process. In this matter, one must not confuse "value" and "pricing"; for instance, the utility of correlated IoT data diminishes quickly if it exhibits no latent value [38–39]. Hence, the monetary value of such data appears low. However, such data can still contribute to assess system-level reliability, such as in sensory networks, where IoT devices constantly transmit their measurement data, or in case the data is used for inference. Such use cases highlight the challenges of a holistic approach in quantifying the data value. However, the value storage and exchange should not be a naive characterization by the single arbitrator/platform but depends entirely on the nature (independent or collective) and the requirements of applications these data can offer. For instance, a more tailored mobile application that benefits users with specific personalized services expects techniques to handle data privacy concerns, for which the value exchange mechanism would be unique.

Arguably, this departure in the understanding of IoT devices, basically confined within sensing/actuation functionalities, to a broader physical and digital world perspective, in principle, impacts how connectivity shall behave and value is added with data exchanges. One can think of emergence of IoT devices that will behave as autonomous sellers and buyers of data in a decentralized data market. An

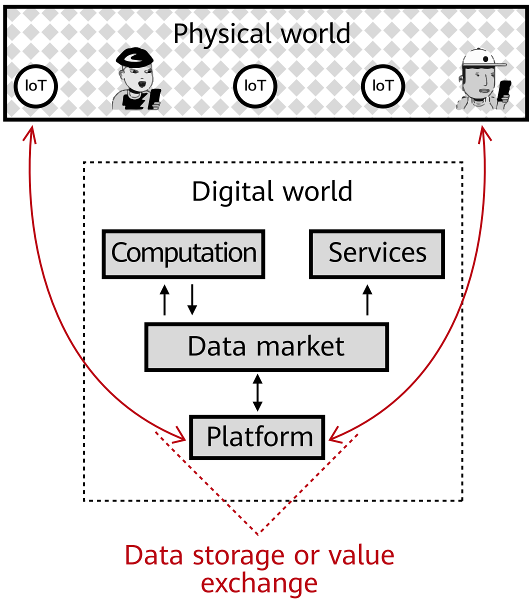


Figure 5 Illustration of data trading in an IoT network

example, the elastic computational operation on data in the digital space, coupled with value storage or exchange technologies, such as distributed ledger technologies (DLTs) [40], quantifies the utility of data as transaction details and provides a different take on communication requirements to operate data trading. Similarly, in a Smart Factory setting, the value of exchanged data between devices during operation also reveals the properties of shared media access patterns, which can be exploited as feedback to tune vital parameters defining the communication resources in general. This explains the rationale behind the need to incorporate frequent interactions between the physical and the digital world, which brings value out of data, its storage and exchange while optimizing connectivity.

8 Towards Complex IoT Environments

Although the initial IoT designs focused on simple applications, the maturity of the technology leads towards more complex systems where the single device model falls short. Rather than isolated and low-capacity devices, we encounter IoT applications that are deployed and executed in several heterogeneous edge devices, interconnected with a network — wireless and/or wired — that dynamically adapts to changes in the environment and with built-in intelligence and trustworthiness. These environments rely on several distributed technologies, such as edge computing [41], edge intelligence [42] and DLT [40], and their complex interactions.

For instance, in a manufacturing plant, we find a number of interconnected industrial robot arms, machinery and automated guided vehicles (AGVs). The accomplishment of a complex manufacturing goal (the IoT application) is based on the autonomous collaboration between the nodes, with very heterogeneous capabilities. This requires the orchestration of the computation and communication resources for an overall reliable, trustworthy and safe operation. Another example is a fleet of e-tractors equipped with sensors and computing resources to perform the mission assigned to them. The computing resources enable each tractor to perform computation tasks on spot, thus acting as an edge-based device, and they collaborate to achieve the common goal. Tasks that cannot be performed on the vehicle will be offloaded to an available cloud infrastructure. Components within the tractor are usually connected with time-sensitive networking (TSN), whereas edge-cloud communication and tractor-to-tractor communication are wireless.

Characterizing the performance and the energy efficiency of these complex systems is a daunting task. The conventional approach has been to characterize every single device or link and technology, but this approach is too simplistic. For example, the energy expenditure of an IoT device will strongly depend on the context in which it is put, in terms of, e.g., goal of the communication or traffic behavior. Therefore, the system performance and the total energy footprint are not just a simple sum of an average per-link or per-transaction contribution of an isolated device. A more accurate picture of the overall performance and energy consumption is obtained by taking the complex IoT system as the basic building block. At the same time, the timing characterization of the system becomes more involved, and the new ecosystem of timing metrics discussed in Section IV must be adapted to capture the distributed interrelations [9].

9 Conclusion

This paper has provided a perspective on the evolution of wireless IoT connectivity in 6G wireless systems. In order to justify the enthusiasm towards developing new 6G systems, we have taken a critical view on 5G IoT connectivity. Specifically, we have illustrated cases that are potentially not captured by the 5G classification of IoT into mMTC and URLLC, respectively. In order to put the IoT evolution in a proper perspective, we have started from a general IoT framework in which we identify three principal uses of the data transferred from/to IoT devices: learning, inference, and data storage/exchange. This general framework has been expanded through different dimensions of wireless IoT evolution: time, space, intelligence, and value. Finally, we have discussed the emergence of complex IoT environments, seen as building blocks that are suitable to analyze the energy efficiency of these systems.

References

[1] M. R. Palattella, M. Dohler, A. Grieco, G. Rizzo, J. Torsner, T. Engel, and L. Ladid, "Internet of Things in the 5G era: Enablers, architecture,and business models," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 3, pp. 510-527, Feb. 2016.

[2] D. Wang, D. Chen, B. Song, N. Guizani, X. Yu, and X. Du, "From IoT to 5G I-IoT: The next generation IoT-based intelligent algorithms and 5G technologies," *IEEE Communications Magazine*, vol. 56, no. 10, pp.114-120, Nov. 2018.

[3] M. E. Porter and J. E. Heppelmann, "How smart, connected products are transforming competition," *Harvard Business Review*, vol. 92, no. 11,pp. 64-88, Nov. 2014.

[4] ITU-R, "Minimum requirements related to technical performance for IMT-2020 radio interface(s)," International Telecommunication Union(ITU), Report ITU-R M.2410-0, Nov. 2017.

[5] 3GPP, "Study on new radio (NR) access technology physical layeraspects," 3rd Generation Partnership Project (3GPP), TR 38.802, Mar.2017.

[6] Y. Polyanskiy, "A perspective on massive random-access," in *2017 IEEEInternational Symposium on Information Theory (ISIT)*, 2017, pp. 2523-2527.

[7] C. Bockelmann, N. K. Pratas, G. Wunder, S. Saur, M. Navarro, D. Gregoratti,G. Vivier, E. De Carvalho, Y. Ji, v. Stefanović, P. Popovski,Q. Wang, M. Schellmann, E. Kosmatos, P. Demestichas, M. Raceala-Motoc, P. Jung, S. Stanczak, and A. Dekorsy, "Towards massive connectivity support for scalable mMTC communications in 5G networks," *IEEE Access*, vol. 6, pp. 28 969-28 992, 2018.

[8] P. Popovski, "Ultra-reliable communication in 5G wireless systems," in *1st International Conference on 5G for Ubiquitous Connectivity*, 2014, pp. 146-151.

[9] P. Popovski, F. Chiariotti, K. Huang, A. E. Kal-r, M. Kountouris,N. Pappas, and B. Soret, "A perspective on time towards wireless 6G," *arXiv preprint arXiv:2106.04314*, Jun. 2021.

[10] R. B. Miller, "Response time in man-computer conversational transactions," in *Fall Joint Computer Conference*. AFIPS, Dec. 1968, pp. 267-277.

[11] A. T. Z. Kasgari, W. Saad, and M. Debbah, "Human-in-the-loop wireless communications: Machine learning and brain-aware resource management,"*IEEE Transactions on Communications*, vol. 67, no. 11, pp.7727-7743, Jul. 2019.

[12] S. Kaul, R. Yates, and M. Gruteser, "Real-time status: How often should one update?" in *International Conference on Computer Communications (INFOCOM)*. IEEE, Mar. 2012, pp. 2731-2735.

[13] X.-M. Zhang, Q.-L. Han, X. Ge, D. Ding, L. Ding, D. Yue, and C. Peng, "Networked control systems: A survey of trends and techniques," *IEEE/CAA Journal of Automatica Sinica*, vol. 7, no. 1, pp. 1-17, Jul.2019.

[14] O. Ayan, M. Vilgelm, M. Klügel, S. Hirche, and W. Kellerer, "Age-of-information vs. value-of-information scheduling for cellular networked control systems," in *10th International Conference on Cyber-Physical Systems (CPS/IoT)*. ACM/IEEE, Apr. 2019, pp. 109-117.

[15] X. Zheng, S. Zhou, and Z. Niu, "Urgency of information for context-aware timely status updates in remote control systems," *IEEE Transactionson Wireless Communications*, vol. 19, no. 11, pp. 7237-7250, Jul.2020.

[16] T. L. Marzetta, "Noncooperative cellular wireless with unlimited numbers of base station antennas," *IEEE Transactions on Wireless Communications*, vol. 9, no. 11, pp. 3590-3600, Oct. 2010.

[17] O. T. Demir, E. Björnson, and L. Sanguinetti, "Foundations of user-centric cell-free massive MIMO," *Foundations and Trends in Signal vProcessing*, vol. 14, no. 3-4, pp. 162-472, Mar. 2021.

[18] E. Björnson, H. Wymeersch, B. Matthiesen, P. Popovski, L. Sanguinetti, and E. de Carvalho, "Reconfigurable intelligent surfaces: A signal processing perspective

with wireless applications," *arXiv preprint arXiv:2102.00742*, 2021.

[19] O. Kodheli, E. Lagunas, N. Maturo, S. K. Sharma, B. Shankar, J. F. M. Montoya, J. C. M. Duncan, D. Spano, S. Chatzinotas, S. Kisseleff, J. Querol, L. Lei, T. X. Vu, and G. Goussetis, "Satellite communications in the New Space era: A survey and future challenges," *IEEE Communications Surveys & Tutorials*, vol. 23, no. 1, pp. 70-109, Mar. 2021.

[20] M. S. Abildgaard, C. Ren, I. Leyva-Mayorga, Čedomir Stefanović, B. Soret, and P. Popovski, "Arctic connectivity: A frugal approach to infrastructural development," *arXiv preprint arXiv:2108.13012*, Aug. 2021, accepted for publication in Arctic Journal.

[21] I. Leyva-Mayorga, B. Soret, and P. Popovski, "Inter-plane inter-satellite connectivity in dense LEO constellations," *IEEE Transactions on WirelessCommunications*, vol. 20, no. 6, pp. 3430-3443, Jun. 2021.

[22] B. Di, L. Song, Y. Li, and H. V. Poor, "Ultra-dense LEO: Integration of satellite access networks into 5G and beyond," *IEEE Wireless Communications*, vol. 26, no. 2, pp. 62-69, May 2019.

[23] S. Dang, O. Amin, B. Shihada, and M.-S. Alouini, "What should 6G be?" *Nature Electronics*, vol. 3, no. 1, pp. 20-29, Jan. 2020.

[24] I. F. Akyildiz, A. Kak, and S. Nie, "6G and Beyond: The future of wireless communications systems," *IEEE Access*, vol. 8, pp. 133 995-134 030, Jul. 2020.

[25] J. Park, S. Samarakoon, A. Elgabli, J. Kim, M. Bennis, S.-L. Kim, and M. Debbah, "Communication-efficient and distributed learning over wireless networks: *Principles and applications," Proceedings of the IEEE*, vol. 109, no. 5, pp. 796-819, May 2021.

[26] P. Zalewski, L. Marchegiani, A. Elsts, R. Piechocki, I. Craddock, and X. Fafoutis, "From bits of data to bits of knowledge — An on-board classification framework for wearable sensing systems," *Sensors*, vol. 20, no. 6, p. 1655, Mar. 2020.

[27] A. Elsts, R. McConville, X. Fafoutis, N. Twomey, R. J. Piechocki, R. Santos-Rodriguez, and I. Craddock, "On-board feature extraction from acceleration data for activity recognition," in *International Conference on Embedded Wireless Systems and Networks (EWSN)*. ACM,Feb. 2018, pp. 163-168.

[28] E. Bourtsoulatze, D. Burth Kurka, and D. Gündüz, "Deep joint source-channel coding for wireless image transmission," *IEEE Transactions on Cognitive Communications and Networking*, vol. 5, no. 3, pp. 567-579, Sep. 2019.

[29] S. Vandenhende, S. Georgoulis, M. Proesmans, D. Dai, and L. Van Gool, "Revisiting multi-task learning in the deep learning era," *arXiv preprint arXiv:2004.13379*, vol. 2, Apr. 2020.

[30] H. Xie, Z. Qin, G. Y. Li, and B.-H. Juang, "Deep learning enabled semantic communication systems," *IEEE Transactions on Signal Processing*, vol. 69, pp. 2663-2675, Apr. 2021.

[31] Z. Chen, K. Fan, S. Wang, L. Duan, W. Lin, and A. C. Kot, "Toward intelligent sensing: Intermediate deep feature compression," *IEEE Transactions on Image Processing*, vol. 29, pp. 2230-2243, Sep. 2019.

[32] P. Popovski, O. Simeone, F. Boccardi, D. Gündüz, and O. Sahin, "Semantic-effectiveness filtering and control for post-5G wireless connectivity," *Journal of the Indian Institute of Science*, vol. 100, no. 2, pp.435-443, Apr. 2020.

[33] M. Kountouris and N. Pappas, "Semantics-empowered communication for networked intelligent systems," *IEEE Communications Magazine*, vol. 59, no. 6, pp. 96-102, Jul. 2021.

[34] E. Uysal, O. Kaya, A. Ephremides, J. Gross, M. Codreanu, P. Popovski, M. Assaad, G. Liva, A. Munari, T. Soleymani, B. Soret, and K. H. Johansson, "Semantic communications in networked systems," *arXiv preprint arXiv:2103.05391*, Mar. 2021.

[35] X. Wang, H. Ji, C. Shi, B. Wang, Y. Ye, P. Cui, and P. S. Yu,"Heterogeneous graph attention network," in *The*

World Wide Web Conference (WWW). ACM, May 2019, pp. 2022-2032.

[36] L. Zhao, X. Peng, Y. Tian, M. Kapadia, and D. N. Metaxas, "Semantic graph convolutional networks for 3D human pose regression," in *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE/CVF, Jun. 2019, pp. 3425-3435.

[37] Q. Lan, D. Wen, Z. Zhang, Q. Zeng, X. Chen, P. Popovski, and K. Huang, "What is semantic communication? A view on conveying meaning in the era of machine intelligence," *arXiv preprint arXiv:2110.00196*, Oct. 2021.

[38] S. N. Ali, G. Lewis, and S. Vasserman, "Voluntary disclosure and personalized pricing," in *21st Conference on Economics and Computation (EC)*. ACM, Jul. 2020, pp. 537-538.

[39] A. Agarwal, M. Dahleh, T. Horel, and M. Rui, "Towards data auctions with externalities," *arXiv preprint arXiv:2003.08345*, Mar. 2020.

[40] T. M. Fernández-Caramés and P. Fraga-Lamas, "A review on the use of blockchain for the Internet of Things," *IEEE Access*, vol. 6, pp. 32 979-33 001, May 2018.

[41] A. Alnoman, S. K. Sharma, W. Ejaz, and A. Anpalagan, "Emerging edge computing technologies for distributed IoT systems," *IEEE Network*, vol. 33, no. 6, pp. 140-147, May 2019.

[42] S. Deng, H. Zhao, W. Fang, J. Yin, S. Dustdar, and A. Y. Zomaya, "Edge intelligence: The confluence of edge computing and artificial intelligence," *IEEE Internet of Things Journal*, vol. 7, no. 8, pp. 7457-7469, Apr. 2020.

6G Native Trustworthiness

Fei Liu [1], Rob Sun [2], Donghui Wang [3], Chitra Javali [1], Peng Liu [3]

[1] Singapore Research Centre

[2] Ottawa Wireless Advanced System Competency Centre

[3] Wireless Technology Lab

Abstract

Since the emergence of digital wireless communication, security mechanisms have been embedded into protocols and functions. The principle of "security by design" is thus well-known. The current security architecture will evolve into a native trustworthiness architecture in 6G. Such an architecture is expected to adapt to holistic networks and meet the diversified requirements from the multi-stakeholder industry ecosystem in the future. In this paper, we first propose a 6G multi-lateral trust model in which blockchain for wireless networks is introduced as a trusted infrastructure. We then analyze the physical layer security technologies and the widely researched quantum-key-distribution techniques. Challenges and technologies of privacy, AI-enabled security, as well as measurement of trust are further discussed and analyzed as potential components for 6G native trustworthiness.

Keywords

6G, trustworthiness, security, privacy, blockchain, physical layer security, QKD, AI

1 Introduction

Trust is a prerequisite for information exchange between parties. Trust establishment is founded not only on mutual identification, but also on the security and privacy preservation capabilities that are embedded into the signaling and data flow throughout the network. A robust network system can proactively identify risks and threats, and take remedial actions in the event of an attack or natural disaster. When all these functions are directly triggered by events, changes, or user requests without manual configuration and scheduling, the trustworthiness is deemed as native. Native trustworthiness can be achieved through a trustworthy architecture design, covering security, privacy, and resilience.

Compared with 5G, 6G networks will be more distributed and provide some unique user-centric services. Such requirements are bound to pose challenges to the current communications network-centric security architecture. A more inclusive trust model is required. It is therefore necessary to propose a native trustworthiness architecture that covers the entire lifecycle of communications networks.

In the following sections, we report our explorations of appropriate and effective trustworthiness-related technologies for 6G.

Wireless communication was first introduced in 1980s, and in the subsequent years, has gone through revolutionary transformation in terms of the security architecture. The first generation, 1G, was based on analog transmission that was prone to eavesdropping, interception, and cloning. 2G introduced the concept of digital modulation technique and was able to provide some basic security mechanism. Figurer 1 illustrates the security architecture evolution from 3G to 5G. 3G introduced two-way authentication and Authentication and Key Agreement (AKA), thus overcoming the limitations of one-way authentication in 2G.

4G features more diversified connection modes compared with its predecessors. The Diameter protocol used in 4G, however, is vulnerable to attacks, including attacks that track user location and intercept voice transmission to access sensitive information. Other security risks with 4G include downgrade attack, intercepting Internet traffic and text messages, causing operator equipment malfunction, and carrying out illegitimate actions [1], among others.

The 5G architecture is service-oriented, with many improvements on security introduced. 5G provides more efficient and secure mechanisms, such as unified

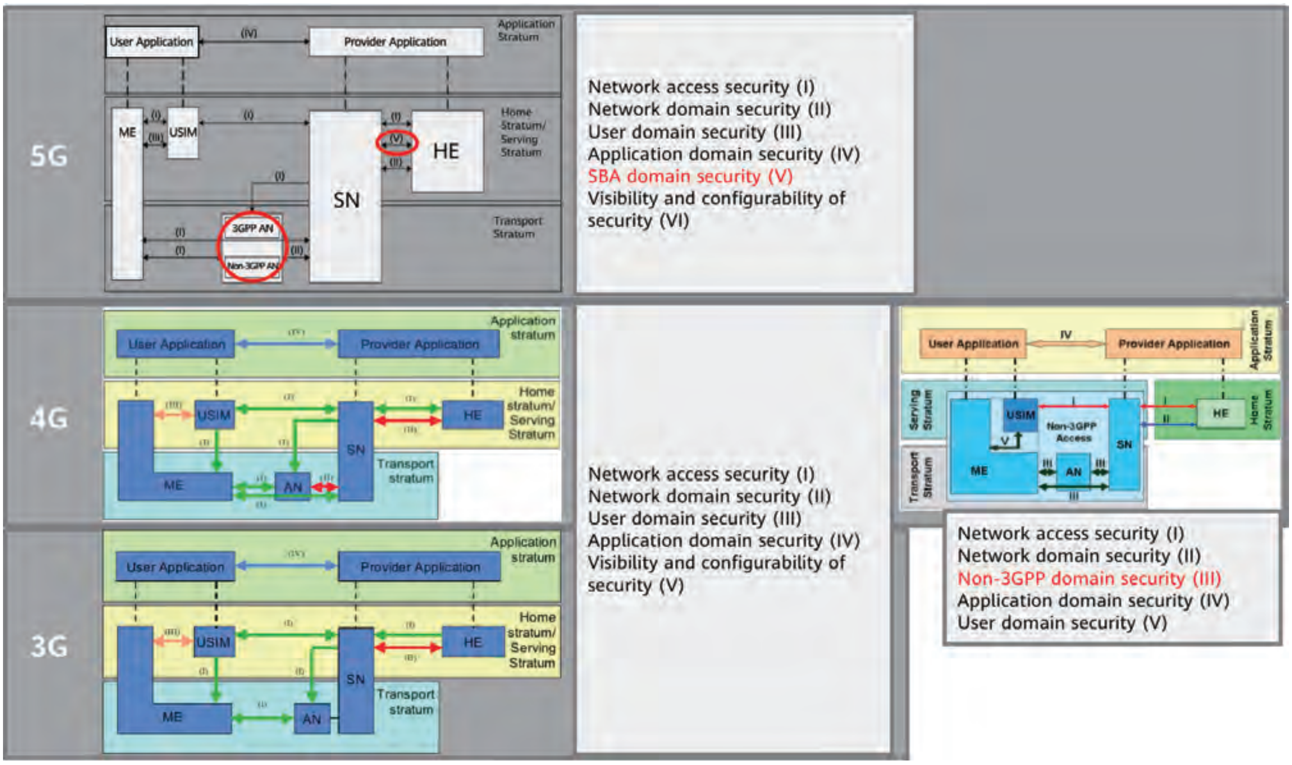

5G
User Application
(IV)
Provider Application
Application Stratum
ME
USIM
(I)
(III)
SN
HE
(V)
(II)
Home Stratum/ Serving Stratum
3GPP AN
Non-3GPP AN
Transport Stratum
Network access security (I)
Network domain security (II)
User domain security (III)
Application domain security (IV)
SBA domain security (V)
Visibility and configurability of security (VI)
4G
User Application
Provider Application
Application stratum
USIM
HE
SN
Home stratum/ Serving Stratum
ME
AN
Transport stratum
3G
User Application
Provider Application
Application stratum
USIM
HE
SN
Home stratum/ Serving Stratum
ME
AN
Transport stratum
Network access security (I)
Network domain security (II)
User domain security (III)
Application domain security (IV)
Visibility and configurability of security (V)
User Application
IV
Provider Application
Application Stratum
Serving Stratum
USIM
Non-3GPP Access
SN
HE
Home Stratum
Transport Stratum
ME
AN
Network access security (I)
Network domain security (II)
Non-3GPP domain security (III)
Application domain security (IV)
User domain security (V)

Figure 1 3GPP security architecture evolution authentication, Subscription Concealed Identifier (SUCI) that hides the subscriber ID during authentication, protocol-level isolation between slices, and secondary authentication serving service providers. The Security Assurance Specifications (SCAS) require that all network functions be tested by accredited evaluators so as to provide reference for operators.

The 5G security architecture is almost perfect. However, it is applicable to a centralized network architecture and the trust relationships between network elements in 5G are established at the protocol level, not involving device and network behavior. In the 6G ecosystem, trusted connections are key for all parties concerned, which extend security and privacy to a more inclusive framework — trustworthiness.

In order to build a 6G trustworthiness architecture that serves distributed networks and is compatible with the existing centralized networks, adopting new design concepts and developing new 6G-oriented trustworthiness capabilities is the top priority of 6G research.

ITU-T Recommendation X.509 defines trust in the ICT domain as follows: “Generally, an entity can be said to ‘trust’ a second entity when it (the first entity) makes the assumption that the second entity will behave exactly as the first entity expects” [2]. The ITU-T has been working on trust standardization focusing on the ICT domain from 2015, and has released several recommendations and technical reports [3–7] that describe the architectural and technical views of trust. The study of trustworthiness started with IoT as the first application domain. There were also reports and standards published that laid out the strategies to explore trust in other application domains like cybersecurity and networks. In 2017, trustworthiness was defined by NIST for the first time in the CPS domain as “demonstrable likelihood that the system performs according to designed behavior under any set of conditions as evidenced by characteristics including, but not limited to safety, security, privacy, reliability and resilience” [8]. Subsequently, in 2018, ITU-T approved the research for a new framework of security that focuses on establishing trust between entities in the 5G ecosystem [9].

Researchers have explored trust relationships extensively, applying different methodologies such as game theory and ontology [10] and analyzing risks in cloud-based modes. On the commercial side, several vendors and operators have been striving to meet consumer demands by continuously upgrading their product design and development, in which the “security by design” concept is emphasized and standardization policies are followed. All in all, it becomes imperative to define trustworthiness for future 6G communication networks.

2 Fundamentals of 6G Trustworthiness

In the following we explain the 6G trustworthiness framework we propose, which encompasses two vital principles, three objectives, and a multi-lateral trust model.

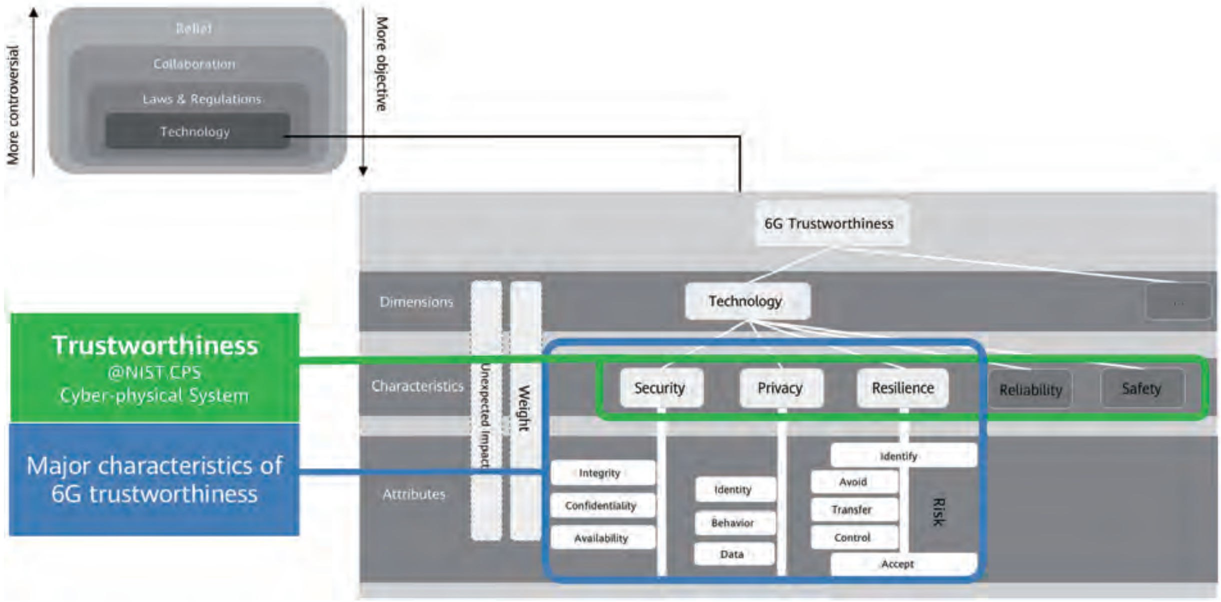


Figure 2 Trustworthiness framework

2.1 Principles

There are two principles to follow in the design of 6G native trustworthiness architecture.

- Principle 1: Trustworthiness of 6G characteristics

Driven by intelligent networks, 6G applications range from sensor networks to critical health-care and satellite communications. 6G trustworthiness must be able to meet the different requirements of holistic networks and diverse applications based on their technical and business domains, and be quickly adaptable for applications that require centralized authority and edge autonomy.

- Principle 2: Trustworthiness inherent in the 6G lifecycle

Trustworthiness requirements must be considered in tandem with network requirements in the entire 6G lifecycle, from design to development, operations, and maintenance. And trustworthiness analysis, assessment, and evaluation must be continuously performed to achieve satisfactory results.

2.2 Objectives

Security, privacy, and resilience are the three pillars of 6G trustworthiness. Each of the pillars are underpinned by unique underlying attributes as shown in Figure 2. To achieve trustworthiness in 6G networks, the 6G network architecture must meet the following objectives with regard to the three pillars:

- Objective 1: Balanced security

Security is supported by three attributes, confidentiality, integrity and availability (CIA). One of the essential criteria for 6G native trustworthiness is the ability to weigh the three attributes adaptively based on the applications and scenarios and ultimately achieve a balance between network quality/user experience and the security capability.

- Objective 2: Everlasting privacy preservation

User identify, user behavior, and user-generated data are the three types of data concerning privacy protection on a network. Only authorized parties can interpret information that reveals a user's identity and behavior. In 6G networks, user identity and user behavior have their uniqueness, which is rooted in the unified definition of user identity and the composition of signaling messages. User-generated data is not stored on the telecom network and is protected in the process of data processing and operations using techniques such as encryption and security management.

- Objective 3: Smart resilience

Resilience centers on risk analysis in a network. There are several stages of risk management. The first stage is to identify the risk factors. The second is to take suitable measures to avoid the risks by leveraging big data analytics. Then if the risks cannot be avoided, they can be transferred to other entities so that the network can be recovered successfully. And the after-effects must be controllable to the minimal level. If all the preceding measures cannot be taken, the final stage is to accept the risks causing only non-fatal damage to the network [11–13].

2.3 Multilateral Trust Model

We introduce a multilateral trust model as shown in Figure 3 in 6G to meet the needs of diversified trust scenarios.

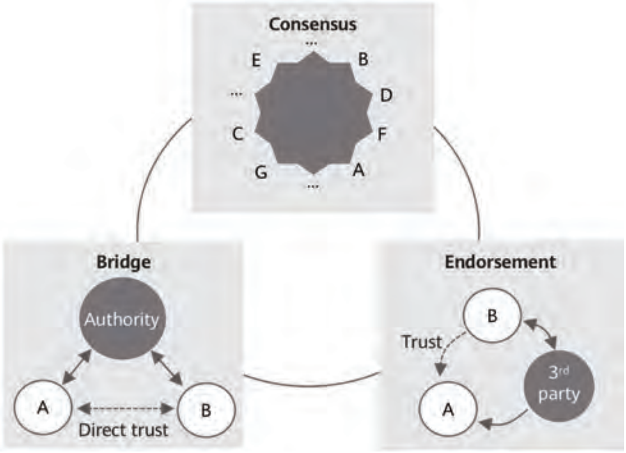


Figure 3 Multilateral trust model

A multilateral trust model includes three modes: bridge, endorsement, and consensus. In the bridge mode, an accreditation authority authenticates and authorizes entities A and B respectively, transfers trust between the two entities, and eventually establishes trust between them. The endorsement mode involves relying on third parties to evaluate an entity's trustworthiness. In this mode, a third party evaluates an entity's trustworthiness and submits the evaluation result to the other entity. The consensus mode is the most significant of the three as it adopts a decentralized architecture where transactions are distributed among entities. The entities involved in the consensus mode can be elements on a network, parties in a supply chain, or organizations in an industrial ecosystem. In this mode, transactions are attestable and responsibilities are shared among multiple parties. This powers this mode with high efficiency and scalability, enabling it to meet the agile and customized access requirements of 6G.

The three modes of this model should all be duly considered in the design of security architectures and mechanisms. And the enabling technologies for the three modes should all be researched and developed, such as the identity management and authorization technologies applicable to the bridge mode, the third-party security evaluation technologies suitable for the endorsement mode, and the blockchain technology ideal for implementing the consensus mode.

3 Enabling Technologies for 6G Trustworthiness

3.1 6G Blockchain

To establish a trust consortium based on which multiple parties can have mutual trust in one another for resources sharing and transactions can be performed autonomously, a customized blockchain for wireless networks is needed. 6G blockchain will serve as the basis for traceability mechanisms that ensure trust.

The following sections describe the convergence of blockchain and communications networks, blockchain technology under the privacy governance framework, and blockchain technology customized for wireless networks.

3.1.1 Convergence of Blockchain and Communications

6G blockchains can be classified into three types: independent blockchain, coupled blockchain, and native blockchain, depending on the degree of coupling between blockchain and the communications network.

- Independent blockchain

Independent blockchains are independent of the communications service and protocol processes, providing data storage and traceability for network O&M and management. Typical applications include roaming billing and settlement. These interactions, though not included in the signaling flows defined by 3GPP, are significant for establishing trustworthy relationships between operators and enhancing efficiency by utilizing smart contracts.

- Coupled blockchain

Coupled blockchains are those that interact with the communications network in the protocol process. The interactions include offline chaining and online checking. Take blockchain-based authentication as an example: The information owner or an authorized operator stores some information, such as a credential, or hash values into a blockchain in advance. When a communication request is initiated, the receiver authenticates the requester by looking up its credential in the blockchain, during which the requester waits for a response. If the authentication is successful, the receiver accepts the connection request and continues with the subsequent process.

- Native blockchain

A native blockchain refers to a blockchain whose algorithms, communication protocols, and enabling functions are all inherent in the communications networks. Writing to the blockchain and searching in it both occur online and in real time, as part of the communication process. However, the real-time application of blockchain technology in communications networks is faced with many new challenges. One of the goals of 6G is to create a real-time and large-scale blockchain system that serves as the foundation for network operational trustworthiness, so that every real-time data session and every real-time signaling transaction will be immutably recorded, for example, on a privilege-based super ledger [14]. Thus, there is the need to design a 6G customized blockchain architecture of low latency and high throughput, which satisfies the potential requirement of wireless communication and networks and also meets the privacy protection objectives.

3.1.2 Blockchain Compliant with Privacy Protection Framework

In recent years, a number of personal data privacy and security laws have been implemented around the globe, for example, “Data Security Law of the People's Republic of China” (PRC) [15], “Act on the Protection of Personal Information (Act No. 57 of 2003)” [16], “CLOUD Act” [17], and "California Consumer Privacy Act (CCPA)" [18] . Among these laws, the EU General Data Protection Regulation (GDPR) [19], is one of the toughest.

According to Article 5 of GDPR, all processing of personal data shall follow the principles of:

- Lawfulness, fairness, and transparency
- Purpose limitation
- Data minimization
- Accuracy
- Storage limitation
- Integrity and confidentiality
- Accountability

Cryptography, if applied appropriately, can help in complying with these principles. Given that, we are working on developing 6G oriented cryptographic solutions for customizing a 6G blockchain.

The following describes the zero-knowledge proof system [23], one of the preliminary ideas we proposed about privacy preservation on a 6G blockchain, which marks the start of our research.

In the Nakamoto model all the transactions in a blockchain are in plain text. Hence, a native privacy algorithm needs to be implemented to ensure that the data storage in a 6G blockchain is compliant with GDPR and other privacy regulations. The state-of-the-art technology zk-SNARK (zero-knowledge succinct non-interactive argument of knowledge) [20] is computationally complex because it requires several iterations to find the arithmetic roots of a polynomial equation so as to attain a soundness error within a threshold. Complexity is also involved in the trusted setup that involves computation of many cryptographic algorithms and time-consuming operations. The variants zkBoo and zkBOO++ [20–21] have eliminated the requirement of trusted setup and have employed garbled circuit [22–24], which is different from the arithmetic circuit used by zk-SNARK. However, they still use the monolithic statement for contract verification and auditing is a drawback and thus cannot be used practically for large systems.

We propose zk-Fabric, a native privacy framework based on the zero knowledge proof system [24]. It has the following features:

- The input parameter size is linear to the input.
- The solution is realized by Boolean gates circuit.
- The semantic statements from the prover are transformed to polylithic syntax.
- A non-interactive oblivious transfer (OT) based multi-party joint verification system is adopted.

Figure 4 shows the zk-Fabric framework consisting of three modules. The objective is to verify the statements of Alice anonymously without revealing secrets. Alice transforms her input statements into a Turing complete Boolean circuit

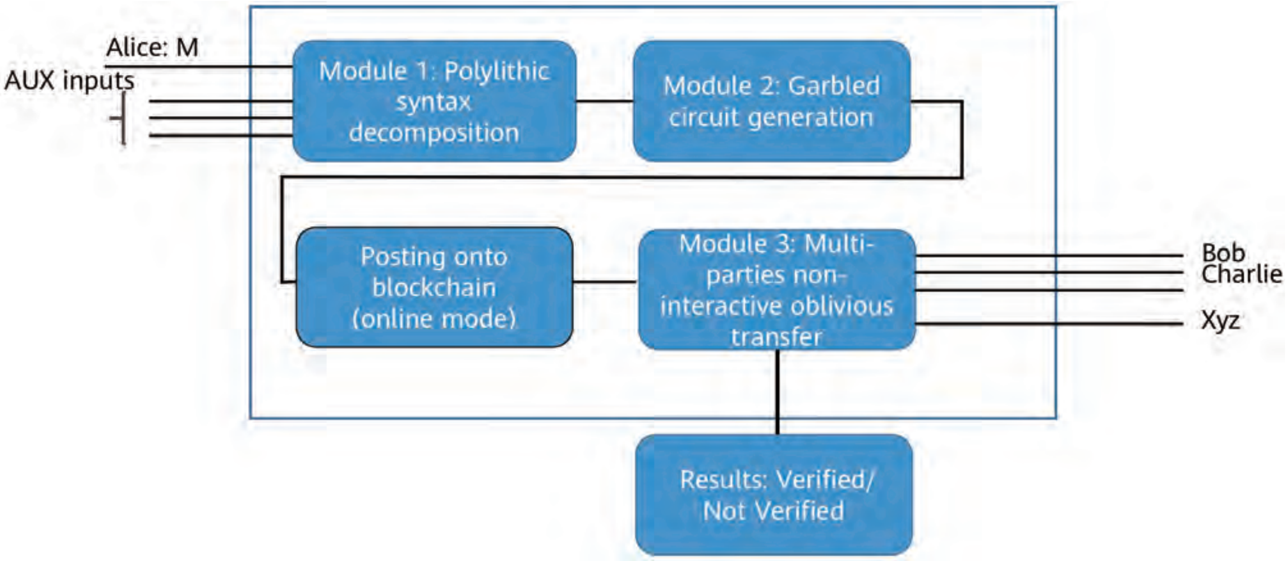

Alice: M
AUX inputs
Module 1: Polylithic syntax decomposition
Module 2: Garbled circuit generation
Posting onto blockchain (online mode)
Module 3: Multi-parties non-interactive oblivious transfer
Bob
Charlie
Xyz
Results: Verified/ Not Verified

Figure 4 zk-Fabric framework

with the decomposition algorithm in Module 1 (polylithic syntax decomposition) and partitioned garbled circuits for multiple verifiers in Module 2 (garbled circuit generation). The information is published on a publicly accessible blockchain. In Module 3 (multi-party non-interactive OT), the multiple verifiers verify the statements through the online public system.

In a nutshell, the zk-Fabric allows a cluster of verifiers to online, anonymously, and jointly compute a succinct digest of garbled circuits *C* which is prepared by a prover, who also practices the partitioning of the garbled circuit and randomly dispatches segments of them to a publicly accessible repository, i.e. the blockchain or a web portal. The goal is to build a more comprehensive public verification system which can validate more complex statements than other technologies that can only perform a monolithic verification, in other words, with which a verification can only conduct a single hashed value in an arithmetic circuit at a time. The zk-Fabric framework also achieves full privacy preservation computation (encrypted computation) based on OT and garbled circuit.

For security evaluation, we demonstrate that zk-Fabric can maintain privacy against the semi-honest threat model (Note: zk-Fabric may not be sufficient in protection against the "Malicious" model). We can formalize this using a generalized Fiat-Shamir's secret sharing scheme, which defines a -secure n-party protocol and packs *l* secrets into a single polynomial. One can run a joint computation for all inputs by just sending a constant number of field elements to the prover. As a result of packing *l* secrets into a single polynomial, we can reduce the security bound *t* of zk-Fabric with multiple verifiers as $t = \frac{n-1}{2}$ to $t' = t - l + 1$. In zk-Fabric, OT is a very useful building block in achieving protection against semi-honest participants.

For computational efficiency, we demonstrate that zk-Fabric can achieve efficiency with two key refinements. First, we employ the Karnaugh Map technique to reduce the number of logical gates with a simplified expression. Second, we build garbled circuits with partitions by tightly integrating the verification procedure with a multi-party OT scheme. This reduces computational costs on the verifiers' side compared with native approaches.

Note that our security definition and efficiency requirement immediately imply that the hash algorithm used to compute the succinct digest must be collision resistant.

Inspired by the security notions of OT-Combiners, we start with the construction of an overall zk-Fabric system that builds on the partitioned OT scheme. Figure 5 shows an example of two polylithic inputs to be “blindly” verified by three offline verifiers with the construction of partitioned garbled circuits.

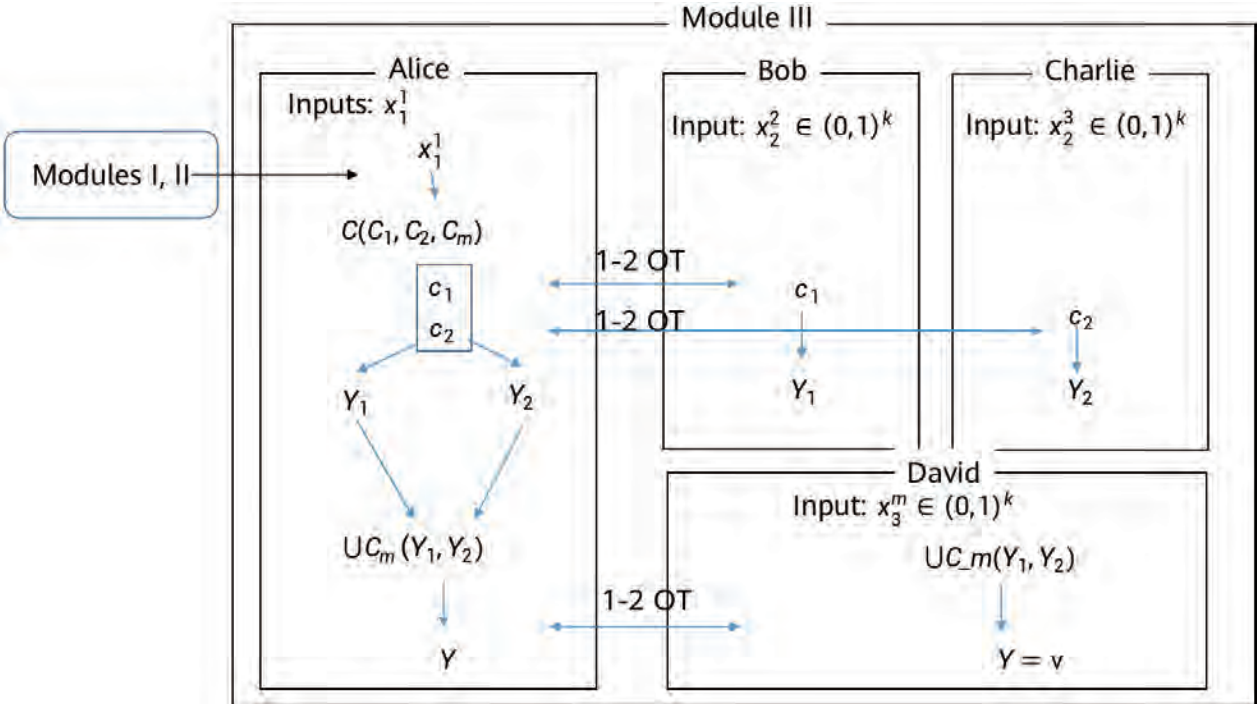


Figure 5 zk-Fabric system

3.1.3 Blockchain Customized for Wireless

6G networks feature faster data transfer rates, lower latency, and more reliable communications than their predecessors. The following are some key data of 6G:

- Peak rate: 100 Gbit/s to 1 Tbit/s
- Positioning accuracy: 10 cm indoors and 1 m outdoors.
- Communication delay: 0.1 ms
- Battery life of devices: up to 20 years
- Device density: ~100 devices per cubic meter
- Downtime rate of devices: one millionth
- Traffic on communication channel: about 10,000 times as much as that of today's networks

However, bitcoin currently has a transaction throughput of 7 transactions per second (TPS), Ethereum has a transaction throughput of 15–20 TPS, and Hyperledger Fabric has a transaction throughput with order of magnitude as high as 10^3. The low throughput of blockchain transactions forms a sharp contrast with the high performance of 6G. In most service scenarios, particularly high-frequency trading scenarios, the current blockchain cannot meet the actual application requirements. Therefore, we need to continue researching on the consensus algorithm in blockchain to improve the consensus efficiency and enhance the scaling techniques. Meanwhile, we need to boost throughput from the aspect of system architecture.

Based on the 6-layer blockchain architecture model, popular capacity expansion technologies can be classified into three schemes depending on the layers.

- Layer-0 scalability optimizes the data transmission protocols at the network and transport layers of the OSI model, without changing the upper-layer architecture of the blockchain. It is a performance improvement solution that retains the blockchain ecosystem rules. Layer 0 scalability involves relay network optimization and OSI model optimization.
- Layer-1 scalability (on-chain scalability) optimizes the structure, model, and algorithms of the blockchain across the data layer, network layer, consensus layer, and incentive layer to improve the blockchain performance.
- Layer 2 scalability (off-chain scalability) executes contracts and complex computing off the chain to reduce the load on the blockchain and improve its performance. Off-chain scalability does not change the blockchain protocol. The current technologies for off-chain scalability include payment channel, sidechain, off-chain, and cross-chain technologies, among others.

The “scalability trilemma” states that any blockchain technology can never feature all three organic properties of blockchain — scalability, decentralization, and security. When scalability is enhanced, decentralization and security will be compromised. Therefore, research on 6G blockchain is not just about improving throughput. It should also cover selecting appropriate technology paths based on the 6G characteristics to strike a balance among the three properties of blockchain and to ensure its adaptability to 6G scenarios.

3.2 Quantum Key Distribution

The first quantum key distribution (QKD) protocol was proposed by C.H.Bennett and G. Brassard [25] in 1984, and is known as BB84 after its inventors and year of publication. In this protocol, the sender (Alice) and the receiver (Bob) wish to agree on a secret key. Alice sends each bit of the secret key in a randomly selected set of conjugate basis through transposition quantum gate transformation, to Bob. An eavesdropper (Eve), unaware of the basis used, cannot decode the quantum bit (qubit) by measuring in the middle, as once a qubit in transposition is being measured by the eavesdropper, it collapses into a state which ultimately introduces errors in Bob’s measurements. This is known as the non-locality theorem [26].

BB84 and its variants are designed for point-to-point (Alice to Bob) setup, which has its limitations, for example, it remains a challenge to deliver entangled qubits to more than two parties. In this paper, we discuss a multi-user (MU) QKD protocol which utilizes two entangled qubits to deliver a secret key to multiple parties with n = 3. In our design, we utilize a centralized trust model in which a key operator (O) can manage the subgroups of nodes, and the subgroups rely on the operator (O) to distribute the key securely through the QKD protocol. In the end, through the quantum correlation routine at the operator over the authenticated classic channel, all three parties obtain the shared key.

As an extension, the MU QKD protocol can be applied to more than three parties by keeping the operator as the trust anchor point and iteratively reusing the three-party MU QKD protocol. Thus the shared key can be obtained by n = 2ℓ + 1 nodes.

The MU QKD protocol can be put into extensive practical use given its broadcast nature, with security ensured by the underlying quantum physics. One of the applications is mobile phone key distribution, where a key operator is able to multicast the pre-shared key for authentication to multiple end nodes. Another application is for quantum repeaters. A prominent challenge in transmitting qubits on quantum Internet is that qubits cannot be copied, which naturally rules out signal amplification or repetition for overcoming transmission losses and bridging great distances. To enable long-distance quantum communication and implement complex quantum applications, most of the current literature models quantum repeater with the "Store and Forward" quantum mechanics, such as Quantum Memory [27]. The "Store and Forward" qubits manipulation essentially breaks down the point-to-point basis of QKD, and it poses challenges to obtain end–to-end provable security.

3.3 Physical Layer Security (PLS)

The higher frequency bands such as the millimetre waves and terahertz waves, higher bandwidth, and larger antenna arrays in 6G networks open up new horizons for the design and development of physical layer security. In this article, physical layer security refers specifically to the use of physical layer technologies for security.

The following key characteristics of 6G wireless signals can be leveraged to provide secure communication between legitimate parties:

- Multi-path fading: As wireless signals are transmitted, they undergo large- or small-scale propagation fading as the result of obstruction by objects such as buildings and hills during transmission. Moreover, reflections and scatterings from various objects cause multi-path fading and the components of the signal vary with distance.
- Time-varying: The wireless signals exhibit time-varying property as both the transmitter and the receiver are on-the-go and the radio waves experience scatterings, reflections and refractions due to the presence of many stationary and moving objects around.
- Reciprocity: Wireless channels are reciprocal in space, implying that the channel responses can be estimated in either direction if measured within the channel coherence time.
- Decorrelation: The channel responses exhibit rapid temporal and spatial decorrelation.

3.3.1 Physical Layer Technology Contributing to Secret Key Generation

The decorrelation and reciprocity of wireless signals can be leveraged to extract secret keys between two legitimate entities. Researchers have leveraged the physical-layer based features for secure device pairing and secret key generation [28–29]. Secret key generation consists of two steps: (i) channel sampling and (ii) key extraction. In the first step, the two legitimate entities exchange a series of probes to measure the channel between them. The channel measurements can either be in the frequency or time domain. In the second step, the channel measurements are converted to a sequence of secret bits through quantization. This extracted key can be combined with higher layer security algorithms for encryption. Given the presence of a passive adversary Eve who eavesdrops all signals transmitted between the two parties, the channel estimate will not be correlated with those of either Alice or Bob as the signals undergo multi-path fading. It is a challenging task for Eve to retrieve the same secret keys as the legitimate parties do. As shown in Figure 6, the legitimate devices observe similar characteristics whereas the eavesdropper gets different channel estimates.

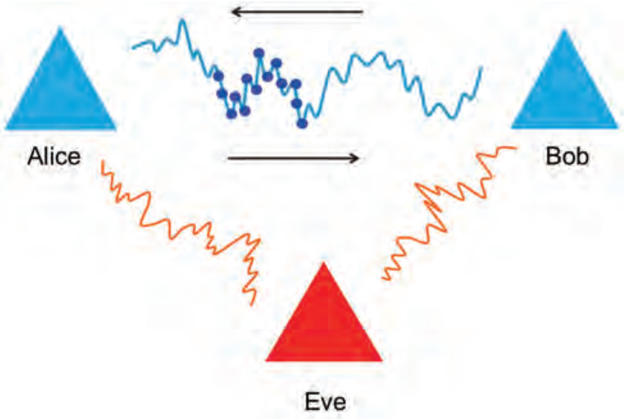


Figure 6 Channel characteristics between legitimate and non-legitimate devices

The following describes the basic physical layer metrics that are essential to measure the security performance:

- Entropy is the amount of randomness in the information content and is defined by:

$$H(M) = -\sum p(m) \log p(m)$$

where $p(m)$ is the probability that takes on the value of message M.

- Mutual information is a quantity indicating how secure a communication channel is. If the mutual information between message M and the encrypted message X intercepted by Eve is zero, the communication channel is considered secure. It can be expressed as:

$$I(M; X) = 0$$

Or, it can be expressed in terms of entropy as:

$$I(M;X) = H(M) - H(M|X)$$

where $H(M|X)$ is the conditional entropy defined as the remaining uncertainty in message M after observing the encrypted message X.

- Secrecy rate is the rate at which a message is transmitted to the legitimate receiver, while being intercepted by Eve. It is expressed as:

$$C_s = C_B - C_E$$

where C_B and C_E are the secrecy rates of Bob and Eve respectively. The secrecy rate can be increased using signal design and optimization techniques.

- Secrecy outage probability is the probability at which a specified value of secrecy capacity C_s cannot be attained by a system. Here limited channel information of Bob and Eve is available to Alice.

Bit error rate (BER) is the number of bit errors received divided by the total number of bits transmitted. The BER for legitimate entities must be lower than that for adversaries.

3.3.2 Physical Layer Technology Contributing to Authentication Protocol

Researches have also been carried out on physical layer technologies used for security authentication. Researchers proposed to use the indoor-based Wi-Fi channel characteristics for generating proof of location for mobile users [30]. Proof of location is evidence that attests a user's presence at a particular time and location. This proof is provided by a trusted entity for mobile users. With this proof, the mobile users can be verified for authenticity by the service providers. The research demonstrates that the proof of location is kept secure, not being tampered by an adversary nor modified or transferred to other users.

4 Privacy Preservation

'Personal data' means any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person [31]. In telecom networks, personal data can be categorized into three types: user IDs, user-generated data, and user behavior.

- User IDs

A telecom network assigns personal IDs such as lifetime network IDs, service IDs, and fine-granularity temporary IDs for users. On a telecom network, personal IDs are fully protected. In 5G, initial IDs used for user authentication on the network are protected by end-to-end encryption.

- User-generated data

User-generated data, such as the contents of a phone call and an application on the

Internet, are neither stored on the network nor analyzed by operators. Such data is encrypted during transmission and cannot be understood by interceptors.

- User behavior

The behavior of UEs accessing, leaving, or performing a handover can be observed on the control plane. To hide user information, the network provides an encryption scheme for signals. If the signal encryption is not implemented on the network, users' habits, such as the frequency of phone calls and the movement between locations, can be estimated by tracing the user and observing the signaling changes.

In the 6G era, it will be a challenging task to preserve privacy and protect personal information. With AI-enabled decision-making for applications, consumers will be able to enjoy services tailored to their preferences, but they may not be aware of the unprecedented amount of personal data that has to be collected for such personalized services. For instance, autonomous driving and smart-home applications will collect sensitive information such as the user's location as a user drives. Using smart appliances will reveal that an individual is present at his/her residence. Cloud-based storage will open up doors for privacy breaches. A report [32] lists several data breaches that took place in the 21st century, where a number of records and accounts related to individuals were exposed.

In order to prevent privacy breaches, privacy preservation must be considered in the design phase of 6G lifecycle and managed in all stages involving data operation. ENISA [33] laid out eight privacy design strategies as explained by Jaap-Henk Hoepman in *Privacy Design Strategies* [34]. The strategies are divided over two categories: data-oriented strategies and process-oriented strategies. The data-oriented strategies (as shown in Table 1) focus on preserving the privacy of the data themselves and the process-oriented strategies (as shown in Table 2) focus on the methodologies/approaches for data processing.

Table 1 Data-oriented strategies

Minimize	Limit the processing of personal data as much as possible.
Separate	Separate the processing of personal data as much as possible.
Abstract	Limit the detail in which personal data is processed as much as possible.
Hide	Protect personal data, or make it unlinkable or unobservable. Make sure it does not become public or known.

Table 2 Process-oriented strategies

Inform	Inform data subjects about the processing of their personal data in a timely and adequate manner.
Control	Provide data subjects adequate control over the processing of their personal data.
Enforce	Commit to processing personal data in a privacy-friendly way, and adequately enforce this.
Demonstrate	Demonstrate you are processing personal data in a privacy-friendly way.

Several privacy enhancing technologies have been researched in depth for more than a decade. These researches focused on minimizing personal data to avoid any unnecessary process-oriented tasks. Following are some of the privacy enhancing technologies:

- Homomorphic encryption (HE) allows computations to be performed on encrypted data. The results are encrypted and do not reveal any information about the data themselves. Users can decrypt the data and analyze the results. HE can be classified into partial HE and full HE. The pioneering work on HE dates back to 2009 when Gentry proposed the first full HE scheme, and several improved schemes have been introduced over the following years. Even so, the implementation of HE was limited as it required a thorough understanding of the HE scheme and the complex underlying mathematics. To address this issue, an open source project SEAL [35–36] was introduced by Microsoft with the intention to make HE schemes available for everyone. SEAL provides a convenient API interface and many illustration examples to show the correct and secure way of using the interface, along with related study materials. Traditionally, in applications involving cloud storage and data processing, the end users need to trust the service provider responsible for storing and managing data, and the service provider must ensure that user data is not exposed to any third parties without the user's consent. SEAL manages this concept systematically by replacing the trust with well-known cryptographic solutions. This not only enables processing of encrypted data, but also guarantees protection of user data. Below we explain the important factors to be considered when designing solutions with HE:

    - The performance overhead is very large, since HE increases the size of original data by several fold. Thus, it is not recommended for all applications.

    - In HE solutions, a single secret key is held by a data owner. Thus, co-computing between multiple data owners requires a multi-key fully homomorphic scheme.

    - The security property of homomorphic encryption determines that it can only provide passive security, and can't guarantee the security of applications using it in the active attack environment.

- Zero-knowledge proof (ZKP) is a cryptographic technique that verifies information without having to reveal the information itself. Researchers at MIT developed this concept. A ZKP protocol must satisfy the following properties:
  - Completeness: If the prover submits legitimate information, then the protocol must allow the verifier to verify that the information submitted by the prover is true.
  - Soundness: If the prover submits false information, then the protocol must allow the verifier to reject the claim by the prover.
  - Zero-knowledge: The method must only allow the verifier to determine the authenticity or falsity of the information submitted by the prover without having to reveal anything.

ZKP can be categorized into interactive ZKP and non-interactive ZKP. As the name suggests, in interactive ZKP there are several interactions between verifier and prover and the verifier challenges the prover several times until the verifier is convinced. However, in non-interactive ZKP there is no interaction between the two parties. zk-SNARK (zero-knowledge succinct non-interactive argument of knowledge) [37] and zk-STARK (zero-knowledge scalable transparent argument of knowledge) [38] are non-interactive ZKP protocols. zk-SNARK was first used in the Zerocash blockchain protocol [39] which enables a participant to prove its possession of particular information without revealing the information itself. zk-STARK was released in 2018 offering transparency i.e., no requirement of trusted setup and poly-logarithmic verification time.

- Secure multi-party computation (SMPC): As an extension of HE, SMPC allows multiple parties to work on the encrypted data, with no party being able to view the other parties' information. This ensures that data is kept private in SMPC. The natural advantage of SMPC has encouraged several research projects on machine learning to ensure privacy. Facebook AI has developed CrypTen [40], a privacy preserving framework based on SMPC. It is a software built on PyTorch, and researchers familiar with machine learning can call the API to build applications for privacy preserving.

- Differential Privacy (DP) protects the privacy of individuals by injecting statistical noise to the dataset using a cryptographic algorithm. The noise layer helps distinguish different groups in a particular dataset. Although the proposed method has very little impact on the accuracy of data, it ensures plausible deniability and hence preserves the privacy of individuals. In DP systems, a user needs to submit "query" to obtain the data of interest. The system then performs an operation known as "privacy mechanism" to add some noise to the data requested. This function returns an "approximation of the data", thereby hiding the original raw data. The output "report" of a query consists of the privacy-protected result along with the actual data calculated and a description about the data calculation. The following are two important metrics in DP:
  - Epsilon: It is a non-negative value that measures the amount of noise or the privacy of output report. It is inversely proportional to noise/privacy, that is, a lower Epsilon means more noise/private data. If the Epsilon value is larger than 1, it indicates that the risk of exposing actual data increases. Hence, the

ML/AI models must aim to limit the value within the range of 0–1.

- Delta: It measures the probability of the report being non-private. It is directly proportional to Epsilon.

DP systems mainly aim for data privacy, however, one must be aware of the underlying tradeoff between data usability and data reliability. If the noise and privacy level increases, the Epsilon value reduces, and the accuracy and reliability of data decreases.

There is a popular open source project called SmartNoise [41], which aims to help the implementation of DP in ML solutions. It has two main components:

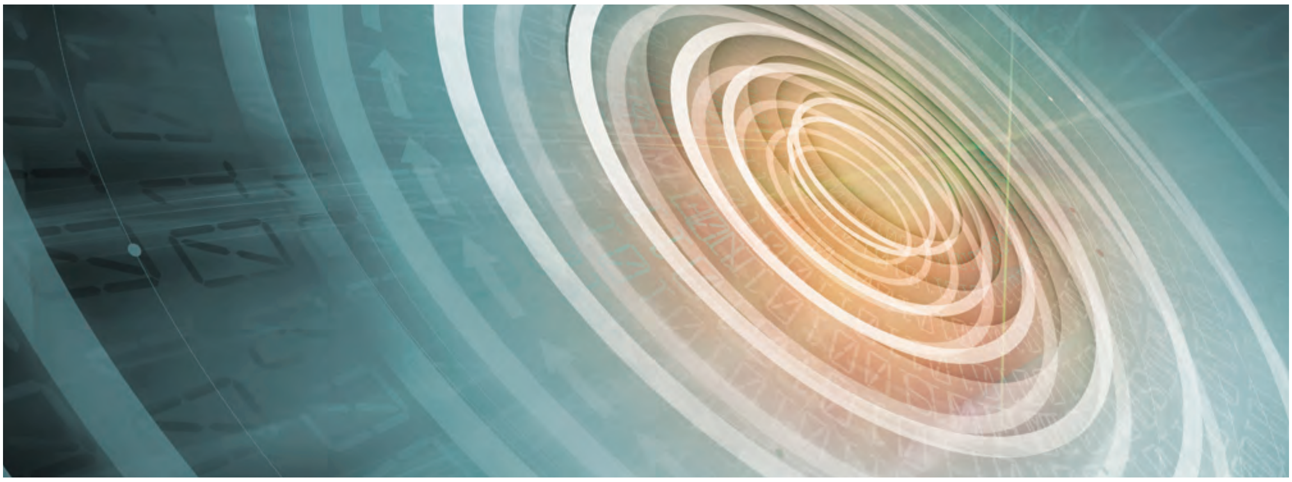
- Core Library, where many privacy mechanisms are stored
- SDK Library, where tools and services required for data analysis are stored

5 AI Security and Trust

In 6G, services and applications will be highly intelligent and autonomous. The E2E architecture of 6G will be based on blockchain and AI. This will involve working on a tremendous amount of data and decision-making of the network will be entirely based on data analytics. Therefore, there is a need to secure the AI systems throughout the machine learning lifecycle, which mainly consists of data acquisition, data curation, model design, software build, training, testing, and deployment and updating. In each stage of the lifecycle, confidentiality, integrity and availability must be ensured to keep the model secure. If security is not prioritized in the early stages of the lifecycle, adversaries can tamper the models for many important applications like healthcare and autonomous driving, leading to severe consequences.

Following are the different types of attacks:

- Poisoning: An AI model is compromised by attackers and does not behave in the way it is designed or intended to perform a specific task, but rather behaves in a way the attackers want it to. An adversary can launch such an attack by (a) poisoning the data set, i.e. injecting incorrect data into the training data set or incorrectly labelling the data, (b) poisoning the algorithm, or (c) poisoning the model.
- Evasion attack: The data input during deployment or testing is tampered so that the learned model deviates from the correct results. An example of evasion attack is modification of traffic signal. A minor modification will lead to an autonomous car misinterpreting the traffic signal and making wrong decisions.
- Backdoor attack: Such an attack is triggered only when a specific pattern is input to a model. The model behaves normally when provided with inputs not in the triggering pattern. Hence, it is a challenging task to validate if a model is subject to or compromised by a backdoor attack. This kind of attacks can happen in both training and testing. If an attacker injects input in the attack-triggering pattern to the model during training, undesired results will be output by the model after it is deployed.
- Model extraction: An attacker analyzes the input, output and other related information of the target model and performs reverse engineering to construct a model same as the target model.



There are different defense mechanisms for the aforementioned attacks. Adversarial training is the approach of feeding adversarial inputs to the training dataset and optimizing the model iteratively until it behaves correctly. This method can improve the robustness of models against predicted attacks. Adversarial training can be used to defend against evasion attacks. Another defense mechanism is defensive distillation, which is based on the concept of transfer learning of knowledge from one model to another. A model is trained to perform the probabilities of another model that is trained to provide accurate outputs.

However, these defense mechanisms may not be able to ensure security against all attacks. Some other methods should also be considered, such as noise injection, enhancing the data quality during the training phase, and inserting an additional layer that detects the attacks on the model. Synergizing multiple defense approaches can help make AI more secure and robust.

6 Measurement, Verification and Attestation

As described previously, trustworthiness embodies three foundation pillars: security, privacy, and resilience. Assessing the trustworthiness of a system inherently involves evaluating the three foundation pillars continuously and iteratively throughout the 6G lifecycle. In this following we explain the methodologies that can be adopted in this regard.

6.1 Security Analysis

Security analysis of network protocols can be conducted in two approaches: (i) logic and symbol computation and (ii) computational complexity theory. The first approach uses cryptographic primitives and is the foundation of many automated tools, whereas the second approach involves reasoning and computational complexity and scores the strengths and vulnerabilities of the protocols. Some of the tools used for security analysis of protocols are: Tamarin prover, ProVerif, AVISPA, and Scyther. The 6G E2E architecture will involve authentication and key agreement protocols between various entities, it is therefore essential to perform security analysis of the protocols to identify vulnerabilities and security loopholes and prevent devastating outcomes.

6.2 Privacy Protection Framework and Privacy Verification

Network privacy must be considered as early as in the design phase. GDPR requires that organizations comply with all the privacy requirements. Data controllers must frequently review and audit the process of data-oriented tasks and ensure that the tasks adhere to the data protection policies. GDPR also offers privacy certification service. It is an indication to customers that certified organizations will adhere to the standards on data privacy and protection. This certificate is already used by some products and websites [33]. Also, GDPR has proposed an initiative — PDP4E that provides software tools and methodologies for organizations to validate whether their applications and products comply with the GDPR policies [42]. The tools mainly focus on four aspects:

- Privacy risk management
- Gathering privacy-related requirements
- Privacy and data protection by design framework
- Assurance framework

6.3 Trustworthiness Measurement and Security Evaluation

To achieve trustworthiness, it is also necessary to measure network resilience continuously. Risks have to be identified and analyzed and timely action must be taken to prevent serious consequences. Similar to the quantitative measurement of quality of service (QoS) and quality of experience (QoE), ITU-T has mentioned that a quantitative method can be employed to measure trustworthiness [13]. This trustworthiness, however, is application specific and dependent on the use scenarios. In addition, security risks can be evaluated either quantitatively or qualitatively. The quantitative measurement, as the name signifies, assigns a numeric value as the risk level, whereas the qualitative approach assigns a rating based on the possible consequences [43]. Risk analysis can be done to re-consider/re-evaluate the security solutions, thereby eliminating threats and mitigating risks.

7 Conclusion

The 6G network shall aim for seamless intelligent connectivity of all the devices that have the network capability. Compared to all the previous generation technologies, 6G will raise the level of user experience. As 6G shall extend the massive machine communication, ultra-low reliable latency and enhanced mobile broadband, the three pillars of 5G to sensing and AI, ensuring a trusted network becomes a challenging task. In this paper, we have first presented the fundamentals of 6G trustworthiness architecture i.e., the principles and objectives and the multi-lateral trust model design for 6G. The enabling technologies for 6G viz., blockchain, quantum key distribution and physical layer security have also been discussed. The 6G networks will collect and process significant amount of data to provide network services, by applying artificial intelligence and machine learning. Hence, preserving the privacy of consumers will be of paramount importance in the future networks. Privacy by design principles that explain the design strategies in the 6G lifecycle have been discussed in the paper along with the potential technologies that preserve the privacy. Furthermore, AI being one of the main enablers in 6G architecture, can act as both a defense and an attack that are covered in the present paper. Finally, the approaches to assess the trustworthiness of a system is presented in the paper. Extensive research is still to be undertaken to meet the challenges of security and privacy issues in tandem with the 6G enabling technologies. We hope that this paper will act as a catalyst and help researchers and scientists to pursue further advanced research that will help in standardizing 6G technologies.

References

[1] "Threats to packet core security of 4G network," *Positive Technologies.*

[2] "Recommendation X.509 Information Technology - Open Systems Interconnection - The Directory: Public Key and attribute certificate frameworks," ITU-T, 2019.

[3] "Standardization of Trust Provisioning Study," ITU-T, 2015.

[4] "Future Social Media and Knowledge Society," ITU-T, 2015.

[5] "Trust Provisioning for Future ICT Infrastructure and Services," ITU-T, 2016.

[6] "The basic principles of trusted environment in ICT infrastructure," ITU-T Recommendation Y.3501, 2017.

[7] "Overview of Trust Provisioning in ICT Infrastructures and Services," ITU-T Recommendation Y.3502, 2017.

[8] E. R. Griffor, C. Greer, D. A. Wollman, and M. J. Burns, "Framework for cyber-physical systems: Volume 2, working group reports, Version 1.0," in *Proc. NIST Special Publication 1500-202*, 2017.

[9] "X.5Gsec-t: Security framework based on trust relationship for 5G ecosystem," ITU-T, Draft Recommendation, 2021.

[10] M. Balduccini, E. Griffor, M. Huth, C. Vishik, M. Burns, and D. Wollman, "Ontology-based reasoning about the trustworthiness of cyber-physical systems," in *Proc. Living in the Internet of Things: Cybersecurity of the IoT,* 2018, pp.10.

[11] Wikipedia. "Law". Wikipedia.org. Available: https://en.wikipedia.org/wiki/Law (Accessed Sept. 21, 2020)

[12] "Measurement frameworks and metrics for resilient networks and services," *Discussion Draft, European Network and Information Security Agency*, 2011.

[13] Trust in ICT, ITU-T, 2017.

[14] Wen Tong, "6G-blockchain: open issues and directions," *2021-2022 Blockchain Research Seminar Series*. Available: https://www.fields.utoronto.ca/talks/6G-Blockchain-Open-Issues-and-Directions

[15] "Data Security Law of the People's Republic of China (《中国人民共和国数据安全法》)." Available: http://www.npc.gov.cn/npc/c30834/202106/7c9af12f51334a73b56d7938f99a788a.shtml

[16] "Act on the protection of personal information (Act No. 57 of 2003)." Available: https://www.cas.go.jp/jp/seisaku/hourei/data/APPI.pdf

[17] "CLOUD Act." Available: https://www.congress.gov/bill/115th-congress/house-bill/4943

[18] "California Consumer Privacy Act (CCPA)." Available: https://oag.ca.gov/privacy/ccpa

[19] GDPR Article 5. Available: https://gdpr-text.com/read/article-5/ (Accessed online 13 Dec 2021)

[20] Jens Groth, "On the size of pairing-based non-interactive arguments." Available: https://eprint.iacr.org/2016/260.pdf

[21] Irene Giacomelli, Jesper Madsen, and Claudio Orlandi, "ZKBoo: faster zero-knowledge for Boolean circuits," in *Proc. of USENIX Security Symposium 2016*, pp 1069-1083.

[22] Yao, Andrew Chi-Chih, "How to generate and exchange secrets," *27th Annual Symposium on Foundations of Computer Science (sfcs 1986), Foundations of Computer Science*, 1986.

[23] Sheng Sun and Tong Wen, "zk-Fabric, a polylithic syntax zero knowledge joint proof system." Available: arXiv:2110.07449

[24] Goldreich, Oded, "Cryptography and cryptographic protocols," in *Distributed Computing - Papers in Celebration of the 20th Anniversary of PODC*, 2003.

[25] C. H. Bennett and G. Brassard, "Quantum cryptography: public key distribution and coin tossing," in *Theoretical Computer Science - TCS*, vol. 560, pp. 175-179, 1984.

[26] B. Cirel' son, "Quantum generalizations of Bell's inequality," *PLetters in Mathematical Physics*, pp. 93-100, 1980.

[27] A. Lvovsky, B. Sanders, and W. Tittel, "Optical quantum memory," *Nature Photon*, vol. 3, p. 4706-714, 2009.

[28] Chitra Javali, Girish Revadigar, Lavy Libman, and Sanjay Jha, "SeAK: secure authentication and key generation protocol based on dual antennas for wireless body area networks," presented at the 10th Workshop on RFID Security (RFIDSec), Oxford, UK, 2014.

[29] Chitra Javali, Girish Revadigar, Ming Ding, and Sanjay Jha "Secret key generation by virtual link estimation," presented at the 10th EAI Conference on Body Area Networks *(BodyNets)*, Sydney, Australia, 2015.

[30] Chitra Javali, Girish Revadigar, Kasper. B. Rasmussen, Wen Hu, and Sanjay Jha, "I am Alice, I was in wonderland: secure location proof generation and verification protocol," in *Proceedings of 41st IEEE International Conference on Local Computer Networks (LCN), Dubai, UAE, Nov 7 - 10*, 2016.

[31] GDPR Article 4. Available: https://gdpr-text.com/read/article-4/ (Accessed online 2021/11/13)

[32] List of data breaches. Available: https://www.csoonline.com/article/2130877/the-biggest-data-breaches-of-the-21st-century.html Accessed online 2021/11/13

[33] "Privacy and data protection by design - from policy to engineering," 2014.

[34] Jaap-Henk Hoepman, "Privacy design strategies - (extended abstract)," *ICT Systems Security and Privacy Protection - 29th IFIP TC 11 International Conference, SEC*, 2014.

[35] https://www.microsoft.com/en-us/research/project/microsoft-seal/ (Accessed online 13 Dec 2021)

[36] https://docs.microsoft.com/en-us/azure/architecture/solution-ideas/articles/homomorphic-encryption-seal (Accessed online 13 Dec 2021)

[37] N. Bitansky, R. Canetti, A. Chiesa, and E. Tromer, "From extractable collision resistance to succinct non-interactive arguments of knowledge, and back again," in *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference on - ITCS '12*, 2012.

[38] Eli Ben-Sasson, Iddo Bentov, Yinon Horesh, and Michael Riabzev, "Scalable, transparent, and post-quantum secure computational integrity," *International Association for Cryptologic Research*, 2018.

[39] Zerocash. Available: http://zerocash-project.org/ (Accessed online 14 Dec 2021)

[40] B. Knott, S. Venkataraman, A.Y. Hannun, S. Sengupta, M. Ibrahim, and L.J.P. van der Maaten, "CrypTen: secure multi-party computation meets machine learning," in *Proceedings of the NeurIPS Workshop on Privacy-Preserving Machine Learning 2020*.

[41] Smart Noise. Available: https://smartnoise.org/ (Accessed online 14 Dec 2021)

[42] GDPR PDP4E. Available: https://cordis.europa.eu/project/id/787034. (Accessed online 2021/11/15)

[43] S. Harris and F. Maymi, "CISSP," *8th Edition, McGrawHill Education*, 2018

Fast Polar Codes for Terabits-Per-Second Throughput Communications

Jiajie Tong [1], Xianbin Wang [1], Qifan Zhang [2], Huazi Zhang [2], Rong Li [1], Jun Wang [1]

[1] Wireless Technology Lab

[2] Ottawa Wireless Advanced System Competency Centre

Abstract

Targeting high-throughput and low-power communications, we implement two successive cancellation (SC) decoders for polar codes. Converted to 16 nm ASIC technology, the area and energy efficiencies are 4 Tbps/mm^2 and 0.63 pJ/bit, respectively, for the unrolled decoder, and 561 Gbps/mm^2 and 1.21 pJ/bit, respectively, for the recursive decoder. To achieve such a high throughput, a novel code construction, referred to as fast polar codes, is proposed and jointly optimized with a highly-parallel SC decoding architecture. First, we reuse existing modules to fast decode additional outer code blocks, and then we modify code construction to facilitate faster decoding for all outer code blocks to a degree of parallelism of up to 16. Furthermore, parallel comparison circuits and bit quantization schemes are customized for hardware implementation. Collectively, they contribute to a 2.66× area efficiency improvement and a 33% energy saving over the state-of-the-art.

Keywords

fast polar codes, Tbps communication, fast decoding, recursive decoder, unrolled decoder

1 Introduction

1.1 Motivations and Background

Higher throughput has always been a primary target during the evolution of mobile communications. Driven by high data rate applications such as virtual/augmented reality (VR/AR) applications, the sixth generation of wireless technology (6G) requires a peak throughput of 1 Tbps [3]. This is roughly a 50×–100× increase over the 10–20 Gbps throughput targeted for 5G standards.

To support such a high data rate, we need to propose a new physical layer design to further reduce implementation complexity, save energy, and improve spectral efficiency. This is particularly true when the peak throughput requirement is imposed on a resource constrained (limited processing power, storage, and energy supply, etc.) device. As channel coding is widely known to consume a substantial proportion of computational resources, it poses a bottleneck for extreme throughput. To this end, channel coding is one of the most relevant physical layer technologies used to guarantee 1 Tbps peak throughput for 6G.

Polar codes, defined by Arikan in [4], are a class of linear block codes with a generator matrix G_N of size N, defined by $G_N \triangleq F^{\otimes n}$, in which $N = 2^n$ and $F^{\otimes n}$ denotes the n-th Kronecker power of $F=\begin{bmatrix}1&0\\1&1\end{bmatrix}$. Successive cancellation (SC) is a basic decoding algorithm for polar codes.

Although the SC decoding algorithm seems unsuitable for high-throughput applications due to its serial nature, state-of-the-art SC decoders [1, 5–8] managed to significantly simplify and parallelize the decoding process such that the area efficiency of SC decoding has far exceeded that of belief propagation (BP) decoding for low-density parity-check codes (LDPC). In particular, these works represent SC decoding as a binary tree traversal [5], as shown in Figure 1a, with each subtree therein representing a shorter polar code. The original SC decoding algorithm traverses the tree by visiting all the nodes and edges, leading to high decoding latency. Simplified SC decoders can fast decode certain subtrees (shorter polar codes) and thus "prune" those subtrees. The resulting decoding latency is largely determined by the number of remaining edges and nodes in the pruned binary tree. Several tree-pruning techniques have been proposed in [5, 9–10]. To achieve 1 Tbps throughput, more aggressive techniques need to be proposed on both the decoding and encoding sides.

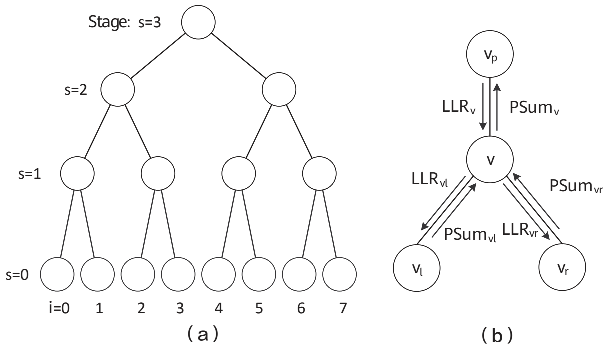


Figure 1 (a) Decoding architecture as a binary tree; (b) Node *v* received/response information

1.2 Contributions

This paper introduces a novel polar code construction method, referred to as "fast polar codes", to facilitate parallelized processing at an SC decoder. In contrast to some existing decoding-only techniques, we take a joint encoding-decoding optimization approach. Similar to existing methods, our main ideas can be better understood from the binary tree traversal perspective. They include (a) pruning more subtrees, (b) replacing some non-prunable subtrees with other fast-decodable short codes of the same code rates, and then pruning these "grafted" subtrees, and (c) eliminating the remaining non-prunable subtrees by altering their code rates. As can be seen, both (b) and (c) involve a modified code construction. Consequently, we are able to fast decode any subtree (short code) of a certain size, without sacrificing parallelism.

The algorithmic contributions are summarized below:

- We introduce four new fast decoding modules for nodes with code rates $\{\frac{2}{M}, \frac{3}{M}, \frac{M\text{-}3}{M}, \frac{M\text{-}2}{M}\}$. Here $M = 2^s$ is the number of leaf nodes in a subtree, where s is the stage number. These nodes are called dual-REP (REP-2), repeated parity check (RPC), parity checked repetition (PCR), and dual-SPC (SPC-2) nodes, respectively. More importantly, these modules reuse existing decoding circuits for repetition (REP) and single parity check (SPC) nodes.
- For medium-code-rate nodes that do not natively support fast decoding, we graft two extended BCH codes to replace the original outer polar codes. As BCH codes enjoy good minimum distance and natively support efficient hard-input decoding algorithms, they strike a good balance between performance and latency.

The extension method is also customized to enhance performance.

- We propose the re-allocation of code rates globally, such that all nodes up to a certain size support the aforementioned fast decoding algorithms. This approach completely avoids traversal into certain "slow" nodes.

For code length N = 1024 and code rate R = 0.875, the proposed fast polar codes enable parallel decoding of all length-16 nodes. The proposed decoding algorithm reduces node visits by 55% and edge visits by 43.5% when compared with the original polar codes, with a performance cost of under 0.3 dB. Two types of decoder hardware are designed to evaluate the area efficiency and energy efficiency.

The implementation-wise contributions are summarized below:

- We design a recursive decoder to flexibly support any code rates and code lengths $N \leq 1024$. We estimate that this decoder layout area is only 0.045 mm^2. For code length N = 1024 and code rate R = 0.875, it achieves a 25.6 Gbps code bit throughput, with an area efficiency of 561Gbps/mm^2.

- We also design an unrolled decoder that only supports one code rate and code length. We estimate that this decoder layout area is 0.3 mm^2. For code length N = 1024 and code rate R = 0.875, it provides a 1229 Gbps code bit throughput, with an area efficiency of 4096 Gbps/mm^2.

2 From Simplified SC Decoding to Fast Polar Codes

Following the notations in [5], node v in a tree is directly connected to a parent node p_v, left child node v_l and right child node v_r, respectively[1]. The stage of node v is defined by the number of edges between it and its nearest leaf node. All leaf nodes are at stage s = 0. The set of nodes of the subtree rooted at node v is denoted by V_v. As such, V_{root} denotes the full binary decoding tree. The set of all leaf nodes is denoted by U, the index of a leaf u [5] is denoted by $i(u)$, and the indices of U is denoted by $i(U)$. Meanwhile, the set of the leaf nodes in subtree V_v is denoted by U_v, and the indices of U_v is denoted by $i(U_v)$. The set of all information bit positions is denoted by I and that of all frozen bits by I^C. The set of the information bit positions in subtree V_v is denoted by I_v and the remaining frozen bit positions therein by I_v^C.

2.1 Simplified SC Decoding

If I_v^C matches patterns, pattern-based simplified decoding can be triggered to process the node in parallel rather than bit-by-bit. From the binary tree traversal perspective, all child nodes of v do not need to be traversed. As a result, decoding latency is reduced.

The existing pattern-based simplified decoding includes 4 different types. Node v is a Rate-1 node [5] if all leaves in the subtree V_v are information bits, and a Rate-0 node [5] if all leaves in the subtree V_v are frozen bits. To improve the decoder's efficiency, [9] defines single parity check (SPC) and repetition (REP) nodes. We can employ pattern-specific parallel processing for each type of node. However, we need to identify and exploit additional special nodes or patterns for improved latency reduction.

In this paper, we present four new types of corresponding nodes:

- Define node v as a dual-SPC (SPC-2) node if V_v includes only two frozen bits, and the frozen bits indices are the two smallest in $i(U_v)$.

- Define node v as a dual-REP (REP-2) node if V_v includes only two information bits, and the information bits indices are the two largest in the $i(U_v)$.

- Define node v as a repeated parity check (RPC) node if V_v includes only three frozen bits, and the frozen bits indices are the three smallest in the $i(U_v)$.

- Define node v as parity checked repetition (PCR) node if V_v includes only three information bits, and the information bits indices are the three largest in the $i(U_v)$.

The corresponding fast decoding methods are described in Section 3.

[1] A leaf node v_{leaf} has no child node, and a root node v_{root} has no parent node.

Pattern-based simplified decoding skips the traversal of certain subtrees when it matches the above patterns.

Currently, there are eight pattern types to cover eight code rates of a subtree: {0, $\frac{1}{M}$, $\frac{2}{M}$, $\frac{3}{M}$, $\frac{M\text{-}3}{M}$, $\frac{M\text{-}2}{M}$, $\frac{M\text{-}1}{M}$, 1}. In other words, nodes with other code rates cannot be fast decoded, and we need to work on the following two parameters.

- Ratio of simplified nodes: currently eight out of the M+1 code rates support simplified decoding, and the ratio is $\frac{8}{M+1}$. Note that only the lowest and highest codes rates can be simplified, meaning code rates between $\frac{3}{M}$ and $\frac{M\text{-}3}{M}$ do not benefit from the fast decoding algorithm. For short and medium length codes, many nodes fall into this range due to insufficient polarization. We hope to further reduce latency by introducing more fast-decodable patterns to cover additional code rates.

- Degree of parallelism: this can be represented by M, since the M bits in a simplified node are decoded in parallel. A larger M means a larger proportion of the binary tree can be pruned due to simplified decoding. We hope to increase M for higher throughput as well.

For M = 8, the ratio of simplified nodes is 8/9, with only one unsupported code rate ($\frac{4}{8}$), but the degree of parallelism is only 8. For M = 16, the ratio of simplified nodes reduces to 8/17, leaving a wide gap of nine unsupported code rates ($\frac{4}{16}$,..., $\frac{12}{16}$), but the degree of parallelism doubles.

2.2 BCH Code

To cover medium code rates, we need to find patterns that can be fast decoded with good BLER performance. Unfortunately, to the best of our knowledge, there exists no parallel decoding method for polar codes with code rates between $\frac{3}{M}$ and $\frac{M\text{-}3}{M}$. The good news, however, is that the outer codes represented by a subtree can be replaced by any codes, as shown in many previous works [11–13]. A good solution involves removing the polar nodes with code rate falling into the gap, and grafting a different code that allows for fast decoding.

BCH codes are ideal candidates due to their good minimum distance property and fast hard-input decoding algorithms. If the error correcting capability is t, it is easy to design BCH codes with a minimum Hamming distance larger than 2 × t. This leads to good BLER performance. Meanwhile, the Berlekamp-Massey (BM) algorithm can decode a BCH code with t = 1 or t = 2 within a few clock cycles. When grafted to polar codes as fast-decodable nodes, hard decisions are applied to the LLRs from the inner polar codes (parent nodes) before being sent to the outer BCH codes (child nodes). Here, the BCH codes are called "BCH nodes".

But BCH codes cannot immediately solve our problem. They only support a few code rates and code lengths, meaning they cannot cover all codes rates within the gap. For the degree of parallelism M = 16, the target code length is 2^4, so the nearest code length of BCH is 15. Meanwhile, BCH codes only support code rates $\frac{7}{15}$ and $\frac{11}{15}$ within the gap, and the corresponding number of information bits are k = 7, k = 11.

To overcome these issues, we must first extend the code length to 16 bits. For BCH codes with k = 7 and t = 2, the original codes can correct two error bits, and we add an additional bit to be the parity check of all BCH code bits. The proposed two-step hard decoding works as follows. When the hard decision incurs three bit errors, and one of the errors has the minimum amplitude, the SPC bit can help correct one error bit first. The remaining two error bits can then be corrected by the BM algorithm. However, the same SPC extension does not work for BCH codes with k = 11 and t = 1. This is because if there are two or more bit errors in the node, the SPC function and the BM algorithm both fail. Alternatively, if there is one error, the SPC decoding failure will lead to further errors during BM decoding. Instead of SPC extension, we repeat one BCH code bit to improve its reliability.

Now that we have grafted two types of BCH nodes, pattern-based decoding can support 10 code rates. The ratio of simplified nodes increases to 10/17, and the maximum gap reduces to $\frac{4}{16}$. Figure 3 shows the code rates supported by pattern-based decoding for a degree of parallelism M = 16.

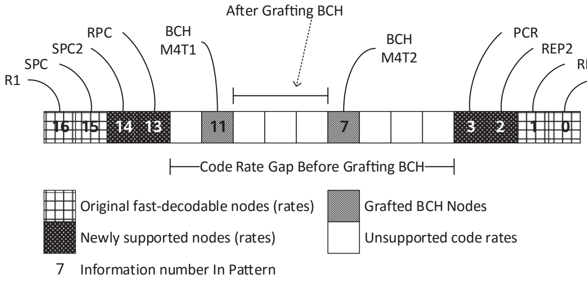

Maximum Code Rate Gap After Grafting BCH
R1
SPC
SPC2
RPC
BCH M4T1
BCH M4T2
PCR
REP2
REP
R0
16
15
14
13
11
7
3
2
1
0
Code Rate Gap Before Grafting BCH
Original fast-decodable nodes (rates)
Grafted BCH Nodes
Newly supported nodes (rates)
Unsupported code rates
7
Information number In Pattern

Figure 2 Nodes (code rates) supporting fast decoding for degree of parallelism M = 16

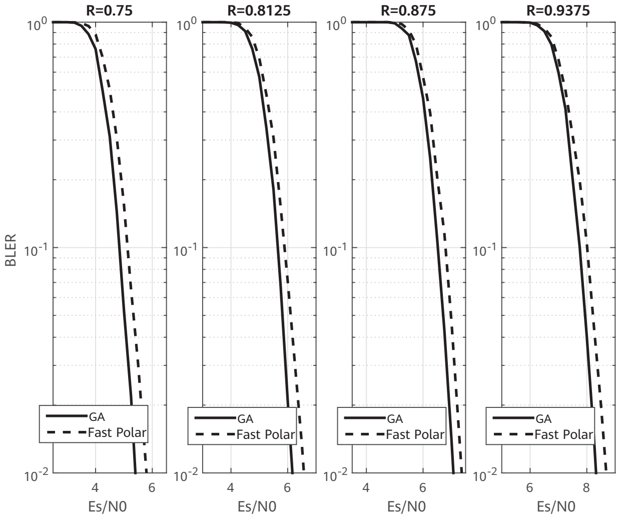


Figure 3 BLER Performance comparison between GA and fast polar code construction

2.3 Fast Polar Codes via Rate Re-Allocation

Even with the inclusion of BCH nodes, the fast decoding algorithm could not cover all the code rates of length-16 subtrees. As the second part of the solution, we propose the construction of fast polar codes to avoid the "slow" nodes, and the use of the existing ten patterns only. Here "fast" resembles the speed of fast SC decoding, but is achieved by altering code construction instead of decoding. Our demonstration shows that it greatly reduces decoding latency and increases throughput with only a slight performance loss.

The following steps show how to construct fast polar codes using only the node patterns of discontinuous code rates:

1. Employ traditional methods such as Gaussian approximation (GA) or polarization weight (PW) to build polar codes with the parameters of code length *N* and code rate *R*.

2. Split all *N* synthesized sub-channels to *N*/16 segments. Each segment constitutes a 16-bit long block code, equivalent to a subtree with 16 leaf nodes.

3. Identify all "slow" segments which do not match the supported code rates or patterns. Re-allocate the code rates among segments to match the nearest supported code rate or pattern, which has *K* information bits.

4. If the number of information bits of the current segment exceeds or falls short of *K*, we remove or add a few information bits according to reliability. Apply this process to the remaining "slow" segments until all segments become fast-decodable.

Algorithm 1 Constructing fast polar codes

```
Input:
  Code length $N$, information length $K$, a set of fastdecodable modes $\Theta$.
Output:
  Re-allocate node-wise code rates so that all nodes support fast decoding.
  Construct an $(N,K)$ polar code based on GA or PW methods.
  Divide the code into segments of length 16, and the number of segments is denoted by $N_s$ .
  Progressively refine the code construction as follows.
  All frozen bit positions are initialized as active states, and "active" bit position can be transformed to an information bit position in the refining process.
  for $t = 1 \cdots N_s$ do
    while the $t$-th segment does not belong to $\Theta$ do
      Denote by $i$ the least reliable information bit position in the $t$-th segment.
      Denote by $j$ the most reliable frozen bit position of active states in the subsequent segments, and denote by $k_j$ the number of information bits in that segment.
      if $k_j \geq 11$ and $k_j < 16$ or $k_j < 3$ then
        Mark $i$ as a frozen bit position and $j$ as an information bit position.
      else
        Mark $j$ as inactive state.
      end if
    end while
  end for
```

The resulting code is coined as "fast polar code". A detailed description of the construction algorithm for fast polar codes can be found in Algorithm 1.

Take code length N = 1024 and code rate R = 0.875 as examples. We count the number of fast-decodable nodes to be visited, $f_{+/-}$-functions [14] to be executed, and edges to be traversed. These numbers provide a good estimate for SC decoding latency [5, 9], and are thus used to compare the construction proposed in this section with the GA construction in Table 1. As we can see, the traversed nodes and edges reduce by 55% and 43.5% , respectively, while the $f_{+/-}$-function executions reduce only by 8.9%. Note that the former two parameters have a greater influence than $f_{+/-}$-functions, because they cannot be parallelized in any form.

Table 1 A summary of current spatial non-stationary channel models and their pros and cons

<table>
<tr><th colspan="8">Distribution of fast-decodable nodes</th></tr>
<tr><th colspan="4">GA Construction</th><th colspan="4">Fast Polar Code Construction</th></tr>
<tr><td>Rate-1</td><td>4</td><td>SPC</td><td>20</td><td>Rate-1</td><td>2</td><td>SPC</td><td>9</td></tr>
<tr><td>SPC-2</td><td>2</td><td>RPC</td><td>0</td><td>SPC-2</td><td>1</td><td>RPC</td><td>0</td></tr>
<tr><td>PCR</td><td>1</td><td>REP-2</td><td>1</td><td>PCR</td><td>3</td><td>REP-2</td><td>1</td></tr>
<tr><td>REP</td><td>11</td><td>Rate-0</td><td>1</td><td>REP</td><td>1</td><td>Rate-0</td><td>1</td></tr>
<tr><td>BCH t = 1</td><td>0</td><td>BCH t = 2</td><td>0</td><td>BCH t = 1</td><td>3</td><td>BCH t = 2</td><td>2</td></tr>
<tr><th colspan="8">Count with respect to binary tree traversal</th></tr>
<tr><th></th><th colspan="2">GA</th><th colspan="2">Fast</th><th colspan="3">Reduction (%)</th></tr>
<tr><td>Nodes</td><td colspan="2">40</td><td colspan="2">22</td><td colspan="3">55%</td></tr>
<tr><td>$f_{+/-}$</td><td colspan="2">4160</td><td colspan="2">3792</td><td colspan="3">8.9%</td></tr>
<tr><td>Edges</td><td colspan="2">76</td><td colspan="2">43</td><td colspan="3">43.5%</td></tr>
</table>

It is worth noting that the proposed fast polar code construction algorithm reallocates the code rates of some nodes against their actual capacity which is derived from channel polarization. This inevitably incurs a BLER performance loss. We run simulations to evaluate the loss, and Figure 3 compares the BLER curves of both constructions under code length N = 1024 and code rates R = {0.75, 0.8125, 0.875, 0.9375}. There is a maximum of 0.3 dB loss at BLER 10^{-2} between GA polar codes and the fast polar codes when adopting QPSK modulation.

3 Fast Decoding Algorithms

In this section, we describe the algorithms used to support fast decoding of the newly defined SPC-2, REP-2, RPC, and PCR nodes. For BCH nodes, we employ the classic BM algorithm which takes hard inputs and supports hardware-friendly fast decoding.

Each fast-decodable node v at stage s can be viewed as an outer code of length $M = 2^s$. The code bits of v as an outer code are denoted by X_v, with M bits.

3.1 SPC-2

For a dual-SPC node v, we divide its code bits X_v into two groups, X_v^{even} with even-numbered indices, and X_v^{odd} with odd-numbered indices. According to the definition of an SPC-2 node, there are two parity-check bits in the subtree V_v, and the corresponding parity functions $p[0]$ and $p[1]$ can be written as

$$\begin{cases} p[0] : \oplus x = 0, x \in X_v \\ p[1] : \oplus x = 0, x \in X_v^{odd} \end{cases}$$

We add the two parity functions to get a parity function $p[2]$:

$$p[2] = p[0] \oplus p[1]: \oplus x = 0, x \in X_v^{even}$$

Since the two parity functions $p[1]$ and $p[2]$ involve two disjointed sets of code bits, the decoding of an SPC-2 node can be parallelized to two SPC nodes, each of which inherits half of the elements from X_v. We can reuse two SPC decoding modules to fast decode the SPC-2 node.

3.2 REP-2

For a dual-REP node v, we divide its code bits X_v into two groups, X_v^{even} with even-numbered indices, and X_v^{odd} with odd-numbered indices. There are two information bits in the subtree V_v according to the definition of an REP-2 node, and these are denoted by u_{M-2} and u_{M-1}.

It can be easily verified that X_v^{odd} are the repetition of u_{M-1}, and X_v^{even} are the repetition of $u_{M-2} \oplus u_{M-1}$. Consequently, we can divide a length-M dual-REP node into two M/2 REP nodes, and we can reuse two REP decoding modules in parallel to fast decode the REP-2 node.

3.3 RPC

For an RPC node v, we divide its code bits X_v into four groups as follows:

$$X_v^i = \{x \in X_v, \mathrm{mod}(l(x),4) = i\},\ i \in \{0,1,2,3\} \qquad (1)$$

According to the definition of an RPC node, there are three parity-check bits in the subtree V_v, and the parity functions p[0], p[1] and p[2] can be written as

$$\begin{cases} p[0] : \oplus x = 0,\ x \in X_v^0 \cup X_v^1 \cup X_v^2 \cup X_v^3 \\ p[1] : \oplus x = 0,\ x \in X_v^1 \cup X_v^3 \\ p[2] : \oplus x = 0,\ x \in X_v^2 \cup X_v^3 \end{cases}$$

We add the latter two parity functions to get parity function $p[3]$:

$$p[3] = p[1] \oplus p[2]:\ \oplus x = 0,\ x \in X_v^1 \cup X_v^2$$

And we add this parity function to the first one to get parity function $p[4]$:

$$p[4] = p[0] \oplus p[3]: \oplus x = 0,\ x \in X_v^0 \cup X_v^3$$

We define $\hat{c}_i = \oplus\ x,\ x \in X_v^i,\ i \in [0,1,2,3]$. According to parity functions $p[1]$ to $p[4]$, we can easily verify that the following relationship holds true:

$$\hat{c}_1 \oplus \hat{c}_3 = \hat{c}_2 \oplus \hat{c}_3 = \hat{c}_1 \oplus \hat{c}_2 = \hat{c}_0 \oplus \hat{c}_3 = 0 \qquad (2)$$

Equation (2) implies the existence of a virtual repetition code of rate , because:

$$\hat{c}_0 = \hat{c}_1 = \hat{c}_2 = \hat{c}_3 = 0$$

or

$$\hat{c}_0 = \hat{c}_1 = \hat{c}_2 = \hat{c}_3 = 1$$

where $\hat{c}_0, \hat{c}_1, \hat{c}_2, \hat{c}_3$ are the virtual repeated code bits.

Given the above knowledge, the decoding algorithm for an RPC node at stage s where $s \le 2$, can be easily derived as Algorithm 2, in which

$$\mathrm{sig}(a) \triangleq \begin{cases} 0, & a \ge 0 \\ 1, & a < 0. \end{cases}$$

Algorithm 2 Decoding a RPC node

```
Input:
    The received signal $a_v$ = { $a_{v_k}$ , $k = 0 \cdots M-1$};
Output:
    The codeword to be recovered: $\hat{x} = \{\hat{x}_k, k = 0 \cdots M-1\}$;
    Initialize: $\Delta_0 = 0$; $\Delta_1 = 0$
    Initialize: $\delta_i = \infty$, $c_i = 0$, $p_i = 0$ for $i = 0 \cdots 3$;
    Initialize: $\hat{x}_k = sig(a_{v_k})$ for $k = 0 \cdots M-1$};
    for $i = 0 \cdots 3$ do
        for $j = 0 \cdots M/4$ do
            $k = j \times 4 + i$;
            $c_i = c_i \oplus sig(a_{v_k})$;
          if $|a_{v_k}| < \delta_i$
            $p_i = k$;
            $\delta_i = |a_{v_k}|$;
        end for
        if $c_i = 1$
          $\Delta_0 = \Delta_0 + \delta_i$
        else
          $\Delta_1 = \Delta_1 + \delta_i$
    end for
    for $i = 0 \cdots 3$ do
        if $((\Delta_0 > \Delta_1) \cap (c_i = 0)) \cup ((\Delta_0 < \Delta_1) \cap (c_i = 1))$
          $\hat{x}_{p_i} =\sim \hat{x}_{p_i}$
    end for
```

3.4 PCR

For a PCR node v, we divide its code bits X_v into four groups, just like in (1). There are three information bits in this node according to the definition of a PCR node, and these are denoted by u_{M-3}, u_{M-2}, and u_{M-1}.

We define c_i, $i \in \{0, 1, 2, 3\}$ according to the following equation

$$[\ c_0\ c_1\ c_2\ c_3\ ] = [\ 0\ \ u_{M-3}\ \ u_{M-2}\ \ u_{M-1}\ ] \times G_4 \qquad (3)$$

It can be easily verified that X_v^0 are the repetition of c_0, X_v^1 are the repetition of c_1, X_v^2 are the repetition of c_2, and X_v^3 are the repetition of c_3. Consequently, we divide the input signal a_v into four groups according the indices, and combine the input signals within each group into four enhanced signals Δ_i, $i \in \{0, 1, 2, 3\}$, as in an REP node.

Equation (3) implies the existence of a virtual single parity check code of rate $\frac{3}{4}$, with virtual code bits c_i, $i \in \{0, 1, 2, 3\}$, so that we can reuse the SPC module to decode it. A detailed description of PCR decoding is provided in Algorithm 3.

Algorithm 3 Decoding a PCR node

```
Input:
    The received signal $\alpha_v = \{\alpha_{v_k},\ k = 0 \cdots N-1\}$;
Output:
    The codeword to be recovered: $\hat{x} = \{\hat{x}_k,\ k = 0 \cdots N-1\}$;
    Initialize: $\Delta_i = 0$ for $i = 0 \cdots 3$;
    for $i = 0 \cdots 3$ do
        for $j = 0 \cdots N/4$ do
            $k = j \times 4 + i$
            $\Delta_i = \Delta_i + \alpha_{v_k}$
        end for
    end for
    $\{\hat{c}_0, \hat{c}_1, \hat{c}_2, \hat{c}_3\}$ = SPC_DEC($\{\Delta_0, \Delta_1, \Delta_2, \Delta_3\}$)
    for $i = 0 \cdots 3$ do
        for $j = 0 \cdots N/4$ do
            $k = j \times 4 + i$
            $\hat{x}_k = c_i$
        end for
    end for
```

4 Hardware Implementation

We designed two types of hardware architectures to verify the performance, area efficiency, and energy efficiency.

- **Recursive Decoder:** This architecture supports flexible code lengths and coding rates of mother code length N from 32 to 1024 with the power of 2. With rate matching, flexible code lengths with $0 < N \leq 1024$ and code rates with $0 < R \leq 1$ are supported. The $f_{+/-}$ functions in nodes are processed by single PE (processing element) logic, and one decision module supports all 9 patterns[1]. The decoder processes one packet at a time.
- **Unrolled Decoder:** This architecture only supports a fixed code length and code rate. In our architecture we hard coded code length $N = 1024$ and code rate $R = 0.875$. This fully unrolled pipelined design combines exclusive dedicated PEs to process each $f_{+/-}$ function in the binary tree. Same to the decision modules that 21 dedicated node specific logic is implemented to support 21 nodes patterns. With 25 packets simultaneously decoding, and thanks to the unrolled full utilization of processing logic and storage, this decoder provides extreme high throughput with high area efficiency and low decoding energy.

Both of the decoder implementations mentioned above adopt successive cancellation algorithms accelerated by pattern-based fast decoding. The maximum degrees of parallelization are 128 for SPC and SPC-2 nodes, and 256 for R1 nodes. All other nodes enjoy a degree of parallelism of 16.

4.1 Parallel Comparison Circuit

We can observe that there are several large SPC nodes in the right half of the binary tree. As described, these SPC nodes need to be processed with a higher degree of parallelism to achieve a higher throughput. The SPC decoding algorithm is very simple, and can be explained as follows. First, obtain the signs of an SPC node's input signals, then find the minimum amplitude of input signals and record its position. Next, perform a parity check on the signs. If it passes, return these signs. Otherwise, reverse the sign of the recorded minimum-amplitude position and return the updated signs.

In order to process a large SPC node, a circuit is required to locate a minimum amplitude from a large number of input signals. The traditional pairwise comparison method requires a circuit of depth $log_2(M)$, where M is the number of amplitudes to be compared. Finding the smallest among, for example, 128 amplitudes requires 7 comparison steps. Considering that clock frequency is set to 1 GHz, it is very challenging to meet the timing constraints and complete all comparisons in one clock cycle.

To address these issues, we advocate a parallel comparison architecture to replace the traditional one. For node v at stage s, its input signals α_v include $M = 2^s$ elements, the amplitudes of which are denoted as $[A_0\ A_1 \cdots A_{M-1}]$. Each amplitude has x-bit quantization, and we fill the x-bit quantized binary vectors into the columns of a matrix as follows:

$$[A_0 \cdots A_i \cdots A_{M\text{-}1}] = \begin{bmatrix} b_0^0 & \cdots & b_i^0 & \cdots & b_{M\text{-}1}^0 \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ b_0^j & \cdots & b_i^j & \cdots & b_{M\text{-}1}^j \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ b_0^{x\text{-}1} & \cdots & b_i^{x\text{-}1} & \cdots & b_{M\text{-}1}^{x\text{-}1} \end{bmatrix}$$

[1] R0 node is bypassed in SC decoding.

If we rewrite the matrix with respect to its row vectors matrix we will have $[B_0 \cdots B_j \cdots B_{x-1}]^T$, in which $B_j = [b_0^j \cdots b_i^j \cdots b_{M-1}^j]$, $j \in \{0, 1 \cdots x-1\}$ is a row vector. We propose Algorithm 4 to determine the minimum-amplitude position through reverse mask D, in which the bit "1" indicates the minimum.The parallel comparison algorithm reduces the comparison logic depth from $log_2(M)$ to 1. However, the reverse mask D may have two or more minimum positions, which means that the input signals a_v include two or more minimum amplitudes, and it must generate an error in such cases. To prevent this error, we can apply an additional circuit to ensure the uniqueness of the selected minimum position.

Algorithm 4 Parallel comparison algorithm

```
Input:
    The received signal $a_v = \{a_{v_k}, k = 0 \cdots M-1\}$;
Output:
    The Reverse Mask: $D$ is an $M$-bit Variable;
    Initialize: $[B_0 \cdots B_j \cdots B_{x-1}]^T$ from $\alpha_v$;
    Initialize: An $N$-bits variable $C = 0$, .
    for $j = x - 1 \cdots 0$ do
        $M$-bit Variable $E = (C|B_j)$
        if(Not all bits in $E$ are "1")
            $C = E$
    end for
    Reverse Mask $D = \sim C$
```

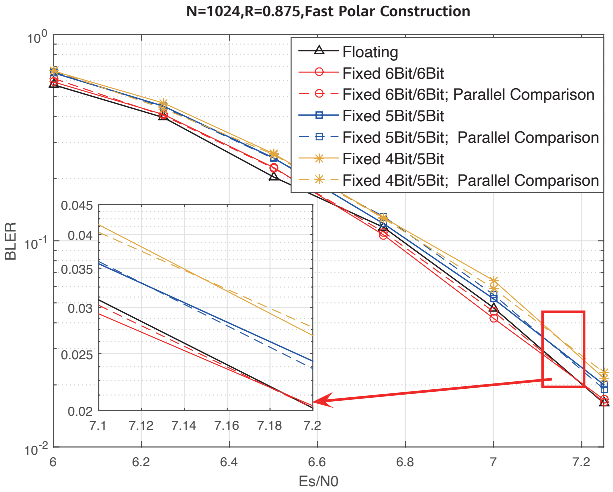


Figure 4 Performance comparison between Floating Point and Fixed Point

4.2 Bit Quantization

An attractive property of polar codes is that SC decoding works well under low-precision quantization (4 bits to 6 bits). Lower precision quantization is the key to higher throughput, as it effectively reduces the implementation area and increases clock frequency.

There are two types of quantization numbers — one for channel LLR and another for internal LLR. We first test a case with 6-bit input quantization and 6-bit internal quantization. According to Figure 4, this setting achieves the same performance as floating-point. The second case is 5-bit quantization and 5-bit internal quantization, which incurs < 0.1 dB loss. Finally, 4-bit input quantization and 5-bit internal quantization incurs < 0.2 dB loss. In this paper, we evaluate the physical implementation result under 5-bit quantization for both input and internal signals to strike a good balance between complexity and throughput.

At the same time, we also compare the BLER performance between the original SPC and parallelized SPC. None of the quantization schemes result in a harmful loss.

4.3 Estimation of Layout Area

We carry out the two FPGA implementations for both the recursive and unrolled architectures and convert the results to the physical implementations.

According the FPGA synthesis results, the recursive decoder has 10170 LUTs and 12772 FFs; meanwhile, the unrolled decoder has 66192 LUTs and 55187 FFs. Both decoders avoid the use of memories, making it is easy to convert the FPGA results to ASIC. Converted to 16 nm technology, the recursive decoder estimated synthesis area and the layout size are 0.032 mm^2 and 0.045 mm^2, respectively, at a clock frequency of 1 GHz. The unrolled decoder's estimated synthesis area and the layout size are 0.17 mm^2 and 0.30 mm^2, respectively, at a clock frequency of 1.20 GHz.

5 Key Performance Indicators

The key performance indicators (KPIs) are reported in this section. First of all, we evaluate the area efficiency using equation:

$$AreaEff\ (Gbps/mm^2) = \frac{InfoSize\,(bits)}{Latency\,(ns)\times Area\,(mm^2)}$$

The recursive decoder takes 40 clock cycles to decode one packet under fast polar code construction with code length $N = 1024$ and code rate $R = 0.875$. As such, the throughput is (1024 bits × 1 GHz)/40 cycles = 25.6 Gbps for coded bits, and ((1024 × 0.875) bits × 1 GHz)/40 cycles = 22.4 Gbps for information bits. Converting to 16 nm process, the area efficiency for coded bits is 561 Gbps/mm^2.

The unrolled decoder takes 25 clock cycles to decode one packet. It is fully pipelined, meaning a new packet of decoded results is generated continuously every cycle after the first 25 clock cycles of the first packet processing time. The throughput is thus 1024 bits × 1.2 GHz = 1229 Gbps for coded bits, and (1024 × 0.875) bits × 1 GHz = 1075 Gbps for information bits. Converting to 16 nm process, the area efficiency for coded bits is 4096 Gbps/mm^2.

Table 2 Comparison with the high throughput polar decoder

Implementation	This Work (Unrolled)	This Work (Recursive)	[1]	[2]	[8]
Construction	Fast-Polar	Fast-Polar	Polar	Product-Polar [15]	Polar
Decoding Algorithm	Fast-SC	Fast-SC	SC	PDF-SC [16]	OPSC
Code Length	1024	1024	32768	16384	1024
Code Rate	0.875	0.875	0.864	0.864	0.83
Technology	FPGA Converted to 16 nm		In TSMC 16 nm		
Clock Frequency (GHz)	1.20	1.00	1.00	1.05	1.20
Throughput/Coded-bit (Gbps)	1229	25.6	5.27	139.7	1229
Throughput/Info-bit (Gbps)	1075	22.4	4.56	120.73	1020
Area/Layout (mm^2)	0.30	0.045	0.35	1.00	0.79
Area Eff/Coded-bit (Gbps/mm^2)	4096	561	15.1	139.7	1555
Power (mW)	784	30.9	-	94	1167
Energy (pJ/bit)	0.63	1.21	-	0.67	0.95

Table 3 Comparison with high throughput polar decoder

Implementation	This Work (Unrolled)	This Work (Recursive)	Polar-5G	LDPC-1K-5G	LDPC-1K-Unrolled	[8]
Construction	Fast-Polar	Fast-Polar	Polar(5G)	LDPC (5G)	LDPC (Unrolled)	Polar
Decoding Algorithm	Fast-SC	Fast-SC	SC	LOMS-3	LOMS-3	OPSC
Code Length	1024	1024	1024	1024	1024	1024
Code Rate	0.875	0.875	0.875	0.875	0.875	0.83
Technology	FPGA Converted to 16 nm		In TSMC 16 nm			
Clock Frequency (GHz)	1.20	1.00	1.00	1.00	1.00	1.20
Throughput/Coded-bit (Gbps)	1229	25.6	10.8	12.44	1024	1229
Throughput/Info-bit (Gbps)	1075	22.4	9.45	10.89	896	1020
Area/Layout (mm^2)	0.30	0.045	0.069	0.154	1.02	0.79
Area Eff/Coded-bit (Gbps/mm^2)	4096	561	157	81	878	1555

We further evaluate the power consumption and decoding energy per bit through a simulation in which 200 packets are decoded. The process, voltage, and temperature (PVT) condition of evaluation is TT corner, 0.8 V, and 20°C, the result of the recursive decoder's power consumption is 30.9 mW, and decoding each bit costs 1.21 pJ of energy on average; while the unrolled decoder's power consumption is 784 mW, and decoding each bit costs 0.63 pJ of energy on average.

We also compare the decoding throughput, area efficiency, and power consumption with several other high-throughput decoders, and present the results in Table 3. From the KPIs, we conclude that unrolled decoders are more suitable for scenarios requiring extremely high throughput but only support fixed code length and rate; recursive decoders are much smaller, which are better for resource constrained devices, and at the same time provide flexible code rates and lengths — a desirable property for wireless communications.

6 Conclusions

In this paper, we propose a new construction method for fast polar codes, which is solely composed of fast-decodable special nodes at length 16. By viewing the decoding process as a binary tree traversal, the fast polar codes can reduce 55% of node visits, 8.9% of $f_{+/-}$ calculation, and 43.5% of edge traversal over the original polar construction at code length N = 1024 and code rate R = 0.875, at the cost of slight BLER performance loss.

We implement two types of decoders for the fast polar codes. The recursive decoder can support flexible code lengths and code rates, with support for code lengths of up to 1024. This decoder layout area is only 0.045 mm^2, and it can provide 25.6 Gbps coded bits throughput with an area efficiency of 561 Gbps/mm^2.

The unrolled decoder only supports one code length N = 1024 and one code rate R = 0.875. However, the fully pipelined structure leads to hardware offering ultra-high area efficiency and low decoding power consumption. The estimated layout area of this decoder is 0.3 mm^2, and it can provide 1229 Gbps code bit throughput with an area efficiency as high as 4096 Gbps/mm^2.

These results indicate that fast polar codes can meet the high-throughput demand of next-generation wireless communication systems, and that recursive and unrolled hardware designs can be adopted to satisfy different system requirements.

References

[1] X. Liu, Q. Zhang, P. Qiu, J. Tong, H. Zhang, C. Zhao, J. Wang, "A 5.16Gbps decoder ASIC for polar code in 16nm FinFET," 2018 *15th International Symposium on Wireless Communication Systems (ISWCS)*, Lisbon, 2018, pp. 1-5.

[2] J. Tong, X. Wang, Q. Zhang, H. Zhang, S. Dai, R. Li, and J. Wang, "Toward terabits-per-second communications: a high-throughput implementation of GN-Coset codes," *IEEE Wireless Communications and Networking Conference (WCNC)*, 2021, pp. 1-6.

[3] W. Saad, M. Bennis, and M. Chen, "A vision of 6G wireless systems: applications, trends, technologies, and open research problems," *IEEE Network*, 2019.

[4] E. Arikan, "Channel polarization: a method for constructing capacity achieving codes for symmetric binary-input memoryless channels," *IEEE Transactions on Information Theory*, vol. 55, no. 7, pp. 3051-3073, Jul. 2009.

[5] A. Alamdar-Yazdi and F. Kschischang, "A simplified successive cancellation decoder for polar codes," in *IEEE Communications Letters*, vol. 15, no. 12, pp. 1378-1380, December 2011.

[6] O. Dizdar and E. Arikan, "A high-throughput energy-efficient implementation of successive cancellation decoder for polar codes using combinational logic," in *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 63, no. 3, pp. 436-447, March 2016.

[7] A. Sral, E. G. Sezer, Y. Ertugrul, O. Arikan, and E. Arikan, "Terabitsper-second throughput for polar codes," *2019 IEEE 30th International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC Workshops)*, 2019, pp. 1-7.

[8] A. Sral, E. G. Sezer, E. Kolagasıoglu, V. Derudder, and K.

Bertrand, "Tb/s polar successive cancellation decoder 16nm asic implementation," Available on http://www.polaran.com/documents/EPIC Polar Code Paper.pdf.

[9] S. A. Hashemi, C. Condo, and W. J. Gross, "Fast and flexible successive cancellation list decoders for polar codes," *IEEE Transactions on Signal Processing*, vol. 65, no. 21, pp. 5756-5769, Nov 2017.

[10] G. Sarkis, P. Giard, A. Vardy, C. Thibeault, and W. J. Gross, "Fast polar decoders: algorithm and implementation," in *IEEE Journal on Selected Areas in Communications*, vol. 32, no. 5, pp. 946-957, May 2014.

[11] Y. Wang and K. Narayanan, "Concatenations of polar codes with outer BCH codes and convolutional codes," *Communication Control and Computing (Allerton) 2014 52nd Annual Allerton Conference on*, pp. 813-819, 2014.

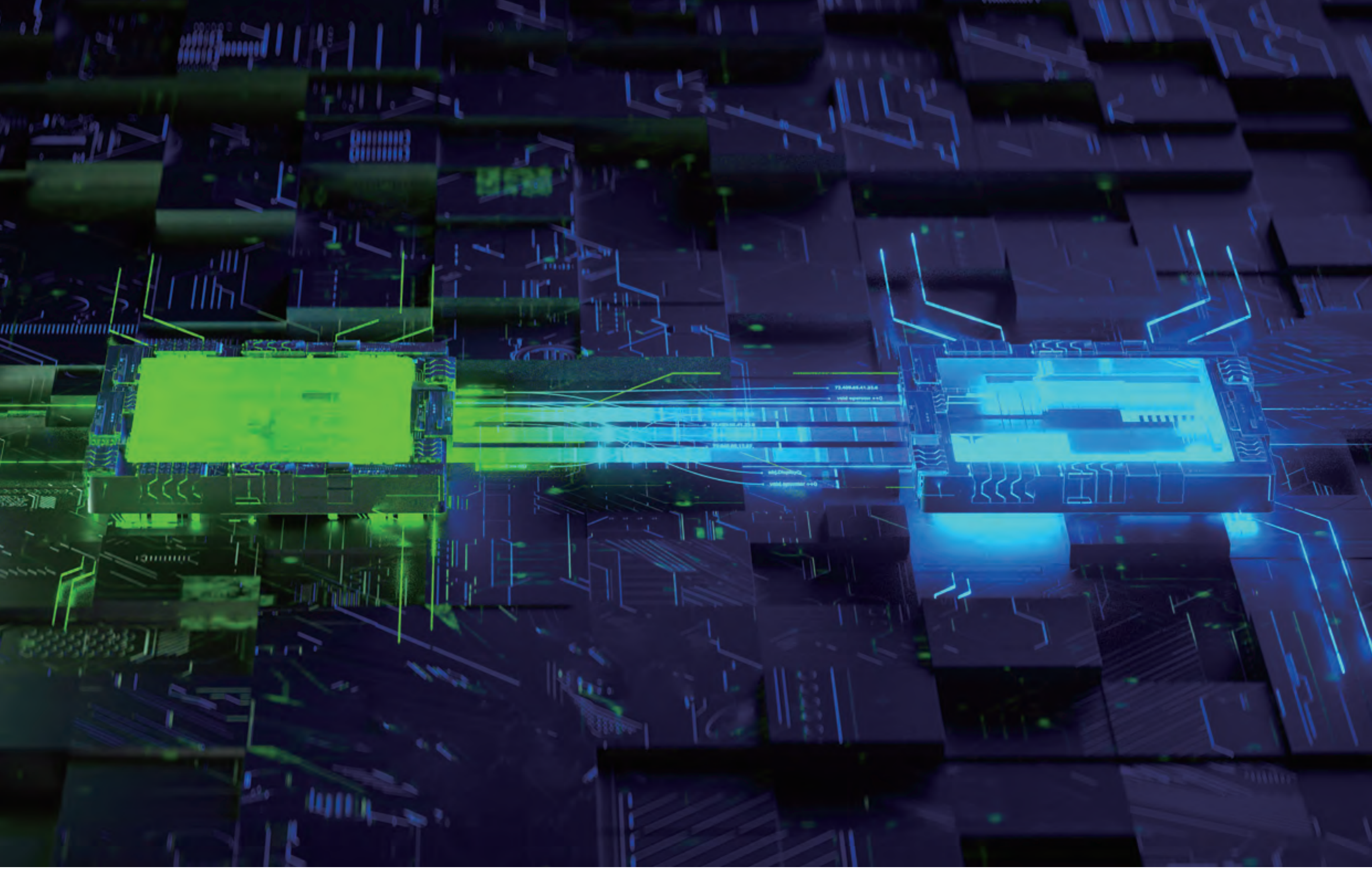
[12] H. Saber and I. Marsland, "Design of generalized concatenated codes based on polar codes with very short outer codes," *Vehicular Technology IEEE Transactions on*, vol. 66, no. 4, pp. 3103-3115, 2017.

[13] D. Goldin and D. Burshtein, "Performance bounds of concatenated polar coding schemes," *Information Theory IEEE Transactions on*, vol. 65, no. 11, pp. 7131-7148, 2019.

[14] A. Balatsoukas-Stimming, M. B. Parizi, and A. Burg, "LLR-based successive cancellation list decoding of polar codes," in *IEEE Transactions on Signal Processing*, vol. 63, no. 19, pp. 5165-5179, Oct. 2015.

[15] X. Wang, H. Zhang, R. Li, J. Tong, Y. Ge, and J. Wang, "On the construction of GN-coset codes for parallel decoding," *IEEE Wireless Communications and Networking Conference (WCNC)*, Seoul, Korea (South), 2020, pp. 1-6.

[16] X. Wang, J. Tong, H. Zhang, S. Dai, R. Li, and J. Wang, "Toward terabits-per-second communications: low-complexity parallel decoding of GN-coset codes," *IEEE Wireless Communications and Networking Conference (WCNC)*, 2021, pp. 1-5.



Joint Message Passing and Autoencoder for Deep Learning

Yiqun Ge [1], Wuxian Shi [1] , Jian Wang [2], Rong Li [2], Wen Tong [1]

[1] Ottawa Wireless Advanced System Competency Centre

[2] Wireless Technology Lab

Abstract

Over the past few years, a lot of effort has been devoted to studying an autoencoder (AE)-based end-to-end or global transceiver. However, the most critical problem with the AE-based transceiver is its poor generalization in the face of variations of random channels. Neurons are fixed once trained, whereas channels continuously change with time. We suggest using a message passing algorithm (MPA) to enable a flexible AE-based transceiver that can be easily generalized for different use cases. This MPA enabling feature would allow not only a coarse learning during the training cycle but also an adaptive inference (reasoning) during the inference cycle.

Keywords

DNN, E2E AE, OOD, transceiver, MPA, SGD, backward propagation

1 Introduction

1.1 AE-based Global Transceiver

There are two different paradigms for applying the deep neural network (DNN) to a wireless transceiver design: block-by-block optimization (shown in Figure 1) and end-to-end (E2E) or global optimization (shown in Figure 2).

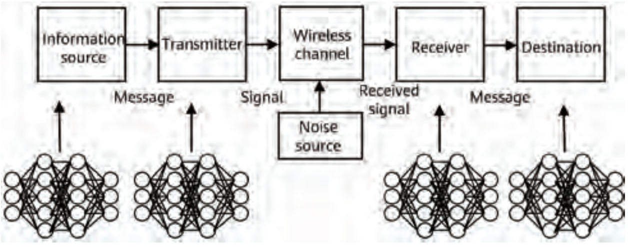


Figure 1 AE-based block-by-block transceiver

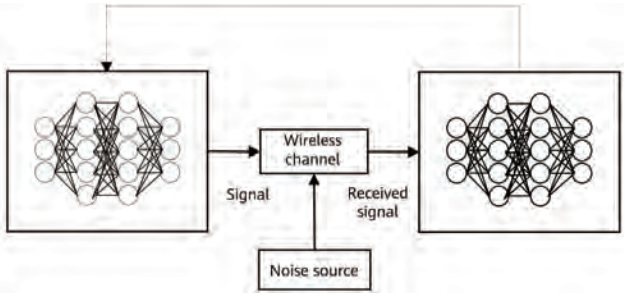


Figure 2 AE-based global transceiver

Since 2017, a lot of effort has been devoted to studying the AE-based global transceiver, which is a natural and straightforward step in applying native AI for wireless communications after observing potential benefits from AE-based block-by-block optimization. Many valuable papers have been published to discuss its feasibility and benchmark its performance in terms of inference against other DNN structures and training methods [1].

A classical physical-layer transceiver assumes reliable information reconstruction at the receiver as a universal transmission goal. To tackle varying hostile channels, it estimates the current channel by reference signals and then equalizes the estimated channel distortions.

In comparison, the AE-based global transceiver simultaneously optimizes both the deep neural layers in the transmitter and those in the receiver for a given source and a given goal in a given channel environment. Specifically, the AE-based global transceiver extracts the most essential information associated to a given goal including reconstruction of information. If the goal was simpler than reconstruction, a smaller amount of essential information would need to be extracted, saving some transmission resources. If the AE-based global transceiver can learn nearly all the possible radio channel realization scenarios into its neurons in a given channel environment, it can save more transmission resources by reducing reference signals and controlling message overhead while achieving the same BLER performance. Both are paramount for a wireless system to achieve higher spectral efficiency. Figure 3 illustrates that the AE-based transceiver bridges three factors: source distribution, goal orientation, and channel environment.

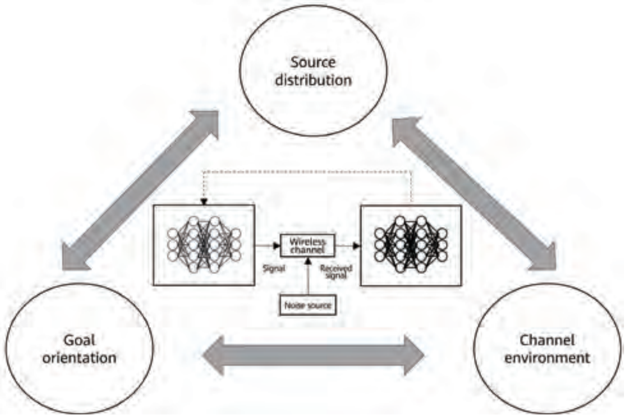


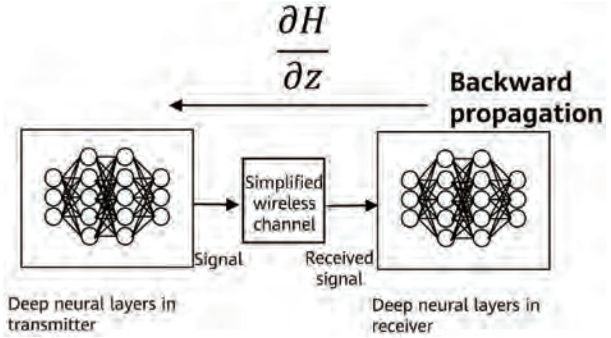
Figure 3 Source-, goal-, and channel-orientation for higher spectrum efficiency

Moreover, the AE-based global transceiver would facilitate an optimal design flow of the transceiver in machine learning (ML)-based data-driven learning methods.

1.2 Current Issues with the AE-based Global Transceiver

Among the technical hurdles that need to be overcome for a simple and straightforward AE-based global transceiver, we underscore three key issues.

Issue 1: Training an AE by a stochastic gradient descent (SGD)-based backward propagation algorithm demands one or more differentiable channel model layers that connect the deep neural layers in the transmitter and those in the receiver as illustrated in Figure 4. Because a true channel must include many non-linear components (such as digital/analog predistortion and conversion) and non-differentiable stages (such as upsampling and downsampling), the deep neural layers in the transceiver are trained for a comprised channel model rather than a true one, which might cause performance loss during the inference cycle with a true channel.



$\frac{\partial H}{\partial z}$
Backward propagation
Simplified wireless channel
Signal
Received signal
Deep neural layers in transmitter
Deep neural layers in receiver

Figure 4 Simplified and differentiable channel model layer(s)

Issue 2: All hidden or intermediate layers are trained according to the posterior probability of its input signal, as shown in Figure 5 [7]. In the AE-based global transceiver, the first layer of the deep neural layers in the receiver is an intermediate layer whose input signal is subject to the current channel distortion. Its impact inevitably penetrates forward to all the deep neural layers in the receiver. If the channel changes drastically beyond the training expectation, the receiver will become obsolete on the inference.

Issue 3: A lack of interpretability from one neural layer to another hides the information about which neurons and connections between the layers are critical for final learning accuracy. Goodfellow et al. [4] gives an example, where a DNN classifier well trained by non-noise images may misclassify a noised image of a panda as a gibbon, as shown on the right of Figure 6. This example implies that a DNN-based classifier heavily relies on some "shortcuts" or "localities" (some pixels in the image of panda) for its final decision. If the shortcuts are intact, the classification will be correct; if the shortcuts are perturbed, the classification will be incorrect. Furthermore, noise-triggered panda-to-gibbon misclassification happens only occasionally with additive random noise, indicating that DNN bets on the intactness of "shortcuts" through a noise channel [5]. The reality that DNN is vulnerable to additive random noise could be disastrous to the application of DNN in wireless transceiver design.

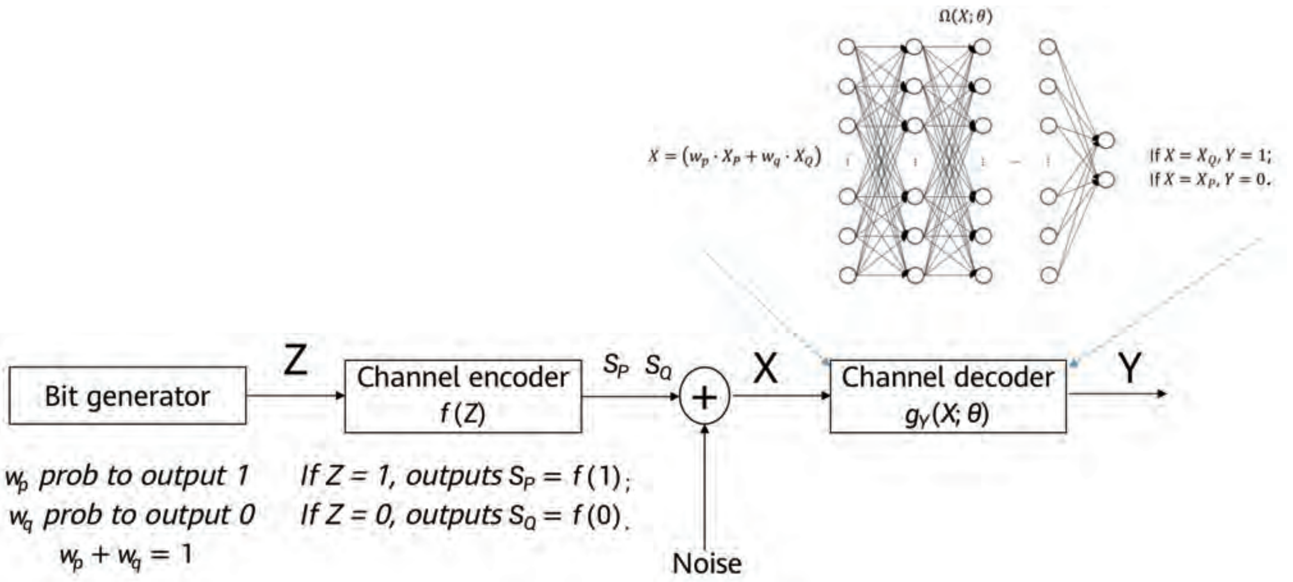


Figure 5 Deep neural layers in the receiver depend on the posterior probability of its input signals [7]

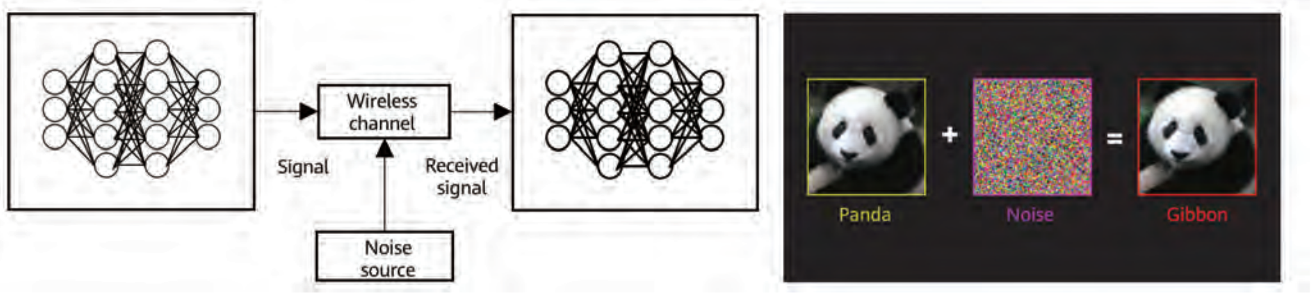


Figure 6 AE-based global transceiver and adversary attack with additive noise [4]

1.3 Out of Distribution (OOD) and Generalization

The existing solutions to these issues are unfortunately hindered by the practical requirements of low-energy, short latency, and decreased overhead in the devices and infrastructures of wireless communications. On the one hand, it is too costly for the AE-based transceiver to accumulate, augment, and re-train itself in a dynamic environment. On the other hand, it is against DNN's "Once-for-All" strategy [6] to meet realistic and energy consumption requirements, i.e., to learn once and to work as long as possible.

All three issues are rooted in the same core problem: DNN's poor generalization against random variation of wireless channels. Because no model (even a superior channel model) is able to exhaustively capture all the possible scenarios of radio propagation, the AE is destined to confront the so-called out-of-distribution (OOD) or outlier problem in reality.

We address this problem in our proposal: adapt the AE-based transceiver to variation of the real-world random wireless channel. An AE-based transceiver framework should come up with sufficient generalization against OOD cases in a data-driven method.

A native-AI contains two different cycles: training and inference (or reasoning). Training needs a massive number of raw data samples and is typically performed with offline computing. Conversely, inference is performed on real-world data sample(s) on the fly.

Inference can be interpreted as either interpolation or extrapolation. Because extrapolation is far less reliable than interpolation, inference accuracy depends on the distributions of true data samples. If a true data sample represented by the blue circles in Figure 7 is within an area to be interpolated by the training data samples represented by the orange circles, the trained AE-based transceiver is likely to handle it properly, as shown on the left of Figure 7. Otherwise, the transceiver would be unlikely to conduct a reliable inference over an outlier data sample, as illustrated by the "unmodeled outlier" dot (in blue) on the right of Figure 7.

In a wireless context, these outliers are often caused by random variation of channels, especially when the channel involved in the inference cycle is shifting away from the channel model used by the training cycle. Along with the inference time, increasing occurrences of outliers would shape the distribution of the received signal, to which Bengio [8] attributes poor generalization of deep learning. Although current remedies may include some extra training, such as transfer training, attention-based recurrent network, or reinforcement learning, none of them are practicable or realistic for the low-energy, low latency, and reduced controlling overhead requirements of future wireless communications.

To make the AE-based global transceiver viable, our primary task is to find an effective and real-time method against outliers or OOD during the inference cycle.

In this paper, we will enable the AE-based global transceiver with a message-passing-algorithm (MPA) based precoder layer to improve generalization performance in dynamic channel variations. In section 2, adhering to the core idea of dimensional transformations, we propose the insertion of a dimension reduction layer into the AE framework. In section 3, we describe how to train the MPA-enabled AE-based transceiver. In section 4, we look at some applications of the MPA-enabled AE-based transceiver. Then finally in section 5, we provide our conclusions.

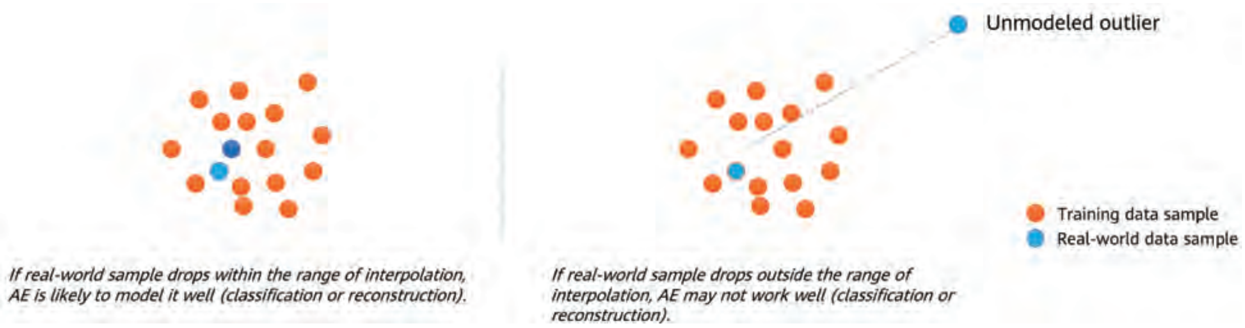

Unmodeled outlier
If real-world sample drops within the range of interpolation, AE is likely to model it well (classification or reconstruction).
If real-world sample drops outside the range of interpolation, AE may not work well (classification or reconstruction).
● Training data sample
● Real-world data sample

Figure 7 Outlier data samples degrade inference accuracy

2 Deep MPA Precoder

In this section, we introduce the dimension reduction layer appended to the transmitter as an example, to explain how the MPA-based precoder works and how it improves the generalization performance against variations of random channels. First, a dimension reduction layer is appended to the deep neural layers in the transmitter to conduct a linear dimension reduction transformation from L (number of extended dimensions by the precedent deep neural layers) to N (degree-of-freedom of the current channel). We assume that current N degree-of-freedom channel measurement is available to the transmitter. In practice, this is realized by frequent channel feedbacks from the receivers or uplink/downlink channel reciprocity. The dimension reduction layer appended to the transmitter is particularly interesting for the downlink, in which the deep neural layers in the terminal receivers are supposed to remain unchanged for as long as possible.

Tuning a linear dimension reduction transformation by MPA enhances generalization against outliers caused by the variations of channels during the inference cycle. Tuning or adjusting the coefficients of the dimension reduction layers is an iterative algorithm that includes forward message passing from functional nodes to variable nodes and backward message passing from variable nodes to functional nodes. Before describing the iterative adjusting procedure over a dimension reduction layer, we will introduce two widely used native-AI technologies: support vector machine (SVM) [12] and attention DNN [13].

To better describe the working mechanism of the introduced transceiver, we first provide the key parameters used in this paper.

2.1 Forward Iteration (Intonation) and Non-linear SVM

An SVM is a supervised machine learning model used for data classification, regression, and outlier detection. In general, an SVM model is composed of a non-linear dimension extension function $\varphi(\cdot)$, a linear combination function $f(\mathbf{x}) = \mathrm{w}\cdot\varphi(\mathbf{x}) + \mathbf{b}$, and a binary classification function $\mathrm{sign}(\cdot)$, where $\mathbf{x}$ is the input data, $\mathbf{w}$ is the weight coefficient vector, and $\mathbf{b}$ is the bias vector, as shown in Figure 8. The objective of SVM is to divide the data samples into classes to find a maximum marginal hyperplane, as shown in Figure 9.

Geometrically, $\boldsymbol{w}\cdot\varphi(\vec{x}) + \vec{b}$ forms an N-dimensional hyperplane. Some hyperplanes are better than others, and this can be measured by margin. In the example shown in Figure 9, the hyperplane on the right is better than the one on the left, because the right one has a larger margin.

Table 1 System parameters

Parameter	Meaning
L	Dimension of the transmitter's output
N	Dimension of the communication channel measurement
$\boldsymbol{h}_k$	k-th channel measurement
$\boldsymbol{n}_k$	k-th additive noise measurement
$\boldsymbol{f}_i$	i-th feature of the transmitter's output
$\boldsymbol{t}_k$	k-th feature vector of the MPA layer's output
$\boldsymbol{r}_k$	k-th received signal
F	Input feature matrix $[\boldsymbol{f}_1,\ldots,\boldsymbol{f}_L]$
H	Channel vector $[\boldsymbol{h}_1,\ldots,\boldsymbol{h}_N]$
N	Noise vector $[\boldsymbol{n}_1,\ldots,\boldsymbol{n}_N]$
R	Received signal vector $[\boldsymbol{r}_1,\ldots,\boldsymbol{r}_N]$
T	Output feature matrix $[\boldsymbol{t}_1,\ldots,\boldsymbol{t}_N]$

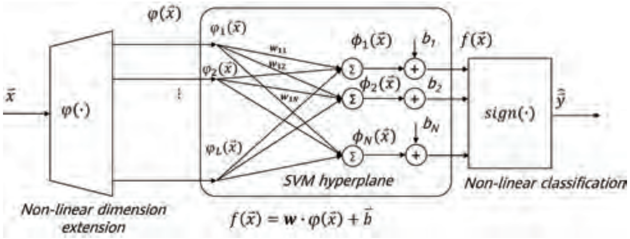
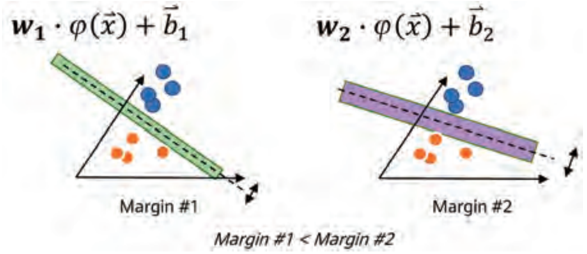


Figure 8 A non-linear SVM for binary classification



$\mathbf{w_1}\cdot\varphi(\vec{x}) + \vec{b}_1$
$\mathbf{w_2}\cdot\varphi(\vec{x}) + \vec{b}_2$
Margin #1
Margin #2
Margin #1 < Margin #2

Figure 9 SVM function defines a hyperplane to classify users

The mathematical description of SVM optimization is as follows:

$$\langle \boldsymbol{w}^*, \vec{b}^* \rangle = \underbrace{\text{argmin}}_{w,\vec{b}}\left(l(\vec{y},\hat{\vec{y}}) + \frac{\|\boldsymbol{w}\|^2}{2}\right),$$

where $l(\vec{y}, \hat{\vec{y}})$ is a given loss measurement function (like a training goal in DNN). To approach the optimal solution $\langle \boldsymbol{w}^*, \vec{b}^* \rangle$, there are several ways such as direct MSE (minimum square error) or SGD. SVM is the predecessor of DNN.

The study of non-linear SVM tells us three things:

- Non-linear dimensional extension transformation $\varphi(\cdot)$ followed by linear dimension reduction transformation $\boldsymbol{w}\cdot\varphi(\vec{x}) + \vec{b}$ can improve classification accuracy, laying a foundation for the MPA-enabled AE-based global transceiver, as shown in Figure 10.
- The linear dimension reduction transformation $\boldsymbol{w}\cdot\varphi(\vec{x}) + \vec{b}$ tunes a hyperplane that separates classes, revealing the mechanism of a forward dimension reduction layer.
- The hyperplane $(\boldsymbol{w}\cdot\varphi(\vec{x}) + \vec{b})$ is tuned with fixed $\varphi(\cdot)$ and $sign(\cdot)$, which indicate a tandem-like training scheme with both the dimension reduction layer and other deep neural layers, as well as an adjustable inference, as shown in Figure 17 later in section 3.1.

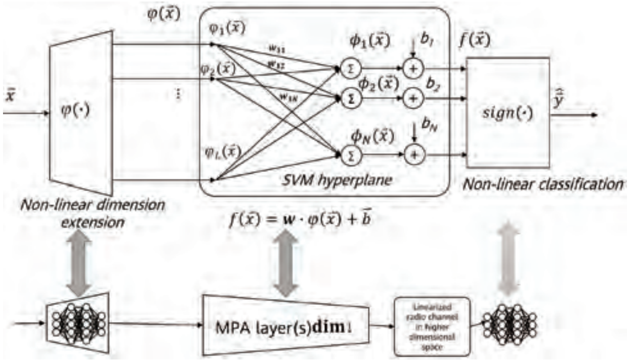


Figure 10 MPA-enabled AE vs non-linear SVM

Hence, the dimension reduction layer in both training and inference cycles keeps adjusting an intermediate hyperplane that helps the final classification by the deep neural layers in the receiver. For wireless context, the hyperplane must be adjusted according to current variations of channels.

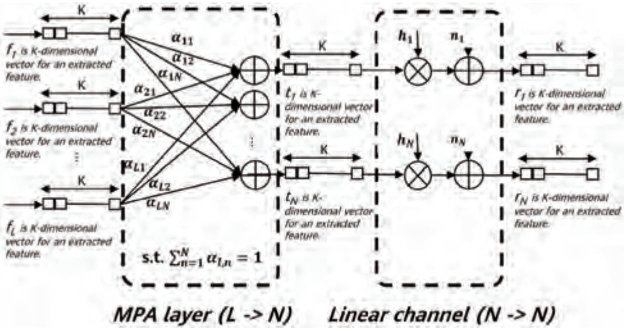

K
f_1 is K-dimensional vector for an extracted feature.
f_2 is K-dimensional vector for an extracted feature.
f_L is K-dimensional vector for an extracted feature.
α_{11}
α_{12}
α_{1N}
α_{21}
α_{22}
α_{2N}
α_{L1}
α_{L2}
α_{LN}
s.t. $\sum_{n=1}^{N}\alpha_{l,n} = 1$
t_1 is K-dimensional vector for an extracted feature.
t_N is K-dimensional vector for an extracted feature.
h_1
n_1
h_N
n_N
r_1 is K-dimensional vector for an extracted feature.
r_N is K-dimensional vector for an extracted feature.
MPA layer (L -> N)
Linear channel (N -> N)

Figure 11 Dimension reduction layer with SVM

Taking advantage of the dimensional transformation of SVM, we can transform the dimension of the transmitter's output, L, to the dimension of the communication channel measurement, N. Figure 11 shows the detailed forward iteration with SVM. In particular, the input of the dimension reduction layer is an L-dimensional feature matrix $\mathbf{F}=[\boldsymbol{f}_1,\boldsymbol{f}_2,\dots,\boldsymbol{f}_L]$, where $\boldsymbol{f}_i$ is the i-th K-dimensional input feature vector. The output of the dimension reduction layer is an N-dimensional feature matrix $\mathbf{T}=[\boldsymbol{t}_1,\boldsymbol{t}_2,\dots,\boldsymbol{t}_N]$, where $\boldsymbol{t}_\mathrm{i}$ is the i-th K-dimensional output feature vector. When the output feature vectors are transmitted via communication channels, the received signal is given by

$$\boldsymbol{r}_\mathrm{i} = \sum_{l=1}^{L} a_{l,i}\cdot \boldsymbol{f}_l \cdot \boldsymbol{h}_i + \boldsymbol{n}_i,\ i = 1,\dots,N,$$

where $a_{l,i}$ is the coefficient of the connection between neuron l and neuron i.

Based on the preceding description, we can conclude that the forward sub-iteration is to keep fine-tuning the hyperplane of the SVM model in both training and inference phases for the given transmitter's feature matrix **F**, channel state information **H**, noise vector **N**, and received signal **R**, as shown in Figure 12.

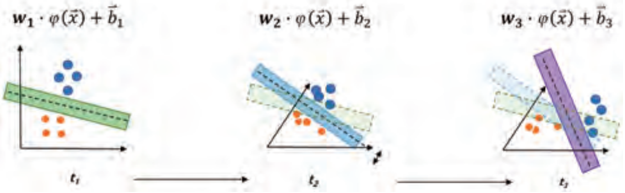

$\boldsymbol{w}_1\cdot\varphi(\vec{x}) + \vec{b}_1$
$\boldsymbol{w}_2\cdot\varphi(\vec{x}) + \vec{b}_2$
$\boldsymbol{w}_3\cdot\varphi(\vec{x}) + \vec{b}_3$
t_1
t_2
t_3

Figure 12 Hyperplane is adjusted over time

2.2 Backward Iteration and Attention DNN

As discussed earlier, the dimension reduction layer needs to be trained by a standalone mode rather than a connection mode with backpropagation from the receiver. In this regard, we consider to use an attention DNN [14] in the backward sub-iteration.

An attention DNN is an efficient approach that measures the similarity of two features with different dimensions. Figure 13 depicts the structure of the attention DNN. The input is the received signal **R**. The attention operation is conducted by computing the inner product of each $\boldsymbol{r}_{\mathrm{i}}$ with an attention coefficient $\boldsymbol{c}_l$, i.e., $\langle \boldsymbol{r}_{\mathrm{i}},\boldsymbol{c}_l\rangle$. This inner product implies the similarity of the signal $\boldsymbol{r}_{\mathrm{i}}$ and the attention coefficient $\boldsymbol{c}_l$, which is normalized by a softmax layer as

$$a_{l,i} = \frac{e^{\langle r_{\mathrm{i}},\, c_l\rangle}}{\sum_{n=1}^{N} e^{\langle r_{\mathrm{n}},\, c_l\rangle}}\ ,\ i = 1,\ldots,N.$$

Then, the output of the attention DNN is given by

$$\boldsymbol{z}_l = \sum_{i=1}^{N} a_{l,i} \cdot \boldsymbol{r}_i\ ,\ l = 1,\ldots,L.$$

We shall note that the number of attentions is smallerthan the number of received signals, i.e., $L < N$.

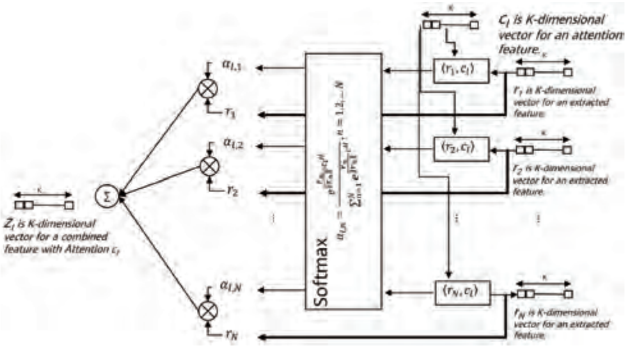

K
c_l is K-dimensional vector for an attention feature.
$\alpha_{l,1}$
r_1
$\alpha_{l,2}$
r_2
$\alpha_{l,N}$
r_N
K
Z_l is K-dimensional vector for a combined feature with Attention c_l
Σ
Softmax
$\alpha_{l,n} = \frac{e^{\frac{r_n}{\|r_n\|} c_l^H}}{\sum_{n=1}^{N} e^{\frac{r_n}{\|r_n\|} c_l^H}}, n = 1,2,\ldots,N$
$\langle r_1, c_l\rangle$
$\langle r_2, c_l\rangle$
$\langle r_N, c_l\rangle$
K
r_1 is K-dimensional vector for an extracted feature
K
r_2 is K-dimensional vector for an extracted feature
K
r_N is K-dimensional vector for an extracted feature

Figure 13 Structure of the attention DNN

The attention DNN can be employed in the dimension reduction layer for back propagation. In particular, each extracted feature vector $\boldsymbol{f}_l$ can be used as an attention coefficient. Then, in the backward subiteration, the coefficient $a_{l,i}$ can be given by

$$a_{l,i} = \frac{e^{\langle r_{\mathrm{i}},\, f_l\rangle}}{\sum_{n=1}^{N} e^{\langle r_{\mathrm{n}},\, f_l\rangle}}\ ,\ i = 1,\ldots,N, l = 1,\ldots,L.$$

The attention DNN allows a number of attentions: $\boldsymbol{c}_l$,l = 1,2,3,…,L. Then, it can generate an L combination $\boldsymbol{z}_l$,l = 1,2,3,…,L. In a practical attention DNN, the number of attentions is smaller than the number of captured features: $L < N$, because, in reality, you cannot hold a great number of attentions simultaneously.

In the MPA layer, in order to measure the similarity between the received signal space $[\boldsymbol{r}_1,\boldsymbol{r}_2,\ldots,\boldsymbol{r}_N]$ and extracted feature space $[\boldsymbol{f}_1,\boldsymbol{f}_2,\ldots,\boldsymbol{f}_L]$, we can borrow the preceding method. Suppose that the deep neural layers in the transmitter are well trained to extract $[\boldsymbol{f}_1,\boldsymbol{f}_2,\ldots,\boldsymbol{f}_L]$ for a specific goal. We can consider one feature $\boldsymbol{f}_l$ as one attention. This is how we compute the coefficients ($a_{l,i}$,l = 1,2,…,L,i = 1,2,3,…,N) of the dimension reduction layer in the backward iteration.

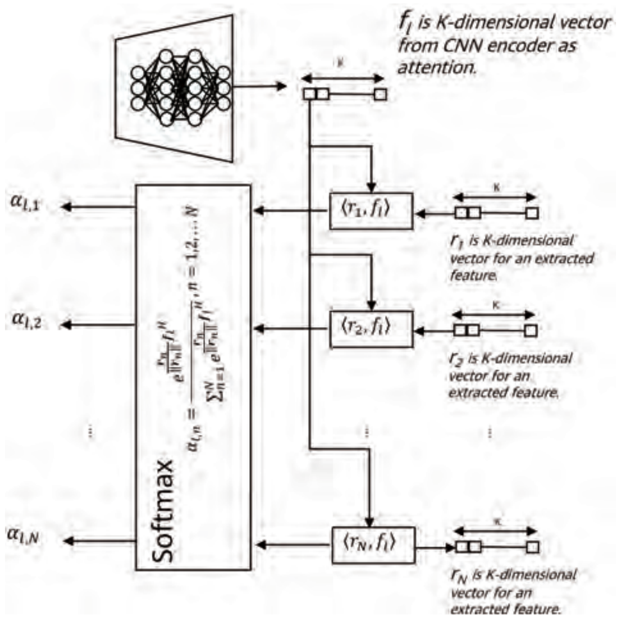

f_l is K-dimensional vector from CNN encoder as attention.
K
$\alpha_{l,1}$
$\alpha_{l,2}$
$\alpha_{l,N}$
Softmax
$\alpha_{l,n} = \frac{e^{\frac{r_n}{\|r_n\|} f_l^H}}{\sum_{n=1}^{N} e^{\frac{r_n}{\|r_n\|} f_l^H}}, n = 1,2,\ldots,N$
$\langle r_1, f_l\rangle$
$\langle r_2, f_l\rangle$
$\langle r_N, f_l\rangle$
K
r_1 is K-dimensional vector for an extracted feature.
K
r_2 is K-dimensional vector for an extracted feature.
K
r_N is K-dimensional vector for an extracted feature.

Figure 14 $a_{l,i}$ represents the weight of each received feature associated to extracted feature $\boldsymbol{f}_l$

Since $a_{l,i}$ tells the similarity between the feature $\boldsymbol{f}_l$ and received feature $\boldsymbol{r}_i$, it can also be the scaling weight from the extracted feature $\boldsymbol{f}_l$ to the transmitted feature $\boldsymbol{t}_i$.

2.3 Standalone MPA Iteration

Figure 15 illustrates the overall MPA iteration. The forward part on the left is equivalent to a non-linear SVM, whereas the backward part on the right is equivalent to an attention DNN.

The MPA iteration is standalone. It is independent of the SGD-based backward propagation of the original AE.

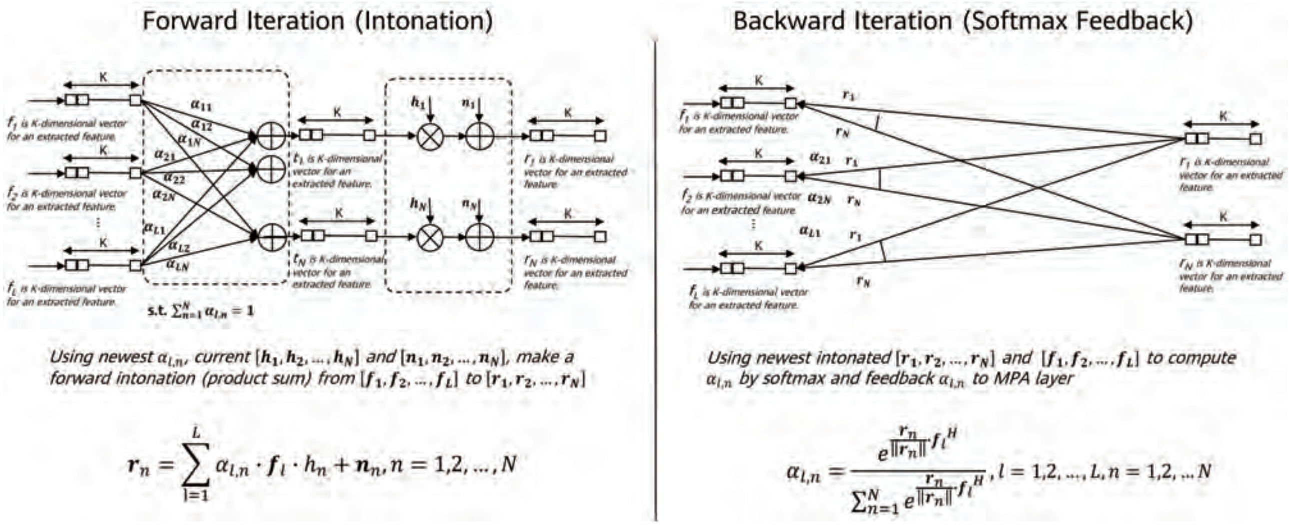

Forward Iteration (Intonation)
Backward Iteration (Softmax Feedback)
K
f_1 is K-dimensional vector for an extracted feature.
f_2 is K-dimensional vector for an extracted feature.
f_L is K-dimensional vector for an extracted feature.
α_{11} α_{12} α_{1N} α_{21} α_{22} α_{2N} α_{L1} α_{L2} α_{LN}
s.t. $\sum_{n=1}^{N}\alpha_{l,n}=1$
t_1 is K-dimensional vector for an extracted feature.
t_N is K-dimensional vector for an extracted feature.
h_1 n_1 h_N n_N
r_1 is K-dimensional vector for an extracted feature.
r_N is K-dimensional vector for an extracted feature.
r_1 r_N
Using newest $\alpha_{l,n}$, current $[h_1, h_2, \dots, h_N]$ and $[n_1, n_2, \dots, n_N]$, make a forward intonation (product sum) from $[f_1, f_2, \dots, f_L]$ to $[r_1, r_2, \dots, r_N]$
$r_n = \sum_{l=1}^{L} \alpha_{l,n} \cdot f_l \cdot h_n + n_n, n = 1,2, \dots, N$
Using newest intonated $[r_1, r_2, \dots, r_N]$ and $[f_1, f_2, \dots, f_L]$ to compute $\alpha_{l,n}$ by softmax and feedback $\alpha_{l,n}$ to MPA layer
$\alpha_{l,n} = \frac{e^{\frac{r_n}{\|r_n\|} f_l^H}}{\sum_{n=1}^{N} e^{\frac{r_n}{\|r_n\|} f_l^H}}, l = 1,2, \dots, L, n = 1,2, \dots N$

Figure 15 Standalone MPA iteration

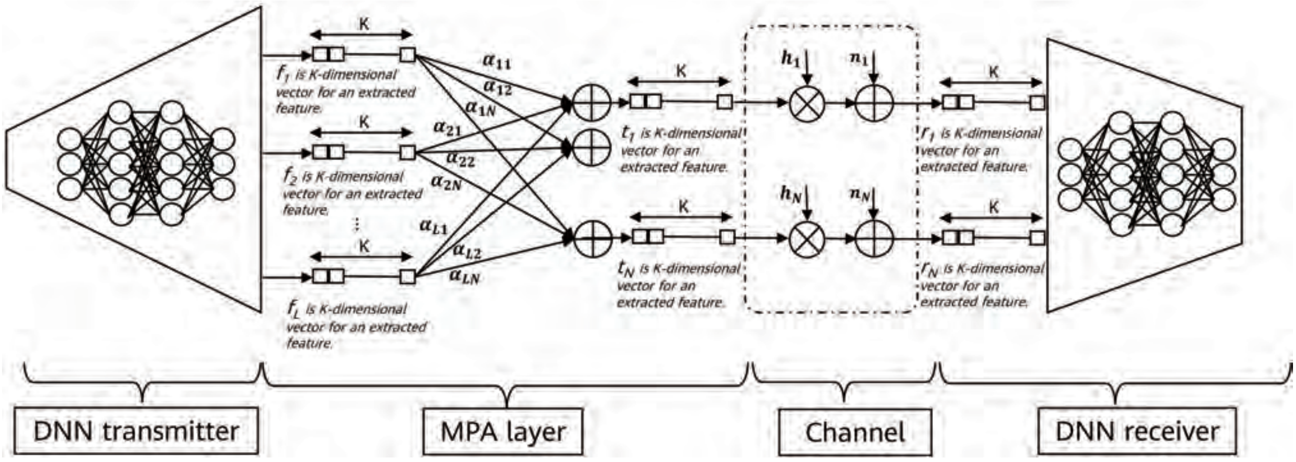

K
f_1 is K-dimensional vector for an extracted feature.
f_2 is K-dimensional vector for an extracted feature.
f_L is K-dimensional vector for an extracted feature.
α_{11} α_{12} α_{1N} α_{21} α_{22} α_{2N} α_{L1} α_{L2} α_{LN}
t_1 is K-dimensional vector for an extracted feature.
t_N is K-dimensional vector for an extracted feature.
h_1 n_1 h_N n_N
r_1 is K-dimensional vector for an extracted feature.
r_N is K-dimensional vector for an extracted feature.
DNN transmitter
MPA layer
Channel
DNN receiver

Figure 16 MPA-enabled AE-based global transceiver

This means that when the MPA iterates its coefficients, it assumes that the remaining deep neural layers in both the transmitter and the receiver are frozen, as shown in Figure 16.

3 Global Tandem Learning

3.1 Coarse Learning

The insertion of a dimension reduction layer divides the training into two training agents: one is a standalone agent for the dimension reduction layer, and the other is backward propagation for the deep neural layers in the transceiver.

The two training agents work in tandem, as illustrated in Figure 17:

- In tandem stage 1, the dimension reduction layers are fixed (as shown on the left of Figure 17). As dimension reduction layers are linear and differentiable in forward intonation, they can pass the gradients backwardly to have deep neural layers trained by backward propagation.
- In tandem stage 2, the deep neural layers are fixed (as shown on the right of Figure 17). The dimension reduction layers are trained by the standalone MPA, just like they are trained by a non-linear SVM.

Tandem stage 1 and stage 2 are iterative until the training converges.The detailed procedure is summarized in Algorithm 1.

Algorithm 1 The training algorithm

```
Initialize the coefficients of the MPA layer, $\alpha_{l,i} = \frac{1}{L}$.
Initialize the batch size $b$.
for steps 1:T do
    In tandem stage 1
        Sample a batch of training messages $\mathbf{X} = [\mathbf{x}_1,...,\mathbf{x}_b]$.
        DNN N-based transmitter computes $\mathbf{F} = [\boldsymbol{f}_1,..., \boldsymbol{f}_L]$
    based on the training messages $\mathbf{X}$.
        Compute
            $\boldsymbol{t}_i = \sum_{l=1}^{L} \alpha_{l,i} \cdot \boldsymbol{f}_l, i=1,...,N,$
        Send $\mathbf{T} = [\boldsymbol{t}_1, \boldsymbol{t}_2,...,\boldsymbol{t}_N]$ to the DNN-based receiver
    via communication channels, as
            $\boldsymbol{r}_i = \boldsymbol{t}_i \cdot \boldsymbol{h}_i + \boldsymbol{n}_i, i=1,...,N.$
        DNN-based receiver inputs the received signals
 $\mathbf{R} = [\boldsymbol{r}_1, \boldsymbol{r}_2,...,\boldsymbol{r}_N]$ into the DNN and computes the decoded
 message.
        Update the transmitter and the receiver by
 backpropagation.
    In tandem stage 2
        for iterations 1:M do
        Compute
            $r_i = \sum_j h_i \alpha_{l,i} \cdot \boldsymbol{f}_l.$
        Compute
            $\|\boldsymbol{r}\| = \sqrt{\boldsymbol{r}_1^2 + \boldsymbol{r}_2^2 + \cdots + \boldsymbol{r}_N^2}.$
        Update
            $\boldsymbol{\beta}_{l,k} = \langle \frac{r_i}{\|r\|}, \boldsymbol{f}_l \rangle, i=1,...,N, l=1,...,L.$
        Update the coefficients of the MPA layer by
            $\alpha_{l,i} = \text{softmax}(\boldsymbol{\beta}_{l,i}), i=1,..., N, l=1,..., L.$
        end for
    end for
Output $\alpha_{l,i}\ l=1,...,L; i=1,...,N.$
```

3.2 Inference Cycle Adaptation

Because it is trained in a standalone mode, the dimension reduction layer can be tuned for each transmission.
It is interesting for the MPA-enabled AE-based global transceiver to adapt its transmitter to the variation of channels during the inference cycle and to keep the receiver unchanged.

On the left of Figure 18 is the AE-based global transceiver without the dimension reduction layers. The AE-based transceiver cannot change its neurons coefficients during the inference cycle, even if the channel environment and/or source distribution has changed already. Its performance relies on the odds that the varying channel is in the interpolation range; otherwise, the transceiver has to accumulate sufficient new data sets to start either a new training or transfer learning to some of the deep neural layers, all of which are detrimentalto apply the AE-based transceiver in a fast-changing environment.

On the right of Figure 18 is the MPA-enabled AE-based global transceiver. The transmitter keeps adjusting the dimension reduction layer(s) for each transmission. Since the dimension reduction layer is an intermediate layer of the transceiver, its coefficient updates in terms of the current channel can significantly enhance generalization. Moreover, the dimension reduction layers are adjusted on the fly by the real-world channel measurement without any labeled data. The detailed procedure is summarized in Algorithm 2.

Algorithm 2 The inference algorithm

```
Input: New messages $\mathbf{X}$.
DNN-based transmitter computes $\mathbf{F} = [\boldsymbol{f}_1,..., \boldsymbol{f}_L]$ based on
the new message $\mathbf{X}$.
for iteration from 1:M do
    Compute
        $\boldsymbol{r}_i = \sum_{l=1}^{L} h_i \alpha_{l,i} \cdot f_l, i=1,...,N.$
    Compute
        $\|\boldsymbol{r}\| = \sqrt{\boldsymbol{r}_1^2 + \boldsymbol{r}_2^2 + \cdots + \boldsymbol{r}_N^2}.$
    Update
        $\boldsymbol{\beta}_{l,k} = \langle \frac{r_i}{\|r\|}, \boldsymbol{f}_l \rangle, i=1,...,N, l=1,...,L.$
    Update the coefficients of the MPA layer by
        $\alpha_{l,i} = \text{softmax}(\boldsymbol{\beta}_{l,i}), i=1,..., N, l=1,...,L.$
end for
Compute
        $\boldsymbol{t}_i = \sum_{l=1}^{L} \alpha_{l,i} \cdot \boldsymbol{f}_l, i=1,...,N,$
Compute
        $\boldsymbol{r}_i = \boldsymbol{t}_i \cdot \boldsymbol{h}_i + \boldsymbol{n}_i, i=1,..., N.$
DNN-based receiver inputs the received signals
$\mathbf{R} = [\boldsymbol{r}_1,...,\boldsymbol{r}_N]$ into the DNN and computes the decoded
message $\hat{\mathbf{X}}$.
Output $\hat{\mathbf{X}}$.
```

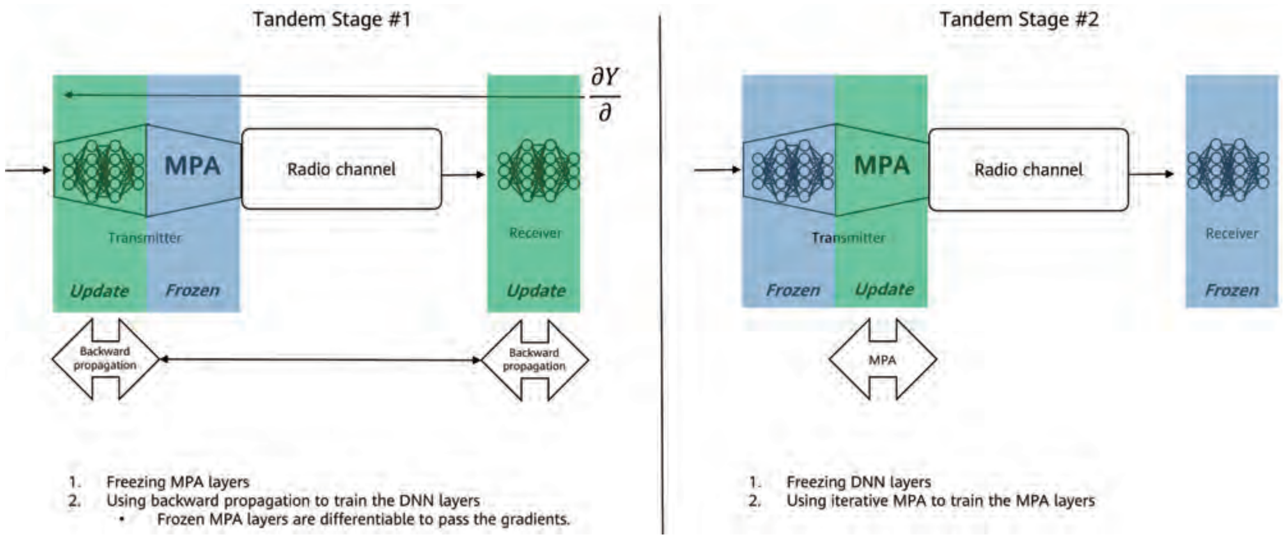


Figure 17 Tandem training cycle
* In each round, the green part is first trained assuming the blue part is fixed,
and then the blue part is trained assuming the green part is fixed.

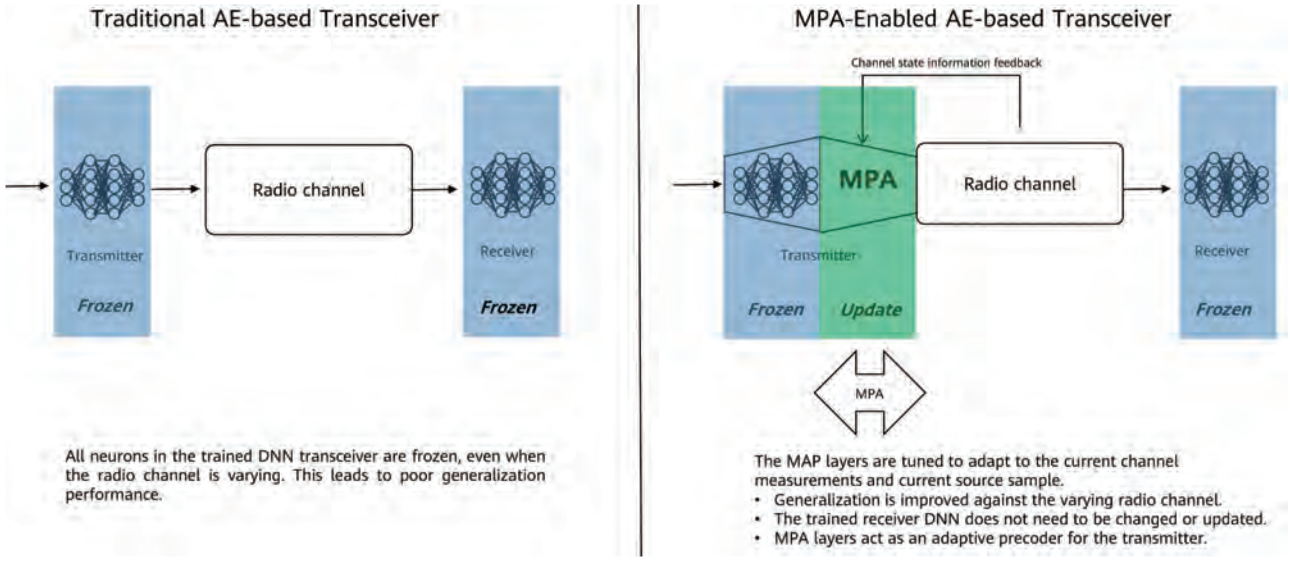

Traditional AE-based Transceiver
Radio channel
Transmitter
Receiver
Frozen
Frozen
All neurons in the trained DNN transceiver are frozen, even when the radio channel is varying. This leads to poor generalization performance.
MPA-Enabled AE-based Transceiver
Channel state information feedback
MPA
Radio channel
Transmitter
Receiver
Frozen
Update
Frozen
MPA
The MAP layers are tuned to adapt to the current channel measurements and current source sample.
• Generalization is improved against the varying radio channel.
• The trained receiver DNN does not need to be changed or updated.
• MPA layers act as an adaptive precoder for the transmitter.

Figure 18 AE-based transceiver, with or without MPA enabled.
The MPA precoder can be flexibly adjusted online to adapt to varying channel conditions.

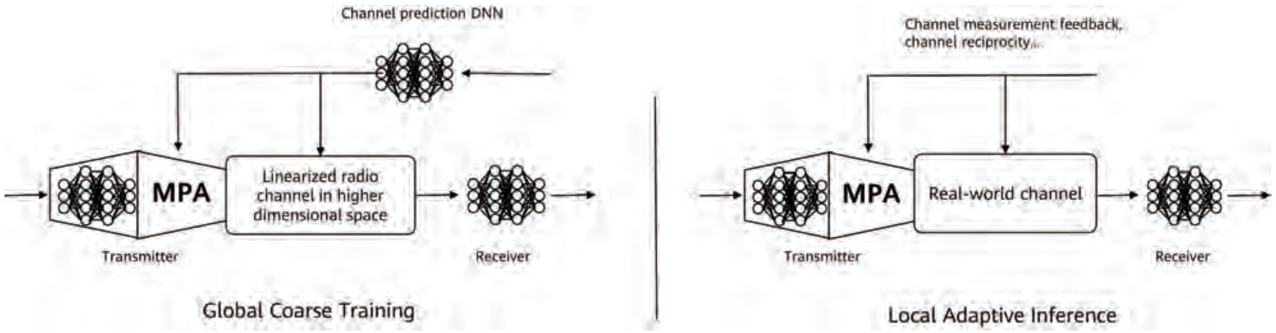


Figure 19 Coarse learning and adaptive inference

3.3 Advantages of MPA-Enabled AE-based Global Transceiver

In addition to the adaptive inference to the changing channel by the dimension reduction layer, the dimension reduction layer can bring about more profound advantages.

First, channel model simplification for being differentiable would harm the performance of the AE-based transceiver, because AE is trained for the simplified channel model rather than a true one. The performance loss is due to the offset between the simplified channel model used during the training cycle and the true channel faced during the inference cycle. If the offset increases beyond expectation, the entire AE-based transceiver would become obsolete. Two remedies are proposed to mitigate the performance degradation. The first one is based on reinforcement learning (RL) that keeps recording the channel states and keeps training the policy DNN and/or evaluation DNN from time to time [2]. However, RL is too complex in a dimensionality such as a wireless system, and actually deals with a dimensionality that is much bigger than AlphaGo does. Therefore, the RL-based adjustment mechanism is impracticable. The second is based on the generative adversary network (GAN) that learns as many channel scenarios as possible into a big DNN model [3]. This is an empirical method that can never be proved to cover all the channel scenarios.

The MPA-enabled AE takes a different technology path. Because MPA adjusts the coefficients of the dimension reduction layer on the fly for each data transmission in the function of current channel measurement during the inference cycle, adaptive inference leverages a coarse channel model during the training cycle. This is what we call "coarse learning". Sometimes, it is hard to demonstrate the advantage by simulations in which the same or similar channel model is emulated in both training and inference cycles [2]. However, the advantage can be demonstrated in real field tests.

Then, the MPA-enabled AE-based transceiver can work with the GAN-based channel model [3]. Our experience tells us that most channels are related to the user's position and environmental topologies such as high buildings, hills, and roads and so on. [3] proposes to use a conditional GAN to model unknown channels and achieves good performance. This method can be used for developing a channel model that supports our training cycle.

In the inference cycle, we suggest relying on the channel estimation on pilots, channel measurements and feedbacks, or channel reciprocity to obtain the most current channel condition. Furthermore, MPA is well known to benefit from the sparsity and tolerate the bias and offsets (this is the reason why LDPC decoder can work). This means that it is not really necessary to measure the channels on full dimensionality but just a part of it, and that even with some estimation error, our scheme can be robust for the overall performance. In addition, the residual errors can be tackled by the receiving deep neural layers that are good at tolerating errors. Thanks to the dimension reduction layers adjusted during the inference and training cycles and used as a precoder for the overall transmitting chain, the receiving deep neural layers can remain untrained, and this is a huge advantage in terms of power saving and extension of battery life of the user's device.

According to Figure 19, we highlight the complete picture of our proposed scheme: For the transceiver algorithm architecture, we choose an appropriate MPA-adjustable dimension reduction, to not only address information bottleneck at the channel input but also minimize the SGD gradient overhead for the training cycle. During the training cycle, we can use a generated channel model. A tandem training scheme is used: back propagate the gradients through the frozen dimension reduction layer,freeze the autoencoder layers, and iterate dimension reduction layers. During the inference cycle, we freeze the autoencoder layers and iterate the dimension reduction layers by true channel measurements or feedbacks.

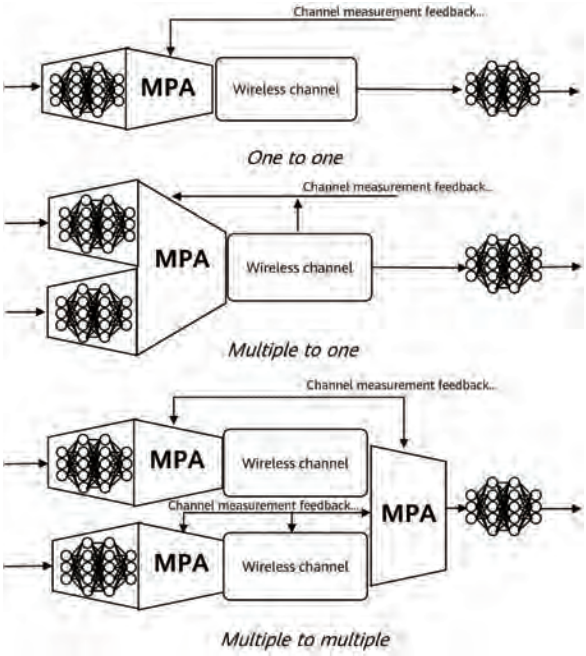

Channel measurement feedback...
MPA
Wireless channel
One to one
Channel measurement feedback...
MPA
Wireless channel
Multiple to one
Channel measurement feedback...
MPA
Wireless channel
Channel measurement feedback...
MPA
MPA
Wireless channel
Multiple to multiple

Figure 20 Flexible insertions of MPA layer(s)

4 Future Directions

4.1 Flexible MPA Enabling Scheme

The dimension reduction layers can be flexibly inserted into an AE-based transceiver for various problems.

In one-to-one single-user communication shown in the top part of Figure 20, we applied this method — flexibly inserting dimension reduction layers into an AE-based transceiver — to design a high order modulation scheme, a massive MIMO scheme, a predistortion scheme,and so on. These designs achieved good results in time-varying channel conditions.

This method can also be used in the multi-user communication scheme. For example, in a downlink MU-MIMO context in the middle part of Figure 20, we can jointly adapt the MPA layer.

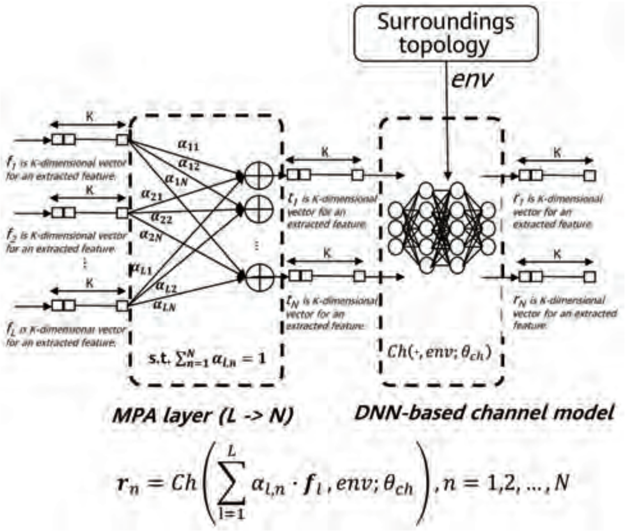
For uplink MU-MIMO, each user transmitter can adapt the MPA layer individually, while the base station receiver can adapt the dimension reduction layer prior to the DNN receiver as shown in the bottom part of Figure 20.

By introducing the dimension reduction layer in the autoencoder structure based on deep learning, we can extend the current DNN framework into a more generalized one, in other words, to enhance the generality of OOD.

4.2 Training with a Complex Channel Model

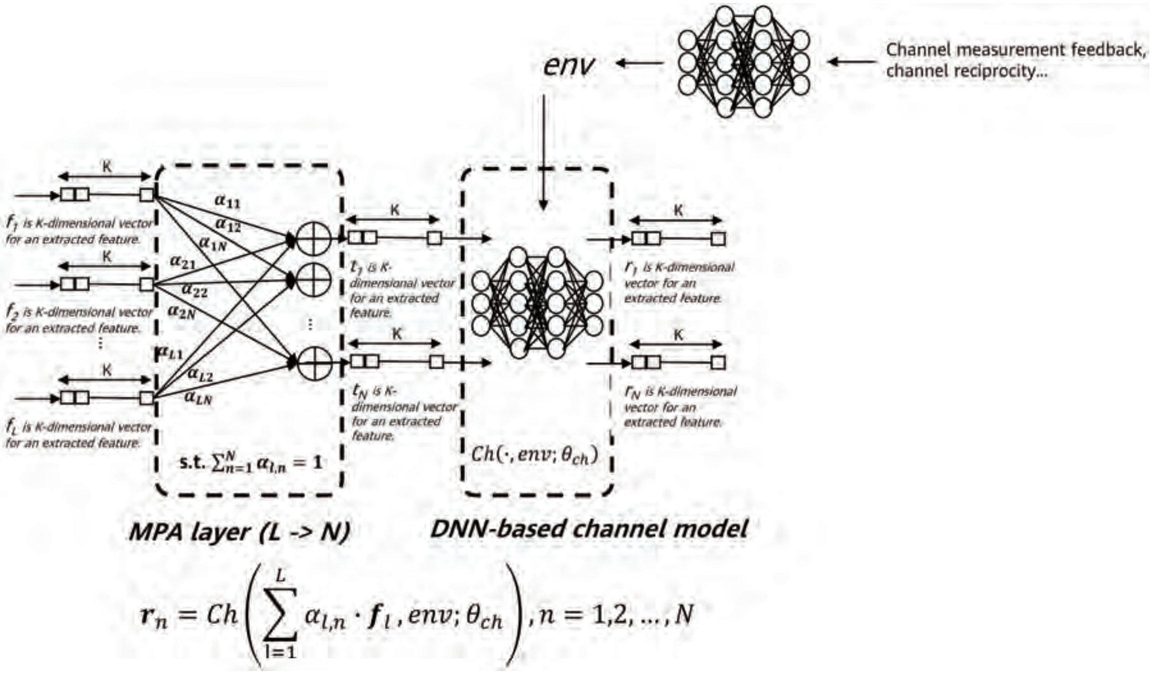
Another research direction is to use a more complex channel model during the training cycle. Most probably, the channel model is generated by a DNN with input of the surrounding topological information, as illustrated in Figure 21.

The MPA iteration is still valid because the inference DNN of channel can be considered as a non-linear function $Ch(\cdot,enh;\theta_{ch})$.



Surroundings topology
env
K
f_1 is K-dimensional vector for an extracted feature
f_2 is K-dimensional vector for an extracted feature
f_L is K-dimensional vector for an extracted feature
α_{11}
α_{12}
α_{1N}
α_{21}
α_{22}
α_{2N}
α_{L1}
α_{L2}
α_{LN}
t_1 is K-dimensional vector for an extracted feature
t_N is K-dimensional vector for an extracted feature
r_1 is K-dimensional vector for an extracted feature
r_N is K-dimensional vector for an extracted feature
s.t. $\sum_{n=1}^{N}\alpha_{l,n}=1$
$Ch(\cdot,env;\theta_{ch})$
MPA layer (L -> N)
DNN-based channel model
$r_n = Ch\left(\sum_{l=1}^{L}\alpha_{l,n}\cdot f_l, env;\theta_{ch}\right), n = 1,2,\dots,N$

Figure 21 Training with a DNN-based channel model



env
Channel measurement feedback, channel reciprocity...
K
f_1 is K-dimensional vector for an extracted feature.
f_2 is K-dimensional vector for an extracted feature.
f_L is K-dimensional vector for an extracted feature.
α_{11}
α_{12}
α_{1N}
α_{21}
α_{22}
α_{2N}
α_{L1}
α_{L2}
α_{LN}
t_1 is K-dimensional vector for an extracted feature.
t_N is K-dimensional vector for an extracted feature.
r_1 is K-dimensional vector for an extracted feature.
r_N is K-dimensional vector for an extracted feature.
s.t. $\sum_{n=1}^{N}\alpha_{l,n}=1$
$Ch(\cdot,env;\theta_{ch})$
MPA layer (L -> N)
DNN-based channel model
$r_n = Ch\left(\sum_{l=1}^{L}\alpha_{l,n}\cdot f_l, env;\theta_{ch}\right), n = 1,2,\dots,N$

Figure 22 During the inference cycle, we suggest using another DNN to generate the current environment parameters.

Nevertheless, if the DNN-based channel model is huge, it will probably be too costly to use this model for inference in an iterative way. Research on how to reduce the size of the DNN by model distillation [15] is required.

For the inference cycle, the environmental parameters can be results from another DNN that deduces the surrounding topological information from the current channel measurements and feedbacks,as shown in Figure 22.

5 Conclusion

In this paper, we address the fundamental issue in the AE-based global transceiver, which is poor generalization against variation of random channels, by introducing an MPA-enabled feature. MPA can adapt the AE-based transceiver during the inference cycle to the current channel measurement. Thanks to MPA, the learning cycle can tolerate more simplification of the channel model. The introduction of MPA into AE architecture is based on solid dimensional transformation technology widely used in classic wireless systems, and the implementation of MPA is a mature and efficient technology in wireless communications.

References

[1] M. Honkala, D. Korpi, and J. M. J. Huttunen, "DeepRx: Fully convolutional deep learning," May 2020, arxiv:2005.01494.

[2] Fayçal Ait Aoudia and Jakob Hoydis, "End-to-End learning of communications systems without a channel model," arXiv:1804.02276v3.

[3] Hao Ye, Le Liang, Geoffrey Ye Li, and Biing-Hwang Fred Juang, "Deep learning based end-to-end wireless communication systems with conditional GAN as unknown channel," arXiv:1903.02551.

[4] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy, "Explaining and harnessing adversarial examples," ICLR 2015, arXiv: 1412.6572.

[5] Melanie Mitchell, "Why AI is harder than we think," April 2021, arxiv:2104.12871.

[6] Han Cai, Chuang Gan, Tianzhe Wang, Zhekai Zhang, and Song Han, "Once-for-All: Train one network and specialize it for efficient deployment," April 2020, arxiv:1908.09791.

[7] Yiqun Ge and Wen Tong, Chapter 9 "Mathematics, Information and Learning," *Mathematics for Future Computing and Communications*, 2021, Cambridge University Press.

[8] Yoshua Bengio, "From conscious processing to system 2 deep learning," July 2021, https://www.youtube.com/watch?v=nE0M3XvaaVU

[9] K. J. Astrom and R. M. Murray, "Feedback systems: An introduction for scientists and engineers," April 12, 2010, Princeton University Press.

[10] R. F. Stengel, "Optimal control and estimation," September 20, 1994, Dover Publications.

[11] John G. Proakis and Masoud Salehi (2008), *Digital Communications*, 5th ed.

[12] Mark A. Aizerman, Emmanuel M. Braverman, and Lev I. Rozonoer (1964), "Theoretical foundations of the potential function method in pattern recognition

learning," *Automation and Remote Control*, 25:821-837.

[13] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin (2017-12-05), "Attention is all you need," arXiv:1706.03762.

[14] Kanika Madan, Rosemary Nan Ke, Anirudh Goyal, Bernhard Schölkopf, and Yoshua Bengio, "Fast and slow learning of recurrent independent mechanisms," ICLR 2021.

[15] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean (2015), "Distilling the knowledge in a neural network," arXiv:1503.02531.

User-Centric Cell-Free Wireless Networks for 6G: Communication Theoretic Models and Research Challenges

Fabian Göttsch [1], Giuseppe Caire [1], Wen Xu [2], Martin Schubert [2]

[1] Faculty of EECS, Technische Universität Berlin

[2] Huawei Technologies Duesseldorf GmbH, Munich Research Center

Abstract

This paper presents a comprehensive communication theoretic model for the physical layer of a cell-free user-centric network, formed by user equipments (UEs), radio units (RUs), and decentralized units (DUs), uniformly spatially distributed over a given coverage area. We consider RUs equipped with multiple antennas, and focus on the regime where the UE, RU, and DU densities are constant and therefore the number of such nodes grows with the coverage area. A system is said scalable if the computing load and information rate at any node in the network converges to a constant as the network size (coverage area) grows to infinity. This imposes that each UE must be processed by a (user-centric) finite-size cluster of RUs, and that such cluster processors are dynamically allocated to the DUs (e.g., as software defined virtual network functions) in order to achieve a balanced computation load. We also assume that the RUs are connected to the DUs through a packet switching network, in order to achieve adaptive routing and load balance. For this model, we define in details the dynamic cluster formation and uplink pilot allocation. As a consequence of the pilot allocation and the scalability constraint, each cluster processor has a partial view of the network channel state information. We define the condition of "ideal partial CSI" when the channel vectors that can be estimated are perfectly known (while the ones that cannot be estimated are not known at all). We develop two attractive cluster-based linear receiver schemes for the uplink, and an uplink-downlink duality that allows to reuse such vectors as precoders for the downlink. Finally, we show that exploiting the channel antenna correlation structure arising from a geometrically consistent model for directional propagation (which is well-motivated for short distance semi-line of sight propagation typical of dense wireless networks), and performing a channel subspace projection of the uplink pilot field, the pilot contamination effect arising from pilot reuse across the network can be effectively reduced to provide only marginal degradation with respect to the ideal partial CSI case. Several system aspects such as initial acquisition of the UEs, UE-RU association, and distributed scheduling for fairness and load balance between uplink and downlink are briefly discussed and identified as challenging research topics for further investigation.

1 Introduction

The ever-growing demand for wireless data and ubiquitous broadband connectivity is pushing industry and standardization bodies to develop and release new generations of wireless systems designed to meet such demands. Since the very beginning, cellular has been the dominant architecture paradigm for outdoor wireless networks (e.g., see [1] and references therein). Namely, a given coverage area A is partitioned into cells, and UEs in any given cell are uniquely associated to the corresponding base station (BS),[1] which implements the full stack of the radio access protocol. BSs are connected via a backhaul network to the rest of the system, and eventually through some gateway to the Internet. The area capacity in bit/s/m^2 is mainly driven by the following three factors: 1) cell density; 2) system bandwidth; 3) physical layer (PHY) and multiaccess schemes.

Taking the Gaussian capacity formula as a rule of thumb to represent the efficiency of the underlying PHY, the area capacity of a cellular system can be expressed as

$$C = \lambda_a \times W \times \eta \log(1 + \text{SINR}) + \text{const.}$$

where λ_a is the cell density (number of BSs per unit area), W is the system bandwidth in Hz, $\eta\log(1 + \text{SINR})$ is the sum spectral efficiency per cell supported by the PHY, and where the constant term accounts for the non-ideal implementation and technology effects. The cell spectral efficiency can be further broken down into the pre-log factor η, usually referred to as the *multiplexing gain*, and the term log(1 + SINR), corresponding to the capacity in bit per channel use of a Gaussian channel with given Signal to Interference plus Noise Ratio (SINR). This expression corresponds to a PHY scheme able to support η virtual parallel Gaussian channels in the spatial domain, i.e., using multiple antenna MIMO technology (e.g., see [2] and references therein). In addition, the term SINR expresses the average *Signal to Interference plus Noise Ratio* of a typical user randomly located in the cell area, where averaging is with respect to large-scale effects (distance dependent pathloss and shadowing/blocking effects), as well as small-scale effects (the fading due to multipath propagation). Taking the average SINR inside the log function yields an upper bound due to Jensen's inequality and the concavity of the logarithm. A refined analysis of the area capacity can be obtained using stochastic geometry approaches (e.g., see [3–5]). In addition, it is often useful to characterize the system performance in terms of the *Cumulative Distribution Function* (CDF) of the per-user rate, instead of the (average) area capacity. In this case, one should consider the rate of a given user randomly placed in the cell, with averaging with respect to the small-scale fading and conditioning on the large-scale effects. As a conditional average, this rate is a random variable, and the corresponding CDF yields the per-user rate distribution (e.g., see [6-7]). Nevertheless, for the sake of the discussion in this introduction section, the above simple formula is sufficient to capture the main factors influencing the system performance.

Historically, the spectrum allocated to cellular systems has steadily grown with the successive system generations. However, the availability of "beachfront spectrum" – namely, licensed spectrum below 6 GHz with broad geographic support – is very limited [8]. Various windows of a few hundreds of MHz are available in different geographic regions, and systems must accommodate for bandwidth aggregation and dynamic spectrum allocation in order to scavenge such scarce available spectrum. Meanwhile, the available spectrum in higher frequency bands (e.g., 7-11 GHz and 24-54 GHz) is certainly more plentiful, although still quite unproven for cellular outdoor and mobile communications as such frequencies struggle to penetrate walls and other blocking objects.

Given that W cannot grow significantly (at least in the spectrum below 6 GHz), the area capacity can be increased by letting the product $\lambda_a \times \eta$ grow, provided that the SINR of the virtual per-user channels does not collapse to too low values. The increase of the multiplexing gain η is achieved through multiuser MIMO (MU-MIMO) technology. In a conventional cellular system, each BS is equipped with an array of M antenna elements, providing effectively a vector channel (namely, a vector Gaussian multiple-access channel (MAC) in the uplink (UL), and a vector Gaussian broadcast channel (BC) in the downlink (DL)). As long as the rank of the channel matrix between the BS antenna array and the served users has rank not smaller than some integer d

[1] For simplicity of exposition we do not distinguish here between cells and sectors, which are conceptually equivalent when each sector handles its own UEs individually.

and is known at the BS side with sufficient accuracy, MU-MIMO receivers (for the UL) and precoders (for the DL) can be designed in order to support d virtually non-interfering data streams in both the UL and the DL. In general, the multiplexing gain η is less than d since other effects must be taken into account, and in particular, the overhead incurred by estimating the UL and DL channel matrix.

Since the first information theoretical studies [9–12] to the inclusion in recent wireless standards [13–15], MU-MIMO is arguably one of the key transformative ideas that have shaped the last 15 years of theoretical research and practical technology development. A successful related concept is *massive MIMO* [2, 16]. This is based on the key idea that, thanks to channel reciprocity and time-division duplexing (TDD) operations, an arbitrarily large number M of BS antennas can be trained by a finite number of UEs using a finite-dimensional UL pilot field of $\tau_p \geq d$ UL symbols per coherence block.

In TDD systems, UL and DL operate on the same frequency band. The channel reciprocity condition is verified if the UL and DL slots occur within an interval significantly shorter than a channel coherence time, and if the Tx/Rx hardware of the BS radio are calibrated. Hardware calibration has been widely demonstrated in practical testbeds, and can be achieved either using particular RF design solutions (e.g., see [17]) or using over-the-air calibration (e.g., see [18]). As far as the channel coherence time is concerned, a typical mobile user at carrier frequency between 2.0 and 3.7 GHz incurs Doppler typical spreads of ~100 Hz corresponding to channel coherence times of ~10 ms. For example, with TDD slots of 1 ms (corresponding to a subframe of a 10ms frame of 5GNR), we are well in the range for which the channel can be considered time-invariant over a UL/DL cycle. For faster moving users or higher carrier frequencies, the phenomenon of "channel aging" [19] between the UL and the DL slot cannot be neglected any longer. Specific channel estimation and short-term prediction (across the UL/DL slot) have been investigated in the literature, in particular for mmWave bands where the propagation happens to be mainly along the line-of-sight and a few discrete reflection paths. However, a thorough discussion of these aspects goes beyond the scope of the present paper.

In this work we assume that the channel is exactly constant over a time-frequency tile of T channel uses in the time-frequency plane, where $T = T_c \times W_c$ and T_c denotes the channel coherence time (in s) and W_c denotes the channel coherence bandwidth (in Hz). For example, a coherence block may coincide (roughly) with a so-called Resource Block (RB) of 5GNR, consisting of 12 subcarriers × 14 OFDM symbols in time, for a total of $T = 168$ channel uses. Under such simplifying assumption, the resulting multiplexing gain is equal to $\eta = d(1-\tau_p/T)$ for $d \leq \min\{\tau_p, M\}$. With massive MIMO (i.e., for M sufficiently large) and per-cell processing [16], eventually the number of UL and DL data streams d is limited by the pilot dimension τ_p and by the coherence block length T. In particular, provided that $M > T/2$, by letting $d = \tau_p$ and maximizing η with respect to τ_p, we find that the maximum multiplexing gain is achieved by letting $\tau_p = T/2$ and yields $\eta = T/4$. This depends only on the propagation and mobility characteristics of the physical channel, which determine T_c and W_c and therefore T.

From the above discussion, it follows that the only other way to increase cellular capacity is cell densification, i.e., increasing λ_a. However, also λ_a cannot increase indefinitely [8], at least as far as a classical cellular architecture is adopted. When the cell density increases (and consequently the cell size reduces), several problems emerge:

- The frequency of handovers increases, with the consequent increase of protocol overhead and delay jitter;
- The inter-cell interference increases to unbearable levels, yielding a too low SINR. This is due to the fact that even if each BS reduces its transmit power as a consequence of the smaller cell size, when the cells are very small, the propagation tends to become closer and closer to free-space conditions, for which the pathloss exponent is small. In large cells with tower-mounted BSs the cell boundaries are enforced by vertical tilting of the BS antenna radiation pattern in the elevation direction. In contrast, small-cell BSs are typically not mounted high on the ground and controlling inter-cell interference by tilting the radiation pattern in the elevation direction is more difficult.
- Specifically in MU-MIMO systems, the total number of users in the system is much larger than the UL pilot dimension τ_p. Hence, UL pilots are reused across the network. This pilot reuse yields the so-called *pilot contamination* effect, i.e., a coherently beamformed inter-cell interference which does not vanish even if the number of antennas per BS M grows to infinity [20-21].

Such effect is more and more dominant when the cell size shrinks, while the number of antennas per BS M remains large.

In view of the above problems, the simple increase of the cell density λ_a, while insisting on a conventional cellular architecture with unique UE-to-BS association and per-cell processing, leads to a number of problems such that, eventually, the user rate will degrade and eventually the cellular area capacity reaches a plateau over which any further increase in the cell density yields only an increase in the cost of the infrastructure, without any benefit in area capacity.

1.1 Beyond the Cellular Architecture

After realizing the intrinsic limitations of the cellular architecture with per-BS non-cooperative processing, a flurry of works advocating the *joint processing* of spatially distributed infrastructure antennas has appeared. This idea can be traced back to the work of Wyner [22], and in the communication theoretic and information theoretic literature has been "re-marketed" several times under different names with slightly different nuances, such as *coordinate multipoint* (CoMP) [23-24], *cloud (or centralized) radio access network* (CRAN) [25–27], and, more recently, as *cell-free massive* MIMO [28–30]. An excellent recent review of this vast literature is given in [31].[1]

Advantages of this approach are the mitigation of pathloss and blocking, introducing proximity between the remote radio units (RUs) and UEs and macro-diversity, and the (obvious) elimination of inter-cell interference, by providing a single giant RU cluster. Both points, though, must be carefully discussed. First, deploying a number of RUs much larger than the number of UEs is practically problematic, very costly, and often infeasible especially for outdoor systems. Then, the joint antenna processing across the whole network does not eliminate the problem of a limited UL pilot dimension $\tau_p \ll K$ and therefore the consequent pilot contamination. Finally, global processing and optimization/allocation of pilots and transmit power across the network yield a non-scalable architecture.

Current research has somehow agreed on a middle-point between conventional per-cell processing and centralized processing of all the RUs in the network. In particular, we may imagine a network formed by K UEs, L RUs and D *Decentralized Units* (DUs), where typically $K > L > D$. The number of antennas per RU is denoted by M, and in the massive MIMO regime we have $ML > K$ (the total number of antennas is larger than the number of simultaneously active users). The RUs are connected to the DUs via a routing fronthaul network. Every user k in the system is associated to a cluster $C_k \subseteq [L]=\{1,2,...,L\}$ of surrounding RUs, and each RU l serves a subset of users $U_l \subseteq [K]$, given by the users k such that $l \in C_k$. Clusters are "user-centric", i.e., each user is associated its own cluster, and different users may have different (partially overlapping) clusters. Each cluster C_k is jointly processed at some DU, which has to collect all signals from/to all the RUs $l \in C_k$. The user-centric cluster processors are dynamically allocated to the DUs and are typically run in software over general purpose hardware. As users move through the network, their clusters evolve and "follow" them. Correspondingly, the cluster processors (virtualized network functions) are dynamically allocated to different DUs, in order to achieve load balancing in the routing fronthaul network.

For such a system, we adopt the definition of scalability given by Björnson and Sanguinetti [32], informally recalled as follows: consider a network as described above, with covering area A on the plane and UE, RU, and DU densities λ_{ue}, λ_{ru}, λ_{du}, respectively, such that we have $K = \lambda_{ue}A$, $L = \lambda_{ru}A$ and $D = \lambda_{du}A$. An architecture is said to be scalable if the complexity of the involved signal processing functions

[1] For the sake of precision, it should be said that CRAN is a network architecture, while CoMP is an LTE feature. However, the differences between these terms in the way they have been used in the theoretical literature are blurred. For example, there is no major conceptual difference between CRAN and CoMP with "full joint processing" from a theoretic viewpoint. In both cases, the signal to (resp., from) multiple spatially distributed users sent from (resp., received by) multiple distributed infrastructure antennas is jointly precoded (resp., jointly decoded) at a central processor. Also, there is no conceptual difference between a cell-free system with fully centralized processing of all antenna sites and a single giant cell with distributed antennas implemented via the CRAN architecture. Since this paper illustrates the problem from a theoretical viewpoint, we shall not discuss further the classification of alternative system proposals from a practical implementation or standard specification viewpoint.

and the data rate conveyed at each point in the network converge to some constant values as $A \rightarrow \infty$.

1.2 Related Work

The first works on cell-free massive MIMO consider a system, in which each UE is connected to all RUs [33–35], investigating different uplink (UL) and downlink (DL) methods. The UL with four levels of cooperation among the RUs is studied in [35], and it is shown that the cell-free architecture can outperform conventional cellular massive MIMO and small cells in terms of per user spectral efficiency (SE). In addition, it is shown that minimum mean-square error (MMSE) combining is needed, as maximum ratio combining (MRC) is not able to outperform the compared network types. The presented methods however are not scalable, since all RUs are connected to all UEs. This issue is addressed in [32, 36] by introducing the formation of finite size clusters such that each UE is only connected to a subset of all RUs. A distributed UL implementation with global large-scale fading decoding (LSFD) is presented in [37], where after forming UE and RU clusters, the global symbol estimate of an RU cluster for a specific UE is formed by a weighted sum of the local estimates with the so-called *LSFD weights*. Distributed DL power allocation at the RUs and a combination of cell-centric and user-centric clusters are investigated in [36]. More user-centric approaches are proposed in [38–40]. The effect of the number of UEs served by an RU on the UL and DL rates is investigated in [38] for different channel estimation techniques. The energy efficiency of two distinct RU selection schemes is analyzed and compared to colocated massive MIMO in [39], considering backhaul power consumption, the number of RUs, and the number of antennas per RU. The spectral efficiency and outage probability using stochastic geometry are studied in [40] in a *fog massive MIMO* system, where UEs with coded UL pilots seamlessly migrate from one RU to another with very low RU association latency.

Focusing on the DL, zero-forcing (ZF) beamforming and maximum ratio transmission (MRT) are studied in cell-free systems and compared to a small-cell architecture in [34], where ZF in the cell-free system outperforms the other methods for most UEs. Different local precoding methods extending the "simple" ZF are proposed in [41] and achieve larger rates compared to the simple ZF in the deployed system with independent Rayleigh fading channels. Exact UL-DL duality for the "true" *use-and-then-forget (UatF)* bounds is studied in [32], assuming the knowledge of all required quantities in the SINR expressions also for not associated RU-UE pairs. Additionally, three scalable precoding schemes (partial MMSE, local partial MMSE and MRT) are investigated for two network topologies with the same number of total RU antennas in the system but with a different level of antenna concentration. As mentioned before, a comprehensive overview of this literature is given in the tutorial monograph [31].

In [42], a model called *ideal partial CSI* is proposed, where each RU only has channel information of the associated UEs. This is due to the fact that, in the cluster association process, only the RUs belonging to cluster C_k are aware of the association between user k and its allocated UL pilot signal. The proposed combining method in [42] aims at maximizing the so-called "optimistic ergodic rate", i.e., the achievable rate when the UE receiver knows the useful signal term and the interference plus noise power. This work has been extended in [43] by considering also the DL and by showing that almost symmetric UL and DL rates can be achieved with a duality concept based on partial channel knowledge and "nominal" SINRs, i.e., assumed SINRs based on available channel information. These models and results will be reviewed in details in the following sections.

1.3 Contributions

In this overview work we present in details the model of a user-centric cell-free scalable system based on TDD reciprocity and MU-MIMO (distributed) precoding given in [42-43]. We consider UL combining and DL precoding schemes, based on ideal partial CSI. We also provide an SINR UL/DL duality based on a definition of SINR that depends only on the partial CSI (i.e., what can be effectively measured at each cluster processor: channel estimates are available for associated RU-UE pairs instead of ideal channel knowledge), and demonstrate that such duality yields almost identical per user SE in terms of the actual SINRs and corresponding ergodic rates. Although not being the focus of this paper, different channel subspace and covariance estimation techniques are investigated in various works (see e.g. [44–46]). In this work, we assume that some user channel covariance estimation technique is used such that the dominant channel subspace can be reliably estimated. Based on this knowledge, we consider a simple approach to pilot decontamination based on dominant subspace projection. It will be demonstrated by simulation, that such approach is sufficient to closely approximate the performance under the ideal partial CSI assumption, i.e., after subspace projection, the estimated channels yield a per-user SE which is essentially equal (up to a small degradation) to the one obtained with ideal (but partial) channel knowledge. We also discuss the differences and similarities of the proposed UL combining and DL precoding schemes with respect to the current state of the art.

Finally, we provide a number of interesting points for further research both at the PHY level (signal processing for channel estimation), and at the MAC/resource allocation level (signaling and algorithms for dynamic cluster formation, resources allocation and fairness scheduling).

2 System Model

We consider a cell-free wireless network with L RUs, each equipped with M antennas, and K single-antenna UEs. Both RUs and UEs are distributed on a squared region on the 2-dimensional plane. As a result of the cluster formation process (to be specified later), each UE k is associated with a cluster $C_k \subseteq [L]$ of RUs and each RU l has a set of associated UEs $U_l \subseteq [K]$. The UE-RU association is described by a bipartite graph G with two classes of nodes (UEs and RUs) such that the neighborhood of UE-node k is C_k and the neighborhood of RU-node l is U_l. An example is given in Figure 1. The set of edges of G is denoted by $\mathcal{E}$, i.e., $G = G([L], [K], \mathcal{E})$.

We assume OFDM modulation and assume that the channel in the time-frequency domain follows the standard block-fading model adopted in a very large number of papers (e.g., see [16, 31-32]), where the channel vectors from UEs to RUs are random but constant over coherence blocks of T signal dimensions in the time-frequency domain, which can be identified here as a RB as already illustrated and discussed in Section I.

Since all our treatment can be formulated on a per-RB basis, we shall neglect the RB index for the sake of notation simplicity. We let $\mathbf{H} \in \mathbb{C}^{LM \times K}$ denote the channel matrix between all the K UE antennas and all the LM RUs antennas on a given RB, formed by $M \times 1$ blocks $\mathbf{h}_{l,k}$ in correspondence of the M antennas of RU l and UE k. Because of the UL pilot allocation (see later), each RU l only estimates the channel vectors of the users in U_l.

As a genie-aided best-case, we define the *ideal partial CSI* regime where each RU l has perfect knowledge of the channel vectors $\mathbf{h}_{l,k}$ for $k \in U_l$. In this regime, the part of the channel matrix $\mathbf{H}$ known at the DU serving cluster C_k is denoted by $\mathbf{H}(C_k)$. This matrix has the same dimensions of $\mathbf{H}$, and contains the channel vectors $\mathbf{h}_{l,j}$ in all the (l,j)-th blocks of dimension $M \times 1$ such that $l \in C_k$ and $j \in U_l$, and all-zero blocks of dimension $M \times 1$ in all other cases.

For example, consider the simple case of $L = 2$ and $K = 6$ as in Figure 2. Let's focus on user $k = 3$, for which $C_3 = \{1, 2\}$. We have $U_1 = \{1, 2, 3, 4\}$ and $U_2 = \{3, 4, 5, 6\}$. The complete channel matrix is given by

$$\mathbf{H} = \begin{bmatrix} \mathbf{h}_{1,1} & \mathbf{h}_{1,2} & \mathbf{h}_{1,3} & \mathbf{h}_{1,4} & \mathbf{h}_{1,5} & \mathbf{h}_{1,6} \\ \mathbf{h}_{2,1} & \mathbf{h}_{2,2} & \mathbf{h}_{2,3} & \mathbf{h}_{2,4} & \mathbf{h}_{2,5} & \mathbf{h}_{2,6} \end{bmatrix}.$$

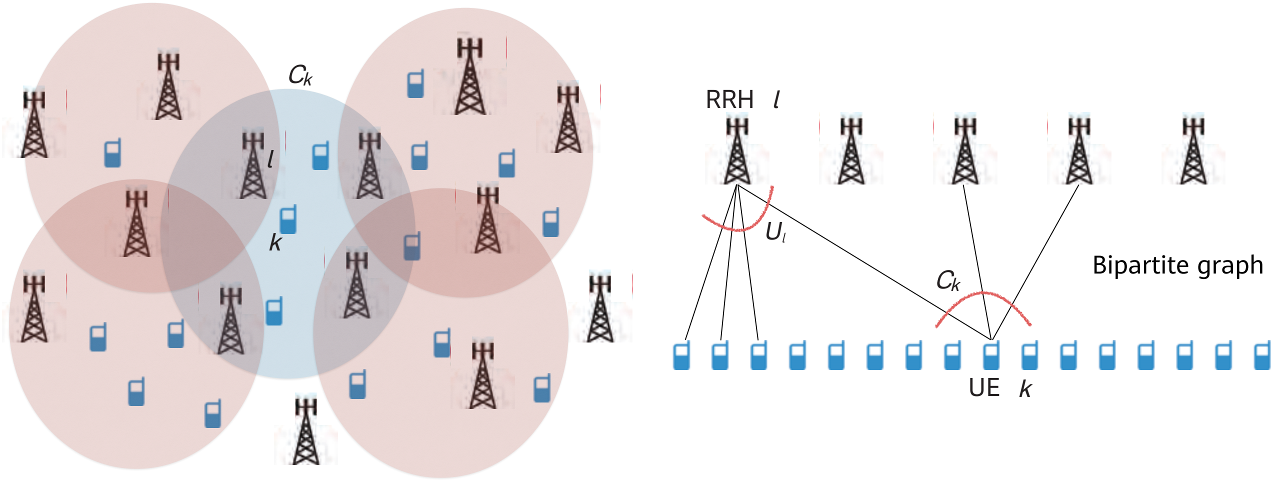


Figure 1 An example of dynamic clusters and the UE-RU association graph. The graph contains a UE-RU edge (k, l) for all $k \in [K]$ and $l \in [L]$ such that $k \in U_l$ and $l \in C_k$.

However, the channel matrix $\mathbf{H}(C_3)$ is given by

$$\mathbf{H}(C_3) = \begin{bmatrix} \mathbf{h}_{1,1} & \mathbf{h}_{1,2} & \mathbf{h}_{1,3} & \mathbf{h}_{1,4} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{h}_{2,3} & \mathbf{h}_{2,4} & \mathbf{h}_{2,5} & \mathbf{h}_{2,6} \end{bmatrix},$$

since users 1 and 2 do not belong to U_2 and users 5 and 6 do not belong to U_1. Hence, the channels of users that are not associated to a given RU cannot be estimated by this RU.

For the individual UE-RU channels, we consider a simplified directional channel model defined as follows. Let $\mathbf{F}$ denote the $M \times M$ unitary DFT matrix with (m, n)-elements $\mathbf{F}_{m,n} = \frac{e^{-j\frac{2\pi}{M}mn}}{\sqrt{M}}$ for $m, n = 0, 1, . . . , M - 1$. Consider the discrete angular support set $S_{l,k} \subseteq \{0, . . . , M - 1\}$. We let the channel $\mathbf{h}_{l,k}$ be a random Gaussian vector in the linear span of the columns of $\mathbf{F}$ indexed by $S_{l,k}$. In particular, we have

$$\mathbf{h}_{l,k} = \sqrt{\frac{\beta_{l,k} M}{|S_{l,k}|}}\mathbf{F}_{l,k}\boldsymbol{\nu}_{l,k}, \qquad (1)$$

where, using a Matlab-like notation, we define $\mathbf{F}_{l,k} \triangleq \mathbf{F}(:,S_{l,k})$. Therefore, $\mathbf{F}_{l,k}$ is a tall unitary matrix of dimensions $M \times |S_{l,k}|$. In particular, for the sake of simplicity we adopt the single ring local scattering model (e.g., see [47]), where $S_{l,k}$ is formed by the integer indices $m \in \{0, 1, ..., M - 1\}$ such that the angle $\frac{2\pi m}{M}$ falls (modulo 2π) in the interval $[\theta - \Delta/2, \theta + \Delta/2]$, and where θ is the angle of the direction between the RU and the UE, and Δ is the scattering ring angular spread. This model captures the directionality of the channel vectors and it is geometric consistent in the sense that two users with the same direction θ with respect to a RU will have channels spanning the same subspace. This directionality aspect is particularly relevant for high frequencies (mmWave) where propagation is dominated by the Line-of-Sight path. The coefficient $\beta_{l,k}$ in (1) represents the large scale fading coefficient (LSFC) including distance-dependent pathloss, blocking effects, and shadowing. The vector $\boldsymbol{\nu}_{l,k}$ in (1) is an $|S_{l,k}| \times 1$ i.i.d. Gaussian vector with components $\sim\mathcal{CN}(0, 1)$. It follows that $\mathbf{h}_{l,k}$ is a Gaussian zero-mean random vector with covariance matrix

$$\Sigma_{l,k} = \frac{\beta_{l,k} M}{|S_{l,k}|}\mathbf{F}_{l,k}\mathbf{F}^{\mathrm{H}}_{l,k}. \qquad (2)$$

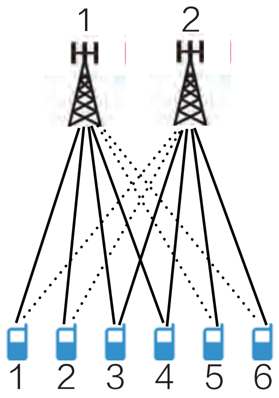


Figure 2 A simple network with L = 2 RUs and K = 6 users used as an example. The dotted edges correspond to channels vectors that cannot be estimated because of the cluster formation mechanism.

2.1 Cluster Formation

We assume that τ_p signal dimensions per RB are dedicated to UL pilots (see [13]), and define a codebook of τ_p orthogonal pilot sequences. The UEs transmit with the same power[1] P^{ue}, and we define the system parameter SNR $\triangleq P^{\mathrm{ue}}/N_0$, where N_0 denotes the noise power spectral density. By the normalization of the channel vectors, the maximum beamforming gain averaged over the small scale fading can be written as the expectation $\mathbf{E}[||\frac{M}{|S_{l,k}|}\mathbf{F}_{l,k}\boldsymbol{\nu}_{l,k}||^2] = M$. Therefore, the maximum SNR at the receiver of RU l from UE k is $\beta_{l,k}M$ SNR. As in [32], we consider that each UE k selects its leading RU l as the RU with the largest channel gain $\beta_{l,k}$ (assumed known) among the RUs with yet a free pilot and satisfying the received SNR condition $\beta_{l,k} \geq \frac{\eta}{M\,\mathrm{SNR}}$, where $\eta > 0$ is a suitable threshold. If such RU is not available, then the UE is declared in outage. In our simulations, the UE to leader RU association is performed in a greedy manner starting from some UE at random. In practice, users join and leave the system according to some user activity dynamics, and each new UE joining the system is admitted if it can find a leader RU according to the above conditions. After all non-outage UEs k are assigned to their leader RU $l = l(k)$

[1] It is customary in communication theory to call "power" the average energy per complex symbol. We follow here this convention. Notice that the physical power for a symbol rate of W symbols/s is given by $W\,P^{\mathrm{ue}}$ and the physical noise power (integral of the noise power spectral density over the system bandwidth of W Hz) is given by $W\,N_0$. Hence, the ratio P^{ue}/N_0 coincides with the ratio of the physical transmit signal power over the noise power at the receiver output.

and therefore have a pilot index $t = t(k) \in [\tau_p]$, the dynamic cluster C_k for each UE k is formed by enrolling successively all RUs l listed in order of decreasing LSFC for which i) pilot $t(k)$ is yet free, ii) the condition $\beta_{l,k} \geq \frac{\eta}{M\,\mathrm{SNR}}$ is satisfied. We also consider a maximum cluster size Q such that if more than Q RUs meet the cluster enrollment condition, only the Q with the largest LSFC are selected. As a result, all UEs $k \in U_l$ make use of mutually orthogonal UL pilots. Furthermore, $0 \leq |U_l| \leq \tau_p$ and $0 \leq |C_k| \leq Q$.

2.2 Uplink Data Transmission

The received $LM \times 1$ symbol vector at the LM RUs' antennas for a single channel use of the UL is given by

$$\mathbf{y}^{\mathrm{ul}} = \sqrt{\mathrm{SNR}}\,\mathbf{H}\mathbf{s}^{\mathrm{ul}} + \mathbf{z}^{\mathrm{ul}}\,, \tag{3}$$

where $\mathbf{s}^{\mathrm{ul}} \in \mathbf{C}^{K \times 1}$ is the vector of information symbols transmitted by the UEs (zero-mean unit variance and mutually independent random variables) and $\mathbf{z}^{\mathrm{ul}}$ is an i.i.d. noise vector with components $\sim CN\ (0, 1)$. The goal of cluster C_k is to produce an effective channel observation for symbol s_k^{ul} (the k-th component of the vector $\mathbf{s}^{\mathrm{ul}}$ from the collectively received signal at the RUs $l \in C_k$). We define the receiver unit norm vector $\mathbf{v}_k \in \mathbf{C}^{LM \times 1}$ formed by $M \times 1$ blocks $\mathbf{v}_{l,k}$: $l = 1,\ldots,L$, such that $\mathbf{v}_{l,k} = \mathbf{0}$ (the identically zero vector) if $l \notin \mathrm{C}_K$. This reflects the fact that only the RUs in C_k are involved in producing a received observation for the detection of user k. The non-zero blocks $\mathbf{v}_{l,k}$: $l \in C_k$ contain the receiver combining vectors. The corresponding scalar combined observation for symbol s_k^{ul} is given by

$$\begin{aligned} r_k^{\mathrm{ul}} &= \mathbf{v}_k^{\mathrm{H}}\mathbf{y}^{\mathrm{ul}} \\ &= \sqrt{\mathrm{SNR}}\,\mathbf{v}_k^{\mathrm{H}}\mathbf{h}_k s_k^{\mathrm{ul}} + \sqrt{\mathrm{SNR}}\,\mathbf{v}_k^{\mathrm{H}}\mathbf{H}_{K\text{-}k}\mathbf{s}_{K\text{-}k}^{\mathrm{ul}} + \mathbf{v}_k^{\mathrm{H}}\mathbf{z}^{\mathrm{ul}}\,, \end{aligned} \tag{4}$$

where $\mathbf{h}_k$ denotes the k-th column of $\mathbf{H}$, $\mathbf{H}_{K\text{-}k}$ is obtained by deleting the k-th column from $\mathbf{H}$, and $\mathbf{s}_{K\text{-}k}^{\mathrm{ul}}$ is the vector $\mathbf{s}^{\mathrm{ul}}$ after deletion of the k-th element.

For simplicity, we assume that the channel decoder has perfect knowledge of the exact UL SINR value

$$\mathrm{SINR}_k^{\mathrm{ul}} = \frac{|\mathbf{v}_k^{\mathrm{H}}\mathbf{h}_k|^2}{\mathrm{SNR}^{-1} + \sum_{j \neq k}|\mathbf{v}_k^{\mathrm{H}}\mathbf{h}_j|^2}\,, \tag{5}$$

where $\mathbf{h}_k$ denotes the k-th column of $\mathbf{H}$. The corresponding UL *optimistic ergodic* achievable rate is given by

$$R_k^{\mathrm{ul}} = \mathbf{E}[\log(1 + \mathrm{SINR}_k^{\mathrm{ul}})]\,, \tag{6}$$

where the expectation is with respect to the small scale fading, while conditioning on the placement of UEs and RU, and on the cluster formation.

Remark 1: The optimistic ergodic rate is achievable when the small-scale fading is a stationary ergodic process in the time-frequency domain and a codeword is transmitted over a sufficiently large number of small-scale fading states, and somehow the useful signal coefficient and the variance of the interference term in (4) are known to the decoder. These, however, are *sufficient conditions* under which the achievability proof of (6) follows easily. In the massive MIMO literature, it is common to consider the UatF lower bound on (6), which contains only the long-term statistics (first and second moments) of the coefficients in $\mathbf{v}_k^{\mathrm{H}}\mathbf{h}_j$ in (5) and therefore is achievable under less restrictive conditions [31, 48]. It has been shown that unless the useful signal coefficient "hardens" such as $|\mathbf{E}[\mathbf{v}_k^{\mathrm{H}}\mathbf{h}_j]|^2$ is large with respect to $\mathrm{Var}(\mathbf{v}_k^{\mathrm{H}}\mathbf{h}_k)$, the UatF bound can be very pessimistic [49]. This is unfortunately the case for typical layouts of cell-free user-centric networks where the channel hardening of very large co-located arrays and rich i.i.d. small-scale fading do not occur. In addition, it cannot be excluded that with some *universal decoding scheme* [50] the rate in (6) can be achieved or at least closely approached. For these reasons, we believe that the optimistic ergodic rate reflects more accurately the actual achievable performance of the system at hand, than the overly conservative UatF lower bound.

2.3 Downlink Data Transmission

The signal corresponding to one channel use of the DL at the receiver of UE k is given by

$$y_k^{\mathrm{dl}} = \mathbf{h}_k^{\mathrm{H}}\mathbf{x} + z_k^{\mathrm{dl}}\,, \tag{7}$$

where the transmitted vector $\mathbf{x} \in \mathbf{C}^{LM \times 1}$ is formed by all the signal samples sent collectively from the RUs. Without loss of generality we can incorporate a common factor $\mathrm{SNR}^{-1/2}$ in the LSFCs, which is equivalent to rescaling the noise at the UEs receivers such that $z_k^{\mathrm{dl}} \sim CN(0, \mathrm{SNR}^{-1})$, while keeping $\beta_{l,k}$ for all (l, k) identical to the UL case. Let $\mathbf{s}^{\mathrm{dl}} \in \mathbf{C}^{K \times 1}$ denote the vector of information bearing symbols for the K users, assumed to be zero mean, independent, with variance $q_k \geq 0$. Under a general linear precoding scheme, we have

$$\mathbf{x} = \mathbf{U}\mathbf{s}^{\mathrm{dl}}\,, \tag{8}$$

where $\mathbf{U} \in \mathbf{C}^{LM \times K}$ is the overall precoding matrix, formed by

$M \times 1$ blocks $\mathbf{u}_{l,k}$ such that $\mathbf{u}_{l,k} = \mathbf{0}$ if $l \notin C_K$. The non-zero blocks $\mathbf{u}_{l,k}: \ l \in C_k$ contain the precoding vectors. Using (8) in (7), we have

$$y_k^{\mathrm{dl}} = \mathbf{h}_k^{\mathrm{H}} \mathbf{u}_k s_k^{\mathrm{dl}} + \sum_{j \neq k} \mathbf{h}_k^{\mathrm{H}} \mathbf{u}_j s_j^{\mathrm{dl}} + z_k^{\mathrm{dl}}, \tag{9}$$

where $\mathbf{u}_k$ is the k-th column of $\mathbf{U}$. The resulting DL (optimistic) SINR, is given by

$$\mathrm{SINR}_k^{\mathrm{dl}} = \frac{|\mathbf{h}_k^{\mathrm{H}} \mathbf{u}_k|^2 q_k}{\mathrm{SNR}^{-1} + \sum_{j \neq k} |\mathbf{h}_k^{\mathrm{H}} \mathbf{u}_j|^2 q_j}, \tag{10}$$

where q_k is the DL transmit (Tx) power for the data stream to UE k. As for the UL, we consider the DL optimistic ergodic rate given by

$$R_k^{\mathrm{dl}} = \mathbf{E}[\log(1 + \mathrm{SINR}_k^{\mathrm{dl}})]. \tag{11}$$

Assuming that the columns of the precoder $\mathbf{U}$ have unit norm, we have that the total DL Tx power collectively transmitted by the RUs is given by

$$P_{\mathrm{tot}}^{\mathrm{dl}} = \mathrm{tr}(\mathbf{E}[\mathbf{x}\mathbf{x}^{\mathrm{H}}]) = \mathrm{tr}(\mathbf{U}\mathrm{diag}(q_k : k \in [K])\mathbf{U}^{\mathrm{H}}) \tag{12}$$

$$= \mathrm{tr}(\mathbf{U}^{\mathrm{H}}\mathbf{U}\mathrm{diag}(q_k : k \in [K])) = \sum_{k=1}^{K} q_k. \tag{13}$$

We assume the UL and DL total transmit power to be balanced. This imposes the condition $\sum_{k=1}^{K} q_k = K$.

3 Uplink Receive Schemes

We illustrate the UL linear receive schemes for the case of ideal partial CSI. In our simulations, we shall use the CSI estimated as described in Section 6 and simply plug the estimated channel vectors in place of the ideally known channel vectors.

In this work, we consider two UL receive schemes referred to as cluster-level zero-forcing (CLZF) and local linear MMSE (LMMSE) with cluster-level combining [42]. The schemes are described in the following.

3.1 Cluster-Level Zero-Forcing (CLZF)

For a given UE k with cluster C_k, we define the set $U(C_k) \triangleq \cup_{l \in C_k} U_l$ of UEs served by at least one RU in C_k. Let $\mathbf{h}_k(C_k)$ denote the k-th column of $\mathbf{H}(C_k)$ and let $\mathbf{H}_k(C_k)$ denote the residual matrix after the k-th column is deleted. The CLZF receive vector is obtained as follows. Let $\bar{\mathbf{h}}_k(C_k) \in \mathbf{C}^{|C_k|M \times 1}$ and $\bar{\mathbf{H}}_k(C_k) \in \mathbf{C}^{|C_k|M \times (|U(C_k)|-1)}$ denote the vector and matrix obtained from $\mathbf{h}_k(C_k)$ and $\mathbf{H}_k(C_k)$, respectively, after all the M blocks of rows corresponding to RUs $l \notin C_k$ and all the all-zero columns corresponding to UEs $k' \notin U(C_k)$ are removed. Consider the singular value decomposition (SVD)

$$\bar{\mathbf{H}}_k(C_k) = \bar{\mathbf{A}}_k \bar{\mathbf{S}}_k \bar{\mathbf{B}}_k^{\mathrm{H}}, \tag{14}$$

where the columns of the tall unitary matrix $\bar{\mathbf{A}}_k$ form an orthonormal basis for the column span of $\bar{\mathbf{H}}_k(C_k)$, such that the orthogonal projector onto the orthogonal complement of the interference subspace is given by $\bar{\mathbf{P}}_k = \mathbf{I} - \bar{\mathbf{A}}_k \bar{\mathbf{A}}_k^{\mathrm{H}}$, and define the unit-norm vector

$$\bar{\mathbf{v}}_k = \bar{\mathbf{P}}_k \bar{\mathbf{h}}_k(C_k) / ||\bar{\mathbf{P}}_k \bar{\mathbf{h}}_k(C_k)||. \tag{15}$$

Hence, the CLZF receive vector $\mathbf{v}_k$ is given by expanding $\bar{\mathbf{v}}_k$ by reintroducing the missing blocks of all-zero $M \times 1$ vectors $\mathbf{0}$ in correspondence of the RUs $l \notin C_k$.

Because of the channel correlation model, it may happen that there are some UEs $j \in U(C_k), j \neq k$, such that the k-th and j-th columns of $\mathbf{H}(C_k)$ are co-linear. In this case, column j is extracted from the channel matrix and it is not included in the CLZF computation. In practice, this means that the interference on user k caused by user j is taken in directly, without mitigation by linear projection.

3.2 Local LMMSE with Cluster-Level Combining

In this case, each RU l makes use of locally computed receive vectors $\mathbf{v}_{l,k}$ for its users $k \in U_l$. Let $\mathbf{y}_l^{\mathrm{ul}}$ denote the $M \times 1$ block of $\mathbf{y}^{\mathrm{ul}}$ corresponding to RU l. For each $k \in U_l$, RU l computes locally

$$r_{l,k}^{\mathrm{ul}} = \mathbf{v}_{l,k}^{\mathrm{H}} \mathbf{y}_l^{\mathrm{ul}}. \tag{16}$$

The symbols $\{r_{l,k}^{\mathrm{ul}} : k \in U_l\}$ are sent to the DU serving cluster C_k, which computes the cluster-level combined symbol

$$r_k^{\mathrm{ul}} = \sum_{l \in C_k} w_{l,k}^{*} r_{l,k}^{\mathrm{ul}} = \mathbf{w}_k^{\mathrm{H}} \mathbf{r}_k^{\mathrm{ul}}, \tag{17}$$

where $w_{l,k}$ is the combining coefficient of RU l for UE k, and $\mathbf{w}_k$ and $\mathbf{r}_k$ are vectors formed by stacking $w_{l,k}$ and $r_{l,k}^{\mathrm{ul}}$ of all RUs $l \in C_k$, respectively.

A classical and effective choice for the receive vector $\mathbf{v}_{l,k}$ is based on LMMSE estimation. In this case, we distinguish between the known part of the interference, i.e., the term $\sum_{j \in U_l : j \neq k} \mathbf{h}_{l,j} s_j^{\mathrm{ul}}$, and the unknown part of the interference, i.e., the term $\sum_{j \notin U_l} \mathbf{h}_{l,j} s_j^{\text{ul}}$ in $\mathbf{y}_l^{\text{ul}}$. The receiver treats the unknown part of the interference plus noise as a white vector with known variance per component. The covariance matrix of the term $\sum_{j \notin U_l} \mathbf{h}_{l,j} s_j^{\text{ul}}$ is given by

$$\begin{aligned} \Xi_l &= \mathbf{E}\left[\left(\sqrt{\text{SNR}} \sum_{j \notin U_l} \mathbf{h}_{l,j} s_j^{\text{ul}} + \mathbf{z}_l^{\text{ul}}\right)\left(\sqrt{\text{SNR}} \sum_{j \notin U_l} \mathbf{h}_{l,j} s_j^{\text{ul}} + \mathbf{z}_l^{\text{ul}}\right)^{\text{H}}\right] \\ &= \mathbf{I} + \sum_{j \notin U_l} \frac{\beta_{l,j} M \text{SNR}}{|S_{l,j}|} \mathbf{F}_{l,j} \mathbf{F}_{l,j}^{\text{H}} , \end{aligned} \tag{18}$$

where $\mathbf{z}_l^{\text{ul}}$ with i.i.d components $\sim CN(0,1)$ is additive white Gaussian noise (AWGN) at RU l. Taking the trace and dividing it by M, we find the equivalent variance per component

$$\sigma_l^2 = \frac{1}{M} \text{tr}(\Xi_l) = 1 + \text{SNR}\left(\sum_{j \neq U_l} \beta_{l,j}\right). \tag{19}$$

Under this assumption, we have that the LMMSE receive vector is given by

$$\mathbf{v}_{l,k} = \left(\sigma_l^2 \mathbf{I} + \text{SNR} \sum_{j \in U_l} \mathbf{h}_{l,j} \mathbf{h}_{l,j}^{\text{H}}\right)^{-1} \mathbf{h}_{l,k}. \tag{20}$$

For the cluster-level combining coefficients, we consider the maximization of the SINR after combining. The effective received signal model at RU $l \in C_k$ relative to UE k can be written as

$$r_{l,k}^{\text{ul}} = \sqrt{\text{SNR}}\left(g_{l,k,k} s_k^{\text{ul}} + \sum_{j \in U_l : j \neq k} g_{l,k,j} s_j^{\text{ul}}\right) + \mathbf{v}_{l,k}^{\text{H}} \xi_l \tag{21}$$

where we define $g_{l,k,j} = \mathbf{v}_{l,k}^{\text{H}} \mathbf{h}_{l,j}$ and let ξ_l the unknown interference plus noise vector, assumed $\sim CN(0, \sigma_l^2 \mathbf{I})$. Stacking $\{r_{l,k}^{\text{ul}} : l \in C_k\}$ as a $|C_k| \times 1$ column vector r_k^{ul}, we can write the output symbols of cluster C_k relative to UE k as

$$\mathbf{r}_k^{\text{ul}} = \sqrt{\text{SNR}}\,(\mathbf{a}_k s_k^{\text{ul}} + \mathbf{G}_k \mathbf{s}_k^{\text{ul}}) + \zeta_k , \tag{22}$$

where $\zeta_k = \{\mathbf{v}_{l,k}^{\text{H}} \xi_l : l \in C_k\}$ has the covariance matrix given by $\mathbf{D}_k = \text{diag}\{\sigma_l^2 ||\mathbf{v}_{l,k}||^2 : l \in C_k\}$ and $\mathbf{a}_k = \{g_{l,k,k} : l \in C_k\}$.

The matrix $\mathbf{G}_k \in \mathbf{C}^{|C_k| \times (|U(C_k)|-1)}$ contains elements $g_{l,k,j}$ in position corresponding to RU l and UE j (after suitable index reordering) if $j \in U_l$, and zero elsewhere. The vector $\mathbf{s}_k^{\text{ul}} \in \mathbf{C}^{(|U(C_k)|-1) \times 1}$contains the symbols of all users $j \in U(C_k) : j \neq k$. Then, given the available CSI, the total interference plus noise covariance matrix is

$$\mathbf{\Gamma}_k = \mathbf{D}_k + \text{SNR}\,\mathbf{G}_k \mathbf{G}_k^{\text{H}}. \tag{23}$$

The corresponding nominal SINR for user k with combining (22) is given by

$$\text{SINR}_k^{\text{cl}} = \frac{\text{SNR}\,\mathbf{w}_k^{\text{H}} \mathbf{a}_k \mathbf{a}_k^{\text{H}} \mathbf{w}_k}{\mathbf{w}_k^{\text{H}} \mathbf{\Gamma}_k \mathbf{w}_k} . \tag{24}$$

The maximization of (24) with respect to $\mathbf{w}_k$ amounts to finding the maximum generalized eigenvalue of the matrix pencil $(\mathbf{a}_k \mathbf{a}_k^{\text{H}}, \mathbf{\Gamma}_k)$. Since the matrix $\mathbf{a}_k \mathbf{a}_k^{\text{H}}$ has rank 1 and therefore it has only one non-zero eigenvalue, the solution is readily given by

$$\mathbf{w}_k = \mathbf{\Gamma}_k^{-1} \mathbf{a}_k , \tag{25}$$

yielding the SINR

$$\text{SINR}_k^{\text{cl}} = \text{SNR}\,\mathbf{a}_k^{\text{H}} \mathbf{\Gamma}_k^{-1} \mathbf{a}_k . \tag{26}$$

For the LMMSE with cluster-level combining scheme, the overall receive vector is obtained by forming the vector $\bar{\mathbf{v}}_k$ by stacking the $|C_k|$ blocks of dimensions $M \times 1$ given by $w_{l,k} \mathbf{v}_{l,k}$ on top of each other and normalizing them such that $\bar{\mathbf{v}}_k$ has unit norm. After expanding $\bar{\mathbf{v}}_k$ to $\mathbf{v}_k$ of dimension $LM \times 1$ by inserting the all-zero blocks corresponding to the RUs $l \notin C_k$, the resulting SINR is again given by (5) and does not coincide in general with (26) since it takes into account the true interference from all users, and not only the known part due to the partial CSI. Nevertheless, the SINR in (24) and its maximization leading to (26) represent the best guess given the available CSI and the statistical information on the overall additional interference plus noise power.

It is interesting to notice that the scheme differs from the distributed large-scale fading decoding (LSFD) in [31], as the LSFD relies on the expected products of the combining and channel vectors. In contrast, we use the instantaneous channel realization estimate and the instantaneous receive vector for the computation of the combining coefficients. We shall provide a more detailed comment on the difference between the two schemes in Section 5.

4 UL-DL Duality and Downlink Precoding Schemes

We shall consider expressions of nominal, i.e., assumed, SINR, where the channels $\mathbf{h}_{l,k}$ with $(l,k) \in \mathcal{E}$ are known, while for $(l,k) \notin \mathcal{E}$, only the LSFCs of these channels are known, and the channel vector spatial distribution is assumed isotropic. This treatment differs from what done in [32] where it is assumed that, if RU l serves UE j but not UE k, $\mathbf{E}[|\mathbf{h}_k^{\mathrm{H}}\mathbf{v}_j|^2]$ can still be estimated accurately at RU l, where $\mathbf{h}_{l,k}^{\mathrm{H}}\mathbf{v}_{l,j}$ is a non-zero component of $\mathbf{h}_k^{\mathrm{H}}\mathbf{v}_j$.

4.1 UL-DL SINR Duality for Partially Known Channels

Consider the UL SINR given in (5). By construction we have that $\mathbf{v}_{l,k} = \mathbf{0}$ for all $l \notin C_k$. Therefore, the term at the numerator $\theta_{k,k} = |\mathbf{v}_k^{\mathrm{H}}\mathbf{h}_k|^2 = |\sum_{l\in c_k}\mathbf{v}_{l,k}^{\mathrm{H}}\mathbf{h}_{l,k}|^2$ contains only known channels and it can thus be used in the nominal SINR expression. The terms at the denominator take on the form

$$\theta_{j,k} = |\mathbf{v}_k^{\mathrm{H}}\mathbf{h}_j|^2 = \left|\sum_{l\in c_k}\mathbf{v}_{l,k}^{\mathrm{H}}\mathbf{h}_{l,j}\right|^2 \tag{27}$$

$$= \left|\sum_{l\in c_k\cap c_j}\mathbf{v}_{l,k}^{\mathrm{H}}\mathbf{h}_{l,j} + \sum_{l\in c_k\backslash c_j}\mathbf{v}_{l,k}^{\mathrm{H}}\mathbf{h}_{l,j}\right|^2 \tag{28}$$

where for $l \in C_k \cap C_j$ the channel $\mathrm{h}_{l,j}$ is known, while for $l \in C_k\backslash C_j$ the channel $\mathrm{h}_{l,j}$ is not known. Taking the conditional expectation of the term in (28) given all the known CSI $\mathcal{E}_k$ at C_k, and noticing that the channels for different (l,j) pairs are independent and have mean zero, we find

$$\mathbf{E}[\theta_{j,k}|\mathcal{E}_k] = \left|\sum_{l\in c_k\cap c_j}\mathbf{v}_{l,k}^{\mathrm{H}}\mathbf{h}_{l,j}\right|^2 + \sum_{l\in c_k\backslash c_j}\frac{\beta_{l,j}M}{|S_{l,j}|}\,\mathrm{v}_{l,k}^{\mathrm{H}}\mathbf{F}_{l,j}\mathbf{F}_{l,j}^{\mathrm{H}}\mathrm{v}_{l,k}\,.$$

Finally, under the isotropic assumption, we replace the actual covariance matrix of the unknown channels with a scaled identity matrix with the same trace, i.e., we have

$$\mathbf{E}[\theta_{j,k}|\mathcal{E}_k] \approx \left|\sum_{l\in c_k\cap c_j}\mathbf{v}_{l,k}^{\mathrm{H}}\mathbf{h}_{l,j}\right|^2 + \sum_{l\in c_k\backslash c_j}\beta_{l,j}||\mathbf{v}_{l,k}||^2\,.$$

Using the fact that $\mathbf{v}_k$ is a unit-norm vector and assuming that the M × 1 blocks $\mathrm{v}_{l,k}$ have the same norm, we can further approximate $||\mathbf{v}_{l,k}||^2 \approx \frac{1}{|C_k|}$. Therefore, the resulting nominal UL SINR is given by

$$\begin{aligned}\mathrm{SINR}_k^{\text{ul-nom}} &= \frac{|\mathbf{v}_k^{\mathrm{H}}\mathbf{h}_k|^2}{\mathrm{SNR}^{-1} + \sum_{j\neq k}\left(\left|\sum_{l\in C_k\cap C_j}\mathbf{v}_{l,k}^{\mathrm{H}}\mathbf{h}_{l,j}\right|^2 + \frac{1}{|C_k|}\sum_{l\in C_k\backslash C_j}\beta_{l,j}\right)}\\ &= \frac{\theta_{k,k}}{\mathrm{SNR}^{-1} + \sum_{j\neq k}\tilde{\theta}_{j,k}}\,,\end{aligned} \tag{29}$$

where we define $\theta_{k,k} = |\mathbf{v}_k^{\mathrm{H}}\mathbf{h}_k|^2$ and for $j\neq k$ as

$$\tilde{\theta}_{j,k} = \left|\sum_{l\in c_k\cap c_j}\mathbf{v}_{l,k}^{\mathrm{H}}\mathbf{h}_{l,j}\right|^2 + \frac{1}{|C_k|}\sum_{l\in c_k\backslash c_j}\beta_{l,j}\,. \tag{30}$$

Notice that the term at the denominator of the nominal UL SINR given by $\frac{1}{|C_k|}\sum_{j\neq k}\sum_{l\in c_k\backslash c_j}\beta_{l,j}$ is the contribution of the interference power per RU caused by other UEs $j\neq k$ to the RUs in C_k, but not also in C_j. We refer to (29) as the nominal UL SINR since this is the SINR that the processor of cluster C_k can estimate from its CSI knowledge.

Next, we consider the DL SINR given in (10) under the assumption that the DL precoding vectors are identical to the UL receive vectors, i.e., $\mathbf{u}_k = \mathbf{v}_k$ for all $k \in [K]$. The numerator takes on the form $\theta_{k,k}q_k$ where $\theta_{k,k}$ is the same as before and contains all known channels. Focusing on the terms at the denominator and taking into account that the vectors $\mathbf{u}_j$ are non-zero only for the M × 1 blocks corresponding to RUs $l \in C_j$, we have

$$\theta_{k,j} = |\mathbf{h}_k^{\mathrm{H}}\mathbf{u}_j|^2 = \left|\sum_{l\in c_j}\mathbf{h}_{l,k}^{\mathrm{H}}\mathbf{u}_{l,j}\right|^2 \tag{31}$$

$$= \left|\sum_{l\in c_j\cap c_k}\mathbf{h}_{l,k}^{\mathrm{H}}\mathbf{u}_{l,j} + \sum_{l\in c_j\backslash c_k}\mathbf{h}_{l,k}^{\mathrm{H}}\mathbf{u}_{l,j}\right|^2, \tag{32}$$

where, as before, for $l \in C_j \cap C_k$ the channel $\mathrm{h}_{l,k}$ is known, while for $l \in C_j\backslash C_k$ the channel $\mathrm{h}_{l,k}$ is not known. Taking the conditional expectation of the term in (32) given all the known CSI, we find

$$\mathbf{E}[\theta_{k,j}|\mathcal{E}_k] = \left|\sum_{l\in c_j\cap c_k}\mathbf{h}_{l,k}^{\mathrm{H}}\mathbf{u}_{l,j}\right|^2 + \sum_{l\in c_j\backslash c_k}\frac{\beta_{l,k}M}{|s_{l,k}|}\mathbf{u}_{l,j}^{\mathrm{H}}\mathbf{F}_{l,k}\mathbf{F}_{l,k}^{\mathrm{H}}\mathbf{u}_{l,j}.$$

Using again the isotropic assumption, we replace the actual covariance matrix of the unknown channels with a scaled identity matrix with the same trace, i.e., we have

$$\begin{aligned}\mathbf{E}[\theta_{k,j}|\mathcal{E}_k] &\approx \left|\sum_{l\in c_j\cap c_k}\mathbf{h}_{l,k}^{\mathrm{H}}\mathbf{u}_{l,j}\right|^2 + \sum_{l\in c_j\backslash c_k}\beta_{l,k}||\mathbf{u}_{l,j}||^2\\ &\approx \left|\sum_{l\in c_j\cap c_k}\mathbf{h}_{l,k}^{\mathrm{H}}\mathbf{u}_{l,j}\right|^2 + \frac{1}{|C_j|}\sum_{l\in c_j\backslash c_k}\beta_{l,k}\,.\end{aligned}$$

Therefore, the resulting nominal DL SINR is given by

$$\begin{aligned}\mathrm{SINR}_k^{\text{dl-nom}} &= \frac{|\mathbf{h}_k^{\mathrm{H}}\mathbf{u}_k|^2 q_k}{\mathrm{SNR}^{-1} + \sum_{j\neq k}\left(\left|\sum_{l\in C_j\cap C_k}\mathbf{h}_{l,k}^{\mathrm{H}}\mathbf{u}_{l,j}\right|^2 + \frac{1}{|C_j|}\sum_{l\in C_j\backslash C_k}\beta_{l,k}\right)q_j}\\ &= \frac{\theta_{k,k}q_k}{\mathrm{SNR}^{-1} + \sum_{j\neq k}\tilde{\theta}_{k,j}\,q_j}\,,\end{aligned} \tag{33}$$

where

$$\tilde{\theta}_{k,j} = \left|\sum_{l\in c_j\cap c_k}\mathbf{h}_{l,k}^{\mathrm{H}}\mathbf{u}_{l,j}\right|^2 + \frac{1}{|C_j|}\sum_{l\in c_j\backslash c_k}\beta_{l,k}\,. \tag{34}$$

Given the symmetry of the coefficients, an UL-DL duality exists for the nominal SINRs. This can be used to calculate the DL Tx power allocation $\{q_k : k\in[K]\}$ that achieves nominal DL SINRs equal to the nominal UL SINRs with uniform UL Tx power per UE.

Remark 2: We focus here on the case of balanced UL-DL SINRs for the following reasons: 1) from a theoretical viewpoint, it is easier and more direct to illustrate the SINR duality in this case, and then generalize it to unbalanced (but proportional) SINRs as done later on in Section 4.2; 2) from a practical viewpoint, in a cell-free architecture the cost, size, and scale manufacturing of RUs play a very important role for these systems to be attractive and effectively deployed. Hence, it is likely to assume that each "antenna element" (including physical antenna, low-noise amplifier (LNA)/ power amplifier (PA) up-down conversion and A/D and D/A conversion) of a multi-antenna RU has (roughly) the same characteristics in a UE. In fact, there are already proposals to reuse UE chipsets to create scalable and economically viable RUs. Therefore, the balanced total transmit power when the total number of users and the total number of RU antenna elements is similar is a meaningful working assumption; 3) achieving balanced UL-DL SINRs has nothing to do with the fact that, typically, the traffic load in the DL is much larger than in the UL. In fact, given balanced UL-DL SINRs, the different traffic load can be matched by scheduling, i.e., allocating a different number of transmission resources to the UL and DL, respectively. The allocation of transmission resources (i.e., time-frequency RBs) to UL and DL is a much more effective way than "power allocation" to match the traffic load requests, since it acts on the prelog factor of the time-averaged rate (throughput) rather than on the term inside the logarithm; 4) working with balanced SINRs has the non-trivial advantage that only the UL receive vectors must be calculated for each new CSI pilot round, while the DL precoding vectors are automatically obtained as a byproduct.

In particular, we choose the target DL SINRs $\{\gamma_k\} \triangleq \{\mathrm{SINR}_k^{\text{ul-nom}}\}$ for all k. The system of (non-linear) equations in the power allocation vector $\mathbf{q} = \{q_k\}$ is given by

$$\mathrm{SINR}_k^{\text{ul-nom}} = \gamma_k, \forall k = 1\ldots,K$$

which can be rewritten in the more convenient linear form (see [10])

$$(\mathbf{I} - \mathrm{diag}(\boldsymbol{\mu})\boldsymbol{\Theta})\mathbf{q} = \frac{1}{\mathrm{SNR}}\boldsymbol{\mu}\ , \qquad (35)$$

by defining the vector $\boldsymbol{\mu}$ with elements

$$\mu_k = \frac{\gamma_k}{(1+\gamma_k)\theta_{k,k}} \qquad (36)$$

and the matrix $\boldsymbol{\Theta}$ with (k,j) elements $\theta_{k,k}$ on the diagonal and $\tilde{\theta}_{k,j}$ in the off-diagonal positions. Since the target (nominal) SINRs $\{\gamma_k\}$ are achievable, it can be shown that the above system of equations has a non-negative solution, given by

$$\mathbf{q}^{*} = \frac{1}{\mathrm{SNR}}(\mathbf{I} - \mathrm{diag}(\boldsymbol{\mu})\boldsymbol{\Theta})^{-1}\boldsymbol{\mu}\,. \qquad (37)$$

In addition, it is immediate to show that this solution satisfies the total power constraint $\sum_{k=1}^{K} q_k^{*} = K$, i.e., the UL and DL have balanced total power. This is shown explicitly as follows. Since in the UL case the transmit symbol energies are all equal to 1, for the choice of target SINRs $\gamma_k = \mathrm{SINR}_k^{\text{ul-nom}}$, the following equation must hold:

$$\mathbf{1} = \frac{1}{\mathrm{SNR}}(\mathbf{I} - \mathrm{diag}(\boldsymbol{\mu})\boldsymbol{\Theta}^{\mathrm{T}})^{-1}\boldsymbol{\mu}$$

Hence, it follows that

$$\begin{aligned} K &= \mathbf{1}^{\mathrm{T}}\mathbf{1} \\ &= \frac{1}{\mathrm{SNR}}\mathbf{1}^{\mathrm{T}}\,(\mathrm{I} - \mathrm{diag}(\boldsymbol{\mu})\boldsymbol{\Theta}^{\mathrm{T}})^{-1}\boldsymbol{\mu} \\ &= \frac{1}{\mathrm{SNR}}\mathbf{1}^{\mathrm{T}}\,(\mathrm{diag}(1/\mu_1,\ldots,1/\mu_K) - \boldsymbol{\Theta}^{\mathrm{T}})^{-1}\mathbf{1} \\ &= \frac{1}{\mathrm{SNR}}\mathbf{1}^{\mathrm{T}}\,(\mathrm{diag}(1/\mu_1,\ldots,1/\mu_K) - \boldsymbol{\Theta})^{-1}\mathbf{1} \\ &= \frac{1}{\mathrm{SNR}}\mathbf{1}^{\mathrm{T}}\,(\mathbf{I} - \mathrm{diag}(\boldsymbol{\mu})\boldsymbol{\Theta})^{-1}\boldsymbol{\mu} \\ &= \sum_{k=1}^{K} q_k^{*}\,. \end{aligned} \qquad (38)$$

In conclusion, we propose to use existing UL receivers/ combiners as DL precoders. This has the advantage of achieving balanced UL and DL (nominal) SINRs as well as total transmit power, and most importantly, that no additional DL precoding computation is required. In particular, in this work we consider CLZF precoding and LMMSE with cluster-level combining precoding.

4.2 The Case of Unbalanced UL and DL

First of all, we should notice that (in line with the observation made in Remark 2), the user rates expressed by (6) have little to do with the actual user throughputs, defined as the long-term averaged rate over a sequence of RBs. In general, the total number of users in the system K_{tot} may be much larger than the number of users K simultaneously active on a single RB, which is the quantity considered here. On top of the PHY described in this paper, there are several other mechanisms such that active users do not need to transmit/receive data on each RB. Considering a long sequence of slots (e.g., formed by RBs) indexed by a slot time s, the throughput of user $k \in [K_{\text{tot}}]$ is given by

$$\overline{R}_k = \lim_{S \to \infty} \frac{1}{S} \sum_{s=1}^{S} R_k[s] \tag{39}$$

where $R_k[s]$ coincides with (6) on the slots s where user k is scheduled, and it is equal to 0 on the slots s where user k is not scheduled. Under mild conditions of stationarity and ergodicity of the scheduling policy and of the fading processes, the throughput in (39) converges to the average quantity

$$\overline{R}_k = \alpha_k \mathbf{E}[R_k] \tag{40}$$

where α_k is the activity fraction of user k, i.e., the fraction of slots where user k is scheduled (active), and $\mathbf{E}[R_k]$ is the rate in (6), further averaged over the active user subset of K - 1 simultaneously scheduled users out of the K_{tot}, active together with user k.

Hence, a way to match the different UL and DL traffic demands consists of scheduling a different number of UL and DL slots per TDD frame. Also, a way to achieve throughput fairness consists of using a fairness-oriented scheduling policy, e.g., following the general theory of network utility function maximization (e.g., see [51]), as applied for example in [52] in the MU-MIMO case. The extension of this type of fairness scheduling approaches to cell-free user-centric networks in a scalable way, i.e., without the need of a fully centralized scheduler, is generally a very interesting problem for further research.

However, it may happen that for some technology reason the total transmit power in the DL is significantly different from the total transmit power in the UL. In this case, it does not make sense to reduce artificially the transmit power of the strongest one to match the weakest. Hence, in the rest of this section we provide the (simple) extension of the UL-DL duality provided before, to the unbalanced total transmit power case.

Let P^{ru} denote the average RU transmit power such that the total transmit power in the DL is LP^{ru}. We consider a virtual UL with per-UE transmit power given by $(L/K)P^{\text{ru}}$ and the corresponding virtual UL SNR parameter $\text{SNR}^{\text{ul}} = \frac{LP^{\text{ru}}}{K\,N_0}$. For such virtual UL with receivers $\{\mathbf{v}_k\}$, the nominal SINR is obtained by replacing SNR with SNR^{ul} in (29). Notice that the vectors $\{\mathbf{v}_k\}$ may or may not correspond to the actual receive vectors for the actual UL. In particular, for the CLZF scheme, the receive vectors are independent of SNR, therefore they are the same for the actual and virtual UL. In contrast, with the LMMSE scheme, the receive vectors depend on SNR and therefore they need to be recomputed for the virtual UL. In any case, using $\mathbf{u}_k = \mathbf{v}_k$ for all $k \in [K]$ for the (actual) DL, and obtaining the DL normalized transmit power according to (37) with the substitution of SNR by SNR^{ul}, we obtain a set of DL transmit power factors of the K DL streams such that 1) the same nominal SINRs of the virtual UL are achieved, and 2) the total average DL transmit power is LP^{ru} as desired.

Of course, in general a combined approach to unbalanced UL and DL traffic can be implemented, where we use the virtual UL approach to take into account the different total transmit available power, and yet use scheduling to adjust the throughput to the individual user demands and fairness.

5 Comparison with Other Proposed UL and DL Schemes

Several UL receive/combining and DL precoding schemes have been proposed in the literature. Here we review a few.

For the UL, the local LMMSE processing in combination with LSFD proposed in [35] and [37] is similar to our local LMMSE with cluster-level combining introduced before. The difference is in the computation of the weighting coefficients, that with LSFD are dependent on the expected products of the combining and channel vectors $\{\mathbf{E}[g_{l,k,j}] : (l,k) \in \mathcal{E}\}$, and $\mathbf{E}[\mathbf{G}_k \mathbf{G}_k^{\text{H}}]$. In contrast, the scheme in this paper uses the instantaneous channel realization for the computation of the combining coefficients. The received signal model at RU l for UE k is given again by (16). The local estimates are sent to the DU processing cluster $\mathcal{C}_k$ and combined according to (17). However, in LSFD the weights are computed in a different way, i.e.,

$$\mathbf{w}_k = (\mathbf{\Gamma}_k^{\text{LSFD}})^{-1} \mathbf{E}[\mathbf{a}_k] \,, \tag{41}$$

where

$$\mathbf{\Gamma}_k^{\text{LSFD}} = \mathbf{D}_k + \text{SNR} \sum_{j \neq k \,:\, j \in \mathcal{U}_l} \mathbf{E}[\mathbf{G}_k(:,j)\mathbf{G}_k(:,j)^{\text{H}}] \,. \tag{42}$$

Recall that the matrix $\mathbf{G}_k$ of dimension $|\mathcal{C}_k| \times (|\mathcal{U}(\mathcal{C}_k)| - 1)$ contains elements $g_{l,k,j}$ in position corresponding to RU l and

UE j if $(l,j) \in \mathcal{E}$, and zero elsewhere, so we have

$$\mathbf{E}[\mathbf{G}_k(:,j)\mathbf{G}_k(:,j)^{\mathrm{H}}]_{m,n} = \begin{cases} \mathbf{E}[\mathbf{v}_{m,k}^{\mathrm{H}}\mathbf{h}_{m,j}(\mathbf{v}_{n,k}^{\mathrm{H}}\mathbf{h}_{n,j})^{\mathrm{H}}], \text{if } (m,k) \in \varepsilon, (n,k) \in \varepsilon, j \in U(C_k) \setminus k \\ 0, \quad \text{otherwise,} \end{cases} \tag{43}$$

and

$$\mathbf{E}[\mathbf{a}_k] = \{\mathbf{E}[g_{l,k,k}] : l \in C_k\}. \tag{44}$$

Instead of the instantaneous channel and combining vectors taking into account small-scale fading as in (25), the expectation (based on large-scale fading) is used. The choice of these combining coefficients is motivated in [35] by maximization of the SINR term resulting from the UatF bound.

As in [35] all RU-UE pairs are associated, the original LSFD scheme for cell-free systems assumes that the elements $\mathbf{E}[\mathbf{G}_k(:,j)\mathbf{G}_k(:,j)^{\mathrm{H}}]_{m,n}$ are known for $m,n \in [L]$ and all $k,j \in [K]$, i.e., the matrix $\mathbf{E}[\mathbf{G}_k(:,j)\mathbf{G}_k(:,j)^{\mathrm{H}}]$ would have dimension $L \times L$ and have only non-zero entries. The scalable LSFD scheme proposed in [37] in combination with dynamic cooperation clustering assumes that the elements $\mathbf{E}[\mathbf{G}_k(:,j)\mathbf{G}_k(:,j)^{\mathrm{H}}]_{m,n}$ are known for $m,n \in C_k$ and $k,j \in U(C_k)$. After removing the elements belonging to RU $l' \notin C_k$ (equal to zero), the matrix $\mathbf{E}[\mathbf{G}_k(:,j)\mathbf{G}_k(:,j)^{\mathrm{H}}]$ is of dimension $|C_k| \times |C_k|$ and has only non-zero entries.

We notice that the LSFD scheme requires knowledge of the expected values $\mathbf{E}[\mathbf{v}_{m,k}^{\mathrm{H}}\mathbf{h}_{m,j}(\mathbf{v}_{n,k}^{\mathrm{H}}\mathbf{h}_{n,j})^{\mathrm{H}}]$ for all $m,n \in [L]$, $k,j \in [K]$, and all $m,n \in C_k$, $k,j \in U(C_k)$, respectively. We argue that while the instantaneous values of these coefficients are easily obtained from the partial CSI available at the cluster processors, these average values require some sort of long-term averaging and thus additional implementation efforts. Therefore, we believe that the scheme proposed in [42] is not only more performant, but also easier to implement in practice.

For the DL, a popular scheme is local partial ZF (LPZF) proposed in [41]. In this scheme, each RU computes the precoding vectors and power allocation locally for its associated UEs. Let us consider the channel matrix between RU l and the associated UEs in U_l, given by

$$\mathbf{H}_l = [\mathbf{h}_{l,k_1}\, \mathbf{h}_{l,k_2} \ldots \mathbf{h}_{l,k_{|U_l|}}] \in \mathbf{C}^{M \times |U_l|}, \tag{45}$$

where $k_1,\ldots,k_{|U_l|}$ are the UEs in the set U_l. In case $M \geq \tau_p \geq |U_l|$ and $\mathbf{H}_l$ is a full-rank matrix, local ZF (LZF) is carried out by computing the pseudo-inverse

$$\mathbf{H}_l^{+} = \mathbf{H}_l(\mathbf{H}_l^{\mathrm{H}}\mathbf{H}_l)^{-1} \tag{46}$$

of $\mathbf{H}_l$. Then, the LZF precoding vector $\mathrm{u}_{l,k}$ is the normalized column of $\mathbf{H}_l^{+}$ corresponding to user $k \in U_l$.

LPZF is a variant of LZF where some users are excluded from the calculation of the pseudo-inverse. In particular, when $M \leq |U_l|$, or that $\mathbf{H}_l$ is rank-deficient ($\tau_p \geq |U_l|$ due to clustering), the RU chooses from U_l the UEs U_l^{ZF} with the largest channel gains (at most M) whose channels are linearly independent, and thus form a full-rank matrix $\mathbf{H}_l^{\mathrm{ZF}}$. The precoding vectors of UEs in U_l^{ZF} are computed by ZF as in (46) and normalized to unit norm. For the remaining UEs $k \in U_l^{\mathrm{MRT}}$, normalized MRT is employed, i.e.,

$$\mathbf{u}_{l,k} = \frac{\mathbf{h}_{l,k}}{||\mathbf{h}_{l,k}||}, \forall k \in U_l^{\mathrm{MRT}}, \tag{47}$$

where $U_l^{\mathrm{ZF}} \cap U_l^{\mathrm{MRT}} = \emptyset$ and $U_l^{\mathrm{ZF}} \cup U_l^{\mathrm{MRT}} = U_l$.

For both cases, the RUs compute the Tx power for each UE locally. In conjunction with ZLF and LPZF two simple schemes are considered: equal power allocation (EPA), where the power allocated to stream k by RU $l \in C_k$ is the same for all streams, i.e.,

$$q_{l,k} = \frac{P^{\mathrm{RU}}}{|U_l|}, \forall k \in U_l, \tag{48}$$

and proportional power allocation (PPA) with regard to the LSFCs such that

$$q_{l,k} = P^{\mathrm{RU}}\frac{\beta_{l,k}}{\sum_{j \in U_l}\beta_{l,j}}, \forall k \in U_l, \tag{49}$$

where $q_{l,k}$ and P^{RU} denote the transmit power allocated at RU l to UE k and the DL power budget at each RU, respectively. Obviously, in all cases, for $k \notin U_l$ we have $q_{l,k} = 0$.

6 CSI Estimation from UL Pilots

In practice, ideal partial CSI is not available and the channels $\{\mathbf{h}_{l,k} : (l,k) \in \mathcal{E}\}$ must be estimated from UL pilots. Thanks to channel reciprocity in the TDD mode, the estimates can be used for both UL combining and DL precoding. We assume that τ_p signal dimensions per RB are dedicated to UL pilots (see [13]), and define a codebook of τ_p orthogonal pilot sequences. The pilot field received at RU l is given by the $M \times \tau_p$ matrix

$$\mathbf{Y}_l^{\mathrm{pilot}} = \sum_{i=1}^{k}\mathbf{h}_{l,i}\phi_{t_i}^{\mathrm{H}} + \mathbf{Z}_l^{\mathrm{pilot}} \tag{50}$$

where ϕ_{t_i} denotes the pilot vector of dimension τ_p used by UE i at the current slot. Since the pilot vectors make use of τ_p symbols, their energy is τ_pSNR, i.e., $||\phi_{t_i}||^2 = \tau_p\,$SNR for all

$t_i \in [\tau_p]$. For all UEs $k \in U_l$, RU l produces the "pilot matching" channel estimates

$$\widehat{\mathbf{h}}^{\text{pm}}_{l,k} = \frac{1}{\tau_p \text{SNR}} \mathbf{Y}^{\text{pilot}}_l \phi_{t_k} \tag{51}$$

$$= \mathbf{h}_{l,k} + \sum_{i \neq k : t_i = t_k} \mathbf{h}_{l,i} + \tilde{\mathbf{z}}_{t_k,l} \tag{52}$$

where $\tilde{\mathbf{z}}_{t_k,l}$ is $M \times 1$ Gaussian i.i.d. with components $\mathcal{CN}(0, \frac{1}{\tau_p \text{SNR}})$. Assuming that the subspace information $\mathbf{F}_{l,k}$ of all $k \in U_l$ is known, we consider also the "subspace projection" (SP) pilot decontamination scheme for which the projected channel estimate is given by the orthogonal projection of $\widehat{\mathbf{h}}^{\text{pm}}_{l,k}$ onto the subspace spanned by the columns of $\mathbf{F}_{l,k}$, i.e.,

$$\widehat{\mathbf{h}}^{\text{SP}}_{l,k} = \mathbf{F}_{l,k}\mathbf{F}^{\text{H}}_{l,k}\widehat{\mathbf{h}}^{\text{pm}}_{l,k} \tag{53}$$

$$= \mathbf{h}_{l,k} + \mathbf{F}_{l,k}\mathbf{F}^{\text{H}}_{l,k}\left(\sum_{i \neq k : t_i = t_k} \mathbf{h}_{l,i}\right) + \mathbf{F}_{l,k}\mathbf{F}^{\text{H}}_{l,k} + \tilde{\mathbf{z}}_{t_k,l} \tag{54}$$

Notice that after the projection, the resulting estimation noise is correlated since it is contained in the channel subspace, in fact the covariance matrix of the projected noise is given by $\frac{1}{\tau_p \text{SNR}}\mathbf{F}_{l,k}\mathbf{F}^{\text{H}}_{l,k}$.

Writing explicitly the pilot contamination term after the subspace projection, we have

$$\mathbf{F}_{l,k}\mathbf{F}^{\text{H}}_{l,k}\left(\sum_{i \neq k : t_i = t_k} \mathbf{h}_{l,i}\right) = \sum_{i \neq k : t_i = t_k} \mathbf{F}_{l,k}\mathbf{F}^{\text{H}}_{l,k}\mathbf{h}_{l,i} \tag{55}$$

$$= \sum_{i \neq k : t_i = t_k} \sqrt{\frac{\beta_{l,i} M}{|S_{l,i}|}} \mathbf{F}_{l,k}\mathbf{F}^{\text{H}}_{l,k}\mathbf{F}_{l,i}\boldsymbol{\nu}_{l,i}\,. \tag{56}$$

This is an $M \times 1$ Gaussian vector with mean zero and covariance matrix

$$\Sigma^{\text{co}}_{l,k} = \sum_{i \neq k : t_i = t_k} \frac{\beta_{l,i} M}{|S_{l,i}|} \mathbf{F}_{l,k}\mathbf{F}^{\text{H}}_{l,k}\mathbf{F}_{l,i}\mathbf{F}^{\text{H}}_{l,i}\mathbf{F}_{l,k}\mathbf{F}^{\text{H}}_{l,k}\,.$$

When $\mathbf{F}_{l,k}$ and $\mathbf{F}_{l,i}$ are nearly mutually orthogonal, i.e., $\mathbf{F}^{\text{H}}_{l,k}\mathbf{F}_{l,i} \approx \mathbf{0}$, the subspace projection is able to significantly reduce the pilot contamination effect.

Note that the previously described UL and DL schemes, and the computation for UL-DL duality with ideal partial CSI can be carried out with channel estimates by replacing the ideal partial CSI $\{\mathbf{h}_{l,k} : (l,k) \in \mathcal{E}\}$ with the channel estimates $\{\widehat{\mathbf{h}}^{\text{pm}}_{l,k} : (l,k) \in \mathcal{E}\}$ or $\{\widehat{\mathbf{h}}^{\text{sp}}_{l,k} : (l,k) \in \mathcal{E}\}$. In practical systems, knowledge of the subspace $\mathbf{F}_{l,k}$ for $(l,k) \in \mathcal{E}$ however is not directly available and requires non-trivial estimation.

How to efficiently estimate the user channel covariance or at least their dominant signal subspace without suffering from the same pilot contamination that appears in the channel estimate themselves is a very relevant and interesting open problem. Notice also that most papers and monographs on the subject (e.g., see [31] and references therein) assume that all channel statistics (in particular, the channel covariance matrices) are "magically" known everywhere. This assumption is extremely unrealistic since although the channel statistics change quite slowly in time with respect to the small-scale fading coherence time, they are generally time-varying and must be continuously tracked in order to enable schemes such as the subspace projection or some forms of MMSE pilot decontamination as proposed in [31] to work properly.

7 Numerical Results

In our simulations, we consider a square coverage area of $A = 225 \times 225$ square meters with a torus topology to avoid boundary effects. The LSFCs are given according to the 3GPP urban microcell pathloss model from [53]. We assume a noise power of -96 dBm, and the UL power P^{ue} is chosen such that $\bar{\beta}M\text{SNR} = 1$ (i.e., 0 dB), when the expected pathloss $\bar{\beta}$ (averaged with respect to the LOS/NLOS probability and shadowing) is calculated for distance $3d_L$, where $d_L = 2\sqrt{\frac{A}{\pi L}}$ is the diameter of a disk of area equal to A/L. We consider RBs of $T = 200$ symbols. The UL (same for DL) spectral efficiency (SE) for UE k is given by

$$\text{SE}^{\text{ul}}_k = (1 - \tau_p/T)R^{\text{ul}}_k. \tag{57}$$

The angular support $S_{l,k}$ contains the DFT quantized angles (multiples of $2\pi/M$) falling inside an interval of length $\triangle$ placed symmetrically around the direction joining UE k and RU l. We use $\triangle = \pi/8$ and the maximum cluster size $Q = 10$ (RUs serving one UE) in the simulations. The SNR threshold $\eta = 1$ makes sure that an RU-UE association can only be established when $\beta_{l,k} \geq \frac{\eta}{M\text{SNR}}$.

For each set of parameters, we generated 50 independent layouts (random uniform placement of RUs and UEs), and for each layout we computed the optimistic ergodic rate by Monte Carlo averaging with respect to the channel vectors. In all figures, the results with subspace projection channel estimates are shown. We compare the DL schemes CLZF and LMMSE with the current state-of-the-art benchmarks represented by the LZF and LPZF schemes of [41] (see Section 5).

We start our evaluation by comparing the sum SE for different K and τ_p in a system with a total of $LM = 640$ antennas, with different arrangements of $L = 10$, $L = 20$ and $L = 40$ RUs (respectively, in Figures 3, 4, and 5). We notice that in most cases LMMSE outperforms LZF, and that for more UEs in the system, we need larger τ_p to maximize the sum SE, as each RU can serve up to τ_p UEs. We notice also that the SE for the case of $K = 200$ users, at the optimal value of τ_p, is quite invariant with respect to the antenna distribution, and in fact the more concentrated antenna distribution ($L = 10$ RUs with $M = 64$ antennas each) yields slightly larger maximum sum SE. In any case, this means that the system is quite flexible (within a reasonable range of parameters) with respect to the antenna distribution, and that a realistic number of RUs with a relatively large number of antennas each represents a practically attractive option with respect to the classical cell-free paradigm, where the number of RUs is predicated to be larger than the number of users.

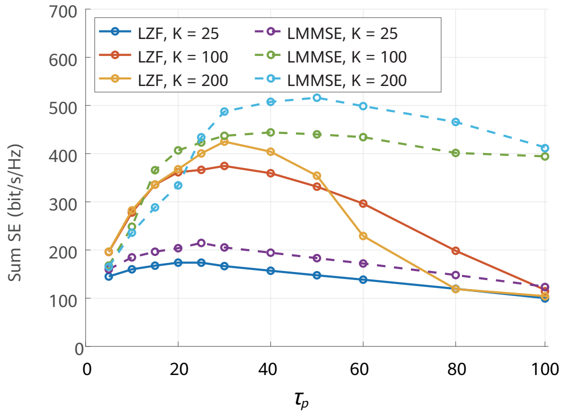


Figure 3 Sum DL SE vs. τ_p for different number of users K for $L = 10$ RUs and $M = 64$ antennas each. LMMSE uses power allocation from duality, while LZF uses PPA.

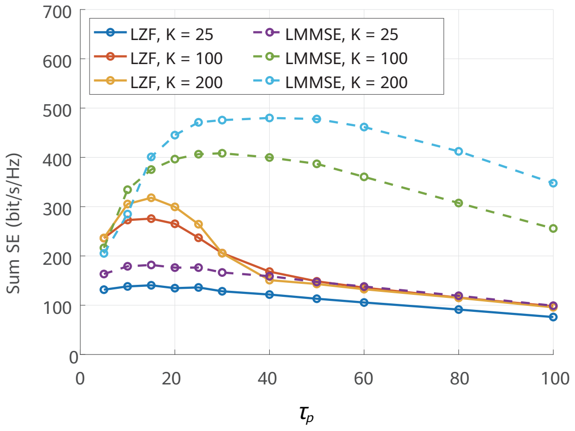


Figure 4 Sum DL SE vs. τ_p for different number of users K for $L = 20$ RUs and $M = 32$ antennas each. LMMSE uses power allocation from duality, while LZF uses PPA.

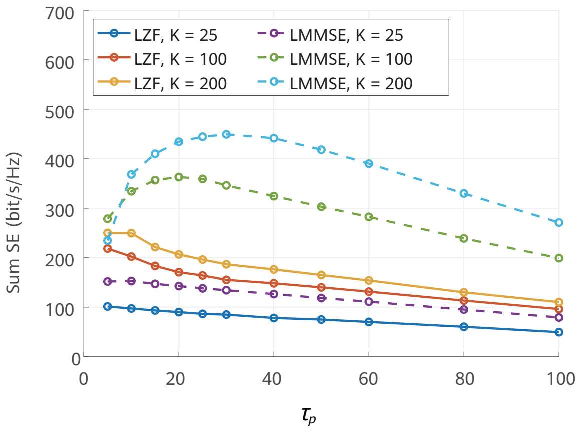


Figure 5 Sum DL SE vs. τ_p for different number of users K for $L = 40$ RUs and $M = 16$ antennas each. LMMSE uses power allocation from duality, while LZF uses PPA.

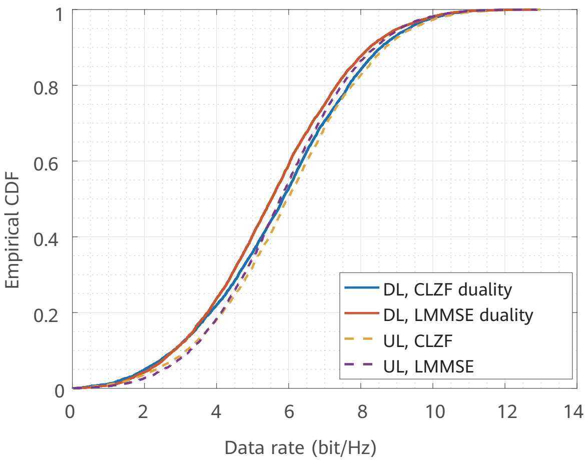


Figure 6 Empirical CDF of the UL and DL per-user data rate (in bit/channel use) where $L = 10$, $M = 64$, and $\tau_p = 40$

We now look at the distribution of the DL rates per UE for the case $K = 100$ users. Figure 6 shows that the proposed UL-DL duality method yields almost symmetric effective ergodic rates for the UL and DL. Also, this figure shows that the CLZF and LMMSE methods perform very similarly. Therefore, since the local LMMSE with cluster-level combining is significantly less computationally intensive, it is definitely our preferred and recommended choice.

8 Conclusion

In this paper we have reviewed the basic model for a scalable cell-free user-centric wireless network architecture, where each UE is served by a cluster of RUs selected dynamically in the vicinity. In our model, RUs have multiple antennas and are capable of local processing. Further processing can be performed at DU nodes at the cluster level. In general, cluster-level decoders are virtualized and are implemented as software-defined network functions, dynamically allocated to DUs in order to maintain a balanced computation load. The scalability of the architecture stems from the fact that, provided that the densities of UEs, RUs, and DUs are constant, the data rate and computation load at any point of the network remains finite, while the network coverage area grows to infinity. In a typical practical operational regime, we expect such networks to have DU, RU and UE densities ordered as $\lambda_{du} < \lambda_{ru} < \lambda_{ue}$, although the number of antennas should be larger than the number of users, i.e., $\lambda_{ue} < M\lambda_{ru}$ where M is the number of antennas per RU.

Following the recent results in [42–43], we presented two types of UL linear processing. The cluster-level ZF at the cluster for user *k* computes a receive vector that sets to zero the interference of all users $j \neq k$ whose channel vectors can be partially estimated by some RU forming the cluster. As an alternative, LMMSE can be applied separately for each RU, producing a local MMSE estimate for the useful symbol of user *k*. These estimates are then combined by the cluster processor, in order to maximize the SINR of the channel "after local MMSE estimation", as seen collectively from all RUs forming the cluster. This second option is much less computationally expensive, and our numerical results show that it provides excellent performance.

We have also shown that under the partial CSI knowledge available at each given cluster processor, there is a notion of "nominal" SINR for which UL-DL duality holds. This motivates the use of the UL receive vectors also as (dual) DL precoding vectors. We have verified numerically that, despite the "nominal SINR" does not correspond exactly to the SINR appearing in the ergodic rate expressions, the rates achieved in UL and DL with this approach are virtually identical. The use of UL receive vectors as DL precoders has also the non-trivial advantage of dramatically reducing the computation load, since only a set of vectors needs to be computed during each new CSI estimation, for both UL and DL.

Finally, we have proposed a simple subspace projection method to (partially) decontaminate the UL pilots, that exploits the fact that with high probability co-pilot users generate channel vectors with different and nearly mutually orthogonal dominant subspaces. In fact, by properly assigning the UL pilots to the users, it is unlikely that two users sharing the same pilot (and therefore, by construction, being served by two disjoint clusters of RUs) are received both at high signal level at a given RU and under the same (or very close) angle of arrival. In [42], it is shown that the pilot projection method is very effective in closely approaching the system performance under the ideal partial CSI assumption, that provides a performance upper bound under the assumptions of cluster-level processing.

Several open problems remain to be investigated. In particular, we mention here the design of an effective scheme for estimating the user channel dominant subspace (to implement the pilot projection scheme), and the careful consideration of the mechanisms of UE-RU association and cluster formation. In particular, in such type of network architectures, a scheme that allows the UEs to detect the presence of the surrounding RUs and a random access procedure to establish the association with at least the cluster leader must be devised. In this respect, an interesting option for possible further investigation consists of letting the UEs send "on-demand" broadcast signals to be identified by the RUs. For example, this may happen at the first time a UE joins the system, and may be repeated if the UE detects that its signal quality is below some minimum service threshold, e.g., due to the fact that the user-centric dynamic cluster is not able to evolve rapidly enough to follow the UE mobility. A UE-triggered association mechanism with embedded pilot collision detection is presented for example in [6]. The scheme in [6] is used at each slot, for a very low-latency UL-DL cycle, but it does not lead to high SE, and a similar scheme could be used only for the initial association and re-association phase. Also, the RUs could apply some form of activity detection and user identification, along the lines of massive random access, as for example treated in [54]. A complementary option (to be considered as an alternative or together with a suitable random access mechanism) consists of operating the cell-free user-centric network not in a stand-alone mode, but as a data rate extension - carrier aggregation option to an existing cellular network operating in a different frequency band. In this case, the coordination operations such as association and cluster formation could be implemented via a control channel handled by the cellular network. These system aspects are left as very promising topics for further investigation.

References

[1] A. Goldsmith, "Wireless communications," Cambridge University Press, 2005.

[2] T. L. Marzetta, E. G. Larsson, H. Yang, and H. Q. Ngo, "Fundamentals of massive MIMO," Cambridge University Press, 2016.

[3] F. Baccelli and B. Błaszczyszyn, "Stochastic geometry and wireless networks," Now Publishers Inc, 2009, vol. 1.

[4] M. Haenggi, "Stochastic geometry for wireless networks," Cambridge University Press, 2012.

[5] Z. Chen, L. Qiu, and X. Liang, "Area spectral efficiency analysis and energy consumption minimization in multiantenna poisson distributed networks," *IEEE Transactions on Wireless Communications*, vol. 15, no. 7, pp. 4862-4874, 2016.

[6] O. Y. Bursalioglu, G. Caire, R. K. Mungara, H. C. Papadopoulos, and C. Wang, "Fog massive MIMO: A user-centric seamless hot-spot architecture," *IEEE Transactions on Wireless Communications*, vol. 18, no. 1, pp. 559-574, 2018.

[7] M. Bayat, R. K. Mungara, and G. Caire, "Achieving spatial scalability for coded caching via coded multipoint multicasting," *IEEE Transactions on Wireless Communications*, vol. 18, no. 1, pp. 227-240, 2018.

[8] J. G. Andrews, X. Zhang, G. D. Durgin, and A. K. Gupta, "Are we approaching the fundamental limits of wireless network densification?" *IEEE Communications Magazine*, vol. 54, no. 10, pp. 184-190, 2016.

[9] G. Caire and S. Shamai, "On the achievable throughput of a multiantenna Gaussian broadcast channel," *IEEE Trans. on Inform. Theory*, vol. 49, no. 7, pp. 1691-1706, 2003.

[10] P. Viswanath and D. N. C. Tse, "Sum capacity of the vector Gaussian broadcast channel and uplink-downlink duality," *IEEE Trans. on Inform. Theory*, vol. 49, no. 8, pp. 1912-1921, Aug. 2003.

[11] H. Weingarten, Y. Steinberg, and S. Shamai (Shitz), "The capacity region of the Gaussian multiple-input multiple-output broadcast channel," *IEEE Trans. on Inform. Theory*, vol. 52, no. 9, pp. 3936-3964, Sept. 2006.

[12] G. Caire, N. Jindal, M. Kobayashi, and N. Ravindran, "Multiuser MIMO achievable rates with downlink training and channel state feedback," *IEEE Trans. on Inform. Theory*, vol. 56, no. 6, pp. 2845-2866, June 2010.

[13] 3GPP, "Physical channels and modulation (Release 16)," 3GPP Technical Specification 38.211, 12 2020, Version 16.4.0.

[14] E. Khorov, A. Kiryanov, A. Lyakhov, and G. Bianchi, "A tutorial on IEEE 802.11ax high efficiency WLANs," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 1, pp. 197-216, 2018.

[15] Q. Qu, B. Li, M. Yang, Z. Yan, A. Yang, D.-J. Deng, and K.-C. Chen, "Survey and performance evaluation of the upcoming next generation WLANs standard - IEEE 802.11ax," *Mobile Networks and Applications*, vol. 24, no. 5, pp. 1461-1474, 2019.

[16] T. L. Marzetta, "Non-cooperative cellular wireless with unlimited numbers of base station antennas," *IEEE Trans. on Wireless Comm.*, vol. 9, no. 11, pp. 3590-3600, 2010.

[17] A. Benzin and G. Caire, "Internal self-calibration methods for large scale array transceiver software-defined radios," in *WSA 2017; 21th International ITG Workshop on Smart Antennas*. VDE, 2017, pp. 1-8.

[18] R. Rogalin, O. Y. Bursalioglu, H. Papadopoulos, G. Caire, A. F. Molisch, A. Michaloliakos, V. Balan, and K. Psounis, "Scalable synchronization and reciprocity calibration for distributed multiuser MIMO," *IEEE Transactions on Wireless Communications*, vol. 13, no. 4, pp. 1815-1831, 2014.

[19] K. T. Truong and R. W. Heath, "Effects of channel aging in massive MIMO systems," *Journal of Communications and Networks*, vol. 15, no. 4, pp. 338-351, 2013.

[20] H. Huh, G. Caire, H. C. Papadopoulos, and S. A. Ramprashad, "Achieving 'massive MIMO' spectral efficiency with a not-so-large number of antennas," *IEEE Transactions on Wireless Communications*, vol. 11, no. 9, pp. 3226-3239, 2012.

[21] J. Hoydis, S. Ten Brink, and M. Debbah, "Massive MIMO in the UL/DL of cellular networks: How many antennas do we need?" *IEEE J. on Sel. Areas on Commun. (JSAC)*, vol. 31, no. 2, pp. 160-171, 2013.

[22] A. D. Wyner, "Shannon-theoretic approach to a Gaussian cellular multiple-access channel," *IEEE Trans. on Inform. Theory*, vol. 40, no. 6, pp. 1713-1727, 1994.

[23] V. Jungnickel, K. Manolakis, W. Zirwas, B. Panzner, V. Braun, M. Lossow, M. Sternad, R. Apelfröjd, and T. Svensson, "The role of small cells, coordinated multipoint, and massive MIMO in 5G," *IEEE Communications Magazine*, vol. 52, no. 5, pp. 44-51, 2014.

[24] H. Huh, A. M. Tulino, and G. Caire, "Network MIMO with linear zero-forcing beamforming: Large system analysis, impact of channel estimation, and reduced-complexity scheduling," *IEEE Transactions on Information Theory*, vol. 58, no. 5, pp. 2911-2934, 2011.

[25] A. Checko, H. L. Christiansen, Y. Yan, L. Scolari, G. Kardaras, M. S. Berger, and L. Dittmann, "Cloud RAN for mobile networks - a technology overview," *IEEE Communications Surveys & Tutorials*, vol. 17, no. 1, pp. 405-426, 2014.

[26] S.-H. Park, O. Simeone, O. Sahin, and S. Shamai, "Joint precoding and multivariate backhaul compression for the downlink of cloud radio access networks," *IEEE Transactions on Signal Processing*, vol. 61, no. 22, pp. 5646-5658, 2013.

[27] I. E. Aguerri, A. Zaidi, G. Caire, and S. S. Shitz, "On the capacity of cloud radio access networks with oblivious relaying," *IEEE Transactions on Information Theory*, vol. 65, no. 7, pp. 4575-4596, 2019.

[28] H. Q. Ngo, A. Ashikhmin, H. Yang, E. G. Larsson, and T. L. Marzetta, "Cell-free massive MIMO versus small cells," *IEEE Transactions on Wireless Communications*, vol. 16, no. 3, pp. 1834-1850, 2017.

[29] J. Zhang, S. Chen, Y. Lin, J. Zheng, B. Ai, and L. Hanzo, "Cell-free massive MIMO: A new next-generation paradigm," *IEEE Access*, vol. 7, pp. 99 878-99 888, 2019.

[30] H. Q. Ngo, A. Ashikhmin, H. Yang, E. G. Larsson, and T. L. Marzetta, "Cell-free massive MIMO: Uniformly great service for everyone," in *2015 IEEE 16th international workshop on signal processing advances in wireless communications (SPAWC)*. IEEE, 2015, pp. 201-205.

[31] Ö. T. Demir, E. Björnson, L. Sanguinetti et al., "Foundations of user-centric cell-free massive MIMO," *Foundations and Trends® in Signal Processing*, vol. 14, no. 3-4, pp. 162-472, 2021.

[32] E. Björnson and L. Sanguinetti, "Scalable cell-free massive MIMO systems," *IEEE Trans. on Comm.*, vol. 68, no. 7, pp. 4247-4261, 2020.

[33] H. Q. Ngo, A. Ashikhmin, H. Yang, E. G. Larsson, and T. L. Marzetta, "Cell-free massive MIMO versus small cells," *IEEE Transactions on Wireless Communications*, vol. 16, no. 3, pp. 1834-1850, 2017.

[34] E. Nayebi, A. Ashikhmin, T. L. Marzetta, H. Yang, and B. D. Rao, "Precoding and power optimization in cell-free massive MIMO systems," *IEEE Transactions on Wireless Communications*, vol. 16, no. 7, pp. 4445-4459, 2017.

[35] E. Björnson and L. Sanguinetti, "Making cell-free massive MIMO competitive with MMSE processing and centralized implementation," *IEEE Transactions on Wireless Communications*, vol. 19, no. 1, pp. 77-90, 2020.

[36] G. Interdonato, P. Frenger, and E. G. Larsson, "Scalability aspects of cell-free massive MIMO," in *ICC 2019 - 2019 IEEE International Conference on Communications (ICC)*, 2019, pp. 1-6.

[37] Ö. T. Demir, E. Björnson, and L. Sanguinetti, "Cell-free massive MIMO with large-scale fading decoding and dynamic cooperation clustering," 2021.

[38] S. Buzzi and C. D'Andrea, "Cell-free massive MIMO:

User-centric approach," *IEEE Wireless Communications Letters*, vol. 6, no. 6, pp. 706-709, 2017.

[39] H. Q. Ngo, L.-N. Tran, T. Q. Duong, M. Matthaiou, and E. G. Larsson, "On the total energy efficiency of cell-free massive MIMO," *IEEE Transactions on Green Communications and Networking*, vol. 2, no. 1, pp. 25-39, 2018.

[40] O. Y. Bursalioglu, G. Caire, R. K. Mungara, H. C. Papadopoulos, and C. Wang, "Fog massive MIMO: A user-centric seamless hot-spot architecture," *IEEE Transactions on Wireless Communications*, vol. 18, no. 1, pp. 559-574, 2019.

[41] G. Interdonato, M. Karlsson, E. Björnson, and E. G. Larsson, "Local partial zero-forcing precoding for cell-free massive MIMO," *IEEE Transactions on Wireless Communications*, vol. 19, no. 7, pp. 4758-4774, 2020.

[42] F. Göttsch, N. Osawa, T. Ohseki, K. Yamazaki, and G. Caire, "The impact of subspace-based pilot decontamination in user-centric scalable cell-free wireless networks," 2021. [Online]. Available: https://arxiv.org/abs/2108.04579

[43] F. Göttsch, N. Osawa, T. Ohseki, K. Yamazaki, and G. Caire, "Uplink-downlink duality and precoding strategies with partial CSI in cell-free wireless networks," 2022. [Online]. Available: https://arxiv.org/abs/2201.04922

[44] S. Haghighatshoar and G. Caire, "Massive MIMO pilot decontamination and channel interpolation via wideband sparse channel estimation," *IEEE Transactions on Wireless Communications*, vol. 16, no. 12, pp. 8316-8332, 2017.

[45] D. Wipf and B. Rao, "Sparse Bayesian learning for basis selection," *IEEE Transactions on Signal Processing*, vol. 52, no. 8, pp. 2153-2164, 2004.

[46] H. Xu, C. Caramanis, and S. Sanghavi, "Robust PCA via outlier pursuit," *IEEE Transactions on Information Theory*, vol. 58, no. 5, pp. 3047-3064, 2012.

[47] A. Adhikary, J. Nam, J.-Y. Ahn, and G. Caire, "Joint spatial division and multiplexing: The large-scale array regime," *IEEE Trans. on Inform. Theory*, vol. 59, no. 10, pp. 6441-6463, 2013.

[48] T. L. Marzetta, E. G. Larsson, H. Yang, and H. Q. Ngo, "Fundamentals of massive MIMO," Cambridge University Press, 2016.

[49] G. Caire, "On the ergodic rate lower bounds with applications to massive MIMO," *IEEE Transactions on Wireless Communications*, vol. 17, no. 5, pp. 3258-3268, 2018.

[50] M. Feder and A. Lapidoth, "Universal decoding for channels with memory," *IEEE Transactions on Information Theory*, vol. 44, no. 5, pp. 1726-1745, 1998.

[51] M. J. Neely, "Stochastic network optimization with application to communication and queueing systems," *Synthesis Lectures on Communication Networks*, vol. 3, no. 1, pp. 1-211, 2010.

[52] H. Shirani-Mehr, G. Caire, and M. J. Neely, "MIMO downlink scheduling with non-perfect channel state knowledge," *IEEE Transactions on Communications*, vol. 58, no. 7, pp. 2055-2066, 2010.

[53] 3GPP, "Study on channel model for frequencies from 0.5 to 100 GHz (Release 16)," 3GPP Technical Specification 38.901, 12 2019, Version 16.1.0.

[54] 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 Transactions on Information Theory*, vol. 67, no. 5, pp. 2925-2951, 2021.

Extremely Large Aperture Arrays and Reconfigurable Intelligent Surface: The Evolution of mMIMO

Rahim Tafazolli [1], Yi Ma [1], Jiuyu Liu [1], Pei Xiao [1], Alexandr Kuzminskiy [1], Fabien Heliot [1], Mohsen Khalily [1], Fan Wang [2]

[1] 5G & 6G Innovation Centers, University of Surrey

[2] Wireless Technology Lab, Huawei Technologies Co., Ltd.

Abstract

Massive multiple-input multiple-output (mMIMO), typically in the form of 64 transmit and 64 receive (64TRX) or beyond at mid-bands, has achieved huge commercial success in 5G. The straightforward question is how mMIMO will evolve in B5G/6G. We believe the answer will be at least twofold: one is extremely large aperture array (ELAA), which brings mMIMO to the next level of extreme spectrum efficiency by significantly increased aperture allowing more TRX and antenna elements, and the other is reconfigurable intelligent surface (RIS), which manipulates the reflected wave to a desired direction using a low cost metasurface. A novel non-stationary statistical channel model for the ELAA system is proposed, which is shown to capture non-stationarity inherent in the ELAA system. Furthermore, signal processing algorithms are examined, and the results show those are scalable (defined as optimal/near-optimal in performance and of low implementation complexity) in conventional mMIMO systems are no longer scalable in ELAA systems due to the channel spatial non-stationarity. Hence new scalable signal processing algorithms tailored for ELAA non-stationary channel shall be investigated. For sub-array ELAA transmit precoding, multi-antenna receiver interference mitigation, in conjunction with widely linear processing, is shown as a very efficient supplement to deal with interference and impairments. Fundamental limits of RIS are provided for understanding the potential benefit of this new technology in comparison with MIMO and MIMO-amplify-and-forward (AF), where RIS improves the eigenvalues of the propagation channel, spectrum efficiency and energy efficiency, thanks to the richer channels resulting from the size of the RIS and the cascaded channel effect of the base station-RIS and RIS- user equipment (UE). We have demonstrated both dynamic and static RISs in real-life indoor environment at C-bands, which enhance the signal level at the receiver side by more than 15 dB and 19 dB, respectively.

1 Introduction

Multiple-input multiple-output (MIMO) is a major technique which significantly boosts the spectrum efficiency as well as coverage in wireless communications. It has been standardized by the 3rd Generation Partnership Project (3GPP) from 3G high speed packet access (HSPA). The first commercial deployment of MIMO in 3GPP standard is 4G Long Term Evolution (LTE), where MIMO of 4 transmit and 4 receive (4TRX) is natively supported by the base station from the first LTE release [1]. Massive MIMO (mMIMO) [2], typically in the form of 64TRX or beyond and even more antenna elements (e.g., 192) at mid-bands, provides significant spectrum efficiency and coverage gain over conventional MIMO techniques, thus standardized by 5G NR from the first release. Up to date, base stations (BS) with mMIMO at C-bands has been the dominant 5G NR commercial deployments all over the world.

Seeking higher spectrum efficiency has been a major design target for wireless communication system from 3G to 5G, and remains the case for B5G/6G [1]. Then the straightforward question is how mMIMO in 5G will evolve to support extreme spectrum efficiency in B5G/6G [3]. We believe the answer will be at least twofold: extremely large aperture array (ELAA) and reconfigurable intelligent surface (RIS).

By significantly increased aperture allowing more TRX and antenna elements, ELAA is able to utilize extreme spatial resolution and deliver superior beamforming gain as well as spatial reuse gain. ELAA boosts the received signal strength (RSS) and reduces the inter-user equipment (UE) interference, therefore improves the spectrum efficiency from both the UE and the network perspective.

ELAA poses some challenges including new channel models and scalable signal processing design. As the ELAA array aperture spans hundreds of wavelengths, the ELAA channel model is fundamentally different from conventional mMIMO. The propagation channel becomes spatial non-stationary where not all the antennas are visible to a UE. Hence a new spatial non-stationary channel needs to be defined. In addition, as the number of TRX goes extremely high, the ELAA signal processing technique has to be scalable, i.e., optimal/near-optimal in performance and of low implementation complexity in order to make ELAA practical.

RIS, different from enhancing the MIMO capability at the BS, produces a smart radio environment by manipulating the reflected wave to a desired direction using a metasurface as a new node [4], which makes RIS very powerful in boosting the link capacity and coverage. The metasurface employed in RIS is composed of a large number of controlling components such as the varactors of PIN diodes to control the radio environment without the need of active components like RF chains as in conventional MIMO systems, which makes RIS very cost-efficient. Due to the promising benefits in terms of low cost, link capacity and coverage, RIS has been recognized as potential key techniques for B5G/6G [1].

The research challenges of RIS include understanding the fundamental capacity limits of RIS and real-world RIS platform verification. The benefit of RIS in comparison with the most relevant existing techniques, i.e., MIMO and MIMO-amplify-and-forward (AF), shall be understood in terms of the eigenvalues of the propagation channel, spectrum efficiency and energy efficiency. The RIS platform, especially dynamic RIS which reconfigures the RIS to dynamically track the change of wireless environment, shall be designed and demonstrated in real world to verify the feasibility of RIS.

2 Extremely Large Aperture Arrays

2.1 Non-stationary Channel Model

ELAA channel models are fundamentally different from conventional mMIMO channel models. They cannot be easily considered as wide-sense stationary uncorrelated scattering (WSSUS) because UEs are usually located in the near field of the ELAA, and for this reason, UE-to-service antenna links can have very different probability density functions (PDFs) [5]. Figure 1 depicts an example of ELAA systems, where the ELAA is a large uniform linear array (ULA) with the aperture of 43 meters and the operating frequency of 3.5 GHz. A UE is communicating to the ELAA through the uplink channel. At the ELAA, some service antennas can see the UE, and the others cannot. Therefore, the ELAA channel has a mix of line of sight (LOS) and non-LOS (NLOS) links. Moreover, shadowing effects can also vary from the link to link. All of these physical characteristics render the ELAA channel spatially non-stationary.

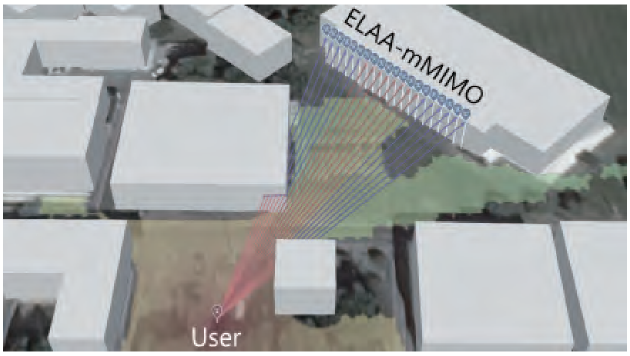


Figure 1 An example of non-stationary ELAA channels from the site view of the University of Surrey, Stag Hill campus

The foundation of wireless research lies in good understanding and modeling of wireless channels. There are mainly two types of channel models: 1) deterministic channel models; 2) statistical channel models. The latter are more suitable for computer simulations because they are usually simple and mathematically trackable. Specifically, for the ELAA non-stationary channel, deterministic models are too complex to implement and not suitable for Monte Carlo simulations. On the other hand, non-stationary statistical channel models can be hardly made as simple as WSSUS channel models. A straightforward approach is to make extensions from WSSUS channel models by incorporating some spatial non-stationarity. For instance, independent and identically distributed (IID) Rayleigh (or Rician) fading channels can be extended to independent and non-identically distributed (IND) Rayleigh (or Rician) fading channels. The concept of visibility region is introduced to describe where the ELAA UE-to-service antenna links have their RSS going beyond a threshold. The distribution of visibility region is dependent on the geometry of wireless environments. For instance, in the literature [7–10], it is assumed that there exists only a single cluster between each UE and service antennas. The length of visibility region obeys a log-normal distribution. When there exist multiple clusters, the distribution of visibility region for each cluster is modeled as the birth-death process or the Markov process. The LOS state of each link is based on the geometry of the wireless environment [11–15]. Table 1 provides a summary of current non-stationary channel models as well as their pros/cons are provided in modeling complexity, generality (i.e., map independency), mathematical trackability, and computer simulation support. After all, we found the single-cluster model too simple to capture the channel spatial non-stationarity, and on the other hand the multi-cluster model too complex for computer simulations. This motivates us to develop a novel non-stationary statistical channel model for the ELAA system.

In our recent work [16], a novel spatially non-stationary channel model is proposed. The proposed model fades out the concept of visibility region and allows UE-to-service antenna links to be NLOS, LOS, or a mix of them. The NLOS/LOS state of each link is described by a binary random variable, which obeys a correlated Bernoulli distribution. More specifically, the NLOS/LOS probability of each link is a function of the two-dimensional (2D) distance between the UE and the corresponding service antenna, which is defined by 3GPP in [17]. Moreover, the probability for two arbitrary links to share the same NLOS/LOS state decreases exponentially with the separation between service antennas. Therefore, the NLOS/LOS state correlation is modeled as an exponentially decaying window. In addition, shadowing effects are carefully incorporated into the proposed non-stationary channel model. Within an NLOS/LOS window, all links share an identical NLOS/LOS state. They also share identical shadowing effects, which follow a log-normal distribution. Both the NLOS/LOS state and shadowing effects vary from the window to window. When a window experiences deep fades, this window can count as an invisible region and vice versa. The pseudo code that is used to generate the non-stationary channel can be found in our recent work [16], which is simple, generic and suitable for link-level Monte Carlo simulations.

Table 1 A summary of current spatial non-stationary channel models and their pros and cons

Channel Model	Complexity	Generality	Trackability	Computer Simulation Support
Map-based deterministic models [6]	High	No	Hard	Very hard
Geometry-based stochastic models [11–15]	High	Yes	Hard	Very hard
Current statistical channel models [7–10]	Low	No	Yes	Yes, but over simplified

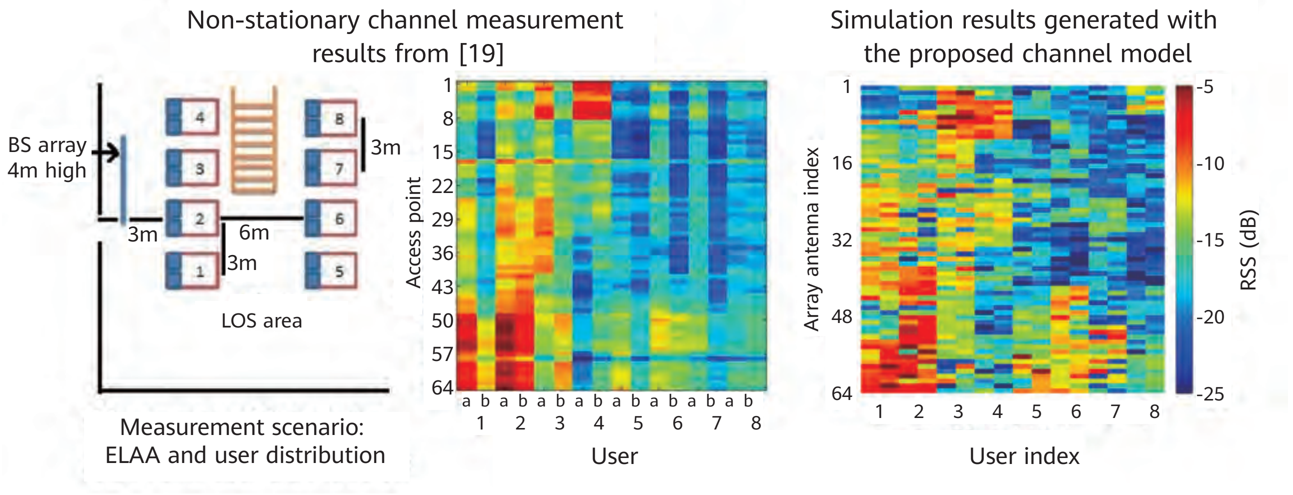


Figure 2 A comparison of RSS maps generated from real-world channel measurement and the proposed channel model

It is perhaps worthwhile to note that our current non-stationary channel model does not include multiuser spatial consistency and service antenna spatial correlation, which are also important and interesting issues for the channel modeling. All of those will be reported in our upcoming journal papers.

To demonstrate the usefulness of the proposed channel model, we use the algorithm (described in the pseudo code) to randomly generate many channel realizations and compare the simulation result with real-world ELAA channel measurements. In order to fit into the real-world indoor environment (ELAA height: 4m, aperture: 3m) in [19] for instance, parameters of the proposed channel model are appropriately configured. In the real-world measurement, RSS data is collected for eight labeled UE places. As shown in Figure 2, a quite similar RSS distribution can be obtained from our channel simulator. This is a clear evidence that the proposed statistical channel model is able to capture non-stationarity inherent in the ELAA system.

2.2 Scalable System Design

mMIMO is recognized as a scalable technology mainly in the sense: 1) it can achieve close-to-optimum performances; 2) its computational complexity scales in the square order of the size of MIMO networks; 3) mMIMO digital transceiver architecture is mainly based on the matched filter (MF) algorithm, which supports the parallel computing technology very well. For a fixed UE-to-service antenna ratio, when the number of service antennas tends to infinity, the instantaneous mMIMO channel capacity shows decreasing fluctuations. This is well known as the channel hardening effect in mMIMO systems [18]. All of these appealing features are based on two hypotheses: 1) the excess use of service antennas over UEs; 2) mMIMO channels are spatially IID. Here, the central question is whether or not these two hypotheses still hold in the ELAA system. If the answer is 'No', what would be the scalable approach for the ELAA system? To this date, this remains an open problem, which has been attracting increasing research investment. In this paper, we provide a couple of technical insights mainly based on some of our recent findings.

Key measures in our study include: 1) the cumulative distribution function (CDF) of the instantaneous channel capacity; 2) the symbol error rate (SER) averaged over ELAA fading channels. The following channel models are studied in our work:

- **Model I**: proposed channel model;
- **Model II**: IID Rayleigh fading model (only suitable for idealistic mMIMO);
- **Model III**: IND Rayleigh fading channel model from [8];
- **Model IV**: IND Rician fading channel model;
- **Model V**: single-cluster geometry statistical channel model from [5]. The visibility region is randomly distributed on the ULA with the length following the log-normal distribution specified in the paper.

In our case study, the ELAA is assumed to have 2000 service antennas with the aperture of 85 meters (on the 3.5 GHz central frequency). The radio propagation environment is the UMi-street canyon. All channel parameter settings follow the 3GPP document [17].

Each UE is assumed to have 4 transmit-antennas. For the proposed channel model (i.e., Model I), antennas belonging to the same UE are assumed to have the same NLOS/LOS states as well as shadowing effects. Moreover, UEs are located in front of the ULA, and they are uniformly distributed on a line that is parallel to the ULA.

There are five UEs and two settings of user density in our first study. One is the high-density case where all UEs are located within a 1-meter range, and the other is the low-density case, where the range is 20-meter long. In the high-density case, UEs share the same NLOS/LOS state for all of their links, and in the low-density case, NLOS/LOS states for different UEs are independently generated. In addition, the shadowing effects are generated independently for users of both densities.

Figure 3 shows the CDF of instantaneous channel capacity in bit/s/Hz when shadowing effects are not considered. For the solid line without shadowing effects, Model III and Model IV exhibit very clear channel hardening effects, which are the same as the conventional mMIMO channel model (i.e., Model II). This is because Model III and Model IV have their non-stationarity mainly coming from the spherical-wave propagation. For fixed user locations, their large-scale fading components are temporally stationary, and their small-scale fading is stationary as well. In contrast, the channel hardening effects are much weaker in Model I (high and low densities) and Model V. This attributes to the presence of mixed NLOS/LOS states in our model and the spatial visibility region in Model V. Due to the same reason, the presence of non-stationary shadowing could also largely weaken the channel hardening effects (see the dashed line in Figure 3). When comparing cases with different user densities, it is found that the users sharing the same LOS can further degrade the channel hardening effects in non-stationary channels. This is because the channel of high-density UEs having a larger variation in channel norm (see Figure 4).

In order to examine the impact of channel non-stationarity on the signal detection, we investigate the single user case in the second study. The UE sends 4 independent data streams to the ELAA and each data stream is modulated with 4QAM. Figure 5 (without shadowing effects) and Figure 6 (with shadowing effects) depict the SER performance of different receivers: linear minimum mean square error (LMMSE), MF, and maximum likelihood sequence detection (MLSD). In conventional mMIMO Rayleigh fading channels, the performance of MF approaching the maximum likelihood performance with such an excess use of service antennas. This fully coincides with the well-known behavior of mMIMO. However, such nice behavior no longer exists in the non-stationary channel. When the channel is IND Rayleigh fading, we can observe slight performance degradation. More critically, the MF algorithm almost fails for Model I and Model IV. This is not surprising because the optimality of MF is based on hypotheses for the conventional mMIMO system, which no longer hold for the ELAA system due to the existence of LOS links.

Let's focus on the SER performance of the LMMSE algorithm. The impact of channel spatial non-stationarity is not as large as that in the MF. However, the performance gap is still considerably large. Strictly speaking, the LMMSE or zero-forcing (ZF) algorithm is not a scalable linear algorithm for mMIMO systems since their complexities grow in a cubic order of the size of MIMO networks. More importantly, the LMMSE or ZF algorithm requires channel matrix inversion, which does not have a reliable parallel computing architecture to this date. Therefore, they cannot take advantage of state-of-the-art high-performance computing technology which relies on the power of parallel computing.

As a conclusion, signal processing algorithms that are scalable in conventional mMIMO systems are no longer scalable in ELAA systems due to the channel spatial non-stationarity. Advanced non-linear algorithms will face even more challenges arising from the system scalability. On the other hand, ELAA non-stationary channel is a sparse channel, and this is the distinctive feature that should be seriously considered in the scalable ELAA system design. In this regard, approximate message passing (AMP) might be an appealing algorithm. However, the application of AMP algorithms in ELAA is not straightforward because they are so far only optimized for stationary processes. A fundamental rethinking and redesign of AMP principles is needed for the future ELAA system research.

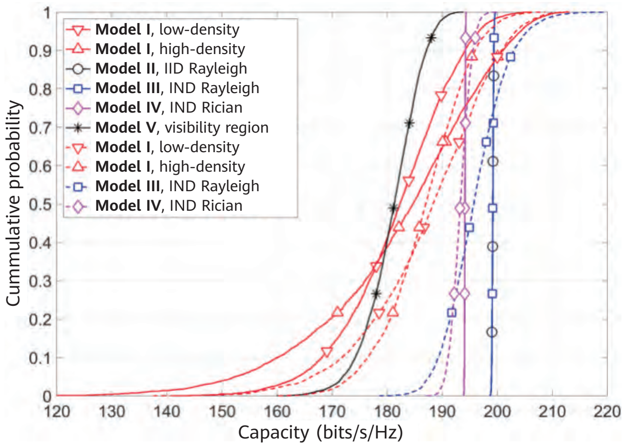


Figure 3 CDF of the instantaneous channel capacity with an average SNR of 10 dB (solid line: without shadowing effects; dashed line: with shadowing effects)

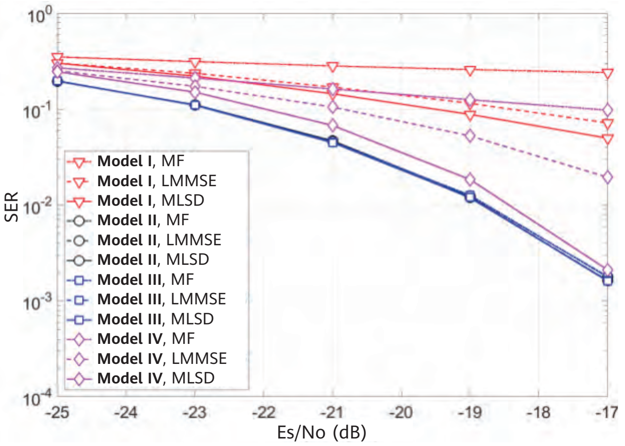


Figure 5 SER vs Es/No for ELAA uplink detection using LMMSE, MF, and MLSD without shadowing effects (single-UE case, 4QAM)

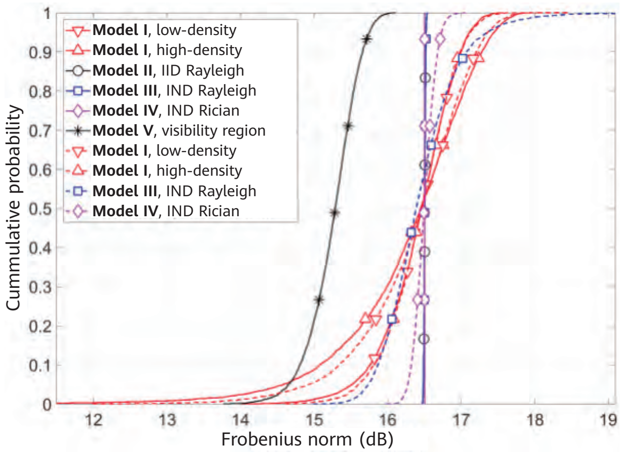


Figure 4 CDF of the instantaneous channel Frobenius norm (**solid line**: without shadowing effects; **dashed line**: with shadowing effects)

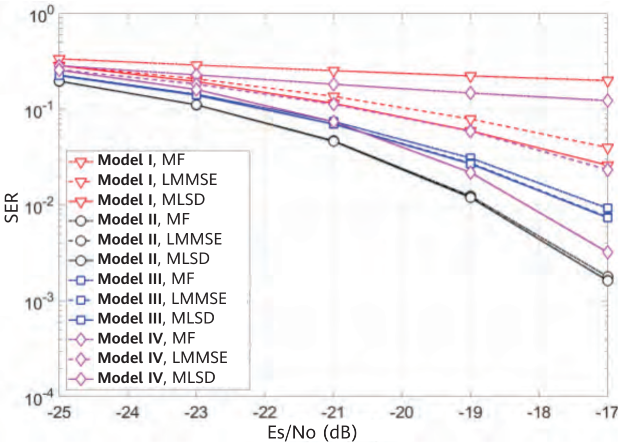


Figure 6 SER vs Es/No for ELAA uplink detection using LMMSE, MF, and MLSD with shadowing effects (single-UE case, 4QAM)

2.3 ELAA Signal Processing Robust to Interference

In this section, we address robust ELAA signal processing on the downlink considering the following ELAA features:

- Spatial non-stationary propagation environment leads to overlapping visibility regions for different users, which creates unavoidable interference for any decentralized sub-array signal processing.
- Exact channel knowledge at the transmitter is impossible, e.g., due to the channel estimation errors and/or ageing effects.
- Although transmit antenna imperfections, e.g., in-phase and quadrature-phase imbalance (IQI), potentially could be compensated at the transmitter for the known imperfection parameters, assuming such knowledge may not be realistic for ELAA.

A typical approach for transmit antenna precoding is to assume some channel knowledge from a multi-antenna transmitter to a single-antenna UE receiver and apply some precoding technique such as maximum ratio or ZF transmission. ELAA sub-array precoding is studied in [20], for single-antenna UEs by means of comparison of the stationary solution with the best and worst interference distributions for sub-array ZF precoding.

In this section, we propose to consider a possibility to address the interference and imperfections for ELAA sub-array processing by means of introducing the multi-antenna interference mitigation capability at UE receivers. Our motivations are summarized as follows:

- Receiver interference mitigation support for sub-array ELAA precoding could address all the interference and imperfection issues listed above.

- A single-antenna receiver is typically just a useful theoretical abstraction. To date, any LTE/5G receiver has at least two receive antennas. Furthermore, downlink data is always transmitted in some time/frequency slots with pilots that can be used for estimation of the signal and interference parameters required for interference mitigation.

We consider a narrowband scenario for an ELAA of M elements serving K users with N receive antennas. The ELAA is divided to B sub-arrays with $M_B = M/B$ antennas each. All users are associated with some sub-arrays with users per sub-array, possibly with $K_b = 0$ for some sub-arrays. Signals are transmitted in slots of L_{tot} symbols containing L_p pilots and L_d data symbols. The $N \times L_{tot}$ signal received by the kth user is defined as follows:

$$Y_k = \Sigma_{b=1}^{B} H_{bk}^{*} IQI\{\sqrt{P_b}\,\hat{G}_b S_b\} + Z_k \tag{1}$$

where H_{bk} is the $M_B \times N$ propagation channel from the bth sub-array to the kth user, $P_b = \mathrm{diag}([p_1,\ldots,p_{k_b}])$ is the $K_b \times K_b$ diagonal matrix of the transmit power on users associated with the bth sub-array, $\hat{G}_b$ is the $M_b \times K_b$ precoder for the bth sub-array, S_b is the $K_b \times L_{tot}$ matrix of the K_b transmitted signals associated with the K_b th sub-array, Z_k is the $N \times L_{tot}$ matrix of AWGN with σ^2 power, $IQI\{A\}$ is the IQI operation applied per row of matrix A with the individual transmit antenna IQI parameters.

To concentrate on the ELAA interference mitigation effects, we consider a simple energy-based user association with sub-arrays and ZF precoding with uniform power allocation of $p_k = 1$ assuming that the $M_b \times 1$ propagation channels to one antenna per UE are known at the transmitter with some mean square error (MSE). $\hat{\bar{h}}_{bk}$, $k = 1,\ldots, K$, $b = 1,\ldots, B$:

$$K_b = \left\{k,b = \underset{q=1,\ldots,B}{\mathrm{argmax}} \left\|\hat{\bar{h}}_{qk}\right\|^2\right\} \tag{2}$$

$$\hat{G}_b = \beta_b \hat{\bar{H}}_b (\hat{\bar{H}}_b^{*} \hat{\bar{H}}_b) \tag{3}$$

$$\beta_b = \sqrt{\frac{K_b}{\mathrm{tr}\left[(\hat{\bar{H}}_b^{*} \hat{\bar{H}}_b)^{-1}\right]}} \tag{4}$$

where $\hat{\bar{H}}_b$ is the $M_b \times K_b$ matrix of $\hat{\bar{H}}_{bk}$ for $k \in K_b$. A single UE is associated only with one sub-array.

At the receiver side, in addition to the conventional single-antenna solution, we consider the conventional interference rejection combining (IRC), e.g., as in [21], widely linear IRC (WLIRC) to address IQI, e.g., as in [22], and their variants with projections to the finite alphabet (FA), e.g., as in [23], IRC with projection to FA (IRCFA), and widely linear version of IRCFA (WLIRCFA).

IRC:

$$\hat{s}_k = \hat{w}_k^{*} Y_k \tag{5}$$

$$\hat{w}_k = (\hat{R}_k + \delta I)^{-1} \hat{h}_k \tag{6}$$

$$\hat{R}_k = \textstyle\sum_{l=1}^{Lp} y_{kl} y_{kl}^{*} \tag{7}$$

$$\hat{h}_k = \textstyle\sum_{l=1}^{Lp} y_{kl} s_{kl}^{*} \tag{8}$$

where $\hat{s}_k$ is the $1 \times L_{tot}$ vector of the estimated signal of the kth user, $\hat{w}_k$ is the $N \times 1$ IRC weight vector, $\hat{R}_k$ and $\hat{h}_k$ are the pilot based second order statistics estimates, δ is the diagonal loading parameter and I is the $N \times N$ unit matrix.

IRCFA:

$$\tilde{s}_k = \tilde{w}_k^{*} Y_k \tag{9}$$

$$\tilde{w}_k = \tilde{R}_k^{-1} \tilde{h}_k \tag{10}$$

$$\tilde{R}_k = \textstyle\sum_{l=1}^{L_{tot}} y_{kl} y_{kl}^{*} \tag{11}$$

$$\tilde{h}_k = \textstyle\sum_{l=1}^{L_{tot}} y_{kl} \mathrm{FA}\{\hat{s}_{kl}\}^{*} \tag{12}$$

where $\hat{s}_{kl}$ are elements of $\hat{s}_k$ in (5), and FA $\{\cdot\}$ is the projector to FA (slicer). Other notations are defined alike in the IRC case.

The WLIRC and WLIRCFA receivers are described by equations (5)–(8) and (9)–(12) respectively, with replacement of Y_k with the $2N \times L_{tot}$ matrix $[\begin{smallmatrix} Y_k \\ \mathrm{conj}(Y_k) \end{smallmatrix}]$, where conj(·) is the complex conjugate operation.

We simulate 200 m long ELAA of $M = 512$ elements with $K = 32$ users and $N = 4$ receive antennas distributed along the array using propagation channel model described in the previous section for 3.5 GHz carrier frequency for micro urban scenario with mixed LOS states and shadowing effects. We use the following definition of the signal to noise ratio (SNR):

$$\mathrm{SNR} = \frac{\mathrm{tr}\,(\hat{H}_1^{*} \hat{H}_1)}{K\sigma^2} \tag{13}$$

We assume 16QAM signaling for a data slot of $L_p = 12$ randomly selected pilots and $L_d = 132$ data symbols, and use $\delta = 0.5\sigma^2$ for diagonal loading in (6).

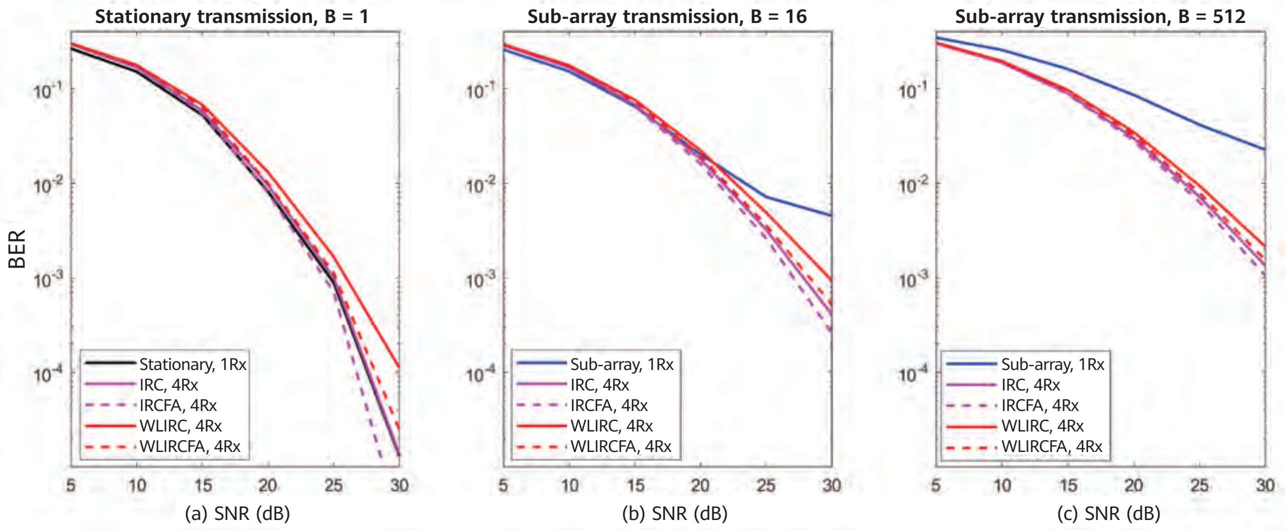


Figure 7 Average BER benchmark for the known channels and without IQI at the transmitter

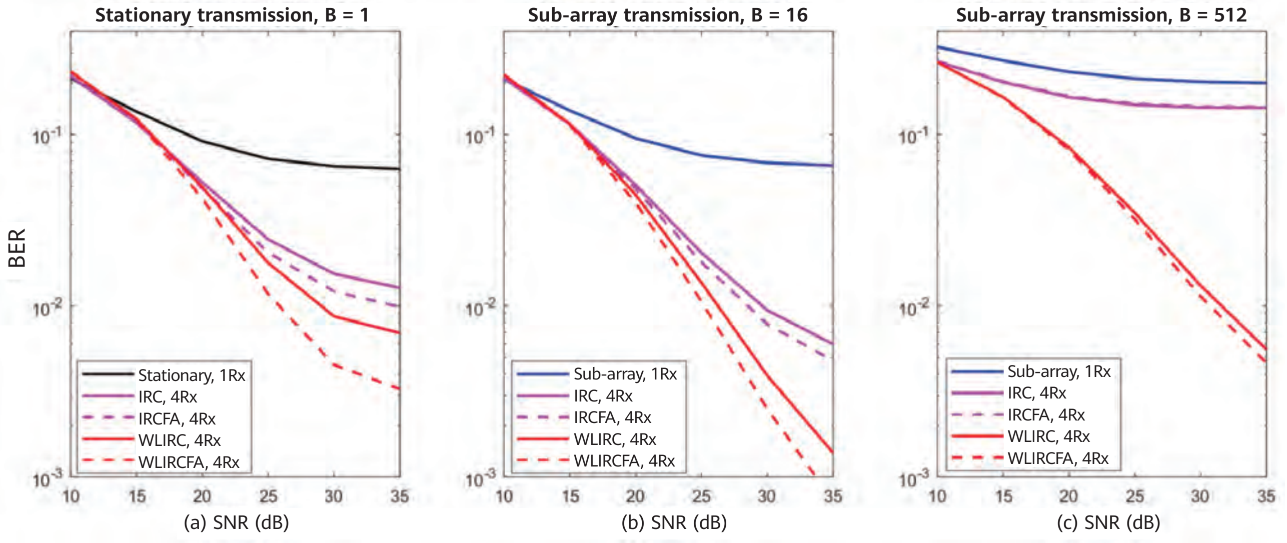


Figure 8 Average BER for 5% channel MSE and IQI of 1–2 dB amplitude and ±5 degree errors at the transmitter

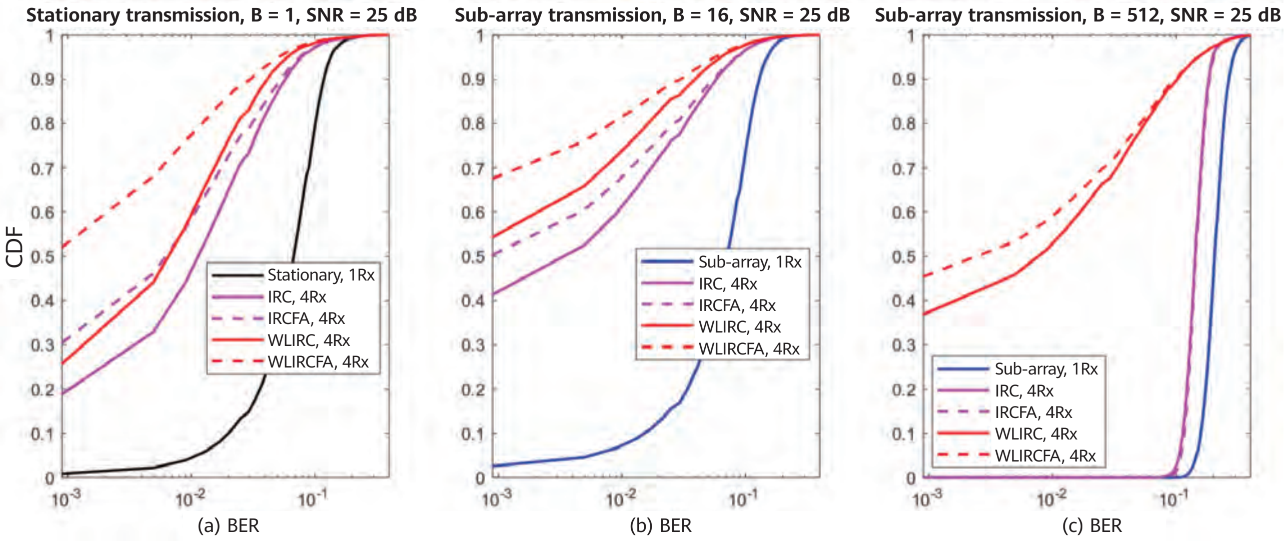


Figure 9 CDF BER results in the same scenarios as in Figure 7 for SNR = 25 dB

For benchmarking purposes, Figure 7 shows the BER results averaged over 500 scenario realizations for the known channels at the transmitter and no IQI for the stationary (unfeasible) precoding for $B = 1$ in Figure 7a, sub-array precoding for $B = 16$ in Figure 7b, and the simplest sub-array precoding for $B = 512$ in Figure 7c. In fact, the last case corresponds to the single transmit antenna selection per user in the considered scenario.

The following observations can be made from Figure 7:

- In Figure 7a, the BER results are practically the same for the single-antenna and IRC receivers. Potentially, some diversity gain for the multi-antenna receiver could be expected in this noise limited scenario, but we practically do not see it here because we use the fixed low diagonal loading assuming that our main scenario is interference limited.
- From Figure 7b and Figure 7c, one can see a significant performance degradation for sub-array processing for the single-antenna receiver, which can be improved with IRC and IRCFA. The overall IRC and IRCFA performance for sub-array transmission is worse than the stationary case with no interference or impairment.
- The WLIRC and WLIRCFA suffer some performance degradation compared to the IRC and IRCFA results because they use a doubled number of the estimated weights for the same training support.

The average BER results in the scenario similar to Figure 7 are shown in Figure 8 for MSE = 5% for the transmit channels and IQI amplitude and phase errors uniformly randomly distributed in the range of 1–2 dB and ±5 degrees independently for each transmit antenna.

CDFs for the BER performance for SNR = 25 dB are plotted in Figure 9 under the same conditions as in Figure 8.

The following observations can be made from Figure 8 and Figure 9:

- Contradictory to the no impairments case shown in Figure 7, multi-antenna receivers significantly improve performance compared to the single-antenna UEs in all scenarios with channel errors and IQI.
- The main gain comes from the widely linear processing because of the significant IQI at the transmitter in the considered scenario.
- The sub-array widely linear processing for $B = 16$ in Figure 8b and Figure 9b outperforms the widely linear processing for the stationary transmission with errors for $B = 1$ in Figure 8a and Figure 9a. Probably, this is because of simpler sub-array interference scenarios for channel errors and IQI compared to the stationary transmission case.

The overall conclusion is that multi-antenna receiver interference mitigation could be considered as a very efficient supplement for sub-array ELAA transmit precoding to deal with interference and impairments.

3 Reconfigurable Intelligent Surface

Wireless communication engineers envision a fully connected world where there is seamless wireless connectivity for everyone and everything. Modern wireless communication networks continue to grow at a very rapid rate which has increased the demand for intelligent and efficient communication networks. However, all dynamic and adaptive features of wireless networks are controlled by either the BS or the UE while the wireless propagation environment remains unaware of various communication processes going through it. It therefore remains an open topic of research and evaluation among the industrial and academic fraternity to impart some level of intelligence to this otherwise passive radio propagation environment.

RIS is a nascent technology that promises to address this demand by enabling the manipulation of various characteristics of the Electromagnetic (EM) waves in the desired direction. However, there are some research gaps that need to be addressed in order to deploy RIS technology in a real wireless communications network. Recently at 6GIC, static RIS and dynamic RIS have been developed and demonstrated in a real-world scenario.

3.1 Fundamental Capacity Limits

The study of the fundamental limits of a new technology or system is usually helpful for understanding the potential benefit of this new technology in comparison with the existing ones. In turn, it helps to figure out if the development, from theory to practice, of such new technology is worth investing money and time. When it comes to RIS surface, its advantages towards existing relay or small cell technology should be significant enough to motivate such an investment [24].

In communication, the fundamental limit of a technology or system usually refers to its capacity, ergodic or outage for a fast or slow time-varying channel, respectively, when the channel state information (CSI) is only known at the receiver, which measures the spectral efficiency (SE). However, with advent of 5G and going into 6G another type of fundamental limits, which has drawn a lot of research interest in the last decade [25], is growing in importance, i.e., the bit-per-Joule capacity that measures the energy efficiency (EE). In this section, we study both the SE and EE of RIS.

Given the theoretical nature of fundamental limits, they are derived based on models and assumptions. Hence, their relevance depends heavily on how accurate/close to the reality these models and assumptions are. In here, we consider the well-established phase-shift model [26] (i.e., appearing in hundreds of work on RIS already) to model the behavior of a RIS. In this model, a RIS is represented by using a diagonal matrix that accounts for the effective phase shifts applied by all RIS reflecting elements. Meanwhile, as far as the consumed power of a RIS is concerned, we assume the power consumption model of [26–27], which is related to phase-shift model.

3.1.1 Propagation Improvement

The main application of RIS in cellular system is to improve the propagation environment (leading to coverage extension and/or enhanced blind spot coverage), which is similar to relay technology. Instead of processing the signal it receives before retransmitting a re-encoded and/or amplified version of it, as in most relaying techniques, RIS passively reflects the signal without amplification but with directionality [28].

A simple way to understand and evaluate the fundamental propagation improvement benefit of RIS against the traditional non-relaying (e.g., MIMO) or relaying (e.g., MIMO-AF) technology is to study how the probability density function (PDF) and/or CDF of the eigenvalues of the propagation channel are modified when RIS is introduced. Let us assume a point-to-point transmission between a base station (BS) equipped with 32 antennas and a user equipment (UE) equipped with four antennas, where the channel between the BS and the UE is assumed to be Rayleigh distributed:

- Case 1 – classic MIMO: no RIS or relay is used;
- Case 2 – MIMO-AF: A relay with 64 antennas, using the one-way AF protocol [29], is placed midway between the BS and the UE;
- Case 3 – two-way (TW)-MIMO-AF: same as case 2, but where the relay uses the two-way AF protocol [30];
- Case 4 – MIMO-RIS: A RIS with 64 phase shifters is placed midway between the BS and the UE;
- Case 5 – same as case 4, but with correlated channel (CC) at the RIS.

Figure 10 depicts the CDF of the eigenvalues of the channel in the five different cases. Note that the curves for MIMO and MIMO-AF almost overlap. The results clearly show the significant improvement in the CDF when the RIS is introduced instead of no RIS (i.e. MIMO) or having relays (i.e., MIMO-AF, TW-MIMO-AF). The RIS improvement in the CDF against MIMO of close to two orders of magnitude is due to the richer channel resulting from the size of the RIS

(64 in this case) and the cascaded channel effect, where the eigenvalues of the overall channel (i.e., BS-RIS channel + RIS-UE channel) are the product of the eigenvalues of these two different channels, especially when these channels are independent. However, if the two channels are correlated at the RIS, the values of the eigenvalues are more spread and the advantage of RIS over MIMO is less significant. Meanwhile, the RIS improvement in the CDF against MIMO-AF of more than one order of magnitude (for independent channels) is caused by noise amplification in MIMO-AF, where the noise of the first hop is amplified in the second hop. If the noise in the first hop where to be close to zero (for the same transmit power used in the simulation), then the CDF of MIMO-AF and MIMO-RIS would overlap. In essence, the CDF of MIMO-AF channel eigenvalues is upper bounded by the one of MIMO-RIS.

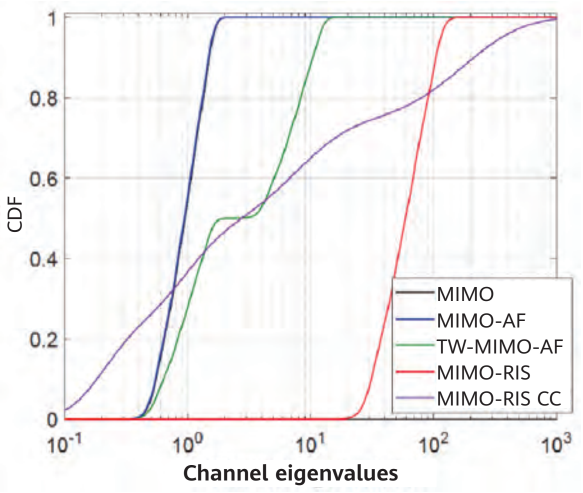


Figure 10 CDF of the channel eigenvalues of different MIMO-based transmission techniques in the point-to-point scenario

3.1.2 Spectral Efficiency Improvement

The PDF of the eigenvalues of the propagation channel is also related to the channel capacity, as it is shown in [31] for the MIMO scenario. Few works have yet attempted to derive the ergodic capacity of a RIS-aided system. Although [32] has derived the PDF of the eigenvalues of the channel, which is as in the MIMO part of the ergodic capacity derivation of a RIS-aided system, it has only provided an approximation of the pdf for the case where both the transmitter and receiver are equipped with one single antenna. Meanwhile, authors in [33] characterized the capacity limit of point-to- point MIMO-RIS, but when assuming that perfect CSI of both parts of the cascaded channel is independently available at the transmitter and receiver, which is different from the ergodic capacity and particularly difficult to achieve in practice. Meanwhile, we have recently derived an exact expression of the ergodic capacity of MIMO-RIS in [34] by first deriving a closed-form expression of the PDF of the eigenvalues of the MIMO-RIS cascaded channel. Given that MIMO-RIS shares some similarities in terms of system model with MIMO relay systems, especially MIMO-AF systems, we have been inspired by the following works [29, 35], on the ergodic capacity of MIMO relay systems for deriving our expression for MIMO-RIS. This expression is useful for investigating the benefit of MIMO-RIS over MIMO and MIMO-AF [29] in terms of SE in Figure 11. Given that propagation improvement usually translates into received SNR improvement and that the capacity increases logarithmically with the SNR, the ergodic capacity of MIMO-RIS is expected to be significantly better than MIMO or MIMO-AF when considering non-correlated channels, based on Figure 10. For instance, given that the CDF of the eigenvalues of MIMO-RIS is 60 to 90 times better and that the maximum of non-zero eigenvalues of the system is four, the maximum SE difference between MIMO and MIMO-RIS is expected to be around log2 (904) = 26 bit/s/Hz, which is not far from the ~24 bit/s/Hz difference seen at 30 dB in Figure 11. As it was expected, the significant propagation improvement of MIMO-RIS against MIMO, MIMO-AF, and TW-MIMO-AF translates into significant SE gain. Even more, against MIMO-AF, which requires two slots of transmission, and hence has a halved SE in comparison with MIMO-RIS on top of the noise amplification penalty. Meanwhile, TW-MIMO-AF, which can transmit information in two directions within two slots of transmission and hence does not suffer from SE, still performs 18 bit/s/Hz away from MIMO-RIS. Finally, even though the channel correlation at the RIS can significantly decrease its SE performance, it is still better than the existing techniques for the settings considered here. Further works need to be undertaken to model and understand the SE performance limitation due to channel correlation at the RIS.

3.1.3 Energy Efficiency Improvement

As far as the EE of RIS is concerned, most of the existing works [26, 36–37] provide signal processing algorithm, e.g., resource allocation, to make RIS-aided systems more energy efficient. For instance, the work of [26] has developed

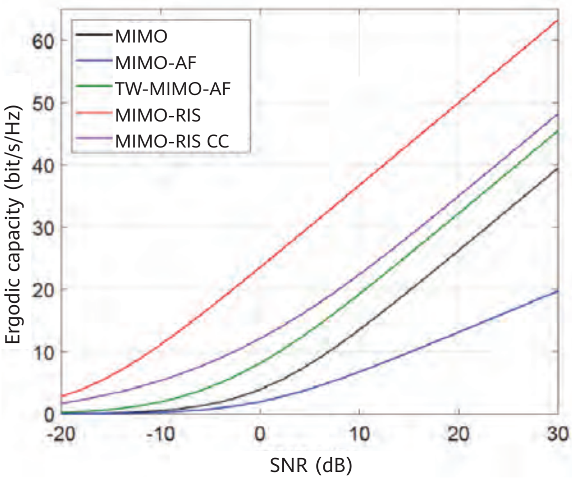


Figure 11 Ergodic capacity of different MIMO-based transmission techniques in the point-to-point scenario

energy-efficient power and phase shift allocation strategies for the downlink of a single-hop multi-user multiple-input single-output (MISO) RIS-aided system. Similar to [26], [36] has designed power and phase shift allocation strategies but for the uplink of a single-hop multi-user MIMO-RIS system, where the trade-off between the EE and SE is studied numerically, i.e., without providing an explicit formulation of the EE as a function of SE. Whereas in [37] the same kind of strategy is developed for single-hop RIS-aided MIMO D2D networks by formulating an EE-based maximization problem. Hardly any works have looked at the fundamental limit of RIS in terms of EE, except perhaps in [38], where upper and lower bounds of the bit-per-Joule capacity are derived, based on an asymptotic expression of the capacity, for a single-hop single-user MISO RIS-aided system with hardware impairment. Consequently, we have recently derived an accurate approximation of the bit-per-Joule capacity of multi-hop MIMO-RIS system in [39], when assuming that all the equipment/device in the system have the same number of antennas, in order to better understand how MIMO-RIS can be useful for improving the EE.

The bit-per-Joule capacity is dependent on both the capacity and the power consumption of a given system, or in other words it represents the trade-off between SE and EE. In comparison with a point-to-point MIMO scenario, where only a BS and UE consume power, a relay-based system or RIS adds an extra node, which also consumes power, such that the power consumption of MIMO-AF or MIMO-RIS is larger than that of MIMO. However, based on the power consumption model of [26–27] for the RIS, the extra consumed power of the RIS is likely to be in the order of Watt, which is way below the power consumption of a large MIMO BS, in the order of hundreds or thousands of Watts. Meanwhile, based on the relay power model of [40], it is clear that the extra power consumption due to RIS will be less than that a traditional relay. Hence, if a RIS is more spectrum-efficient than MIMO as well as MIMO-AF and it consumes less power than MIMO-AF but maybe few percent more than MIMO, the RIS would be more energy-efficient than these techniques, as it is confirmed in Figure 12.

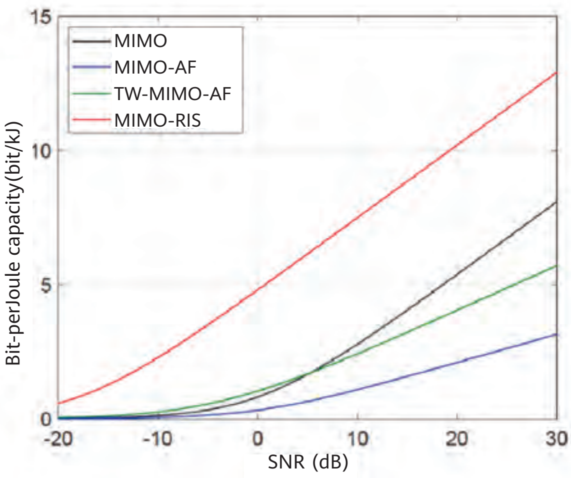


Figure 12 Bit-per-Joule capacity of different MIMO-based transmission techniques in the point-to-point scenario

3.2 Real-World RIS Platform

3.2.1 General Considerations

Future wireless networks are required to interconnect a huge number of online devices with ever-increasing demands for higher data rates especially in a dense urban environment where the presence of a large number of buildings and infrastructure gives rise to harsh propagation conditions for the EM waves. Moreover, to deliver these user demands, we need to progress towards higher frequency bands due to the availability of more bandwidth in these regions of the spectrum. This lands us in the mmWave range which has the potential to facilitate such capacity demands in the order of multi-Gbps data rates. However, mmWave frequencies suffer from high path loss and poor diffraction that lead to poor network coverage especially in the absence of a LOS. Moreover, existing wireless network operators face three significant challenges:

- Lack of seamless connectivity leading to poor quality of service (QoS) especially in harsh propagation environments;

- Supporting billions of online devices with such high data rates which ultimately results in a higher carbon footprint of the network; and
- Uneven user distribution due to various practical challenges in the urban environment leading to an unequal resource utilization at the BSs.

A major reason behind these issues is the lack of control over the wireless radio environment due to the nature of current radio network operation, as only the BS and UE are responsible for handling the effects of the propagation environment. Sensing the environment, recycling existing radio waves and enabling NLOS communication can play a major role in the transformation of existing wireless networks towards a highly efficient and coverage enhanced network which can deliver high QoS and seamless connectivity to a huge number of subscribers. RIS is a core component to address the aforementioned challenges as it offers the capability of manipulating the EM waves towards the direction of interest, as shown in Figure 13. RIS is composed of a large metasurface sheet backed by a phase control unit. The metasurface consists of a number of conductive printed patches (scatterers), where the size of each scatterer is a small proportion of the wavelength of the operating frequency [41]. The macroscopic effect of these scatterers defines a specific surface impedance and by controlling this surface impedance, the reflected wave from the metasurface sheet can be manipulated. Each individual or a cluster of scatterers can be tuned by different phase values in such a way that the whole surface can reconstruct EM waves with desired characteristics without emitting additional radio waves. Moreover, multiple RISs can be easily coated on the building walls, ceilings and windows as required due to their adaptive design [42]. RIS technology can build upon the concept of reconfigurable and software-controlled metasurface that can dynamically manipulate EM waves, resulting in not only reduced coverage holes but also optimized energy consumption [43]. Although, ultra-dense networks can be a solution for coverage enhancement, they can increase the interference level and require backhaul planning along with higher infrastructure management costs. Using co-operative BSs would also require higher density while switching to sub-6 GHz during mmWave coverage outage might solve the coverage issue but would compromise the throughput and reduced QoS due to switching between radio access technologies (RATs). On the contrary, RIS does not suffer from these issues and does not require intense backhaul planning. Another disadvantage of the existing system, which relies on active elements such as relay, is the heightened power consumption and reduced network efficiency. RISs, on the other hand, can be made of smart elements that are not impaired by noise amplification. They are thus capable of controlling the state of individual elements and can sense the environment to cut down power consumption.

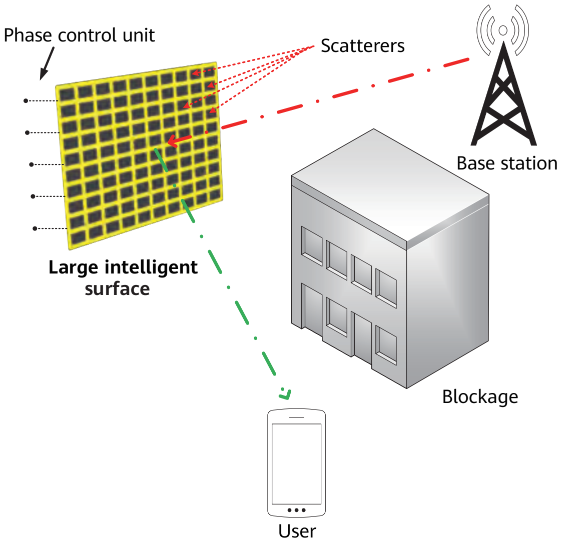


Figure 13 RIS-assisted smart radio network

3.2.2 RIS Design

A RIS is capable of redirecting EM waves to the direction of interest as shown in Figure 14. Similar to the holographic metasurface concept [44], a RIS synthetizes a radiation pattern of interest in a holographic manner [45]. The role of the RIS is to modulate the incident wave on the aperture into a desired aperture field that radiates the radiation pattern of interest. This is achieved by altering the phase of the incident wave with the phase response of each unit cell across the RIS upon reflection. As a result, a RIS can reconfigure the desired radiation pattern in an intelligent and automated manner and eliminate the need for mechanical scanning.

6G networks require much higher data rate with robust and meaningful coverage. RISs will play a pivotal role in this regard where the sporadic waves in an environment can be purposefully recycled and redirected to the network's blind spots.

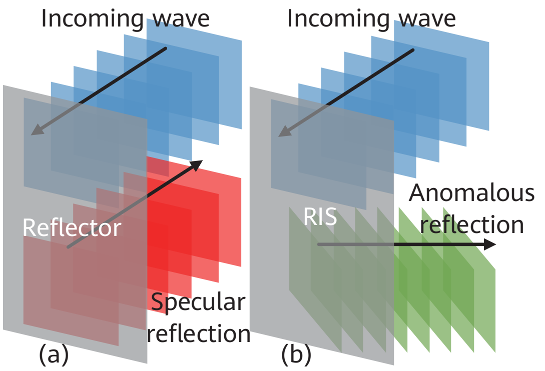


Figure 14 The mechanism of wave reflection in an RIS-enhanced environment. (a) Specular reflection from a conventional reflector; (b) Engineered reflection from an RIS to the desired angle

The design procedure of a RIS is directly linked to the geographical properties of the region where it is going to be deployed. The locations of BSs and the directions of reflected waves will have a direct influence on the constructed patterns of unit cells within the RIS structure. Consequently, the design of a RIS should be customized for each case, which is in contradiction with mass production. One solution to sort this issue out is to equip the RIS with controlling components such as the varactors of PIN diodes to dynamically change the macroscopic response of the RIS without changing its physics. Two types of RISs have been prototyped at 6GIC, as shown in Figure 15 [47]. The structure on the left side of Figure 15 is capable of dynamically changing the reflected beam, thus suitable to be employed in urban areas wherever beam scanning is required. Whereas, the one on the right side of Figure 15 has fixed engineered reflected beam directions but needs no electrical power and works as a stand-alone equipment. It is thus useful either when the direction of the reflected beam does not need to be changed or the provisioning of electrical power is challenging for network operators.

The dynamic RIS contains 2430 unit cells similar to the one illustrated in Figure 16a. The empty spot on the metallic patch (with a size of $w_i \times l_p$) causes the structure to be reactively loaded. This can provide flexibility to manage the unit cell's electrical response by examining different values of the corresponding physical parameters to achieve an optimum response. The varactor diode is placed in the gap between the patches while with a relatively low reverse voltage, this diode provides a specified high capacitance ratio, which makes it an appropriate candidate to regulate the phase response of the unit cell. By varying the reactance value of the varactor diode, the phase of the reflected wave can be tuned in order to reconstruct the beam towards the angle of interest. The response of the proposed unit cell considering the diode is simulated in CST Microwave Studio (CST-MWS) with the results showing in Figure 16b at *f* = 3.5 GHz for capacitance range of *C = 0.5 pF–2 pF*. As shown in this figure, the phase range of variation is 341 degrees with reflection loss of less than -3 dB throughout the entire range of study.

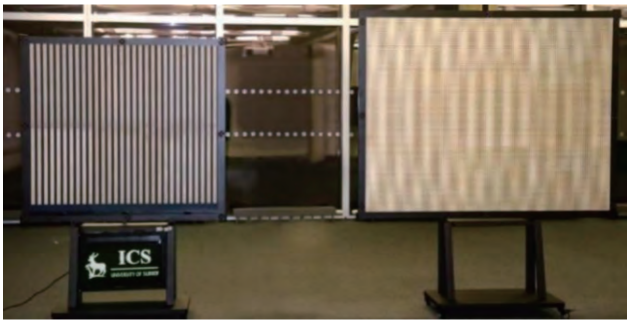


Figure 15 The world's first RIS demo at 6GIC. The left-side one can dynamically change the beam while the one on right has fixed reflected beam with no need of electrical power.

We have demonstrated both dynamic and static RIS at 6GIC building in indoor environment. A video of that demo can be found at [46]. It is observed that the signal level at the receiver side was enhanced by more than 15 dB and 19 dB for dynamic and static RISs respectively.

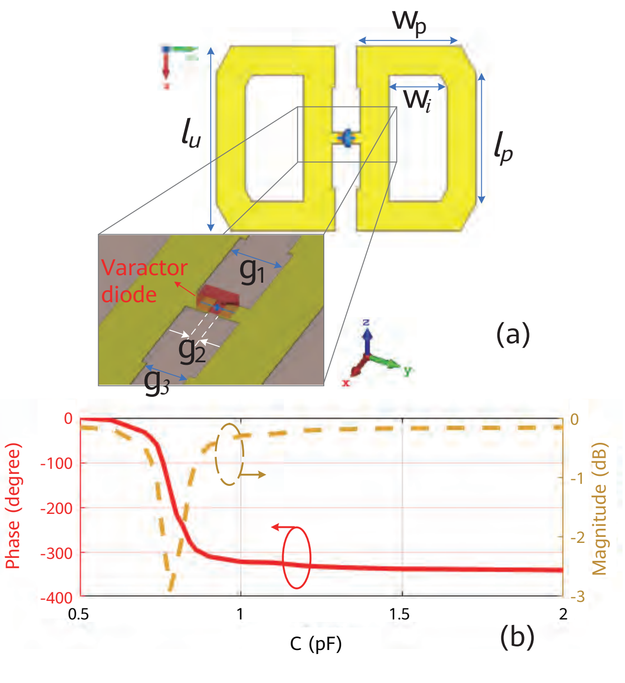


Figure 16 (a) The proposed unit cell for dynamic RIS; (b) Unit cell reflection response at *f* = *3.5* GHz

4 Conclusion

ELAA and RIS, as evolution of mMIMO in 5G NR, will be two important techniques which result in 5–10 times spectrum efficiency of 5G and ultimate coverage for B5G/6G, by significantly increased aperture at the base station and manipulating the reflected wave to a desired direction using a low-cost metasurface at a new node.

The spatial non-stationary statistical channel of ELAA is captured in a novel channel model, where not all the antennas at the base station are visible to UEs. The RSS data generated from the channel model matches well with that from real-world ELAA channel measurements.

It is further shown that the signal processing techniques, which are scalable in mMIMO, are no longer scalable in ELAA due to the spatial non-stationary channel. The channel hardening effect could be largely weakened due to the presence of non-stationary shadowing. Hence there is a clear performance gap for the MF compared over the optimal decoder in the ELAA. The performance gap to the optimal receiver is considerably large even for LMMSE receivers, which may be already complexity-impractical in ELAA. Hence, one future research topic would be new scalable signal processing techniques tailored for ELAA non-stationary channels, close to optimal performance while practical in implementation complexity.

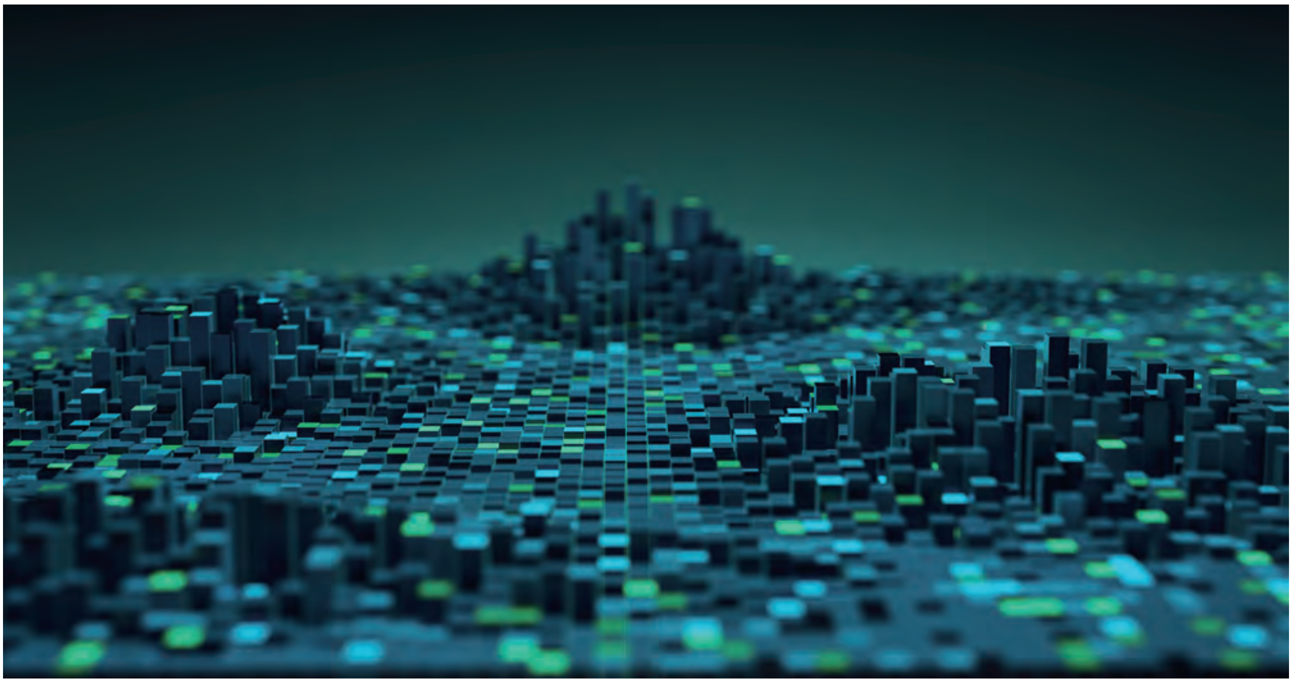
In case sub-array ELAA transmit precoding is applied where a single UE is associated with only one strongest ELAA subarray for simplicity, multi-antenna receiver interference mitigation, in conjunction with widely linear processing, offers clear performance gain especially in case of IQI.

Thanks to the richer channel and the cascaded channel effect, RIS brings substantial gains over MIMO and MIMO-AF based on the analysis of fundamental limits. The improvement of RIS in terms of the channel eigenvalue CDF against both MIMO and MIMO-AF is of more than one order of magnitude in case of independent channels. The spectrum efficiency improvement of RIS over MIMO is ~24 bit/s/Hz at 30 dB SNR, and even larger over MIMO-AF. The energy efficiency improvement if RIS over MIMO and MIMO-AF is 5 bit/KJ and 10 bit/KJ at 30 dB SNR, respectively.

Both dynamic and static RIS at C-band have been demonstrated in a real-world indoor environment, which enhances the signal level at the receiver side by more than 15 dB and 19 dB respectively. This helps recover the video connection that is lost due to blockage and keeps the video connection stable even with mobility.

Acknowledgment

The work was supported by Huawei-the University of Surrey Joint Lab.



References

[1] W. Tong and P. Zhu, "6G: The next horizon: From connected people and things to connected intelligence," *Cambridge University Press*, 2021

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

[3] E. Björnson, L. Sanguinetti, H. Wymeersch, J. Hoydis, and T. L. Marzetta, "Massive MIMO is a reality—What is next? Five promising research directions for antenna arrays," *Digit. Signal Process.*, vol. 94, pp. 3-20, Nov. 2019.

[4] E. Bjornson, H. Wymeersch, B. Matthiesen, P. Popovski, L. San-guinetti, and E. de Carvalho, "Reconfigurable intelligent surfaces: A signal processing perspective with wireless applications," *arXiv preprint arXiv:2102.00742, 2021*

[5] Y. Han, S. Jin, C. Wen, and X. Ma, "Channel estimation for extremely large-scale massive MIMO systems," *IEEE Wireless Commun. Lett.*, vol. 9, no. 5, pp. 633-637, May 2020.

[6] METIS, "METIS channel models," Mobile and wireless communications enablers for twenty-twenty information society (METIS), 02 2015.

[7] A. Amiri, M. Angjelichinoski, E. De Carvalho, and R. W. Heath, "Extremely large aperture massive MIMO: Low complexity receiver architectures," in *Proc. IEEE GLOBECOM Workshops (GC Wkshps)*, 2018, pp. 1-6.

[8] A. Ali, E. De Carvalho, and R. W. Heath, "Linear receivers in non-stationary massive MIMO channels with visibility regions," *IEEE Wireless Commun. Lett.*, vol. 8, no. 3, pp. 885-888, Jun. 2019.

[9] X. Yang, F. Cao, M. Matthaiou, and S. Jin, "On the uplink transmission of extra-large scale massive MIMO systems," *IEEE Trans. Veh. Technol.*, vol. 69, no. 12, pp. 15 229-15 243, Dec. 2020.

[10] V. C. Rodrigues, A. Amiri, T. Abrao, E. De Carvalho, and P. Popovski, "Low-complexity distributed XL-MIMO for multiuser detection," in *Proc. IEEE Int. Conf. Commun. Workshops (ICC Wkshps)*, 2020, pp. 1-6.

[11] S. Wu, C. Wang, e. M. Aggoune, M. M. Alwakeel, and X. You, "A general 3-D non-stationary 5G wireless channel model," *IEEE Trans. Commun.*, vol. 66, no. 7, pp. 3065-3078, Jul. 2018.

[12] J. Flordelis, X. Li, O. Edfors, and F. Tufvesson, "Massive MIMO extensions to the COST 2100 channel model: Modeling and validation," *IEEE Trans. Wireless Commun.*, vol. 19, no. 1, pp. 380-394, Jan. 2020.

[13] J. Wang, C. Wang, J. Huang, and H. Wang, "A novel 3D space-time frequency non-stationary channel model for 6G THz indoor communication systems," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC), 2020*, pp. 1-7.

[14] A. Amiri, S. Rezaie, C. N. Manchon, and E. De Carvalho, "Distributed receivers for extra-large scale MIMO arrays: A message passing approach," *arXiv: Signal Process.*, 2020.

[15] J. Bian, C.-X. Wang, X. Gao, X. You, and M. Zhang, "A general 3D non-stationary wireless channel model for 5G and beyond," *arXiv: Signal Process.*, 2021.

[16] J. Liu, Y. Ma, J. Wang, N. Yi, S. Xue, R. Tafazolli, and F. Wang, "A non-stationary channel model with correlated NLoS/LoS states for ELAA-mMIMO", *IEEE Globecom* 2021, p 1-6.

[17] 3GPP, "Study on channel model for frequencies from 0.5 to 100 GHz," 3rd Generation Partnership Project (3GPP), Technical Specification (TS) 38.901, Dec. 2019, version 16.1.0.

[18] M. Hochwald, T. L. Marzetta, and V. Tarokh, "Multiple-antenna channel hardening and its implications for rate feedback and scheduling," *IEEE Trans. Inf. Theory*, vol. 50, no. 9, pp. 1893-1909, Sep. 2004.

[19] À. O. Martínez, E. De Carvalho and J. Ø. Nielsen, "Towards very large aperture massive MIMO: A measurement based study," in *Proc. IEEE GLOBECOM Workshops*, 2014, pp. 281-286.

[20] A. Ali, E. de Carvalho, and R. W. Heath Jr., "Linear receivers in non-stationary massive MIMO channels with visibility regions," IEEE Wireless Communications Letters, vol. 8, no. 3, pp.885-888, June 2019.

[21] D. Astely and B. Ottersten, "Spatiotemporal interference rejection combining," in Smart Antennas: State of the Art. Ed. By T. Kaiser, et al., Hindawi Publishing Corp., 2005.

[22] B. Picinbono and P. Chevalier, "Widely linear estimation with complex data," *IEEE Trans. Signal Processing*, vol. 43, pp. 2030-2033, Aug. 1995.

[23] M. C. Wells, "Increasing the capacity of GSM cellular radio using adaptive antennas," *IEEE Proc. Communications*, vol. 143, n.5, pp. 304-310, 1996.

[24] O. Ozdogan, E. Bjornson, and E. Larsson, "Intelligent reflecting surfaces: Physics, propagation, and pathloss modeling," IEEE Wireless Communications Letters, vol. 9, pp. 581-585, 2020.

[25] L. M. Correia *et al.*, "Challenges and enabling technologies for energy aware mobile radio networks," *IEEE Commun. Mag.*, vol. 48, no. 11, pp. 66-72, Nov. 2010.

[26] C. Huang, A. Zappone, G. C. Alexandropoulos, M. Debbah, and C. Yuen, "Reconfigurable intelligent surfaces for energy efficiency in wireless communication," *IEEE Trans. Wireless Commun.*, vol. 18, no. 8, pp. 4157-4170, Aug. 2019.

[27] L. You, J. Xiong, D. W. K. Ng, C. Yuen, W. Wang, and X. Gao, "Energy efficiency and spectral efficiency tradeoff in RIS-aided multiuser MIMO uplink transmission," *IEEE Trans. Signal Process.*, vol. 69, pp. 1407- 1421, 2021.

[28] E. Bjornson, O. Ozdogan, and E. Larsson, "Intelligent reflecting surface versus decode-and-forward: How large surfaces are needed to beat relaying?" IEEE Wireless Communications Letters, vol. 9, pp. 244-248, 2020.

[29] S. Jin, M. R. McKay, C. Zhong, and K.-K. Wong, "Ergodic capacity analysis of amplify-and-forward MIMO dual-hop systems," IEEE Transactions on Information Theory, vol. 56, no. 5, pp. 2204-2224, 2010.

[30] K.-J. Lee, K. Lee, H. Sung, and I. Lee, "Sum-rate maximization for two-way MIMO amplify-and-forward relaying systems," in *Proc. IEEE VTC-Spring*, Barcelona, Spain, Apr. 2009.

[31] I. E. Telatar, "Capacity of Multi-antenna Gaussian Channels," *Europ. Trans. Telecommun. and Related Technol.*, vol. 10, no. 6, pp. 585-596, Nov. 1999.

[32] A. A. A. Boulogeorgos and A. Alexiou, "Ergodic capacity analysis of reconfigurable intelligent surface assisted wireless systems," 2020 IEEE 3rd 5G World Forum, 5GWF 2020 - Conference Proceedings, vol. 0, pp. 395-400, 2020.

[33] S. Zhang and R. Zhang, "Capacity characterization for intelligent reflecting surface aided MIMO communication," IEEE Journal on Selected Areas in Communications, vol. 38, pp. 1823-1838, 2020.

[34] M. A. Mosleh, F. Héliot, and R. Tafazolli, "Ergodic capacity analysis of large intelligent surface assisted MIMO systems," in IEEE Globecom Workshop on RIS for Future Wireless Com., Madrid, Spain, Dec. 2021.

[35] H. Shin, M. Win, J. Lee, and M. Chiani, "On the capacity of doubly correlated MIMO channels," IEEE Transactions on Wireless Communications, vol. 5, no. 8, pp. 2253-2265, 2006.

[36] L. You, J. Xiong, D. W. K. Ng, C. Yuen, W. Wang, and X. Gao, "Energy efficiency and spectral efficiency tradeoff in RIS-aided multiuser MIMO uplink transmission," *IEEE Trans. Signal Process.*, vol. 69, pp. 1407-1421, 2021.

[37] S. Jia, X. Yuan, and Y.-C. Liang, "Reconfigurable intelligent surfaces for energy efficiency in D2D communication network," *IEEE Wireless Commun. Lett.*, vol. 10, no. 3, pp. 683-687, Mar. 2021.

[38] Y. Liu, E. Liu, and R. Wang, "Energy efficiency analysis of intelligent reflecting surface system with hardware impairments," in *Proc. IEEE Globecom*, Taipei, Taiwan, Dec. 2020.

[39] F. Héliot, M. A. Mosleh, and R. Tafazolli, "Closed-form approximation of the EE-SE trade-off for multi-hop MIMO-LIS communication systems," submitted to Globecom Workshop on RIS for Future Wireless Com., Jul. 2021.

[40] Y. Qi, F. Héliot, and M. A. Imran, "Green relay techniques in cellular systems," in *Green communications and networking*, F. R. Yu, X. Zhang, and V. C. Leung, Eds. Boca Raton, FL: CRC press, Dec. 2012, ch. 3.

[41] K. Achouri, M. A. Salem, and C. Caloz, "General

Metasurface Synthesis Based on Susceptibility Tensors," *IEEE Trans. Antennas Propag*, vol. 63, no. 7, pp. 2977-2991, July 2015.

[42] C. Liaskos, S. Nie, A. Tsioliaridou, A. Pitsillides, S. Ioannidis, and I. Akyildiz, "A new wireless communication paradigm through software-controlled metasurfaces," *IEEE Commun. Mag*. vol. 56, no. 9, p. 162-169, 2018.

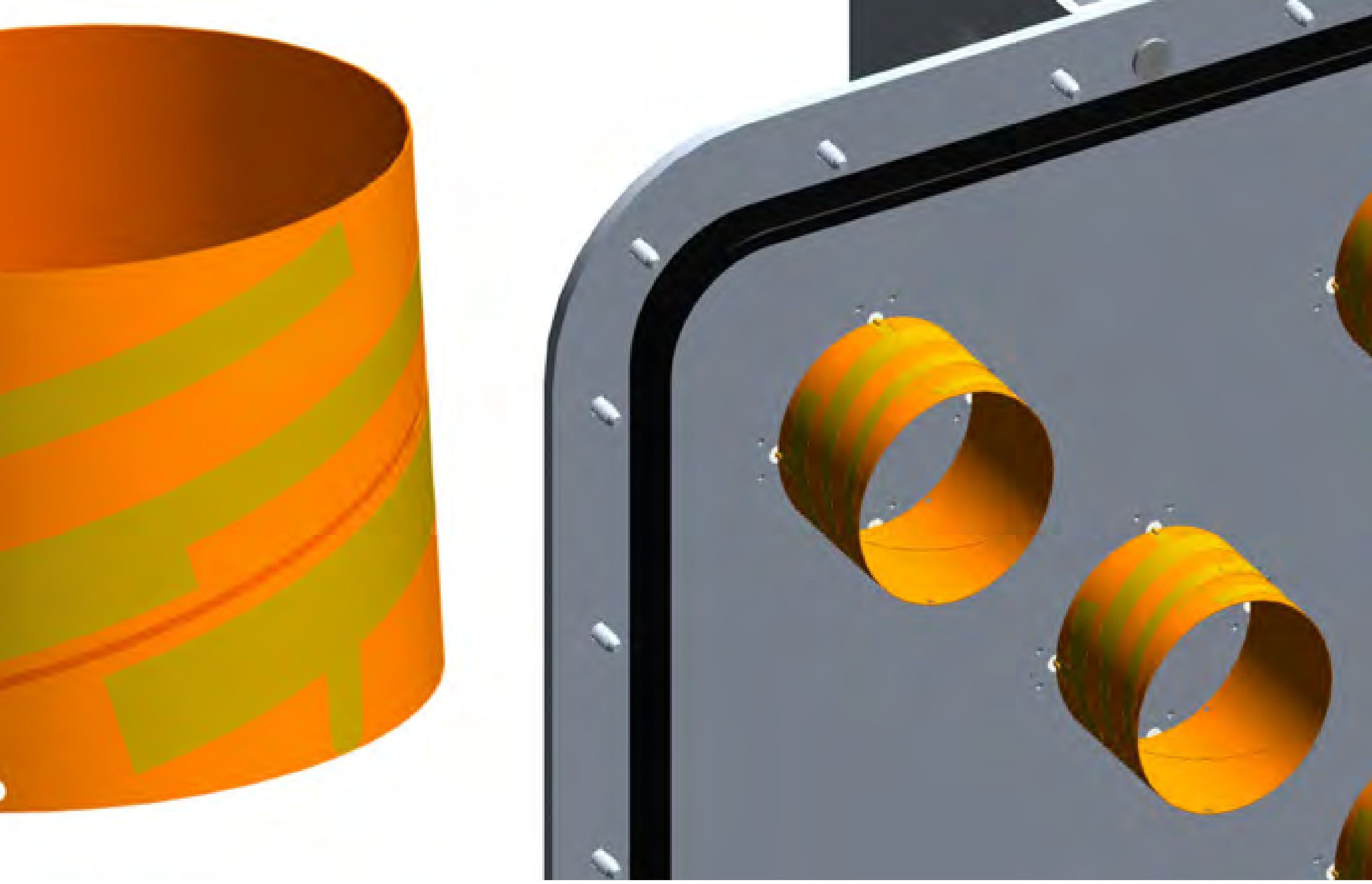
[43] M. Renzo *et al*., "Smart radio environments empowered by reconfigurable AI meta-surfaces: An idea whose time has come," *EURASIP J Wirel Comm*, 2019.

[44] A. Araghi, M. Khalily, P. Xiao, and R. Tafazolli, "Holographic-based leaky-wave structures: Transformation of guided waves to leaky waves," IEEE Microwave Magazine, vol. 22, no. 6, pp. 49-63, 2021.

[45] A. Araghi, M. Khalily, P. Xiao, and R. Tafazolli, "Holographic-based mmw-wideband bidirectional frequency scanning leaky wave antenna," in 14th European Conference on Antennas and Propagation (EuCAP), 2020.

[46] University of Surrey, "Reconfigurable reflecting surfaces for 5G/6G," https://www.youtube.com/watch?v=PoLEWaDg8f4

[47] A. Araghi, Mohsen Khalily, M Safaei, A Bagheri, V Singh, F Wang, and R Tafazolli, "Reconfigurable intelligent surface (RIS) in the sub-6 GHz band: Design, implementation, and real-world demonstration," in IEEE Access, vol. 10, pp. 2646-2655, 2022, doi: 10.1109/ACCESS.2022.3140278.



Quadrifilar Helix Antenna (QHA) — Enabling Highly Efficient Massive MIMO for 5G and Beyond

Fayez Hyjazie [1], Tong Wen [1], Dageng Chen [1], Bijun Zhang [2], Bojie Li [3], Huangping Jin [2], Guanxi Zhang [3], Weihong Xiao [4]

[1] Ottawa Wireless Advanced System Competency Centre

[2] Wireless Network Research Dept

[3] Wireless Network System Technology Innovation Research Dept

[4] Base Station Platform Dept

Abstract

With a compact and MIMO antenna design, Quadrifilar Helix Antenna (QHA) provides spatial multiplexing gain with more ports and more polarization diversity and offers great potential to improve the performance of massive MIMO with limited antenna size. This paper attempts to investigate QHA from multiple aspects, including antenna design, array mapping, polarization analysis, and system performance evaluation. This paper also proposes a novel analysis method to establish the association between antenna array design and system performance. The simulation results and field trials show that QHA can potentially enable a large system capacity improvement and enrich the massive MIMO antenna design collection.

Keywords

Quadrifilar Helix Antenna (QHA), Massive MIMO, 5G

1 Introduction

Massive multiple input and multiple output (MIMO) is a key technology for 5G and beyond. The merits of spatial degrees of freedom leads to large system capacity compared with single input and single output (SISO) [1–2]. Both the academia and the industry have already conducted lots of research on this [2]. Antenna design technology and phased array are milestones in the MIMO system development. For a wireless multi-user (MU) system, spatial resolution is the key factor in exploring the performance boundary of the MIMO system [3–8]. However, in wireless industries, the physical antenna size is limited considering the wind load, weight and other deployment-related issues. As a key part of the wireless transmission link, antenna design has significant impact on the entire system design. Generally, the antenna aperture size is proportional to its beaming gain or directionality, which also represents its spatial resolution [9–15]. It is a well-known fact that the physical antenna aperture size constrains the MIMO system performance.

In addition to the antenna aperture size, polarization diversity and antenna pattern diversity have been exploited in order to decrease the signal correlation of different antenna elements [6–7]. A traditional antenna unit consists of two antenna elements with ±45 polarizations.

Quadrifilar Helix Antenna (QHA), as a compact and multiport antenna, provides more ports and more polarization diversity and offers great potential to push the performance boundary of MIMO system even further [16]. QHA is already being widely used in many applications, including long distance communication, astronomy, and GPS. The main benefits of QHA antenna are its wide band and excellent circular polarization (CP), with easily obtained high antenna gain due to four ports being properly fed. Featuring an antenna design most suited for CP-demand applications, it is rarely mentioned in wireless industries.

This paper discusses the reconstruction of QHA, with its four ports being fed separately and each port behaving like a meandered monopole, providing more polarization diversity that is beneficial for the MIMO system.

This paper is organized into five parts. In part II, the original QHA design will be reviewed, and the newly designed QHA will be proposed and its design principle provided. Based on the newly designed QHA, the MIMO antenna array is then reconstructed. In part III, different antenna array designs are explored and evaluated. In part IV, we will look at the field trials that were conducted to prove the performance improvement by the newly designed QHA. Some key test configurations, results and observations are discussed. Part V of this paper contains the summary and an introduction to future work.

2 Fundamentals of QHA

2.1 Helix Antenna

QHA antenna originates from the helix antenna, which was invented by John D. Kraus in 1947. Helix antenna is a simple implementation of a helical wire that is above a ground plane, as shown in Figure 1. It has many advantages, including stable polarization, wideband operation, and high gain. For a given helix antenna, its input impedance remains nearly consistent over wideband frequencies and it also radiates like an end fire antenna, as shown in Figure 2 and Figure 3. Its performance is not sensitive to the radius of the helical wire and pitch spacing. It has small mutual impedance that can be neglected, which makes it suitable for arrays with minimum consideration of its mutual coupling effect among array elements.

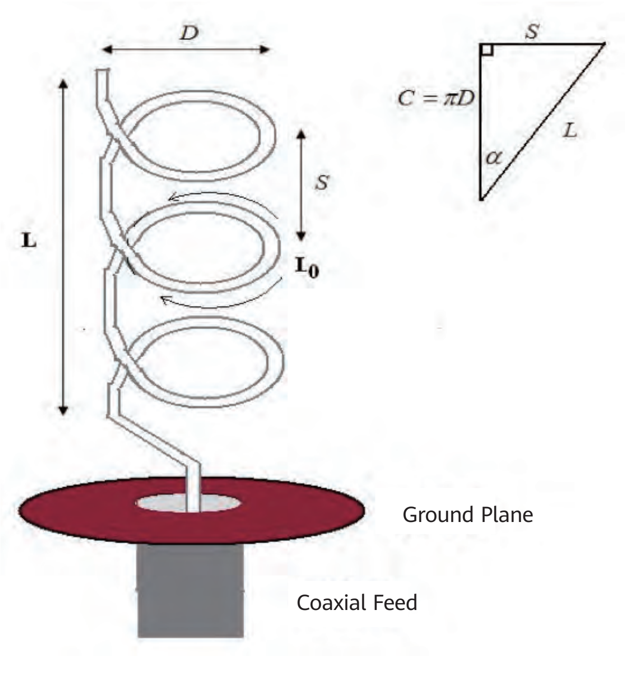
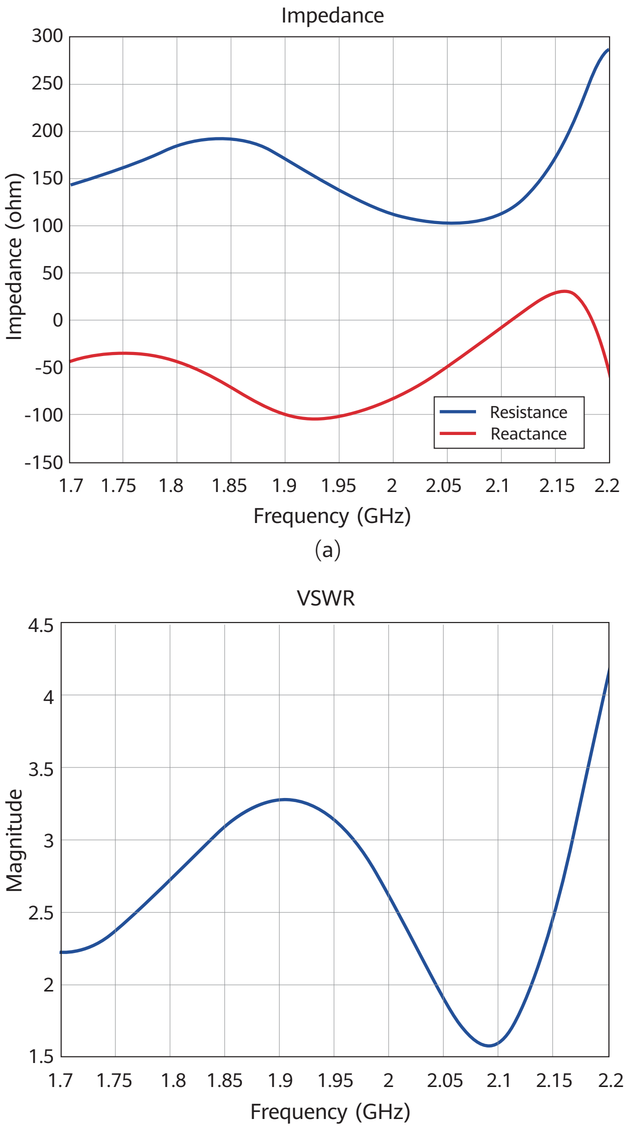


Figure 1 Helix antenna



(a)

(b)

Figure 2 (a) Impedance of helix antenna; (b) VSWR of the helix antenna

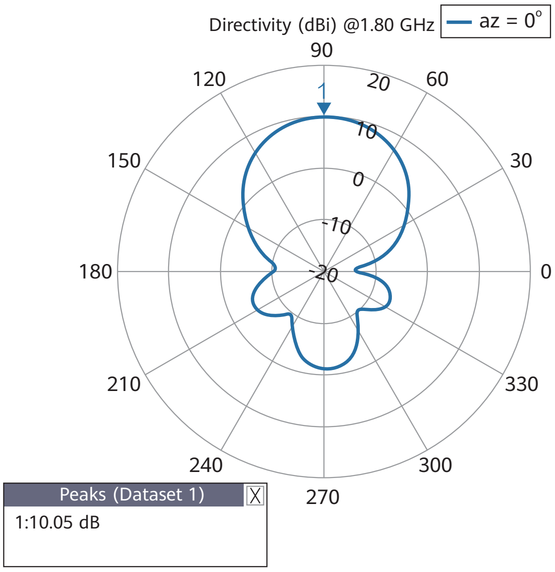
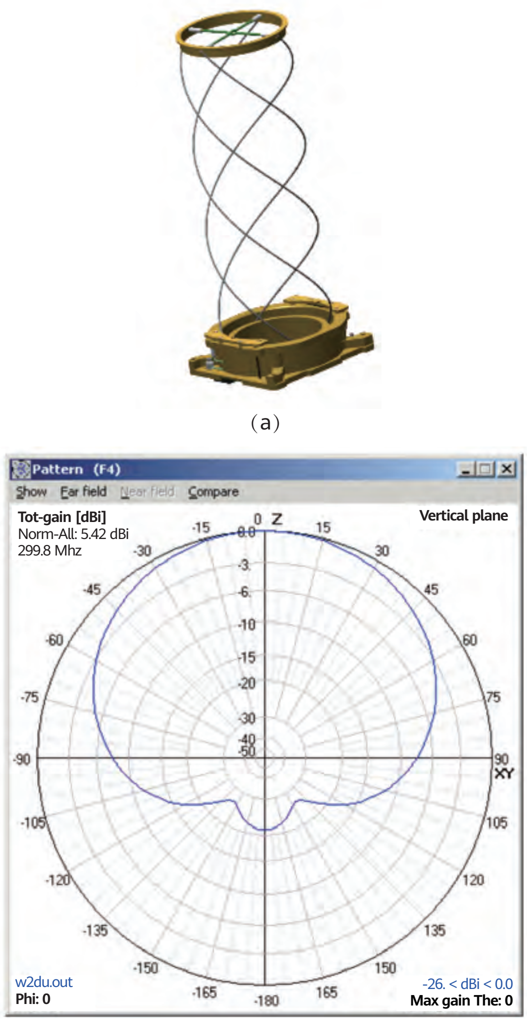


Figure 3 Pattern of the helix antenna

Due to its circular polarization property, high antenna gain, and mechanical simplicity, helix antenna has been widely used in many applications, including radio telescope antenna arrays, in nearly all kinds of satellite antenna solutions, and even as a powerful receiving antenna in ground terminals and earlier mobile phones.

2.2 QHA Antenna Evolution

Original QHA consists of four ports (arms) that are fed jointly, as shown in Figure 4a. Considering its high antenna gain, purity of circular polarization, and wideband operation, QHA is a natural evolution in design from the single helix antenna. Four arms provide more feeding balance than a single arm, and form a more stable pattern with clear circular polarization. Figure 4 shows a QHA antenna with four arms combined featuring wideband, high antenna gain and circular polarization operation.



(a)

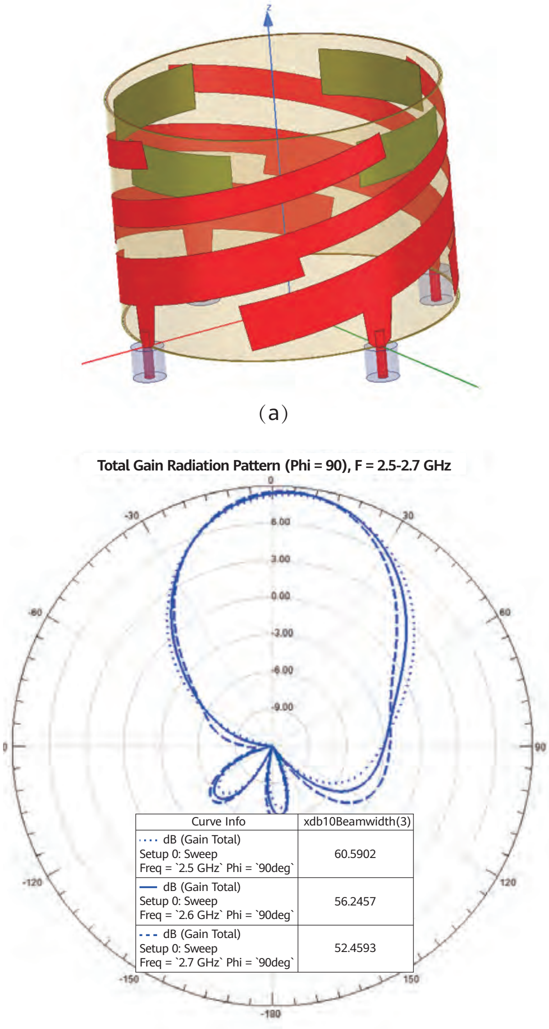
(b)

Figure 4 (a) QHA antenna with four arms combined;
(b) QHA radiation pattern

However, the original QHA design is not suitable for wireless communication applications. As mentioned earlier, the physical antenna aperture size for wireless communication is limited, and the ultimate goal of antenna design is to improve the freedom of the MIMO system.

With proper design, the four arms of a QHA can be fed independently with a broader radiation pattern. Figure 5 shows an exquisitely designed QHA antenna with four arms working separately. The pattern of each arm of the newly designed QHA (with four arms fed separately) is not a good shape compared with the conventional antenna pattern.

For the design in Figure 5, the polarization and beam direction properties are different for different arms, and the design provides more freedom than the conventional MIMO antenna array (with a unified property for each element), providing a high volume of independent paths in a rich scattering wireless channel.



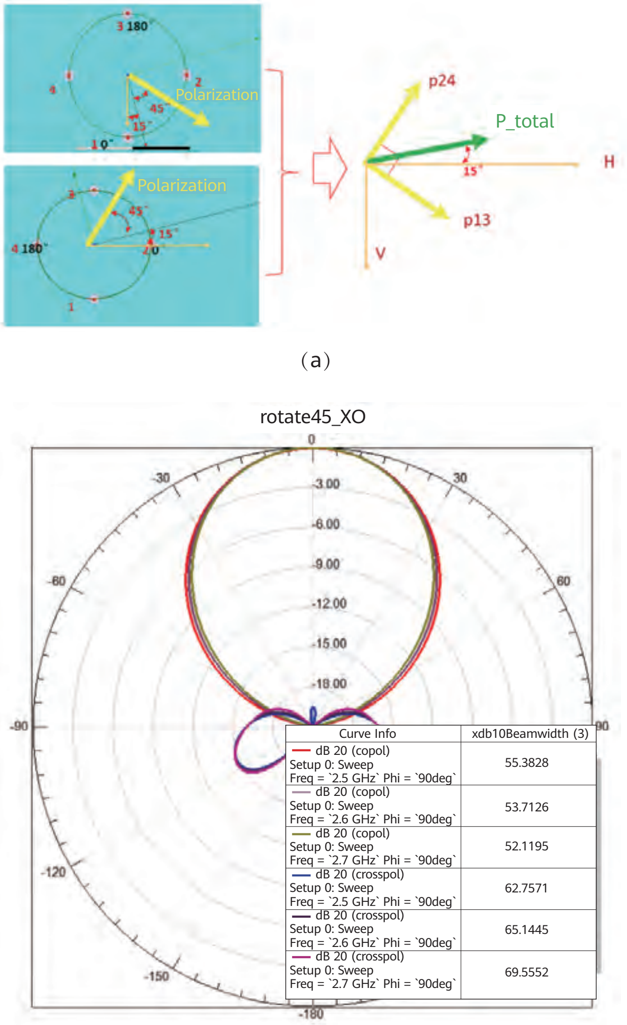
(a)

Curve Info	xdb10Beamwidth(3)
····· dB (Gain Total) Setup 0: Sweep Freq = `2.5 GHz` Phi = `90deg`	60.5902
—— dB (Gain Total) Setup 0: Sweep Freq = `2.6 GHz` Phi = `90deg`	56.2457
- - - dB (Gain Total) Setup 0: Sweep Freq = `2.7 GHz` Phi = `90deg`	52.4593

(b)

Figure 5 (a) QHA with four arms working independently;
(b) Broad side radiation pattern

Regarding the new QHA design shown in Figure 5, if we take advantage of its multi-port and flexibility, we can obtain a really high capacity of different polarizations. Figure 6a shows a polarization analysis of the QHA with the excitation of different ports, where each port or corresponding helix antenna behaves like a broadside radiation antenna with line polarization. We can merge two of the four ports to form a pure linear polarization, as shown in Figure 6b. It can be seen that with different excitations the QHA will have different polarization directions. If we consider two orthogonal polarizations as the basis, then we can synthesize any polarization we want without changing the design of the QHA. This is quite useful when the antenna works in different applications, which will suit specific polarization properties.



(a)

Curve Info	xdb10Beamwidth (3)
—— dB 20 (copol) Setup 0: Sweep Freq = `2.5 GHz` Phi = `90deg`	55.3828
—— dB 20 (copol) Setup 0: Sweep Freq = `2.6 GHz` Phi = `90deg`	53.7126
—— dB 20 (copol) Setup 0: Sweep Freq = `2.7 GHz` Phi = `90deg`	52.1195
—— dB 20 (crosspol) Setup 0: Sweep Freq = `2.5 GHz` Phi = `90deg`	62.7571
—— dB 20 (crosspol) Setup 0: Sweep Freq = `2.6 GHz` Phi = `90deg`	65.1445
—— dB 20 (crosspol) Setup 0: Sweep Freq = `2.7 GHz` Phi = `90deg`	69.5552

(b)

Figure 6 (a) Polarization analysis of the QHA;
(b) Two ports excited merged pattern

MIMO system performance essentially depends on the antenna radiation pattern, which consists of the amplitude pattern and phase pattern. Amplitude pattern is related to the beamforming gain of MIMO antenna array, whereas the phase pattern is related to the angular resolution of the MIMO antenna array.

QHA was not designed to be used in base stations, but it can be made more suitable for wireless commutation systems. Figure 7 shows the evolution process of QHA, from a 4-port combined circularly polarized antenna to a split 4-port helix antenna. With the aim of improving the pattern directivity, isolation and wideband operation, a totally different design is proposed, which we call the 'QHA-inspired antenna'. The optimized design uses a novel spoof structure to realize wideband and high isolation, and has been tested in different scenarios yielding very good, stable performance.

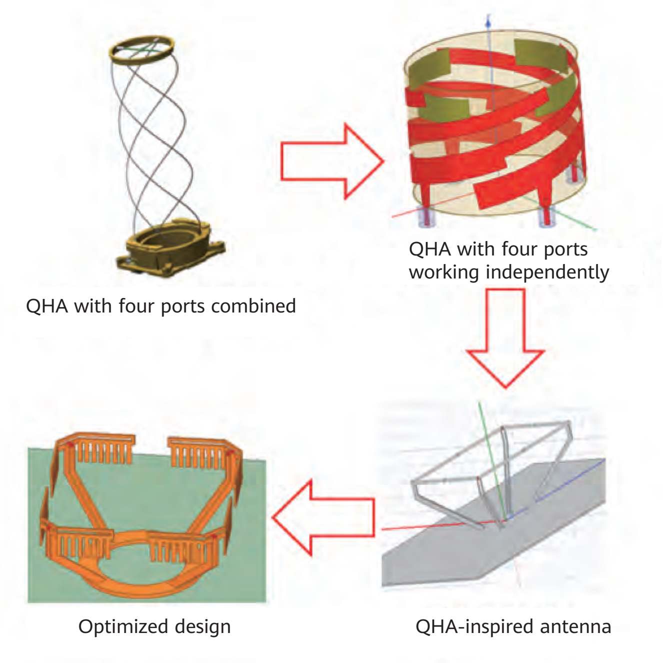

QHA with four ports combined
QHA with four ports working independently
Optimized design
QHA-inspired antenna

Figure 7 Evolution of QHA

2.3 Analysis on the Vector Radiation Pattern of QHA

Figure 8 shows a simple diagram of a 4-element array. We can get the phase lag between different elements produced by the scanning angle, as shown in Figure 9 with amplitude pattern in (a) and phase pattern in (b). If we consider one element within the array as the far-field phase center for the phase patterns of all the elements, we find that each phase pattern will have a different phase plane at a certain direction, and this is where the phase difference in Figure 9 originates. For each direction, the far-field radiation gains for different elements are almost the same. From the angular resolution perspective, the bigger the phase difference of each phase pattern, the higher the array spatial resolution. Correspondingly, this could bring more spatial multiplexing gain to the MIMO system.

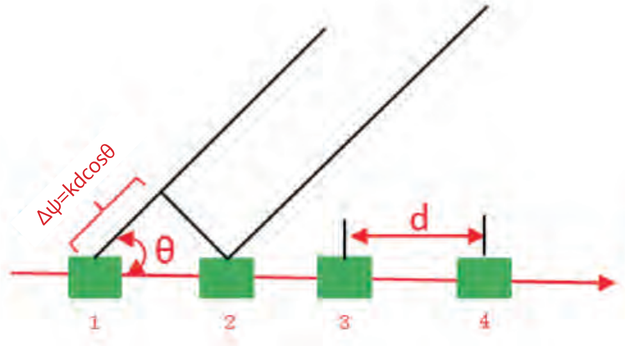


Figure 8 Schematic diagram of a 4-element array

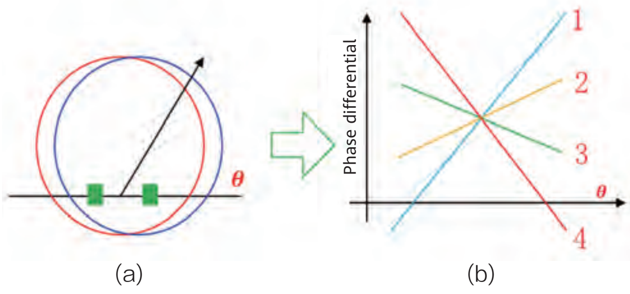


Figure 9 Relative phase differential of four elements

Generally, we consider the beam width of the transmitted signal as the benchmark of angular resolution of the array. We can analyze how the phase and amplitude pattern impact the beam width of the transmitted signal.

We assume the LOS scenario and 2-element array as shown in Figure 9. The amplitude pattern is A and the radiating power for angle θ is A(θ). The phase pattern is ϕ and the radiating phase for angle θ is ϕ(θ). Then we can get the transmitted signal for target angle θ_0:

$$\|hp\|^2 = \left\|hh(\theta_0)^H\right\|^2 = \left\| \left[\sqrt{A(\theta)}e^{j\phi_1(\theta)} \quad \sqrt{A(\theta)}e^{j\phi_2(\theta)}\right] \frac{1}{\sqrt{P}} \begin{bmatrix} \sqrt{A(\theta_0)}e^{j\phi_1(\theta_0)} \\ \sqrt{A(\theta_0)}e^{j\phi_2(\theta_0)} \end{bmatrix}^H \right\|^2$$
$$= \frac{1}{P}\left\| \sqrt{A(\theta)A(\theta_0)} \cdot e^{j\left(\phi_1(\theta) - \phi_1(\theta_0)\right)} + \sqrt{A(\theta)A(\theta_0)} \cdot e^{j\left(\phi_2(\theta) - \phi_2(\theta_0)\right)} \right\|^2 \quad (1)$$

Which also can be derived as

$$\|hp\|^2 = \frac{1}{P}A(\theta)A(\theta_0) + \frac{2}{P}A(\theta)A(\theta_0)\cos\left[\left(\phi_1(\theta) - \phi_2(\theta)\right) - \left(\phi_1(\theta_0) - \phi_2(\theta_0)\right)\right] \quad (2)$$

If we set $\triangle(\theta)=\phi_1(\theta)-\phi_2(\theta)$ as the phase differential between different array elements, then we get

$$\frac{\partial\|hp\|^2}{\partial\theta}= \frac{1}{P}A(\theta_0)\frac{\partial A(\theta)}{\partial\theta}+\frac{2}{P}A(\theta_0)\frac{\partial A(\theta)}{\partial\theta}\cos(\Delta(\theta)-\Delta(\theta_0)) -\frac{2}{P}A(\theta)A(\theta_0)\sin(\Delta(\theta)-\Delta(\theta_0))\frac{\partial\Delta(\theta)}{\partial\theta} \tag{3}$$

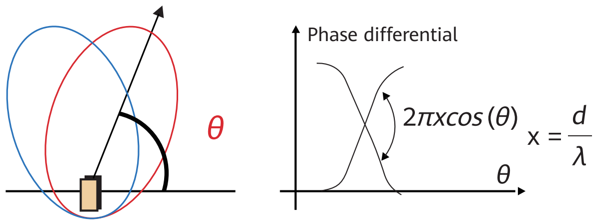
Without loss of generality, if we take $\triangle(\theta_0)=0$ as an example, which is the broadside direction of the array, then we get

$$\frac{\partial\|hp\|^2}{\partial\theta}=\frac{1}{P}\quad(\theta_0)\frac{\partial A(\theta)}{\partial\theta}(1+2\cos(\Delta(\theta))) -\frac{2}{P}A(\theta)A(\theta_0)\sin(\Delta(\theta))\frac{\partial\Delta(\theta)}{\partial\theta} \tag{4}$$

In the preceding equation (4), the first term in (4) indicates the beam direction and width of the amplitude pattern, and the second term in (4) indicates the phase differential of the two phase planes. This indicates that both the amplitude and phase have an impact on the array resolution. Normally, the amplitude pattern should be carefully designed to guarantee the cell coverage performance and cannot be modified arbitrarily. The most likely solution for improving the array resolution is to increase the effective phase difference between antenna elements.

2.4 QHA Radiation Pattern Impact on System Performance Evaluation

In order to fully illustrate the impact of QHA antenna radiation pattern on system performance, simulation for several typical cases was implemented. As stated previously, the amplitude pattern radiation direction/beam width and phase pattern are the key aspects that impact system performance. Figure 10a shows the illustration of beam direction/width, and Figure 10b shows the illustration of phase differential calculation.

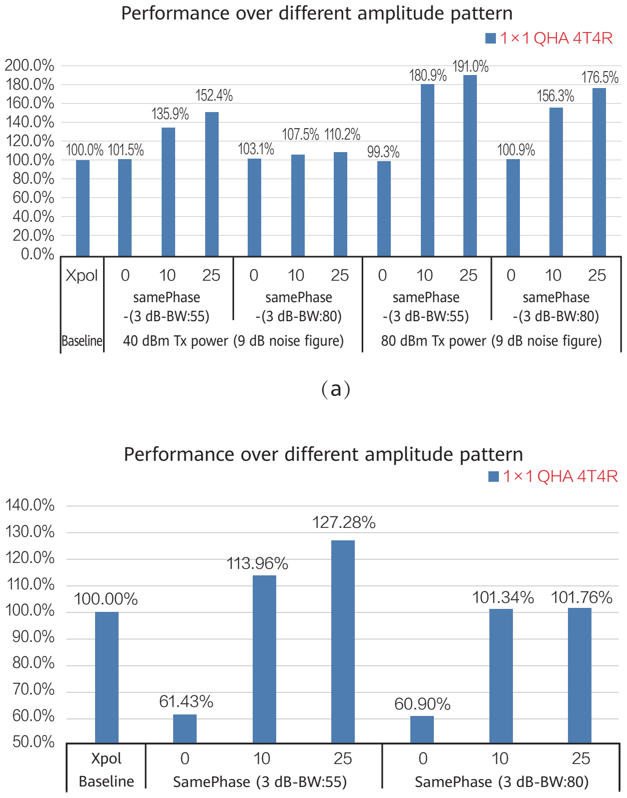


(a) (b)

Figure 10 (a) Diagram of amplitude radiation pattern with max radiation direction shifted; (b) Phase differential calculation description

Figure 11 is an illustration of the performance difference of different amplitude pattern radiation directions with fixed beam width, noise level, and radiation power. In Figure 11, 3 dB-BW means the beam width within 3 dB loss. Only one QHA element was investigated with 4T4R for a base station. It can be seen that the amplitude radiation direction shift contributes to system performance, even in the case of significantly different beam width. In the case of Figure 11a, only single cell was considered. In the case of Figure 11b, multi-cell simulation was conducted, which introduced significant interference between different or adjacent cells.

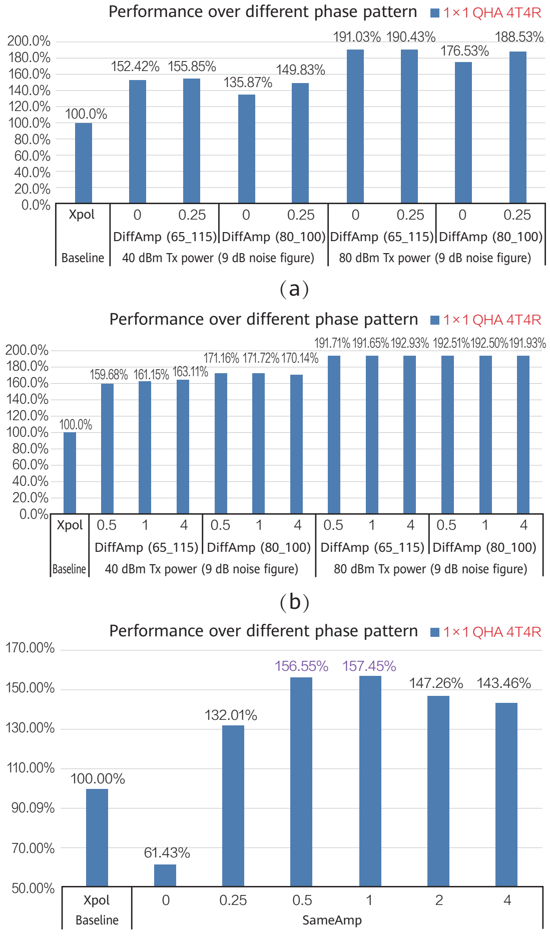
In the single cell scenario, the bigger the shift in the radiation, the higher the throughput performance. However, wider beam width makes the performance less sensitive to the shift in the radiation direction. The conclusion for the multi-cell scenario is the same as that for the single cell scenario. This implies that the amplitude pattern may not necessarily be exactly broadside radiation from the perspective of network capacity. However, as stated earlier, the amplitude pattern should be carefully considered to guarantee the coverage performance.



(b)

Figure 11 (a) Investigation of QHA amplitude pattern over single cell and multiple users scenario; (b) Multi-cell scenario

Figure 12 is an illustration of the performance difference of different phase patterns with fixed beam width, noise level, and radiation power. In Figure 12, DiffAmp (X_Y) means the radiation directions of 2 port groups within QHA are X and Y. Only one QHA element was investigated with 4T4R for a base station. It can be seen in Figure 12a and Figure 12b that, the bigger the phase differential, the higher the throughput performance with equivalent port spacing of half wavelength. However, when the equivalent port spacing is larger than half wavelength, the performance will stop increasing and may even start to decrease. Figure 12c shows the performance in the multi-cell scenario, which considers the interference between different or adjacent cells. It shows that the system throughput is highly related to the equivalent port spacing, where half wavelength is the most optimal configuration of the antenna. Both the single cell and multi-cell scenarios indicate that ports with equivalent spacing of half wavelength have the best throughput performance from the perspective of phase differential and array configuration.



(a)

(b)

(c)

Figure 12 (a) Investigation of QHA phase pattern over single cell scenario with small phase differential; (b) Single cell with big phase differential; (c) Multi-cell scenario

2.5 QHA Implemented in Different Arrays

QHA is a totally new antenna technique in cellular networks, especially in base station application scenarios, where, traditionally, there are only two ports with two orthogonal polarizations. Four-port QHA will provide more freedom for channels and flexibility in a limited space or aperture. Several key array typologies were tried and analyzed, such as a purely QHA array (Figure 13), QHA + X array design (Figure 14), and QHA-inspired array (Figure 15).

For TDD MM 64T or above, the antenna size becomes more and more limited. Therefore, with ports doubled, QHA or its inspired design are quite suitable for size-limited arrays, making space for elements that are quite small or compact. With this advantage, QHA can double the channels while retaining the same array size. This really is a breakthrough for compact array design and makes many future application scenarios possible. With the flexibility of its polarization reconfiguration, a QHA array can be designed with more freedom. The antenna is also a key factor in the wireless channel and allows the QHA array design to be integrated with the entire system. Additionally, based on the axial symmetry of the QHA element, one can also use the QHA element to form a rotational array which can provide more intensive polarization diversity.

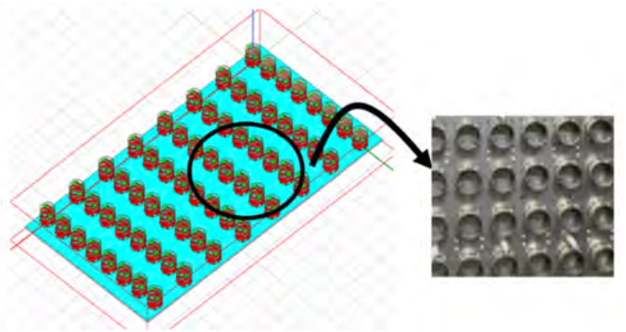


Figure 13 Pure QHA array simulation model and prototype

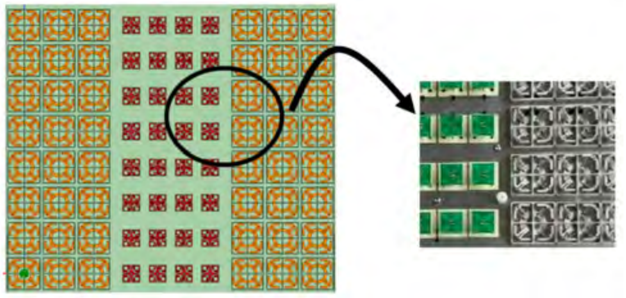


Figure 14 QHA+X mixed array in TDD 128T application (simulation model and prototype)

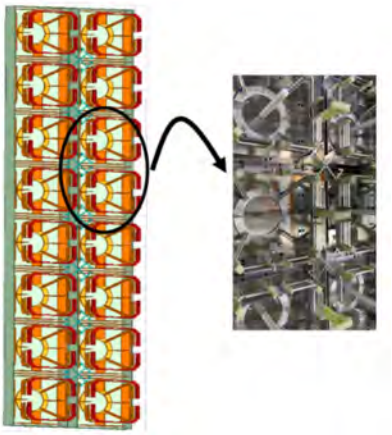


Figure 15 QHA in Sub-1G array (simulation model and prototype)

The QHA-inspired design splits the four ports helix antenna apart, as shown in Figure 15, making each port have pure linear polarization and shifting the phase center outwards. This phase center shifting effect is meaningful in the case of few Transceivers (TRs) scenarios, because it can considerably improve the aperture efficiency of a small array. This means we can improve the array gain without increasing the physical size of the array. This will lead to the array obtaining a higher angular resolution for the system, which means a higher channel capacity.

3 System Design and Performance Evaluation

Based on the new concept of the antenna, we proposed several compact array solutions for the system to achieve higher gain with limited array aperture. In compact arrays, different antenna elements with more interactions through mutual coupling lead to distortion in radiation patterns and active impedance matching. With the new mapping and architecture solutions, the sub-array configurations should also be modified. Because the code book and channel state information assessment are closely related to the phase center of each sub-array, the phase center and beam design of the new antenna should be studied.

3.1 TDD 128T MM System

According to the traditional ±45 polarization element, if one wants to achieve a 128T MM array, the array should be arranged as shown in Figure 16, which is 16 columns with 4 sub-arrays vertically (as an example). With this configuration, array size will be very large, especially in the horizontal direction. Taking C band (3.5 GHz carrier frequency) as an example, with 16 columns of half wavelength spacing, the horizontal width of the array is nearly 688 mm, which is hardly acceptable for current base station applications. If one uses the original QHA element and the improved version, the array width is dramatically reduced by 50% (Figure 17a), and 43% (Figure 17b) respectively. This is a significant reduction in the array size, while retaining the array gain as much as possible.

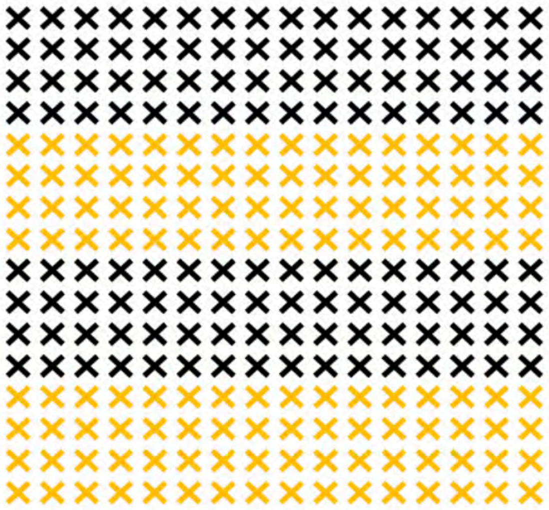
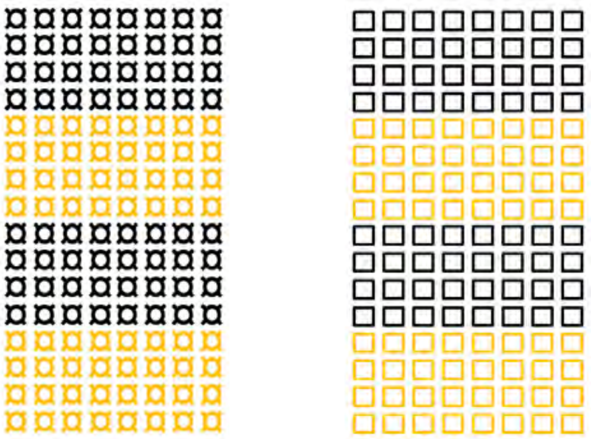


Figure 16 ±45 polarization dipole array with 128T channels (688 mm in width)



(a) Original versions
(344 mm in width)

(b) Improved version
(480 mm in width)

Figure 17 QHA arrays

3.2 Sub-1G MM System

The 690–960 MHz band is the golden band of the wireless communication system due to its excellent wall penetration capability and long distance coverage. But for lower carrier frequencies, the antenna elements are quite big, and as a result, only limited elements can be placed in the Sub-1G area design. With QHA application, as shown in Figure 18, we can achieve a more compact array design with limited array aperture and achieve a higher gain.

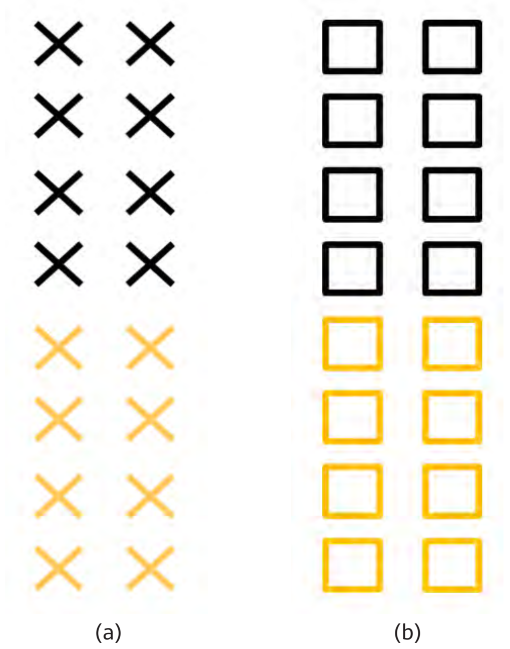


Figure 18 (a) Traditional ±45 polarization 2 columns dipole array; (b) QHA applied Sub-1G array with more channels and higher gain

3.3 Simulation Evaluation and Observation

Based on the analyses about QHA impact on system performance, many test cases were performed to verify. All the simulation results are consistent with previous illustrations of the relationship between the antenna phase center and system performance.

For massive MIMO scenarios, we went through many different cases to investigate the potential of QHA. Figure 19 shows that within the same aperture size, the QHA array achieves nearly 20% percent higher performance, and it can reach 35% gain when the improved QHA is implemented. For Sub-1G cases, we also conducted multiple performance investigations by comparing different system configurations, especially with different number of rows, columns, and polarizations. Figure 20 shows that in different scenarios, QHA can provide nearly 20% improvement in system performance, which is consistent with previous results.

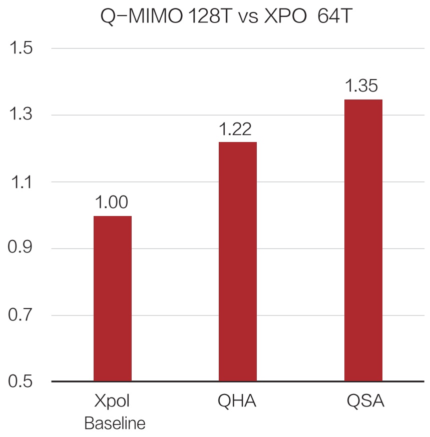


Figure 19 System throughput of 64T X-pol, 128T QHA and improved 128T QHA with the same aperture size

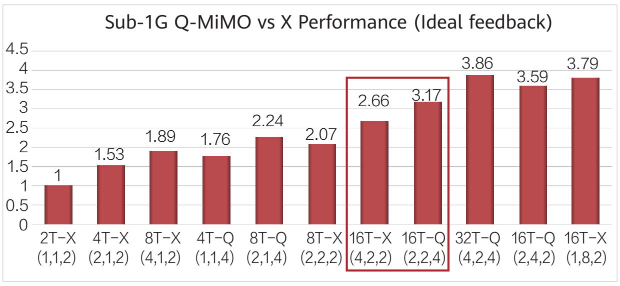


Figure 20 Sub-1G performance comparison between QHA and X-pol

Figure 21 is a view of different cases, from 2T X-pol to 128T QHA . It shows that the relative gain in scenarios with fewer channels is much higher, nearly 50%. As the number of channels increases, the performance gain of the system becomes relatively low, but still remains at around 26%. This is because the QHA phase center expansion effect becom

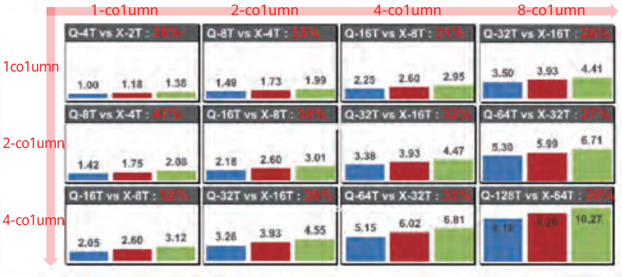


Figure 21 Overall comparison between QHA and X-pol with the same antenna array size

4 Field Trial Verification

Based on the research and investigation of QHA in the wireless system, several typical system prototypes were implemented and tested. The implementation and testing are performed at customer premises. We conducted an intensive field trial test on the new technology with FDD

32T of X-polarization and 64T of QHA on 2.6 GHz. We implemented Sub-1G field trial test on 2T of X-polarization and 16T of QHA.

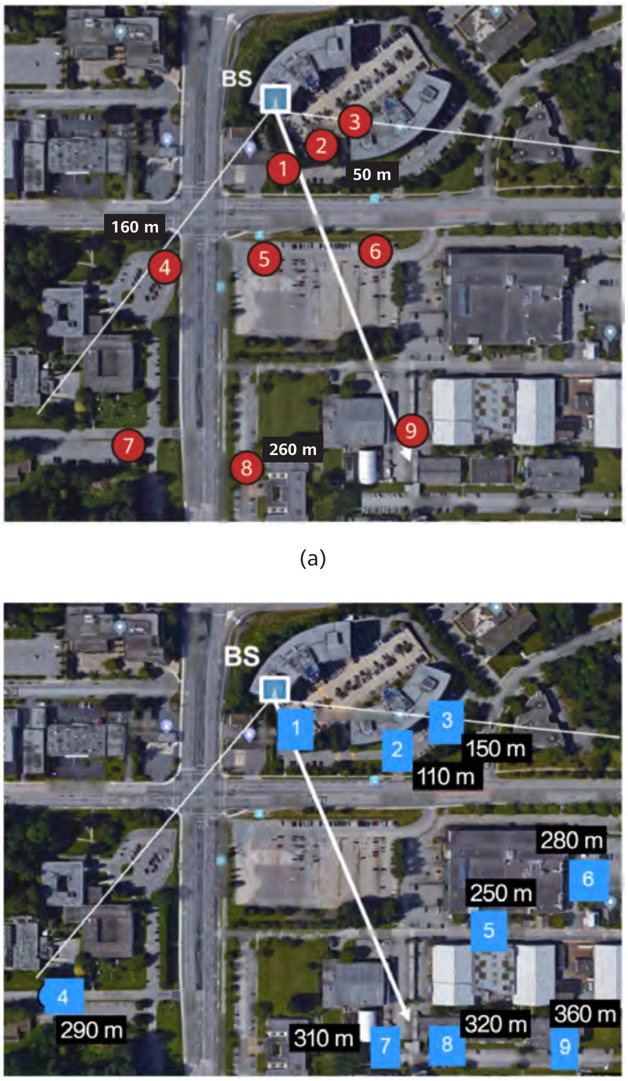
4.1 FDD MM System Configuration and Test Results

4.1.1 Test Configuration of Field Trial

Table 1 shows the configuration of the tested system, and Figure 22a shows the LOS points in the field trial, while Figure 22b shows NLOS points. The test point selection considers the cell coverage limit and the UE accessibility. Figure 23 shows the two-antenna array architecture under system test, which has the same aperture size but with different number of channels, 32T X-pol and 64T QHA.

Table 1 System configuration

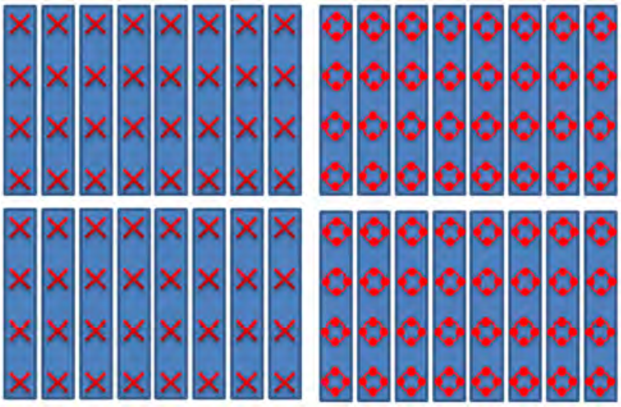
Configuration Item	Value
Operating band	2.67 GHz to 2.69 GHz (DL)
Subcarrier bandwidth	15 kHz
TTI length (slot)	1 ms
OFDM symbols per TTI	14
CP length	Long CP: 5.2 μs (160 samples) Short CP: 4.17 μs (128 samples)
Modulation and coding	NR R15
System bandwidth	20 MHz
Component carrier bandwidth	20 MHz
Number of subcarriers within each carrier	1320 (110 RBs)
DMRS	NR R15 Type2
CSI-RS	NR R15 and plus
Traffic load	Full buffer
Scheduling algorithm	Rank1 or Rank2



(a)

(b)

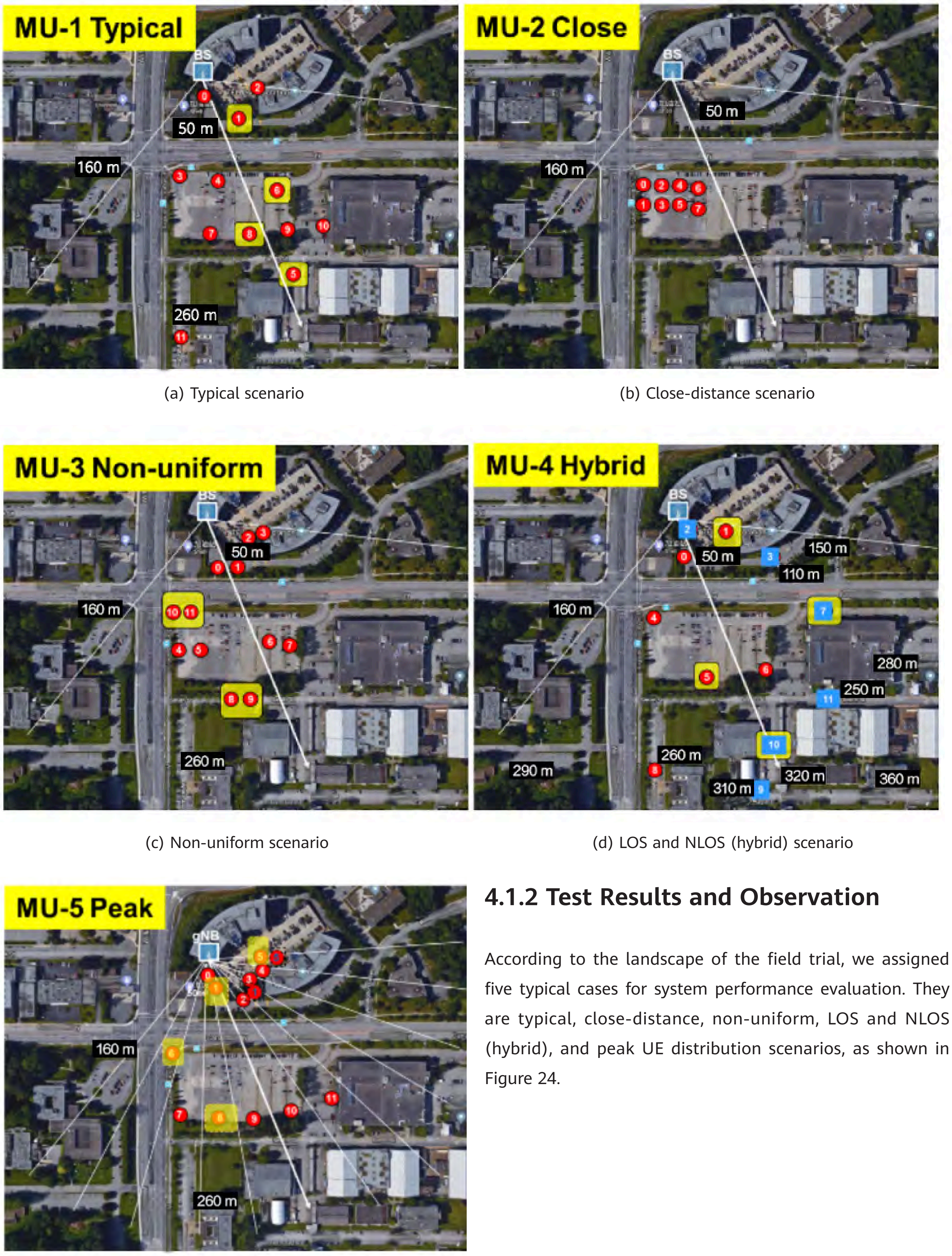
Figure 22 (a) LOS; (b) NLOS candidate positions for high/middle/low SNR scenarios



(a) X-pol array with 32T

(b) QHA array with 64T

Figure 23 Tested antennas



(a) Typical scenario

(b) Close-distance scenario

(c) Non-uniform scenario

(d) LOS and NLOS (hybrid) scenario

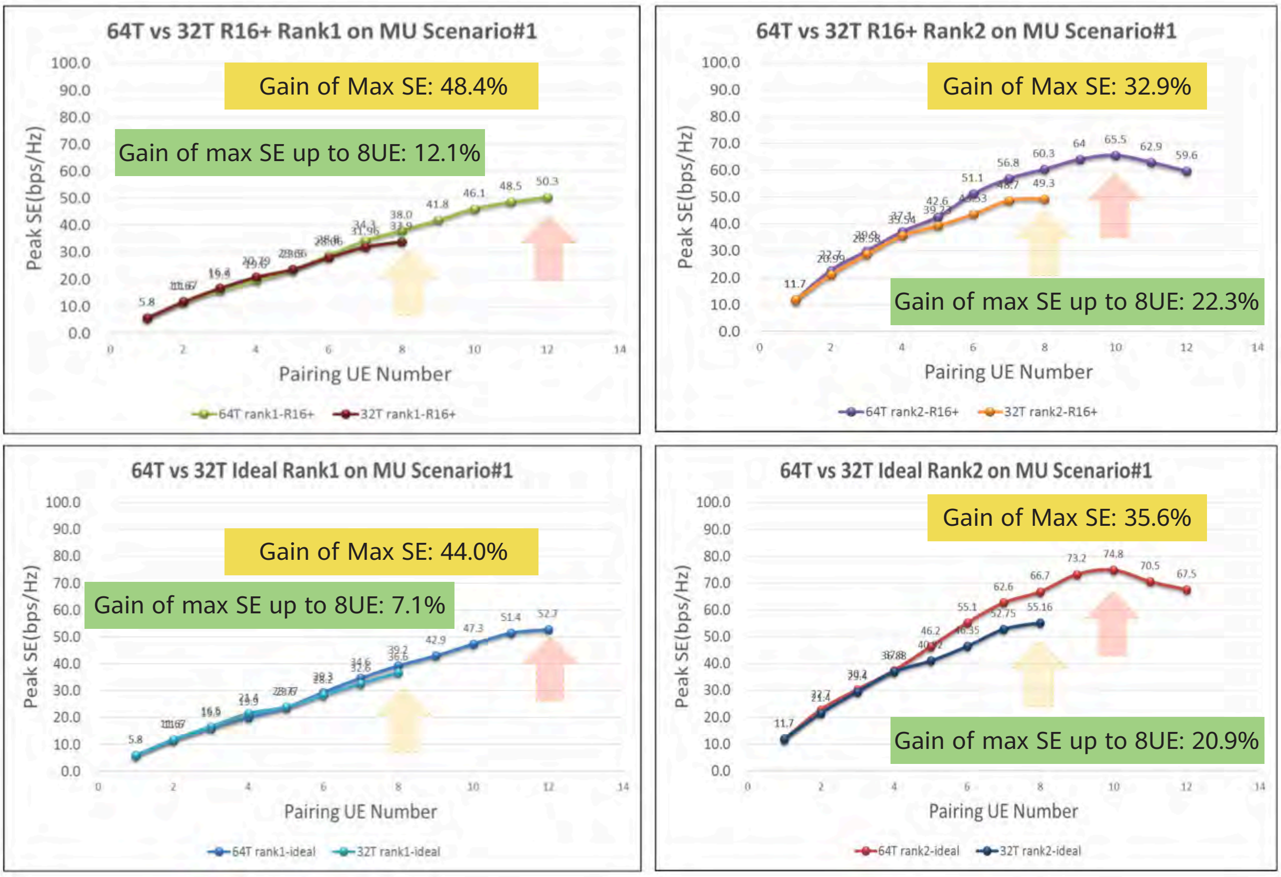
(e) Peak UE distribution scenario

Figure 24 MU test scenarios

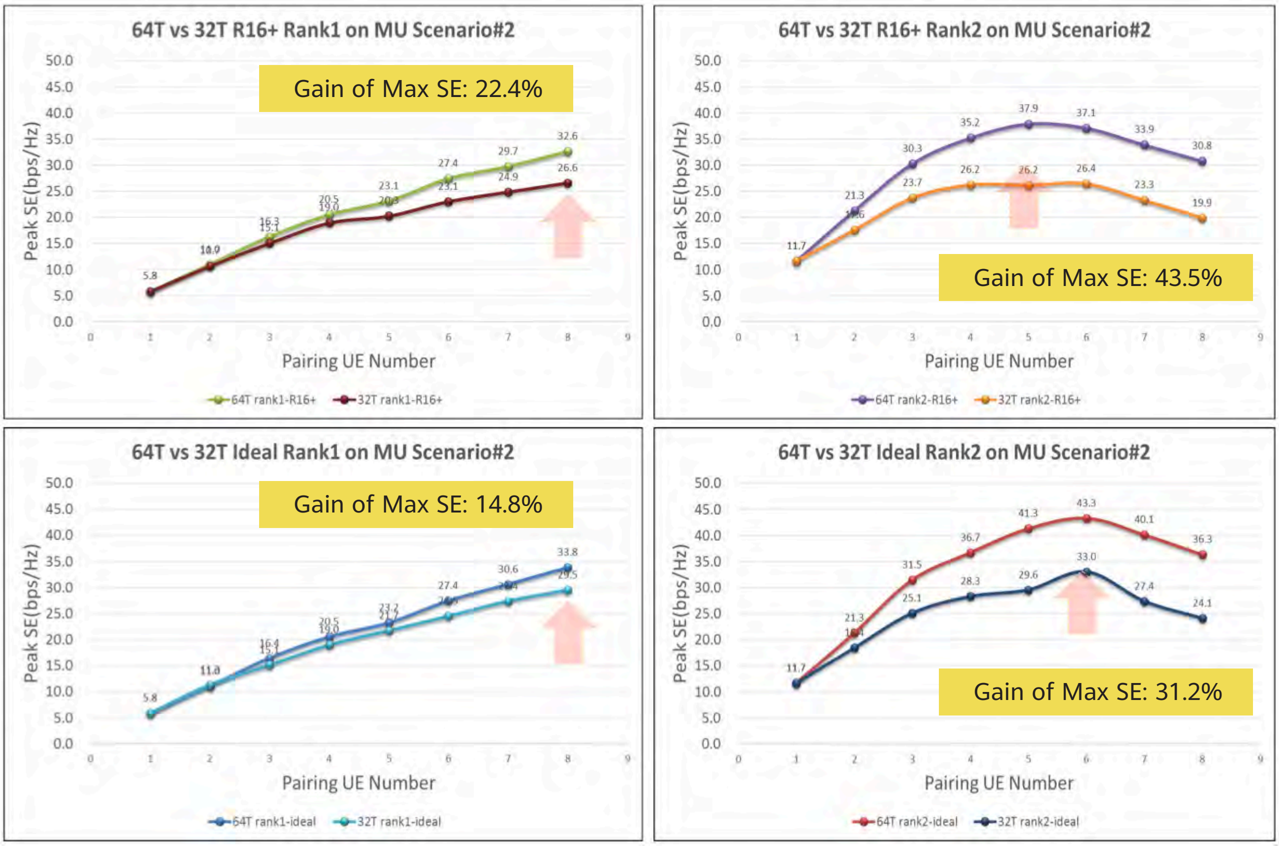
*blue block for NLOS UEs, red block for LOS UEs, yellow block only for 64T test cases.

4.1.2 Test Results and Observation

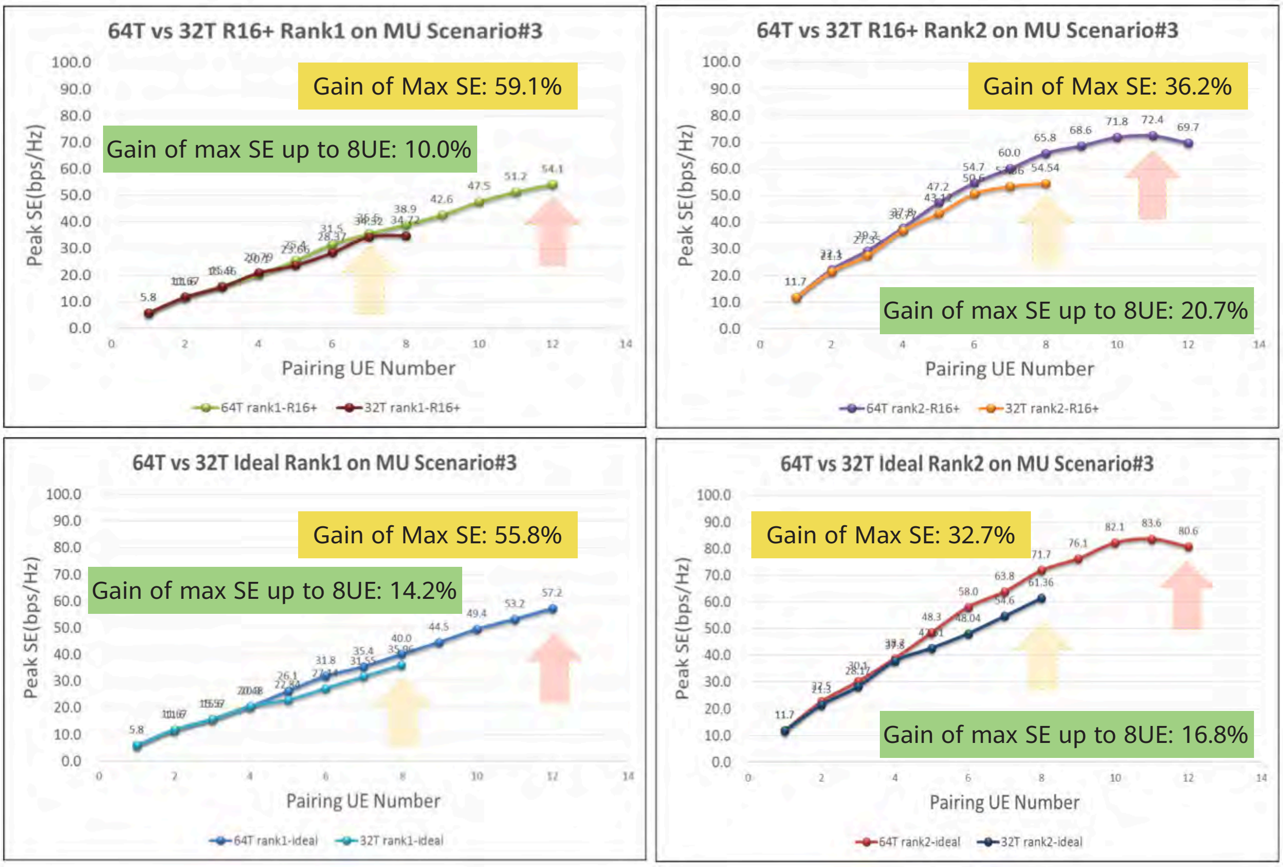
According to the landscape of the field trial, we assigned five typical cases for system performance evaluation. They are typical, close-distance, non-uniform, LOS and NLOS (hybrid), and peak UE distribution scenarios, as shown in Figure 24.


64T vs 32T R16+ Rank1 on MU Scenario#1
Gain of Max SE: 48.4%
Gain of max SE up to 8UE: 12.1%
Peak SE(bps/Hz)
Pairing UE Number
5.8
50.3
48.5
46.1
41.8
38.0
64T rank1-R16+
32T rank1-R16+
64T vs 32T R16+ Rank2 on MU Scenario#1
Gain of Max SE: 32.9%
Gain of max SE up to 8UE: 22.3%
11.7
51.1
56.8
60.3
64
65.5
62.9
59.6
49.3
64T rank2-R16+
32T rank2-R16+
64T vs 32T Ideal Rank1 on MU Scenario#1
Gain of Max SE: 44.0%
Gain of max SE up to 8UE: 7.1%
5.8
39.2
42.9
47.3
51.4
52.7
64T rank1-ideal
32T rank1-ideal
64T vs 32T Ideal Rank2 on MU Scenario#1
Gain of Max SE: 35.6%
Gain of max SE up to 8UE: 20.9%
11.7
46.2
55.1
62.6
66.7
73.2
74.8
70.5
67.5
55.16
64T rank2-ideal
32T rank2-ideal

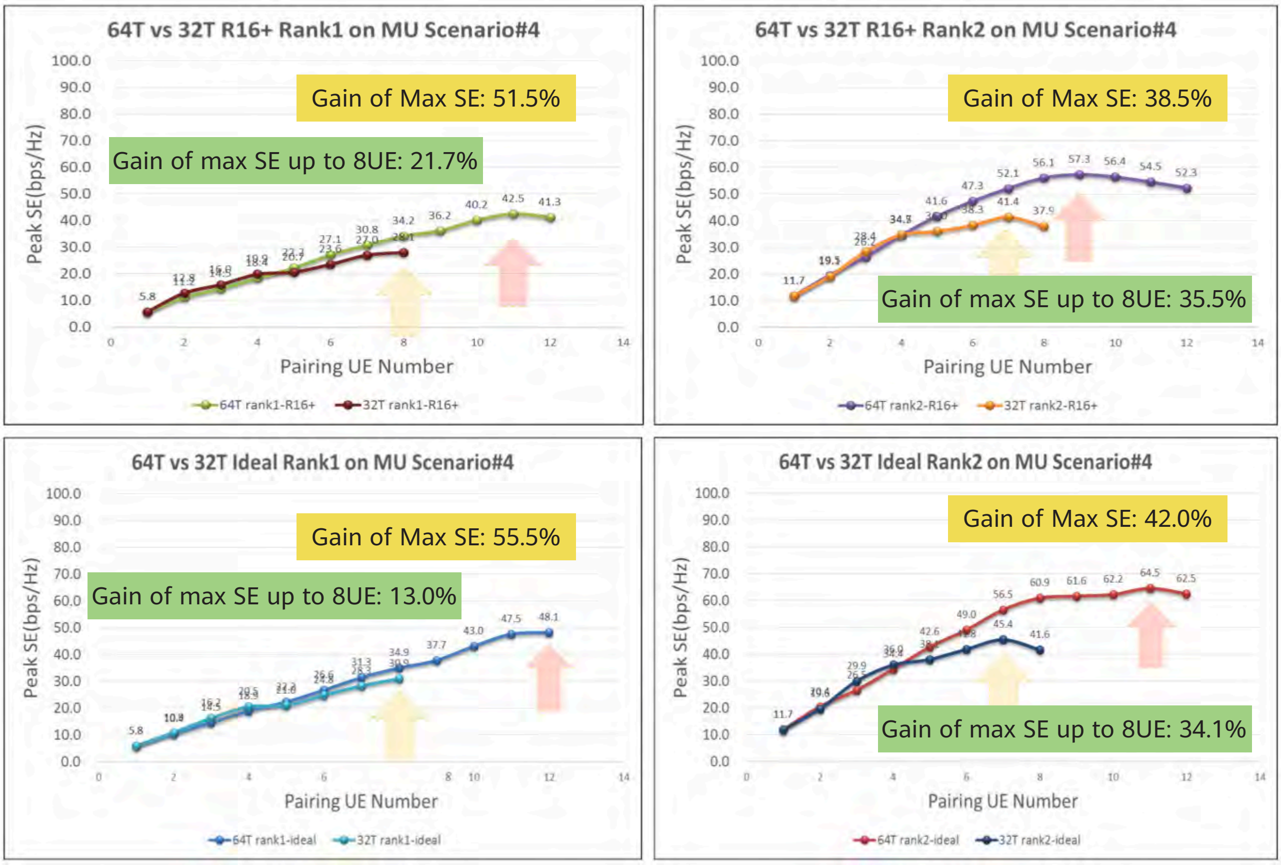
(a)


64T vs 32T R16+ Rank1 on MU Scenario#2
Gain of Max SE: 22.4%
Peak SE(bps/Hz)
Pairing UE Number
5.8
10.9
20.5
23.1
27.4
29.7
32.6
24.9
26.6
64T rank1-R16+
32T rank1-R16+
64T vs 32T R16+ Rank2 on MU Scenario#2
Gain of Max SE: 43.5%
11.7
21.3
30.3
35.2
37.9
37.1
33.9
30.8
23.7
26.2
26.2
26.4
23.3
19.9
64T rank2-R16+
32T rank2-R16+
64T vs 32T Ideal Rank1 on MU Scenario#2
Gain of Max SE: 14.8%
5.8
11.0
16.4
20.5
27.4
30.6
33.8
29.5
64T rank1-ideal
32T rank1-ideal
64T vs 32T Ideal Rank2 on MU Scenario#2
Gain of Max SE: 31.2%
11.7
21.3
31.5
36.7
41.3
43.3
40.1
36.3
25.1
28.3
29.6
33.0
27.4
24.1
64T rank2-ideal
32T rank2-ideal

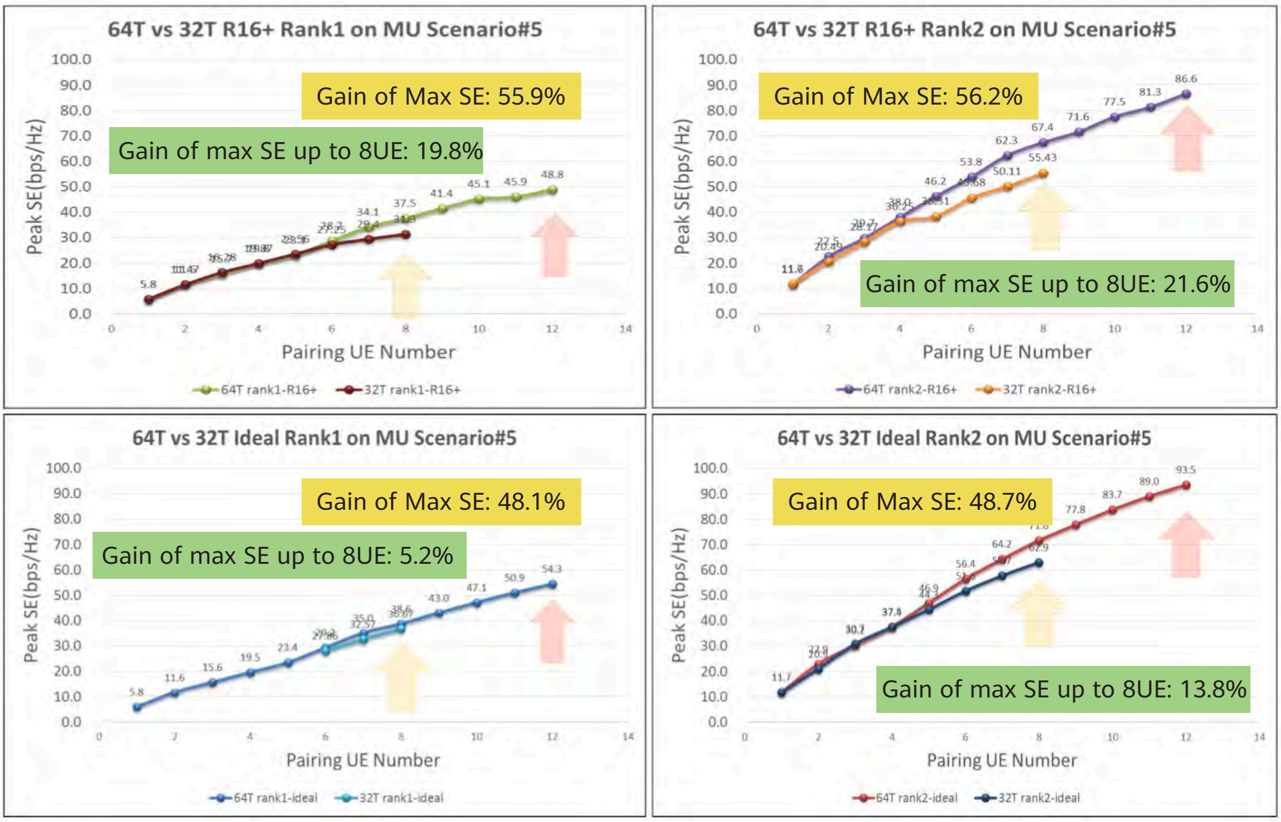
(b)


64T vs 32T R16+ Rank1 on MU Scenario#3
Gain of Max SE: 59.1%
Gain of max SE up to 8UE: 10.0%
Peak SE(bps/Hz)
Pairing UE Number
64T rank1-R16+
32T rank1-R16+
5.8
42.6
47.5
51.2
54.1
38.9
64T vs 32T R16+ Rank2 on MU Scenario#3
Gain of Max SE: 36.2%
Gain of max SE up to 8UE: 20.7%
64T rank2-R16+
32T rank2-R16+
11.7
47.2
54.7
60.0
65.8
68.6
71.8
72.4
69.7
54.54
64T vs 32T Ideal Rank1 on MU Scenario#3
Gain of Max SE: 55.8%
Gain of max SE up to 8UE: 14.2%
64T rank1-ideal
32T rank1-ideal
5.8
31.8
35.4
40.0
44.5
49.4
53.2
57.2
64T vs 32T Ideal Rank2 on MU Scenario#3
Gain of Max SE: 32.7%
Gain of max SE up to 8UE: 16.8%
64T rank2-ideal
32T rank2-ideal
11.7
48.3
58.0
63.8
71.7
76.1
82.1
83.6
80.6
61.36
48.04

(c)


64T vs 32T R16+ Rank1 on MU Scenario#4
Gain of Max SE: 51.5%
Gain of max SE up to 8UE: 21.7%
Peak SE(bps/Hz)
Pairing UE Number
64T rank1-R16+
32T rank1-R16+
5.8
27.1
30.8
34.2
36.2
40.2
42.5
41.3
64T vs 32T R16+ Rank2 on MU Scenario#4
Gain of Max SE: 38.5%
Gain of max SE up to 8UE: 35.5%
64T rank2-R16+
32T rank2-R16+
11.7
19.5
34.8
41.6
47.3
52.1
56.1
57.3
56.4
54.5
52.3
41.4
37.9
64T vs 32T Ideal Rank1 on MU Scenario#4
Gain of Max SE: 55.5%
Gain of max SE up to 8UE: 13.0%
64T rank1-ideal
32T rank1-ideal
5.8
10.8
31.3
34.9
37.7
43.0
47.5
48.1
64T vs 32T Ideal Rank2 on MU Scenario#4
Gain of Max SE: 42.0%
Gain of max SE up to 8UE: 34.1%
64T rank2-ideal
32T rank2-ideal
11.7
29.9
42.6
49.0
56.5
60.9
61.6
62.2
64.5
62.5
45.4
41.6

(d)



(e)

Figure 25 64T QHA vs 32T X-pol test results under different scenarios, respectively

Figure 25 shows the spectrum efficiency (SE) comparison results using two different codebooks in the case of two systems (64T QHA and 32T X-pol with the same aperture). It can be seen that the 3GPP R16+ based feedback has similar performance gain with ideal feedback. This indicates that the QHA array with the same aperture performs better than X-pol array. In addition, more users are paired when the rank of users is increased from rank1 to rank2.

Table 2 Sub-1G system configuration

Configuration Item	Value	Remarks
Operating band	770 MHz for DL	730 MHz for UL
Subcarrier bandwidth	15 kHz	
TTI length (slot)	1 ms	
OFDM symbols per TTI	14	
CP length	Long CP: 5.2 μs Short CP: 4.68 μs	
Modulation and coding	NR R15	
System bandwidth	10 MHz	
Component carrier bandwidth	10 MHz	
Number of subcarriers within each carrier	660 (55 RBs)	
DMRS	NR Type2 DMRS	
Traffic load	Full buffer	
Scheduling algorithm	Rank1 or Rank2	
FDD feedback algorithm	R15/R16/Ideal	

4.2 Sub-1G System Configuration and Test Results

4.2.1 Sub-1G System Configuration

System configuration is shown in Table 2. For performance investigation and comparison, two different antenna array configurations are introduced, namely, one-column X-polarized array with 2T and two-column QHA array with 16T, as shown in Figure 26. Because the 16T QHA array has two columns, the aperture size is doubled against the one-column X-pol array.

Figure 27 shows the landscape of the test field. The cell radius is almost 1 km. There are two UEs nearby numbered as No. 1 and No. 2, and other UEs are mapped into the field for both coverage and throughput capacity test. Specifically, the green line routine is for the driving outage test.

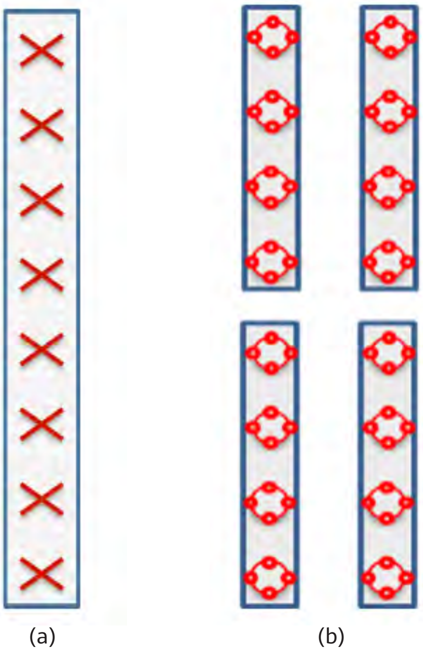


Figure 26 (a) One column X-polarization with 2T; (b) Two columns QHA with 16T

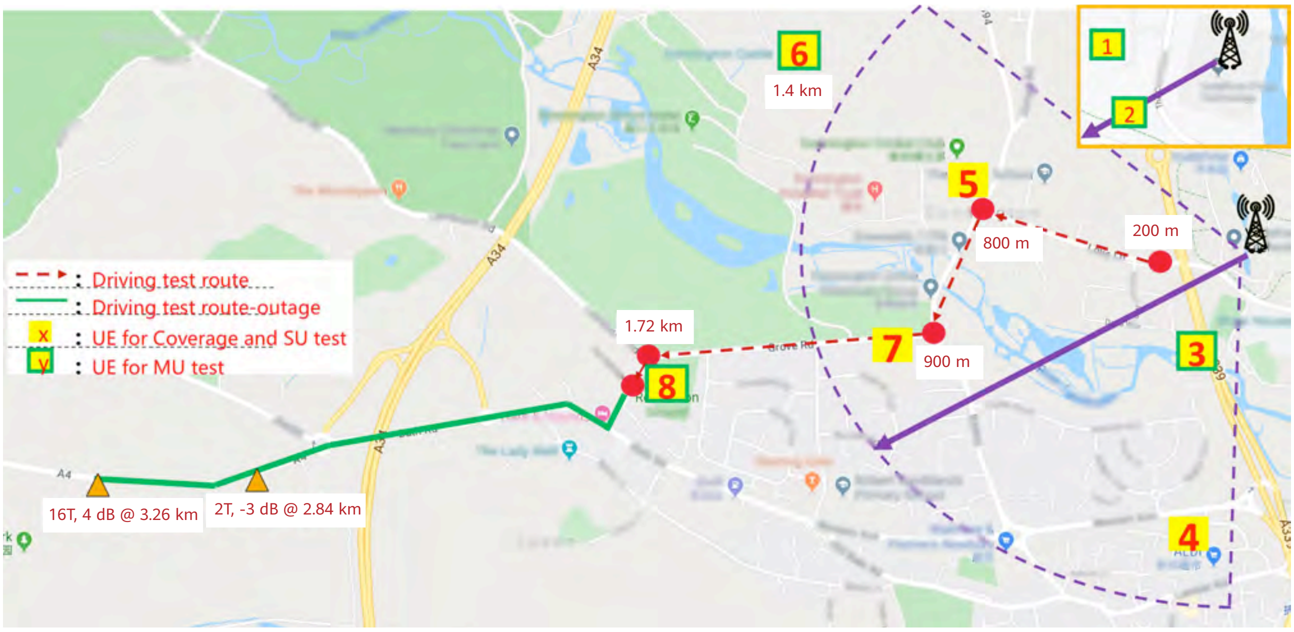


Figure 27 Coverage field trial configuration

4.2.2 Test Results and Observation

Figure 28 shows the heat map of the driving coverage test, with the heat indicating the elevation of the location. The driving routine test was conducted mainly within the valley to avoid severe geographical obstacles. Figure 29 shows the coverage outage test results, 16T QHA with SSB SNR 4 dB UE lost at distance of 3.26 km and 2T X-pol with –3 dB UE lost at distance of 2.84 km.

Figure 30 shows the relative gain of 16T QHA over 2T X-pol array. It can be seen that 16T QHA is much higher in terms of SSB coverage and DL DMRS.

Figure 31 shows the throughput performance of 16T QHA and 2T X-pol. In the SU condition, the average throughput of 16T QHA is about 1.4 times that of 2T X-pol. In the MU condition, different number of UEs are paired and the average throughput of 4 UEs of 16T QHA is about 3 times that of 2T X-pol for Type II codebook and 3.8 times for ideal feedback.

The results indicate that the performance of QHA array architecture is especially optimal in scenarios involving a limited number of channels.

5 Conclusion and Future Work

This paper provides an in-depth introduction to QHA, from theory to application. The QHA-inspired super resolution effect was discussed and the impact of the amplitude and phase patterns on system performance were specifically investigated. From these discussions, we can see that the phase pattern has an obvious impact on system performance, and can be deliberately designed. Based on this consideration, a new type of QHA was proposed. Using the new design, some good results were obtained in several different key scenarios. Specifically, the QHA based antenna solution was applied in 64T and Sub-1G 16T solutions and these two new architectures are very attractive for the practical deployment.

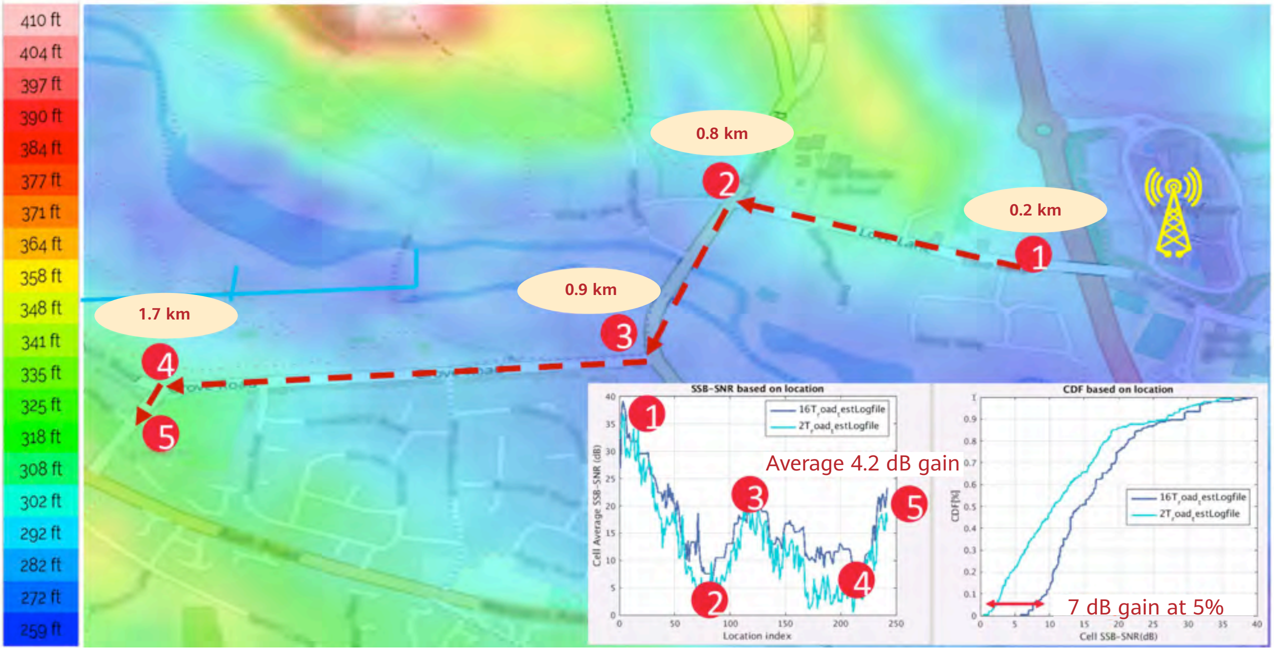


Figure 28 Elevation map of the driving coverage test

As new application scenarios continuously emerge, the evolution of wireless communication will never end. With MIMO and electromagnetic field manipulation technology becoming more important and popular, antenna technology is playing a more critical part in wireless systems and it will become a new driving force for the coming generations. The degree of freedom that antenna technology can provide will determine system performance. New research on the antenna resolution capability and electromagnetic field manipulation will open new doors to improving the performance of future wireless systems.

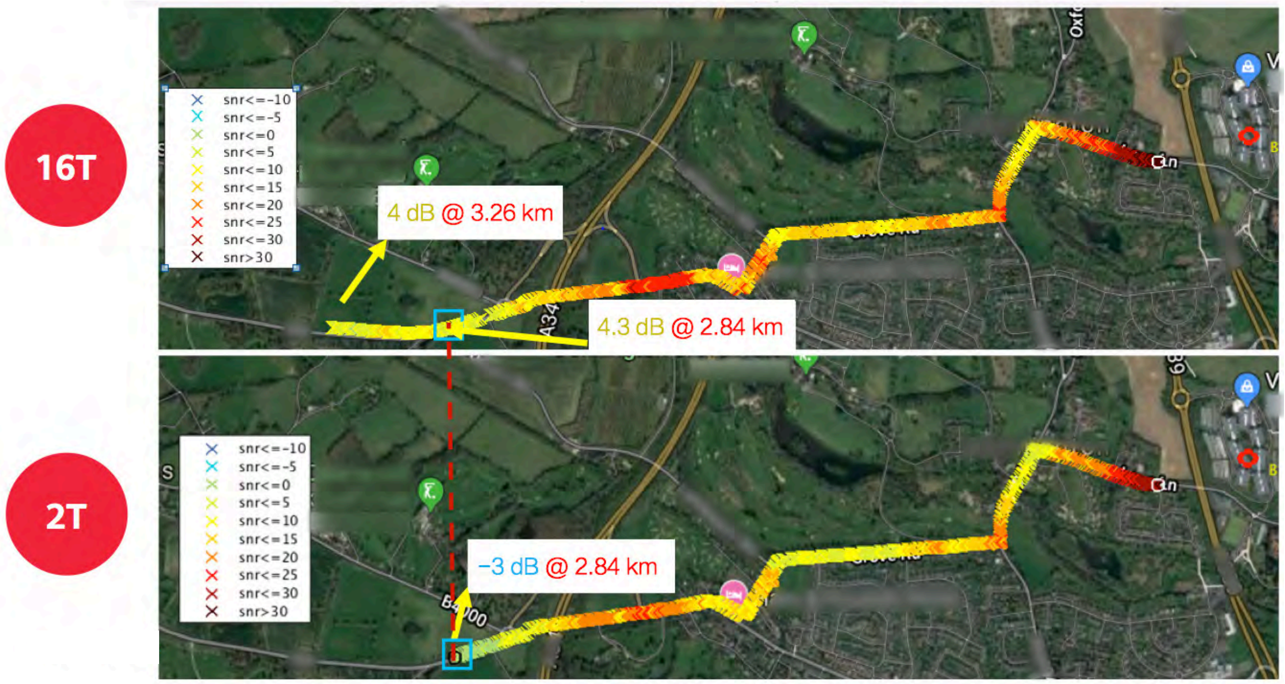


Figure 29 Comparison of coverage test (outage)

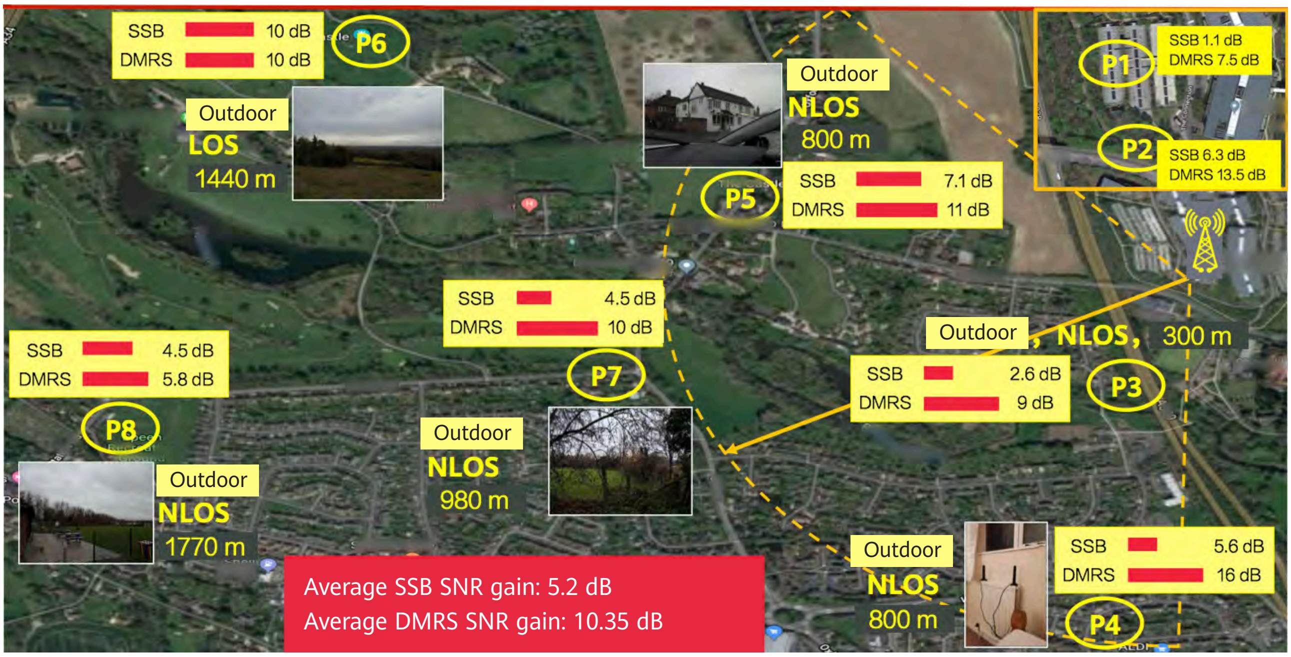


Figure 30 Gain of fix point coverage test (16T type II vs 2T type I)

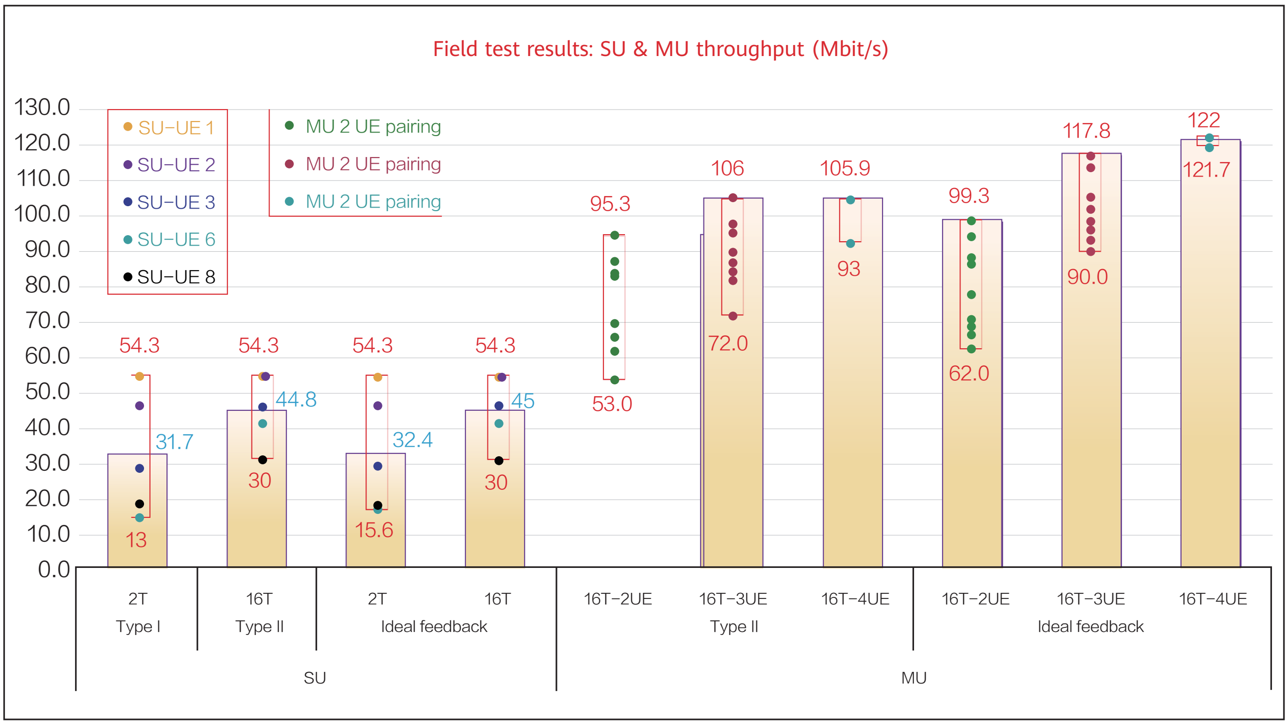


Figure 31 SU and MU throughput field test results

References

[1] G.J. Foschini and M.J. Gans, "On limits of wireless communications in a fading environment when using multiple antennas," in Wireless Personal Communications vol. 6, pp. 311-335, 1998.

[2] Fredrik Rusek, Daniel Persson, Buon Kiong Lau, Erik G. Larsson, Thomas L. Marzetta, Ove Edfors, and Fredrik Tufvesson, "Opportunities and challenges with very large arrays," in IEEE Signal Processing Magazine, pp. 40-60, January 2013.

[3] Zhongwei Tang, and Ananda S. Mohan, "Experimental Investigation of Indoor MIMO Ricean Channel Capacity," in IEEE Antennas and Wireless Propagation Letters, vol. 4, pp.55-58, 2005.

[4] Yongping Wang and Hanqiang Cao, "Capacity Bounds for Rayleigh/Lognormal MIMO Channels with Double-Sided Correlation," in IEEE Communications Letters, vol. 19, pp. 1362-1365, 2015.

[5] Sergey Loyka, and Ammar Kouki, "New Compound Upper Bound on MIMO Channel Capacity," in IEEE Communications Letters, vol. 6, pp. 96-98, 2002.

[6] Liang Dong, Hosung Choo, Robert W. Heath, Jr., and Hao Ling, "Simulation of MIMO Channel Capacity with Antenna Polarization Diversity," in IEEE Transactions on Wireless Communications, vol. 4, pp. 1869-1873, 2005.

[7] Carl B. Dietrich, Jr., Kai Dietze, J. Randall Nealy, and Warren L. Stutzman, "Spatial, Polarization, and Pattern Diversity for Wireless Handheld Terminals," in IEEE Transactions on Antennas and Propagation, vol. 49, pp.1271-1281, 2001.

[8] Jose-Maria Molina-Garcia-Pardo, Martine Lienard, Pierre Degauque, Eric Simon, and Leandro Juan-Llacer, "On MIMO Channel Capacity in Tunnels," in IEEE Transactions on Antennas and Propagation, vol. 57, pp. 3697-3701, 2009.

[9] Na Wu, FangQi Zhu, and QiLian Liang. "Evaluating Spatial Resolution and Channel Capacity of Sparse Cylindrical Arrays for Massive MIMO," in IEEE Access, vol.5, pp. 23994-24003, 2017.

[10] Jørgen Bach Andersen, and Klaus Ingemann Pedersen, "Angle-of-Arrival Statistics for Low Resolution Antennas," in IEEE Transactions on Antennas and Propagation, vol. 50, pp. 391-395, 2002.

[11] Oleksandr Malyuskin and Vincent F. Fusco, "Experimental Study of Electrically Compact Retro directive Monopole Antenna Arrays," in IEEE Transactions on Antennas and Propagation, vol. 65, pp. 2339-2347, 2017.

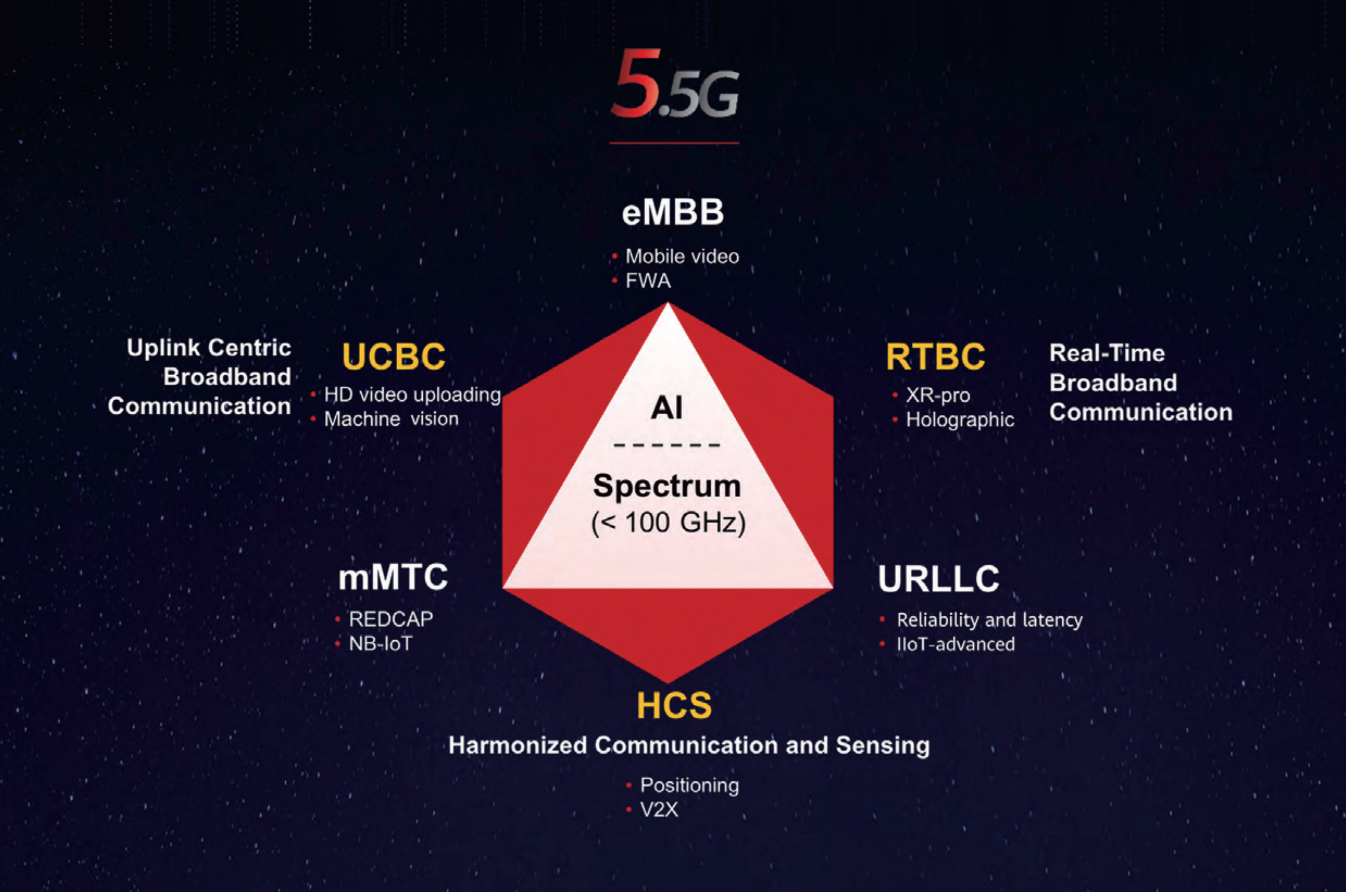
[12] Yashi Zhou, Wei Wang, Zhen Chen, Qingchao Zhao, Heng Zhang, Yunkai Deng, and Robert Wang, "High-Resolution and Wide-Swath SAR Imaging Mode Using Frequency Diverse Planar Array," in IEEE Geoscience and Remote Sensing Letters, vol. 18, pp. 321-325, 2021.

[13] Makoto Sano, Manuel Sierra-Castañer, Tamara Salmerón-Ruiz,Jiro Hirokawa, and Makoto Ando, "Reconstruction of the Field Distribution on Slot Array Antennas Using the Gerchberg–Papoulis Algorithm," in IEEE Transactions on Antennas and Propagation, vol. 63, pp. 3441-3451, 2015.

[14] Chenxi Hu, Yimin Liu, Huadong Meng, and Xiqin Wang, "Randomized Switched Antenna Array FMCW Radar for Automotive Applications," in IEEE Transactions on Vehicular Technology, vol. 63, pp. 3624-3641, 2014.

[15] Oleg A. Iupikov , Marianna V. Ivashina, Niels Skou, Cecilia Cappellin, Knud Pontoppidan, and Cornelis G. M. van't Klooster, " Multibeam Focal Plane Arrays With Digital Beamforming for High Precision Space-Borne Ocean Remote Sensing," in IEEE Transactions on Antennas and Propagation, vol. 66, pp. 737-748, 2018.

[16] T.W.C. Brown, S.R. Saunders, "The intelligent quadrifilar helix: a compact MIMO antenna for IEEE 802.11n," in The Second European Conference on Antennas and Propagation, EuCAP 2007.



5G-Advanced: Uplink Centric Broadband Communication (UCBC)

Jiyong Pang, Zhiheng Guo, Huangping Jin, Jinlin Peng, Zhenfei Tang, Shaobo Wang
RAN Research Dept, Wireless Network

Abstract

As 5G commercialization accelerates globally, the technology is increasingly regarded as key to enhanced user experience and digital industrial transformation. As part of this process, larger volumes of uplink (UL) traffic are created as massive amounts of data are uploaded to 5G networks, resulting in significant challenges to 5G UL capability. As such, continuous research on the follow-up evolution of 5G networks, known as 5G-Advanced, is required with a special focus on uplink centric broadband communication (UCBC). This paper begins by analyzing the requirements and challenges involved with 5G UL communication, and then investigates potential UL technology enhancements geared towards 5G-Advanced.

Keywords

5G, 5G-Advanced, UCBC

1 Introduction

Wireless communications have traditionally focused more on downlink (DL), as the vast majority of data traffic — such as video and music streaming — takes place in that direction. This has been directly reflected in the development of wireless standards over the years. However, following the arrival of 5G, all of this is beginning to change.

The number of 5G applications in use around the world today is growing at an unprecedented rate. Modern society is experiencing digital transformation in fundamental ways, where smart connectivity of everything is leading to massive volumes of data being sent to clouds over wireless networks. This poses a huge challenge for both wireless network UL capacity and UL coverage.

In this regard, 3GPP 5G NR has already introduced UL enhancement technologies in its first two releases, release 15 (Rel-15) and release 16 (Rel-16), and also as part of the near-final release 17 (Rel-17) — for example, higher transmit power of terminals, more symbols for long physical UL control channels (PUCCHs), supplementary ULs (SULs), UL transmission chain (Tx) switching, and more.

However, following the rapid expansion of uplink centric broadband communication (UCBC) — as represented by live uploading of high-definition (HD) videos in extended reality (XR) applications by end users, and in vertical industry applications using remote cameras — current 5G capabilities will be not sufficient to satisfy UL requirements in the foreseeable future (for example, 10-fold UL capacity). Consequently, 5G UL capability needs to be continuously enhanced in subsequent releases.

In the following sections, we will address the driving forces and challenges involved with UL enhancement in 5G-Advanced, as well as the corresponding enabling technologies.

2 Driving Forces and Challenges

With the acceleration of the digital transformation of society as a whole, the demand for UL business in the to-customer (toC) and to-business (toB) fields, as well as the Internet of Things (IoT) field, has exploded. This poses great challenges to 5G networks and in turn promotes technology upgrades.

2.1 ToC Experience Improvement

Network experience improvement is the driving force behind the continuous development of the toC business, which is an essential part of 5G commercialization. Innovative services, such as interactive XR and social media, generate far higher volumes of upstream traffic.

Figure 1 Interactive XR for end consumers

Figure 2 Video upload in smart manufacturing

- Interactive XR

Interactive and immersive XR services, as shown in Figure 1, require HD images and videos to be sent from local consumer devices to the cloud for further rendering. The deployment of 8K or even 12K cameras for front-end acquisition will be rapidly popularized, leading to UL 100 Mbps or even higher rate applications [2].

- Social media

Social media is becoming increasingly popular, resulting in the proliferation of user-generated content to a wide range of platforms and representing yet another catalyst for a boost in network UL capacity.

2.2 ToB Industrial Digitalization

5G is regarded as the cornerstone of the industry's digital transformation. The digital exploration in steel, mining, port, manufacturing, and education shows that video surveillance, remote control, and machine vision are typical industrial applications where uploading HD and ultra-high definition (UHD) photos and videos places high requirements on the network's UL capacity [3], as shown in Figure 2. Consequently, 5G networks must further improve UL capabilities to cope with digital transformation across industries.

- Remote operation

Remote operations such as video surveillance and remote control require real-time transfer of 4K and 8K videos. This means that the current single-point UL rate of 3 Mbit/s will increase to 20 Mbit/s and even 60 Mbit/s in some cases. Given that concurrent video uploads are required in most scenarios, the single-cell UL capacity must be increased exponentially [3].

- Machine vision

Machine vision is becoming a must-have for smart factory architecture in the Industry 4.0 wave. This technology can implement automatic inspection using HD industrial cameras, reducing the time required by 80% while also ensuring quality and improving efficiency. The high precision of machine vision requires single-frame image transfer to be completed with an exact latency and free of any quality loss. For example, the UL bandwidth must be 350–600 Mbit/s in order for industrial cameras to send over 60 HD photos captured per second to the platform.

2.3 Broadband IoT

In the smart city system, various broadband and narrowband IoT devices collect and send real-world data to the cloud [4], as shown in Figure 3. Specifically, massive broadband IoT devices such as wireless cameras deployed for wide/outdoor area surveillance, vehicle monitoring, and unmanned vehicle distribution exert pressure on the overall network UL capacity.

Taking a world expo park as an example, 600 cameras are deployed in the 2 km^2 area, with up to 18 cameras per cell. A single camera has an average UL rate of 12 Mbit/s, and a single cell has an UL capacity of 220 Mbit/s. In the future, XR videos offering higher resolution and frame rates will be expected, leading to the emergence of additional smart services. As a result, the UL capacity of a single cell will have to reach between 500 Mbit/s and 1 Gbit/s.

Figure 3 IoT wireless backhaul in Smart City

3 Promising Technology Directions

To address the scenarios and business needs mentioned in Section 2, a number of promising UL evolution directions and potential enhancement technologies towards 5G-Advanced have emerged, as shown in Figure 4. We divide these technologies into four domains: frequency, time, space and power. These domains are explained in detail below.

3.1 Frequency Domain

Today's commercial 5G terminals offer only a maximum of 2Tx on sub-6 GHz spectrum, which limits the UL capability of carrier aggregation (CA) and SUL as explained below.

- CA/SUL needs at least 1Tx on each band semi-statically. As a result, UL multiple-input multiple-output (MIMO) is disabled on one band for 2Tx user equipment (UE).
- A UE's band configuration and concurrent transmission capabilities are strictly coupled, and more than two bands cannot be configured for one UE if it is equipped with up to 2Tx.
- The power amplifier (PA) capabilities on one band cannot be fully exploited in conjunction with another band.

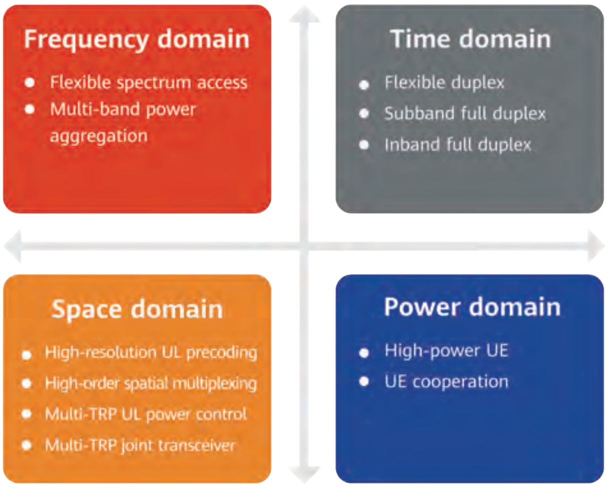


Figure 4 Four technology domains of UL enhancement

In this section, we will discuss how to overcome the above frequency-domain constraints in order to properly maximize UE capabilities for UL transmission.

3.1.1 FSA

Flexible spectrum access (FSA) is proposed as a flexible spectrum utilization mechanism which allows a UE (e.g., a 2Tx UE) to be configured with more UL bands than its concurrent transmission capability, and supports dynamic carrier selection as well as Tx switching between n (n ≥ 2) configured bands, as shown in Figure 5.

Detailed implementation architecture is illustrated in Figure 6, where 2Tx, including 2 power supplies and radio-frequency integrated circuits (RF ICs), is equipped with additional switches for 4 bands. Through this approach, a UE can flexibly access more than two bands with only 2Tx, while intelligently selecting one or two of those bands for concurrent transmission or allocating 2Tx to the most suitable band.

FSA benefits UL capacity and experience in a number of ways, including the following:

- **Enabling UL MIMO.** FSA enables the use of 2Tx for UL

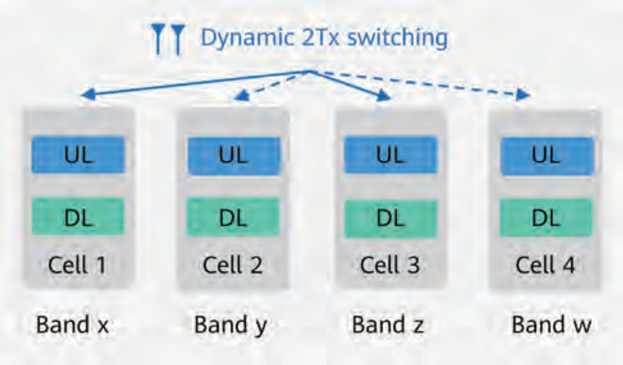


Figure 5 Illustration of the FSA concept

MIMO on any one of the access bands in a switching manner.

- **Efficient utilization of time division duplex (TDD) UL timeslots.** As shown in Figure 7a, when one of the TDD bands is DL, the UE can be switched to another TDD band which is UL according to the TDD configurations.

- **Better adaptation to channel conditions.** As shown in Figure 7b, the network can schedule the UE on a band offering improved channel conditions.

- **Higher trunking efficiency.** As shown in Figure 7c, when a band is congested with traffic, FSA can dynamically allocate a part of the traffic load to another band in order to maximize the use of unoccupied resources.

We evaluate the UL throughput gain of FSA via system-level simulation under the 3GPP Dense Urban scenario. As shown in Figure 8 and Figure 9, when compared with the Rel-17 baseline, FSA offers a 27%–35% gain in average user-perceived throughput (UPT) for burst traffic and a 46% capacity gain (for instance, the number of UEs that can be accommodated per cell) for XR traffic.

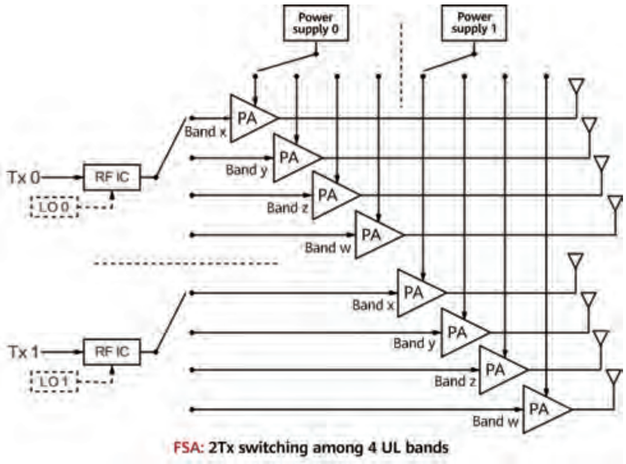


Figure 6 Implementation of an FSA terminal

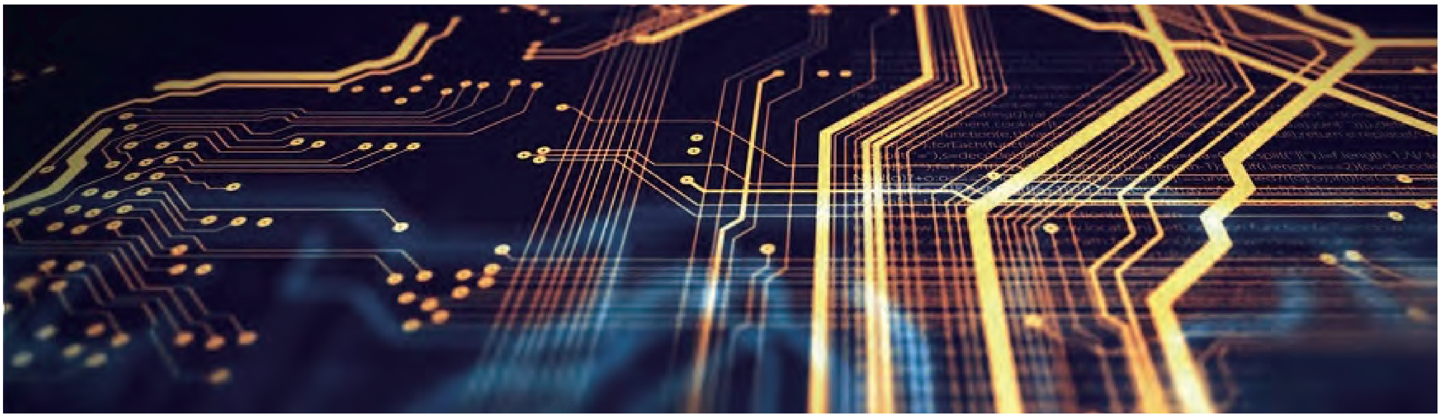
It is worth pointing out that although 2Tx UEs are taken as an example, UEs with higher capabilities (for example, 3Tx or 4Tx) can also obtain similar benefits by employing FSA via Tx and carrier switching or capability sharing among multiple bands.

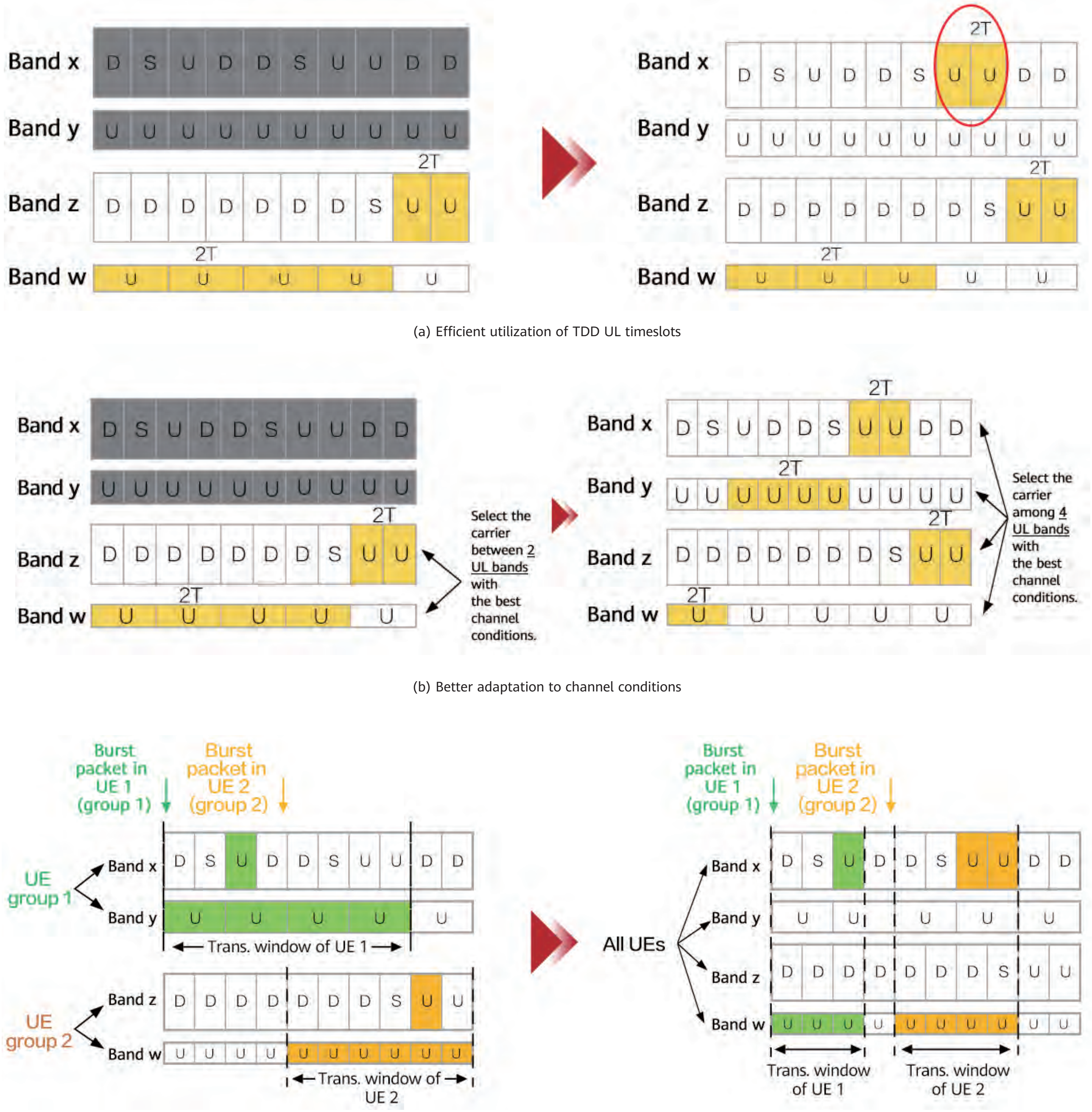
However, to actually achieve the above benefits and gains offered by FSA, several directions should be further studied.

- **Hardware.** UE hardware should be designed to minimize the switching latency or cost. For example, caching can be introduced to reduce the latency.

- **Measurement.** The cost or overhead of channel measurement increases with the number of spectrum bands. On-demand measurement, channel extrapolation and channel prediction can be used to address this issue.

- **Scheduler.** Intelligent selection and allocation of spectrum resources is critical for FSA, which requires an advanced scheduler with balanced tradeoff between the algorithm complexity and optimized performance.




Band x
Band y
Band z
Band w
D S U D D S U U D D
U U U U U U U U U U
D D D D D D D S U U
U U U U U
2T
Select the carrier between 2 UL bands with the best channel conditions.
Select the carrier among 4 UL bands with the best channel conditions.
Burst packet in UE 1 (group 1)
Burst packet in UE 2 (group 2)
UE group 1
UE group 2
All UEs
Trans. window of UE 1
Trans. window of UE 2

(a) Efficient utilization of TDD UL timeslots

(b) Better adaptation to channel conditions

(c) Higher trunking efficiency

Figure 7 FSA benefits

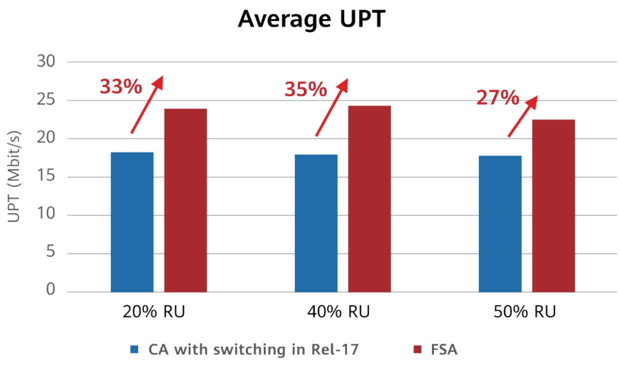


Figure 8 Average UPT performance of FSA (2Tx UE)

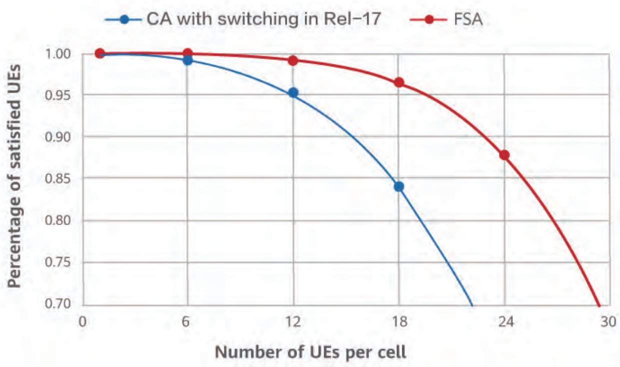


Figure 9 XR uplink capacity performance of FSA (2Tx UE)

3.1.2 Multi-band Power Aggregation

For UEs supporting CA, instead of aggregating spectrum resources, Tx power is aggregated among multiple bands based on PA capabilities to achieve higher transmit power on one band. For example, in 2 UL-band TDD CA, UEs support at least 1 PA and 1Tx on each band. As such, UEs can support a maximum of 26 dBm plus 26 dBm — for example, 29 dBm transmit power after aggregation, as shown in Figure 10.

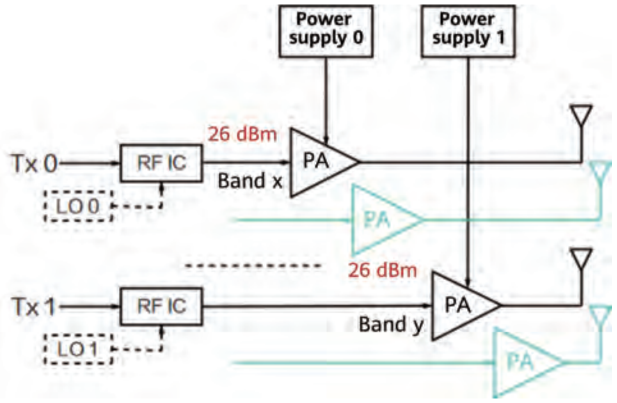
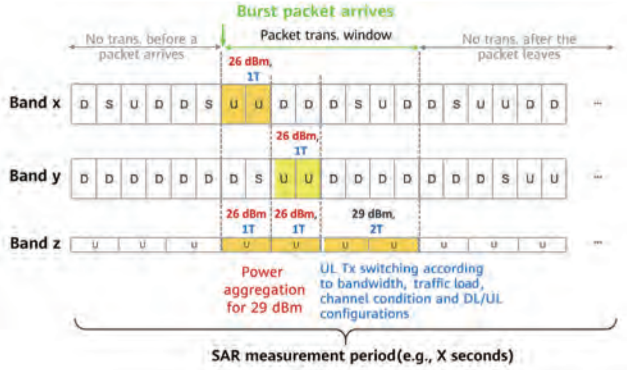


Figure 10 UE power aggregation among multiple bands

Moreover, power aggregation can work together with FSA. To begin, Tx and PAs are dynamically selected, and the selected PAs transmit at their respective maximum power. For example, as shown in Figure 11, one 2Tx UE (concurrent transmission with two PAs) transmits in bands x and z with two-band power aggregation in the first timeslot when a burst packet arrives, and then switches to bands y and z with two-band power aggregation in the second timeslot, according to bandwidth, traffic load, channel condition and DL/UL configurations. This mechanism is especially useful for burst traffic, since it can boost UL peak data rate and correspondingly increase UL UPT with boosted bandwidth and power.


Burst packet arrives
No trans. before a packet arrives
Packet trans. window
No trans. after the packet leaves
26 dBm, 1T
26 dBm, 1T
26 dBm 1T
26 dBm, 1T
29 dBm, 2T
Power aggregation for 29 dBm
UL Tx switching according to bandwidth, traffic load, channel condition and DL/UL configurations
SAR measurement period(e.g., X seconds)

Band x	D	S	U	D	D	S	U	U	D	D	D	S	U	D	D	S	U	U	D	D	...
Band y	D	D	D	D	D	D	D	S	U	U	D	D	D	D	D	D	D	S	U	U	...
Band z	U		U		U		U		U		U		U		U		U		U		...

Figure 11 Multi-band power aggregation with FSA

3.2 Time Domain

With the rapid increase of communication applications, including personal data service and Industry 4.0, a significant gap has widened between the diverse performance demands of different communication applications on latency and throughput and the fixed supply of existing synchronized TDD (Sync-TDD), which requires exploration of more flexible schemes. Promising duplex schemes which aim to cope with the above challenges include flexible duplex (i.e., dynamic TDD), subband full duplex, and inband full duplex.

3.2.1 Flexible Duplex

High-density small cells have become a trend of 5G communications, leading to varying requirements for traffic in different cells. In addition, there is a high possibility that a macro cell with DL-dominant TDD configuration is deployed nearby small cells. Applying the Sync-TDD scheme in such scenarios results in a waste of UL and DL timeslots in some cells due to the fixed timeslot allocation. As a result, the throughput is below expectations and the latency increases.

To cope with this challenge, flexible duplex is regarded as an effective method as it configures UL and DL timeslots for cells independently, as shown in Figure 12. Flexible duplex can improve throughput and latency. However, due to the change of UL and DL channels in different cells, severe cross-link interference (CLI) may occur. As long as the UL signal can combat the CLI from the DL, this scheme can provide significant UL capacity.

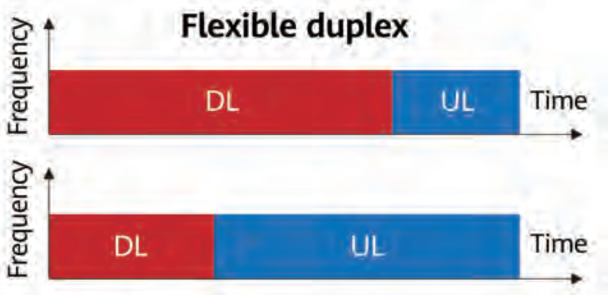


Figure 12 Illustration of flexible duplex

In order to demonstrate the performance of this scene, we provide the following simulation results via system-level evaluation. In scenarios where 3 macro cells coexist with 18 small cells in a factory, the TDD configurations of the macro cells and the small cells are DDDSU and DSUUU, respectively. It is assumed that the interference rejection combining (IRC) receiver and coordinated multi-point (CoMP) reception are applied by the small cells. More detailed parameter settings are included in Table 1.

By analyzing the simulation results shown in Figure 13, it is clear that flexible duplex offers more significant boosts to both average and edge cell throughputs compared to Sync-TDD, though CLI exists. It also shows that CLI mitigation can further unlock the potential of flexible duplex.

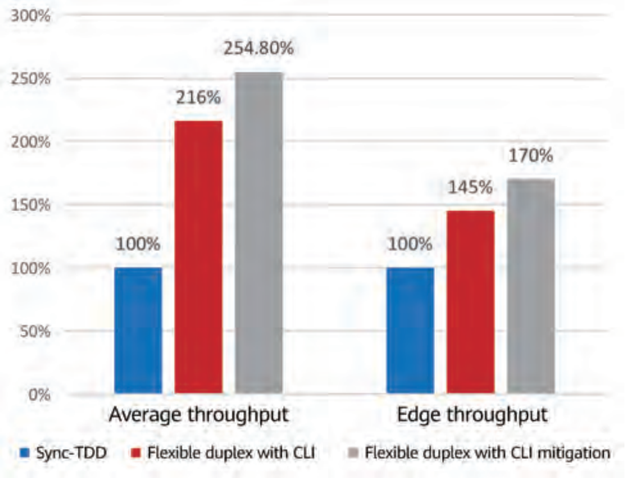


Figure 13 UL throughput gain of flexible duplex

3.2.2 Subband Full Duplex

To achieve both high UL capacity and the required low latency, subband full duplex or complementary TDD can be adopted, as shown in Figure 14. In this scenario, some of the small cell's frequencies are used for UL traffic, while the remainder is used in the same timeslot for DL traffic. From the network's perspective, both DL and UL resources are always available at the same time, significantly reducing latency when compared with the DL-dominant TDD mode. Of course, several CLIs still exist, such as macro-to-small cell CLI in the UL subband, adjacent-channel CLI between macro and small cells, adjacent-channel CLI between small cells, adjacent-channel self-interference within small cells, and adjacent-channel CLI between UEs. Throughput can be

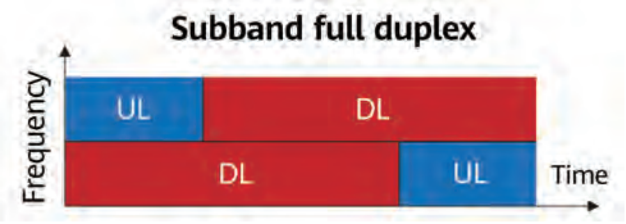


Figure 14 Illustration of subband full duplex

Table 1 Detailed simulation parameters

Parameter	Details
Multiple access	Orthogonal frequency-division multiple access
Duplex	TDD Macro: DDDSU Pico: DSUUU
Carrier frequency	4.9 GHz
Inter-site distance	Macro: 300 m Pico: 20 m
Modulation	Up to 256QAM
Numerology	30 kHz
Channel model	Macro: Uma Pico: TS 38.901 for IIoT
UE distribution	Macro: 80% indoor with 3 km/h, 20% outdoor with 30 km/h Pico: 100% indoor with 3 km/h
Simulation bandwidth	20 MHz
Antenna configuration	Macro: BS@32 Tx RUs, UE@4 Tx RUs Pico: BS@8 Tx RUs, UE@4 Tx RUs
Transmission scheme	Macro: up to 12 layers Pico: up to 12 layers (6 cooperative TRPs)
Scheduling granularity	4 RB
Maximum UE transmit power	26 dBm
Scheduling	Proportional fairness
Receiver	MMSE-IRC
Power control parameter	P0 = -60, alpha = 0.6
TRP quantity	Macro: 3 Pico: 18
SRS transmit period	10 transmission time intervals
Scheduling granularity in the time domain	1 timeslot
UE quantity	Macro: 10 Pico: 4
BS transmit power	60 dBm

greatly enhanced if the co-channel self-interference and co-channel UE-to-UE CLI are carefully studied.

Taking the foregoing factory scenario as an example, calculating the interference link budget shows that the UL signal-to-interference-plus-noise ratio (SINR) can be larger than 10 dB even when considering the co-channel and adjacent-channel CLIs. This proves that subband full duplex is feasible in the factory scenario. Note that self-interference in small cells is assumed to be mitigated via Tx/Rx separation.

A more challenging scenario is subband full duplex in macro deployment, where each macro base station (BS) is a subband full-duplex BS with three sectors in most cases. Our analysis shows the following:

- For one interfering BS, the adjacent-channel interference is about 22 dB above the noise and may be partially suppressed by the IRC receiver. However, since this interference is nonlinear, more advanced interference measurement and suppression technologies should be further studied.
- Another challenging aspect is the blocking interference at a level of -20 dBm in cases where the UL receiver of subband full duplex is equipped with a wideband front-end filter before its low-noise amplifier. This may block the UL.

As subband full duplex represents a potential study point in Rel-18, the above challenges require urgent attention.

3.2.3 Inband Full Duplex

Inband full duplex promises a higher spectral efficiency in future communication networks than both flexible duplex and subband full duplex. For inband full duplex, both BSs and UEs, or at least BSs, are transmitting and receiving on the same frequency band at the same time, which has the potential to double the system throughput and reduce communication latency. However, interference will become even more complex for communication networks, as both self-interference and CLI will have to be considered. Further study is required to overcome these challenges.

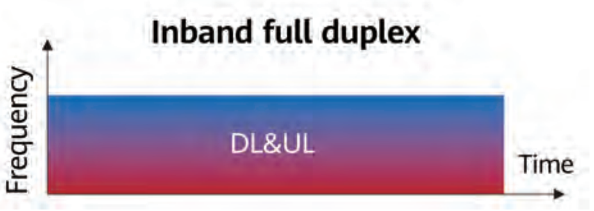


Figure 15 Illustration of inband full duplex

3.3 Space Domain

UL MIMO could potentially offer significant improvements to UL capacity. UL capacity can be further improved by adopting higher-resolution UL precoding and enabling higher-order spatial multiplexing as well as optimizing power control. We will also discuss the ambitious optimization objective of UL capacity, which considers practical scheduling and transmission factors in multiple transmission reception points (multi-TRP) scenarios.

3.3.1 High-Resolution UL Precoding

Indicating UL precoding to UEs occupies DL time-frequency resources and requires careful design to retain the accuracy of UL precoding with a low overhead. Two UL transmission modes are supported in current 5G specifications: codebook-based (CB) and non-codebook-based (NCB).

- For the CB mode, UL precoding is derived based on the sounding reference signal (SRS) sent by a UE and then indicated back through DL control signaling. Due to limited DL control signaling overhead, only the UL wideband coarse codebook (composed of just ±1 and ±j) is employed. Furthermore, the current codebook design cannot match different types of UEs, such as those with irregular antenna shapes or patterns.
- For the NCB mode, UEs measure DL channel state information reference signals (CSI-RS) and calculate high-resolution precoding by virtue of channel reciprocity. UEs then send weighted SRSs using individually preferred UL precoders back to the BS or central processing unit (CPU) for final determination, taking multi-user (MU) interference into consideration.

The drawbacks of both CB and NCB modes are summarized as follows:

- The current UL precoder determination is wideband-wise, with one precoder applied to the entire scheduled resource of one UE.

- A powerful UE may be equipped with more than two antennas, particularly in the case of certain industrial applications. Higher-resolution precoding considering MU interference is required to improve the overall network capacity.

As a result, frequency-selective and high-resolution precoding is preferred and required to retain the accuracy of UL precoding. Our initial system-level simulation shows that an additional 20% of cell average throughput gain can be achieved compared to current CB and NCB modes. Correspondingly, the following two directions can be considered for UL precoding enhancement in specifications:

- UL precoding indication via weighted DL CSI-RSs
- UL precoding indication via multi-level DL signaling for overhead reduction

3.3.2 High-Order Spatial Multiplexing

A maximum of 12 orthogonal antenna ports are supported in current 5G specifications. Generally, non-orthogonal demodulation reference signal (DMRS) ports can be configured for scenarios involving more than 12 UL layers, but this degrades the accuracy of UL channel estimation due to relatively high cross-correlation among DMRS ports. To support additional potential UL transmission layers with less impact on UL channel estimation, two possible directions can be studied and specified: a higher maximum number of orthogonal DMRS ports, or low-correlation DMRS ports. As illustrated in Figure 16, the performance of up to 24 layers with DMRS enhancement can achieve a cell average gain of 68% compared with the non-orthogonal DMRS of current specifications in our initial simulation results in the industrial IoT (IIoT) scenario.

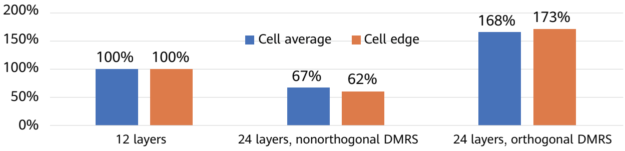


Figure 16 Performance gain of enhanced UL DMRS

3.3.3 Multi-TRP Power Control

More potential UL layers and joint UL multi-TRP processing introduce higher requirements on UL power control, while the optimal values of transmit power are usually arbitrary and differ from layer to layer. However, the existing power control in current specifications restricts the UL performance due to the following:

- One UE may be paired with different UEs in various timeslots for MU-MIMO, and this dynamic pairing can

lead to a large UL transmit power variation which can be larger than the existing power control adjustment steps. Both closed-loop and open-loop power control schemes can be enhanced to match this large power variation.

- For multi-TRP reception, it is recommended to perform UE power control based on the path loss from all the TRPs involved in joint processing. However, current UL power control in standards only relies on the path loss from one serving TRP.

Our initial simulation results show that an approximate cell edge performance gain of 20% can be achieved via enhanced power control.

3.3.4 Multi-TRP Joint Transceiver

Multi-TRP joint transceiver is regarded as the preferred technology for improving UL capacity. As per Figure 17, consider an UL cellular network with a set of deployed TRPs and a set of candidate UEs. Each UE is associated with a subset of deployed TRPs and transmits over a subset of frequency units. For each scheduling timeslot in a practical system, the BS would choose a number of UEs (scheduled UEs) for UL transmission on the candidate frequency units within the working bands. For each frequency unit, one or more scheduled UEs would be spatially multiplexed and their transmitted signals can be demodulated due to different channel directions. Statistically, the BS would fully consider the scheduling possibility of each UE to guarantee the fairness and/or minimum quality of service (QoS) requirements.

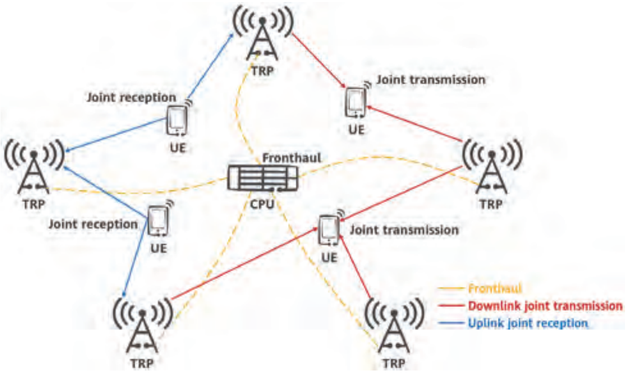


Figure 17 Multi-TRP joint transceiver

The target for UL scheduling and transmission is maximizing the sum rate over all possible receive and transmit weights based on all possible scheduled user sets, frequency unit sets, and the chosen TRPs set for each UE. However, the sum-rate maximization requires joint optimization of the transmit weight, receive weight, and UE scheduling in the frequency and time domains, where its non-convex and combinatorial characteristics prevent a global optimal solution with affordable complexity from being attained. A number of major works already study weight design under a given UE scheduling and TRP association, including:

- Given the scheduling result and transmit weight, the optimal linear receive weight is actually equal to the linear minimum mean square error (LMMSE) weight, which can be easily demonstrated by reformulating the objective function as a Rayleigh quotient with respect to the receive weight.

- When the successive interference cancellation receiver is used, iterative water-filling is an effective technique for transmit weight design [5].

- When the receive weight is fixed as the LMMSE weight, the problem of achievable-rate maximization with respect to the transmit weight of each UE is equivalent to the quadratic form applying the weighted minimum mean square error (WMMSE) algorithm [6] or matrix fractional programming [7], which allows the close-form solution in each iteration to be obtained. The authors of [8] proposed jointly optimizing the transmit and receive weights in an alternating manner.

If UE scheduling is further considered as a part of optimization, we can derive an equivalent reformulation of the original problem by introducing a binary indicator variable in the objective function indicating if a UE will be scheduled in a given frequency and time unit. This reformulation leads to a challenging mixed integer programming problem. Sparse beamforming is an emerging technique that introduces an L0 norm constraint for the transmit weight energy of each UE to force the transmit power of a UE at some frequency and time units to be zero, while also implicitly determining UE scheduling [9].

In addition, two major challenges exist regarding the implementation of multi-TRP joint transceivers:

- The first challenge relates to intractable centralized processing due to the unaffordable computational complexity of high-dimensional matrix operations. Decentralized implementation, which utilizes local channel state information (CSI) for weight design at each TRP along with simple aggregation of the local processing results among TRPs at the CPU, is a potential scheme [10]. However, performance degradation is inevitable due to inadequate exploitation of information from cooperative TRPs. As such, a low-complexity solution with guaranteed performance remains desirable.

- Another challenge stems from limited backhaul and fronthaul overhead, which impedes ideal cooperation. A potential technique for addressing this issue involves exploiting proper over-the-air (OTA) signaling for UEs to feed back information from cooperative TRPs. Each TRP carries out weight design individually, utilizing local CSI and information feedback from each UE. To approach the level of performance possible under ideal cooperation, multiple-round iterations of information exchanges are required via feedback and feedforward [11]. However, reducing the overhead of OTA feedback and the number of iterations in such cases is still an open issue.

3.4 Power Domain

UL capacity can also be improved in the power domain. This can be done either by directly increasing UEs' transmit power or by virtually raising the transmit power through UE cooperation.

3.4.1 Higher Maximum Power

Current commercial PA hardware for frequency division duplex (FDD) and TDD supports a maximum of 23 dBm and 26 dBm transmit power, respectively. In other words, a higher-power UE is adopted only for TDD. The key reason for this is because, in reality, DL-dominant TDD configuration is used with a smaller number of UL timeslots in one TDD configuration period. In addition, most of the timeslots are DL timeslots, and UEs will not transmit in such timeslots. When UEs transmit at the maximum power in the UL timeslots, the average UL power over the TDD configuration period remains small and the specific absorption rate (SAR) requirements are met.

However, in FDD, UL timeslots are continuous in the time domain, and if UEs transmit at the maximum power (for example, 26 dBm), the SAR requirements may not be met. To increase the maximum power of UEs beyond 23 dBm, a longer maximum power averaging window (i.e., long-term duty-cycle control) can be considered in cases where higher UL power can transmit UL traffic in a shorter duration, enabling UEs to rest at other points within the window in order to satisfy the SAR requirements, as shown in Figure 18.

3.4.2 UE Cooperation

UE cooperation (or UE aggregation) via sidelink (SL) device-to-device communications can be used to improve the overall UL experience of both capacity and coverage. As illustrated in Figure 19, a cooperative UE (CUE) can be leveraged to operate cooperatively with the source UE (SUE) in the scenarios where data rate or coverage requirements cannot be fully met [12].

- The CUE and SUE can be linked through an SL to form a virtual UE featuring higher capabilities due to the aggregated transmit power and distributed transmit antennas.
- When the SUE is out-of-coverage, the CUE can act as a relay to forward the SUE's UL data to the network.
- In addition, the CUE can be used as a diverse UE or a backup UE of the SUE to ensure reliability and latency requirements, especially in the case of IIoT applications.

As illustrated in Figure 20, the performance of UE cooperation with 1 CUE can achieve cell average gains of 21% and cell edge gains of 96% when compared with SUE alone in our initial system-level simulation. Moreover, larger gains can be obtained with additional CUEs.

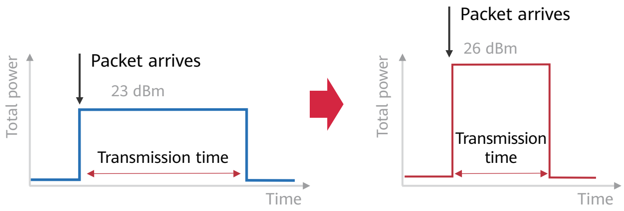


Figure 18 Higher power in a shorter Tx duration

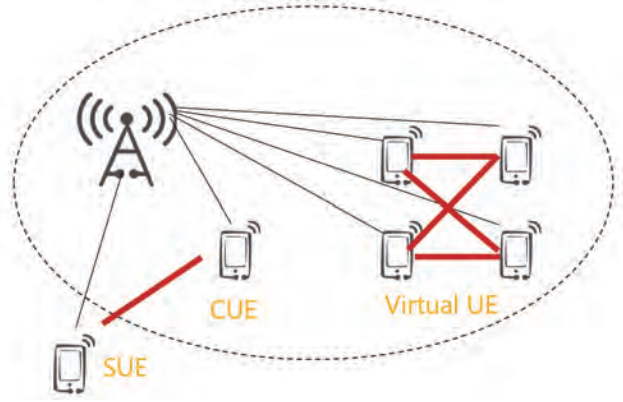


Figure 19 UE cooperation

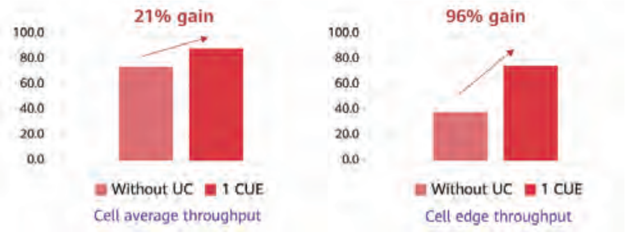


Figure 20 Performance gains via UE cooperation

4 Conclusion

With the rapid development of wireless UCBC applications, particularly those that rely on HD video upload, network UL capabilities have never been so highly valued — and that value is set to further increase in the future. While 5G has established higher UL capabilities than previous generations, further enhancement and evolution are still urgently required. We believe that UL enhancement will become an important research and standardization topic in 5G-Advanced — one that can be carried out from the following four domains: frequency, time, space and power.

References

[1] https://www.3gpp.org/ftp/PCG/PCG_46

[2] https://www.huaweicentral.com/huaweis-5g8k-3d-vr-solution-has-been-released-brings-enhancement-in-5-5g-uplink-ultra-broadband/

[3] GTI white paper. Value of 5G high uplink in industrial digitalization. https://www.gtigroup.org/zx_images/lichunyu/Value%20of%205G%20High%20Uplink%20in%20Industrial%20Digitalization.pdf

[4] Syed AS, Sierra-Sosa D, Kumar A, and Elmaghraby A, "IoT in smart cities: A survey of technologies, practices and challenges," Smart Cities. 2021; 4(2):429-475. https://doi.org/10.3390/smartcities4020024

[5] W. Yu, W. Rhee, S. Boyd, and J. M. Cioffi, "Iterative water-filling for Gaussian vector multiple-access channels," in IEEE Transactions on Information Theory, vol. 50, no. 1, pp. 145-152, Jan. 2004.

[6] Q. Shi, M. Razaviyayn, Z.-Q. Luo, and C. He, "An iteratively weighted MMSE approach to distributed sum-utility maximization for a MIMO interfering broadcast channel," IEEE Trans. Signal Process., vol. 59, no. 9, pp. 4331-4340, 2011.

[7] K. Shen and W. Yu, "Fractional programming for communication systems—part II: Uplink scheduling via matching," IEEE Trans. Signal Process., vol. 66, no. 10, pp. 2631-2644, 2018.

[8] R. Mosayebi, M. M. Mojahedian, and A. Lozano, "Linear interference cancellation for the cell-free C-RAN uplink," in IEEE Transactions on Wireless Communications, vol. 20, no. 3, pp. 1544-1556, March 2021.

[9] B. Dai and W. Yu, "Sparse beamforming and user-centric clustering for downlink cloud radio access network," in IEEE Access, vol. 2, pp. 1326-1339, 2014.

[10] E. Björnson and L. Sanguinetti, "Making cell-free massive MIMO competitive with MMSE processing and centralized implementation," in IEEE Transactions on Wireless Communications, vol. 19, no. 1, pp. 77-90, Jan. 2020.

[11] A. Tolli *et al.*, "Distributed coordinated transmission with forward-backward training for 5G radio access," in IEEE Communications Magazine, vol. 57, no. 1, pp. 58-64, January 2019.

[12] Ma C X, Liu R K, Liao S, *et al.*, "User cooperation scheduling in cellular systems," in Proceedings of 2020 IEEE Globecom Workshops, 2020, Taipei. 1-6.

Technologies for 800 Gbit/s and 1.6 Tbit/s Data Center Modules

Maxim Kuschnerov [1], Talha Rahman [1], Youxi Lin [1], Nebojsa Stojanovic [1], Stefano Calabro [1], Jinlong Wei [1], Wenjun Shi [2], Nian Cai [2], Zhiwei Li [2], Jianyu Zheng [2], Meng Zhou [2], Lihui Hu [3], Fei Yu [4], Jinlin Zeng [5], Qinhui Huang [6], Huixiao Ma [6], Raymond Leung [6], Changsong Xie [1], Lewei Zhang [7]

[1] Optical & Quantum Communications Laboratory, Munich Research Center

[2] Optical System and Algorithm Development Dept, Optical Business Product Line

[3] High-Speed and High-Frequency Lab, Central Hardware Engineering Institute

[4] Board Planning and Architecture Design Dept, Central Hardware Engineering Institute

[5] Precision Manufacturing Laboratory, Manufacturing Dept

[6] B&P Laboratory, Central Research Institute

[7] Data Communication Chip & Optical Architecture Group (Module), Data Communication Product Line

Abstract

Future 800 Gbit/s and 1.6 Tbit/s Ethernet standards will require 200 Gbit/s per lane optical transmission to achieve lower cost and lower power implementation compared with 4 × 100 Gbit/s interface technology. We demonstrate advanced integrated components with higher bandwidth together with high performance digital signal processing using filtering tolerant timing recovery and reduced state sequence estimation in order to demonstrate the technical feasibility of 200 Gbit/s per lane optical transmission. We analyze the modulation format selection and system impairments to arrive at a recommendation for future standards.

Keywords

data center networks, Ethernet, connectivity, 800 GbE, 1.6 TbE

1 Introduction

The build out of cloud infrastructure and the pace of innovation of cloud-based applications has accelerated in recent years, in part fueled by the pandemic. Businesses world-wide have been continuing to migrate their IT infrastructure into the cloud, which is driven by the roughly two year cadence of Ethernet switching capacity doubling with 25.6 Tbit/s switches shipping in 2021. The optical interfaces formed the backbone of data center networks, interconnecting leaf, spine and core switching layers. With 400 Gbit/s modules based on 4 × 100 Gbit/s PAM4 already shipping in higher volumes, the industry is already sampling higher density 100G per lane technology with 8 × 100 Gbit/s modules. Standardization activities turn to 800 Gbit/s and 1.6 Tbit/s interfaces based on 200 Gbit/s per lane optical transmission, being addressed in the 800G Pluggable MSA and IEEE. Figure 1 shows the projected standardization timeline of 800 GbE and 1.6 TbE in IEEE.

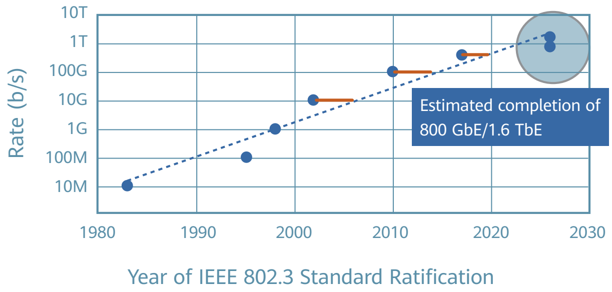


Figure 1 Interactive XR for end consumers

200 Gbit/s per lane optical transmission using intensity modulation with direct detection (IMDD) is seen as the enabler for these future Ethernet rates for up to 2 km transmission, and is competing with coherent optics at 10 km reaches. The technological step towards 200 Gbit/s per lane is technically challenging and requires several enabling technologies on the analog component side and for digital signal processing.

In this paper, we will present the technical verification and discuss enablers for 200 Gbit/s lane, analyzing the modulation choice, timing recovery (TR) for bandwidth limited channels, reduced complexity maximum likelihood sequence estimation and the industry first 200 Gbit/s transmitter and receiver optical subassembly (TOSA/ROSA) end-to-end transmission at 224 Gbit/s (112 Gbaud).

2 200G Optical Intensity Modulation

Higher-level modulation in Ethernet was introduced for the first time for 50 Gbit/s per lane signaling using PAM4. For optical transmission, 50 Gbit/s PAM4 is used in 200 Gbit/s (4 × 50 Gbit/s) optics as well as for 400 Gbit/s LR8. With the transition to 100 Gbit/s per lane, the industry converged on maintaining the same modulation scheme as for 50 Gbit/s per lane, leveraging the design and testing methodology of PAM4. On the host side, electrical 100 Gbit/s PAM4 faced more serious design challenges for use cases such as backplane transmission. The main question for 200 Gbit/s per signaling is whether PAM4 is feasible or alternative modulation formats needs to be considered, including PAM6, PAM8 or discrete multi-tone (DMT).

Higher order modulation formats such as PAM8 or even PAM16 are usually very limited due to their high error floor and so far never have been practical choices in optical standards. DMT, in principle, could offer a more efficient use of the spectrum, however it comes at the cost of higher DSP power at the transmitter and receiver due to the need for frequency domain processing and is generally compromised for short reach interconnections due to its high peak-to-average power ratio (PAPR), which severely limits the link budgets for passive transmission. In standardization discussions on 200 Gbit/s SerDes, PAM4 and PAM6 are being discussed as the most likely modulation choices. Historically, the host side modulation was also used on the line side, leveraging the host FEC.

We performed an initial comparison of PAM4 and PAM6 using an arbitrary waveform generator (AWG) with 33 GHz 3 dB bandwidth, a driver amplifier with 50 GHz 3 dB bandwidth, an externally modulated laser (EML) with 40 GHz, and a receiver side a 75 GHz PIN diode and a 63 GHz digital sampling scope with 160 Gsamples/s [2]. While the receiver side has clearly sufficient bandwidth, the transmitter is still quite limited and not compliant with future needs for PAM4. At the receiver side a semiconductor optical amplifier (SOA) was used as a preamplifier. The gross rates, including overhead for forward error correction, were 224 Gbit/s for PAM4 using 112 Gbaud and 225 Gbit/s for PAM6 using 90 Gbaud. The small difference enabled a more straight forward resampling at the transmitter side, but was not due to any assumptions regarding the actual overhead for future Ethernet.

As shown in Figure 2, it was demonstrated that PAM4 can achieve better overall link budget and lower error floor despite the initially limited transmitter fidelity, which relaxes the burden of FEC dimensioning and can allow for a low power and low latency solution. Moreover, as shown in Figure 3, the multipath interference (MPI) tolerance of PAM4 is much higher than PAM6, thus enabling the crucial double and triple links for the FR reach class, which are the basis for IEEE campus connectivity standards.

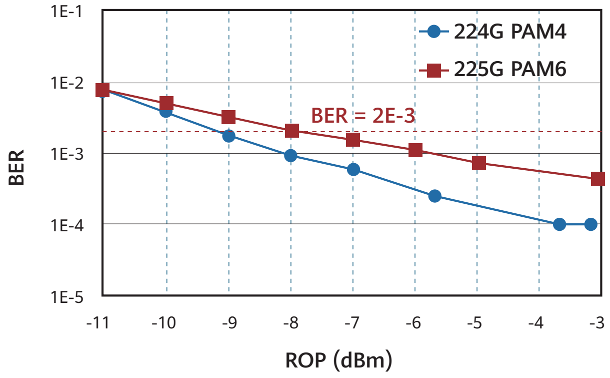


Figure 2 Comparison of PAM4 and PAM6 using an EML

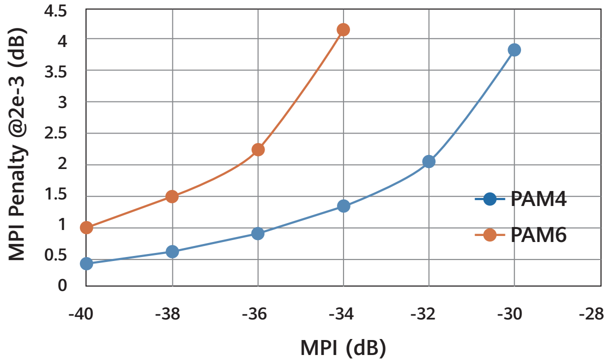


Figure 3 MPI tolerance of 224 Gbit/s PAM4 and PAM6

For a triple link configuration, common for FR interfaces, an MPI penalty of -35 dB worst case has to be assumed. In this case, the MPI penalty of PAM6 reaches 3 dB. The findings on the modulation format choice are consistent with the discussions in OIF on 800LR/ZR, where similar to the 800G Pluggable MSA, a four level signaling format 16QAM based on four level electrical modulation equivalent to PAM4, will be used for optical transmission ranges of 2 km and above. Thus, PAM4 is the optimal choice in the optical domain, given the performance of optoelectronics devices, ASICs and allocation penalties, and it also enables an easier adoption of existing testing methodology at 400 GbE.

3 CD Tolerance of PAM4

Increasing the baud rate of IMDD systems makes the signals more susceptible to chromatic dispersion (CD), as the inter-symbol interference (ISI) originating from CD, scales with the square of the baud rate. The square law function of direct detection leads to the loss of the signal phase and thus makes it only possible to partially compensate for the distortion. Thus, increasing the baud rate automatically puts basic limitations on the maximum reach and/or on the channel spacing in O-band, due to the non-zero dispersion at and around 1310 nm. Figure 4 analyses the CD tolerance of 224 Gbit/s PAM4 assuming an EML transmitter with a chirp of 0.5. A positive frequency chirp interacts with chromatic dispersion either constructively or destructively and shifts the CD tolerance window in this case, towards the shorter wavelengths.

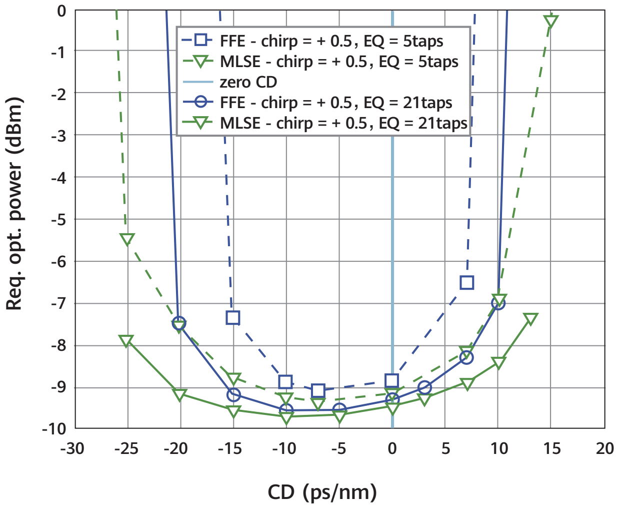


Figure 4 Chromatic dispersion tolerance of 224 Gbit/s PAM4 for different equalizer configurations

For 800 Gbit/s modules using 4 × 200 Gbit/s line signaling over wavelength division multiplexing (WDM) for the 2 km transmission class, a coarse WDM grid (CWDM4) with 20 nm channel spacing would be the preferred choice in order to achieve uncooled transmission and thus lower the cost of the module. On the other hand, reducing the channel spacing could reduce the CD penalty but increase the module cost by requiring thermo-electric cooling (TEC).

ITU-T G.652 specification defines the ranges of chromatic dispersion variation which is shown in Figure 5. For a CWDM4 grid starting at 1270 nm and going to 1330 nm with ±6.5 nm variation, the worst case CD values at 2 km are below -12 ps/nm for the shortest wavelength and below

6 ps/nm for the longest one. When analyzing the CD penalty for the different configurations in Figure 4, it becomes clear that a 5-tap feed forward equalizer (FFE), (as defined in IEEE for the 400 GbE reference receiver), is insufficient, and chirp management at the transmitter becomes necessary, unless a longer FFE is used in combination with a maximum likelihood sequence detection (MLSE).

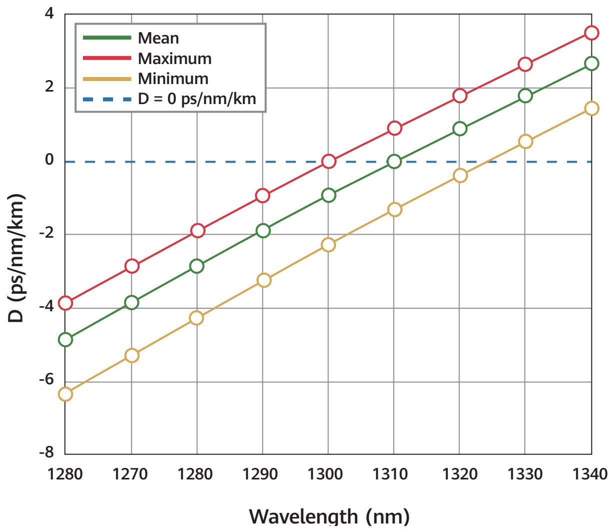


Figure 5 Chromatic dispersion coefficient of SSMF specified in ITU-T G.652

The first analysis of 224 Gbit/s PAM4 optical transmission thus points to technical difficulties when increasing the overall end-to-end bandwidth and a need for more advanced detection. Thus, digital signal processing (DSP) will require a timing recovery for bandwidth limited channels, especially when using PAM4 not only for optical, but also for electrical backplane interconnects and a power efficient MLSE, which likely will have to become part of the reference receiver in future standards.

4 Timing Recovery for Bandwidth-Limited Channels

The use of advanced DSP in faster than Nyquist (FTN) systems seems to be a promising solution to minimize the requirements on component bandwidth [2]. However, in such systems, DSP normally relies upon advanced algorithms to enable data recovery, and the TR design becomes very critical. In particular, the TR requires very sophisticated algorithms such as advanced phase detection, TR equalization, advanced lock detection (LD), and accurate TR loop design.

Basic data center transceiver TR blocks are presented in Figure 6. An analog-to-digital convertor (ADC) samples the input analog signal at a symbol rate (one sample per symbol ADC; 1 sps) to save power that is almost proportional to the ADC sampling rate. A single clock source (an oscillator) may support more channels, e.g., four channels in 4 × 200G systems, to prevent crosstalk and decrease the transceiver cost. A phase interpolator (PI) adjusts frequency and phase of the ADC clock signal. The phase-frequency information of the received signal clock is obtained by a phase detector (PD). The PD output signal is averaged by a TR filter (TRFIL) that is often realized via a proportional-integral filter so that the TR can be modeled by a second-order loop. When the received signal suffers from serious intersymbol interference (ISI), the PD may be incapable to provide sufficient PI driving signal, and a PD filter (PDFIL) is necessary for improving clock extraction. TR control and optimization (TR C&O) algorithms support and optimize the TR functionality. Their tasks include PDFIL and TRFIL taps setting, sampling phase optimization, and detecting TR acquisition and TR instability. It should be mentioned that the TR lock detector design in 1 sps systems with severe ISI is extremely difficult and should be done with care.

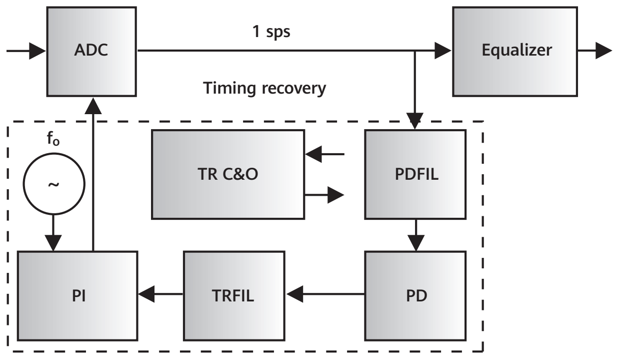


Figure 6 Timing recovery block diagram

High-baud rate ASIC-implemented 1 sps PDs are mainly based on the so-called Mueller and Müller PD (MMPD) and its variants [3]. The sign MMPD is more convenient for 2-level signals, whereas these variants may cause TR instabilities for higher-order modulation formats. This PD uses two samples, x_1 and x_2, taken at a symbol period distance, and the TR is controlled by a signal $x_1\text{sign}(x_2)-x_2\text{sign}(x_1)$.

An advanced PD, the so-called abs PD (ABSPD), supports all modulation formats and has the best performance in ISI channels [4]. It uses a signal abs(x_1+x_2)(abs(x_1)−abs(x_2)) to adjust the ADC clock. These two PDs have been compared in 112GB PAM4 experiments with the total system 3-dB bandwidth of 41 GHz and 44 GHz without and with a digital predistortion (DPD), respectively. A closed-loop root-mean-square jitter (J_{rms}) versus a number of PDFIL taps is presented in Figure 7. The MMPD is at least 1 dB inferior to the ABSPD and this difference reaches 3 dB for the 5-tap PDFIL (no DPD used).

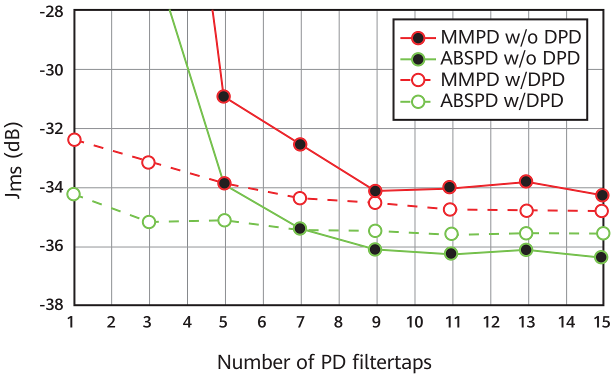


Figure 7 Coot-mean-square jitter in 224 Gbit/s PAM4 experiments

The PDFIL is a linear filter that may have five or more taps. A blind PDFIL taps setting can become difficult as ISI and clock offset are not known in advance. A gradient algorithm has to be selected to enable a reliable and fast taps setting. However, TRFIL taps may significantly vary during the acquisition phase and thus change a loop bandwidth, causing TR instabilities. To prevent this problem, the TR may combine several PDFILs to emulate a phase-frequency detector behavior which keeps the loop bandwidth constant. Using the PFD structure enables the fastest TR acquisition, stable TRFIL taps setting, and very accurate TR parameters estimation such as a phase detector gain (Kpd) and offset. The Kpd estimation results during the acquisition phase are shown in Figure 8 and the Kpd is more accurate in a longer estimation period.

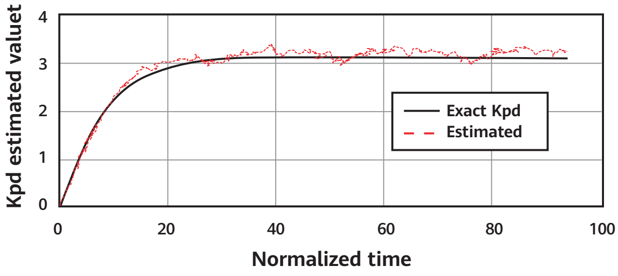


Figure 8 Phase detector gain estimation

Some components such as digital-to-analog convertors (DAC) could have a nonlinear phase transfer characteristic and the aforementioned PDs will not always work at the optimum sampling phase. Therefore, the DSP must be able to find the best sampling phase, and control the TR to remain at this phase. The best sampling phase depends on the equalizer structure, which can be for Volterra filter for example. For such a system, BER at different sampling phases for four different equalizers is presented in Figure 9, where *VF*(*n*,*k*) denotes a Volterra filter with *n* linear taps and *k* symbols used in the second-order part of the filter. The equalizer can work in a sampling phase dithering mode to find the minimum BER sampling location. The sampling phase can be adjusted by manipulating the PDFIL taps.

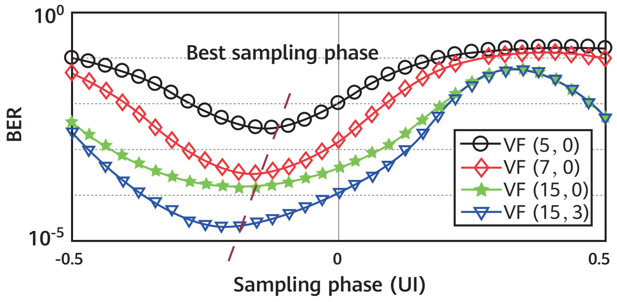


Figure 9 BER versus sampling phase with different equalizers

5 Maximum-Likelihood Sequence Estimation

If the channel is limited in bandwidth and/or exhibits gain dips in the passband, a linear full-response equalizer tends to enhance the power spectral density of the noise in the frequency regions that require amplification. The criterion of the minimum mean squared error (MMSE) sets the weights of the equalizer according to the best possible trade-off between noise enhancement and residual ISI in terms of the overall mean squared error. Nevertheless, if the amplitude response is not flat over frequency or, more generally, if the channel response is not unitary, this receiver architecture is not optimal.

In [5], Forney introduced a receiver structure consisting of a linear filter, called a whitened matched filter, a symbol-rate sampler, and a recursive nonlinear processor based on the Viterbi algorithm (VA). He proved that this type of receiver implements a maximum-likelihood estimator (MLSE) of the entire transmitted sequence. The whitened matched filter consists of the cascade of the matched filter and a symbol-spaced whitening filter obtained by spectral factorization of the noise autocorrelation function. In this receiver, noise whitening is crucial because the branch metrics for the VA are computed as the squared Euclidean distance between the symbols of the constellation and the symbol-spaced samples, which is only correct if the noise samples are uncorrelated.

In a practical situation, Forney devised that a near-optimum procedure is to use a linear equalizer to shape the actual channel to some desired target channel whose symbol-spaced impulse response $f(D)$ is short and whose spectrum is similar to the actual channel spectrum, and then use a VA that is appropriate for $f(D)$. Ideally, $f(D)$ reproduces exactly the amplitude response of the actual channel, so that the linear equalizer does not enhance the noise power. On the other side, the length of the impulse response $f(D)$ determines the complexity of the VA and, therefore, it is clear that under a complexity constraint a certain approximation error must be tolerated.

This approach implements receiver-side partial-response equalization and maximum-likelihood sequence detection (MLSD) based on the VA. In contrast to a full-response equalizer, a partial-response equalizer assumes as the shaping target a channel with a specific amount and structure of ISI.

If $f(D)$ is conveniently chosen as a binomial of the type 1±D^n, the partial response channel has one tap of channel memory and the VA has only M states, where M is the number of constellation symbols. A common choice, already proposed by Forney, is the duobinary response $f(D) = 1 + D$, which approximates a severe bandwidth limitation. It must be noted that the VA does not require channel estimation because the channel response is shaped into a known profile by the linear equalizer.

In [6], Li et al. propose to implement Forney's partial-response equalizer as the combination of an adaptive full-response equalizer and a fixed symbol-spaced post-filter with impulse response $f(D)$.

In [7] the authors introduce a tap coefficient α ($0 \leq \alpha \leq 1$) in the post-filter transfer function to obtain the generalized duobinary (GDB) response $f(D) = 1 + \alpha \cdot D$, while maintaining the one-symbol memory. This additional degree of freedom permits to tune the receiver-side spectral shaping target according to the actual channel response, thereby improving the overall performance.

A VA is known to generate error bursts at its output, which could impact the performance of the subsequent FEC decoder. As already noted by Forney in [5], duobinary (DB) precoding can be used at the transmitter to limit the length of the most likely error events after the MLSD to exactly two symbols. In this case the VA is configured to yield an estimate of the received (rather than the transmitted) sequence, which is thereafter reduced modulo of the cardinality of the PAM alphabet. In general, for small-to-moderate values of the coefficient α, DB precoding is dispensed with since it results in some performance penalty. In this case the FEC code shall be dimensioned to cope with the expected error distribution.

In Figure 10 we illustrate the type of channel response that can be ideally equalized using a post-filter of the type $f(D) = 1 + \alpha \cdot D$ for PAM4, assuming 224 Gbit/s transmission with 12% FEC overhead and RRC shaping with roll-off 0.2. As α increases, a more severe bandwidth limitation is emulated. For α = 1 a spectral notch at the symbol rate is emulated. However, in practice such a notch should be avoided because of its detrimental effects on all the receiver modules before the post-filter.

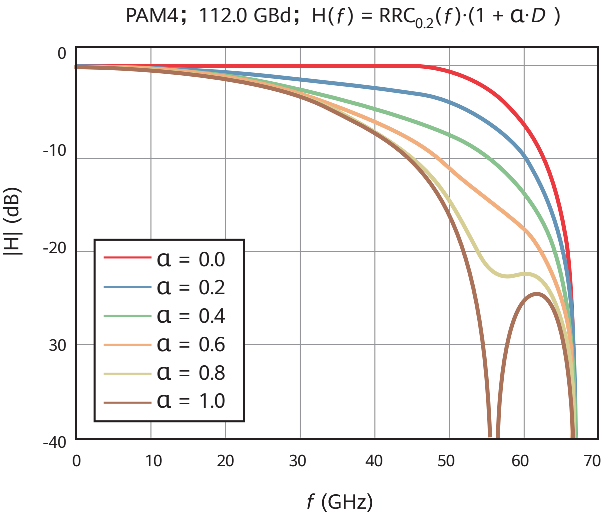


Figure 10 Analog spectrum of a 224 Gbit/s PAM4 signal, which results in ideal partial response signal of the type $1 + \alpha \bullet D$ after RRC matched filtering with roll-off 0.2

We observe that the parameter α can be adapted during operation according to the spectrum of the noise process after the full-response equalizer. Finding the optimal whitening post-filter is tantamount to estimating the autoregressive model 1/ (1 + α·*D*), which fits the sequence of noise samples. For this purpose Burg's technique is the recommended approach (the Yule-Walker method is also feasible but is known to be less stable). Burg's technique can be implemented on the microcontroller and the result can be used to configure both the post-filter and the MLSE.

This approach makes it apparent that the post-filter can be optimized only on the base of the symbol-spaced samples without precise knowledge of the actual response of the analog channel. In Figure 11 we show the spectrum of the useful signal at 1 sample per symbol (sps) after the post-filter for PAM4. Again, we assume 200G transmission with 12% FEC overhead (the roll-off is irrelevant).

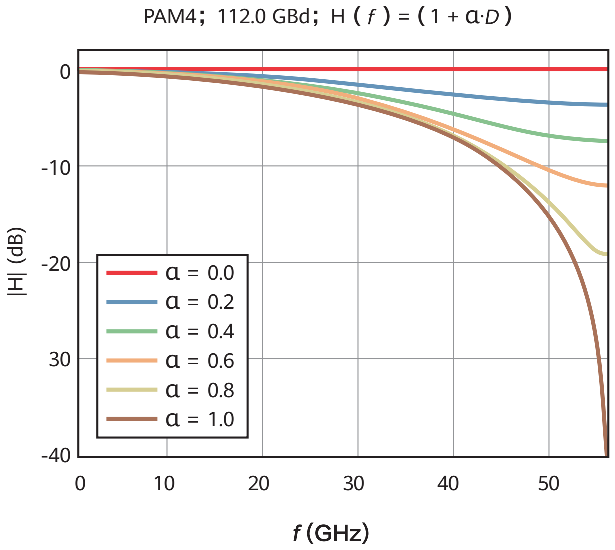


Figure 11 Spectrum of post-processed 224 Gbit/s PAM4 at 1 sps

Finally, if a soft-decision FEC decoder is used, the VA must be replaced by a soft-output recursive nonlinear processor that provides log-likelihood ratios (LLRs) for the encoded bits. In this case a soft-output Viterbi algorithm (SOVA) or the BCJR maximum a posteriori (MAP) algorithm can be employed.

6 Reduced State Sequence Estimation

Eyuboğlu and Qureshi proposed in [8], reduced-state sequence estimation (RSSE) as a technique to simplify MLSE.

As illustrated in Figure 12 for bipolar PAM4 over the generalized duobinary (GDB) channel 1 + α·*D*, the algorithm reduces to the following steps:

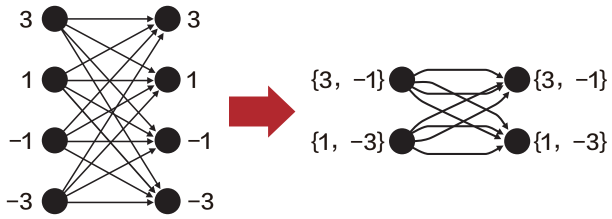


Figure 12 Simplification of the MLSE trellis according to the RSSE algorithm for the PAM4 case

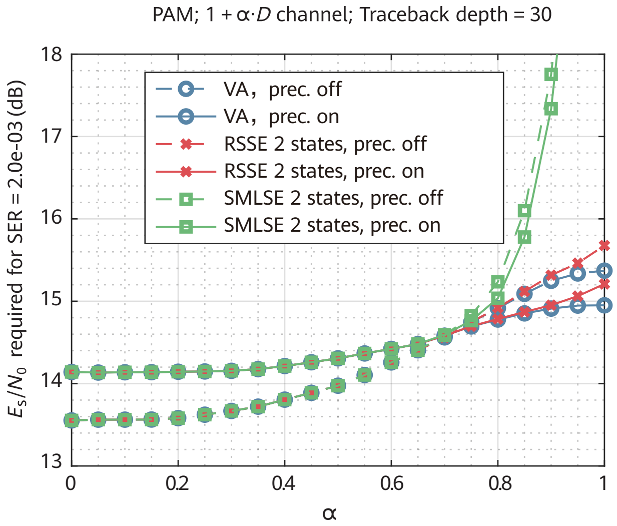


Figure 13 Performance comparison between VA, RSSE, and SMLSE for PAM4 over the GDB channel

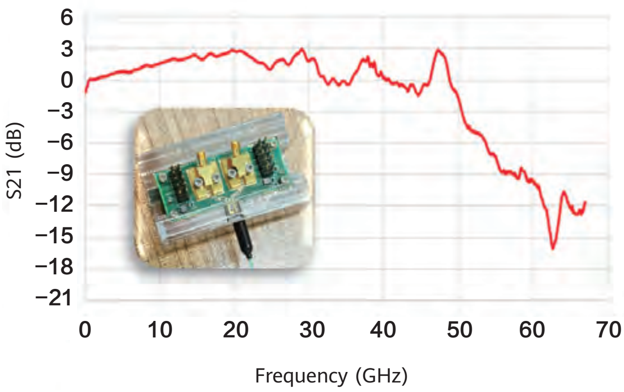


Figure 14 S21 of the TOSA

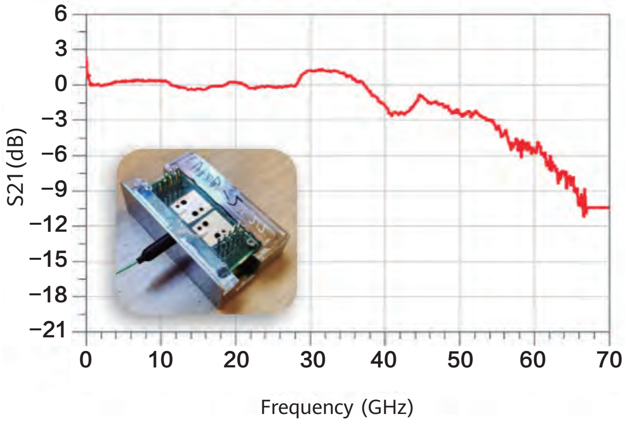


Figure 15 S21 of the ROSA

- Pairs of states corresponding to symbols at twice the minimum distance are merged.
- Since there is no 1-to-1 correspondence between states and symbols, each state must keep track of the surviving symbol.

At each stage:

First, the parallel transitions are processed, making a delay-free decision between the symbols 3 and 1 or 1 and 3 by simple slicing.

Then, non-parallel transitions are processed, selecting at each state the path with the lower accumulated metric (2 paths per state are compared).

Figure 13 illustrates a comparison of VA, RSSE and simplified MLSE (SMLSE) in case of bipolar PAM4 transmission over the generalized duobinary (GDB) channel 1 + α·*D* in the presence of AWGN. SMLSE is a simplification of the MLSE algorithm, which, like RSSE, operates on a 2-state trellis [9], but, different from RRSE, uses pre-decisions instead of a state-merging technique. The figure reports the required value of the electrical signal-to-noise ratio to reach a symbol error rate (SER) of 2e-3, roughly corresponding to a bit error rate (BER) of 1e-3 in case of Gray mapping. The three algorithms achieve similar performance for low and moderate values of the parameter α. However, for α > 0.8, SMLSE fails, whereas RSSE achieves a similar performance as the optimal VA.

In conclusion, for the equalization of PAM4 over the channel 1 + α·*D*, optimal detection requires trellis processing with 4 states and is based on the Viterbi algorithm (VA). Simplified trellis processing with 2 states is possible using either reduced-state sequence estimation (RSSE) or SMLSE. The 4-state VA achieves overall the best performance; RSSE achieves almost everywhere the optimal performance and suffers only from a minor penalty at very high values of α. SMLSE performs well at low to moderate values of α, but fails for α > 0.8.

7 224 Gbit/s Optical End-to-End Verification

So far, the verification of 224 Gbit/s PAM4 was performed on discrete components, which suffered from bandwidth limitations and reflections due to the non-optimized RF connections. In this paper, we present for the first time an end-to-end verification with prototypes of TOSA and ROSA modules.

The receiver optical subassembly uses a photodiode with > 56 GHz 3 dB bandwidth and a SiGe transimpedance amplifier (TIA) with > 50 GHz 3 dB bandwidth, responsivity > 0.7 A/W and a noise of < 17 pA/√Hz. TIA and PD are co-packaged in 2.5D assembly to achieve a higher overall bandwidth. As shown in Figure 15, the ROSA 3 dB bandwidth is at 50 GHz.

It is clear that the 3 dB bandwidth of both TOSA and ROSA are slightly below the Nyquist frequency of 224 Gbit/s PAM4, which requires the aforementioned MLSE with partial response equalization to achieve a recovered signal with good signal fidelity.

The transmission experiment is performed using an AWG with 65 GHz nominal analog 3 dB bandwidth, up to 6 bits ENOB, intrinsic jitter < 60 fs, 1.4 Vpp differential output voltage at 128 GBaud, channel-to-channel skew adjustment with 15 fs resolution, and < 150 dBc wideband phase noise for frequencies > 1 MHz. The resulting bit error rate vs. received optical power (ROP) is shown in Figure 16.

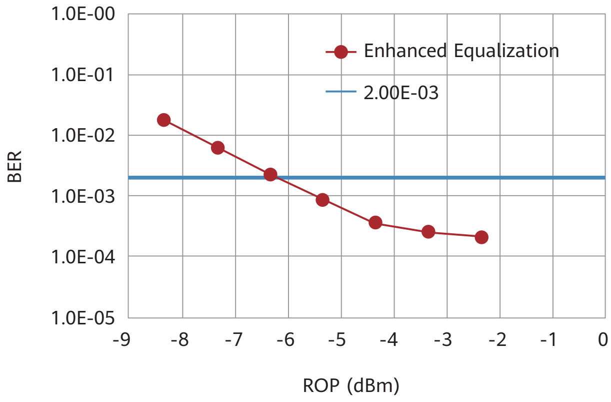


Figure 16 Optical transmission performance of 224 Gbit/s PAM4 using the TOSA and ROSA

Assuming an FEC limit of 2e-3, which is subject to future standardization in IEEE, an ROP of -6 dBm is demonstrated at an output power of 3 dBm. For a CWDM4 4 × 200 Gbit/s transmission with additional multiplexing and demultiplexing, ca. 3 dB additional loss would need to be allocated, leading to a total measured link budget of 6 dB in the experiment. 400 Gbit/s CWDM4 defines 4 dB fiber insertion loss for 2 km, meaning that 2 dB remains as a demonstrated margin for MPI, differential group delay (DGD) and transmitter and dispersion penalty, as well as end-of-life aging. However, a majority of transmitter penalties are already included in this measurement and would not have to be accounted for anymore in this assessment.

8 Conclusion

This paper demonstrated the initial feasibility of 200 Gbit/s PAM4 for up to 2 km, although it is anticipated that the performance of the optoelectronics will have to be further improved to enhance receiver sensitivity and decrease the pre-FEC error floor.

200 Gbit/s per lane optical transmission using PAM4 is analyzed to be the prime candidate for future 800 GbE and 1.6 TbE standards, including 200 Gbit/s copper interconnects, in order to achieve a converged transmission standard similar to 400 GbE.

Latest advances on module design lead to the conclusion that 800 GbE and 1.6 TbE can be supported in pluggables using 100 Gbit/s and 200 Gbit/s per lane SerDes, with power envelopes for QSFP-DD800 or OSFP-XD modules of above 30 W.

References

[1] https://www.ieee802.org/3/B400G/public/index.html

[2] Jinlong Wei *et al.*, "Experimental demonstration of advanced modulation formats for data center networks on 200 Gb/s lane rate IMDD links," Optics Express, Volume 38, Issue 23, 2020.

[3] K. H. Mueller and M. S. Müller, "Timing recovery in digital synchronous data receivers," IEEE Trans. Commun., vol. 24, pp. 516-531, May 1976.

[4] N. Stojanovic *et al.*, "Baud-rate timing phase detector for systems with severe bandwidth limitations," in Optical Fiber Communication Conference (OFC) 2020, OSA Technical Digest (Optical Society of America, 2020), paper M4J.5.

[5] G.D.Forney, Jr., "Maximum-likelihood sequence estimation of digital sequences in the presence of intersymbol interference," IEEE Transactions on Information Theory, Vol. IT-18, No. 3, pp. 363-378, May 1972.

[6] J. Li, E. Tipsuwannakul, T. Eriksson, M. Karlsson, and P.A. Andrekson, "Approaching Nyquist limit in WDM systems by low-complexity receiver-side duobinary shaping," IEEE/OSA Journal of Lightwave Technology, Vol. 30, No. 11, pp. 1664-1676, June 1, 2012.

[7] J. Li, M. Karlsson, and P.A. Andrekson, "1.94Tb/s (11x176Gb/s) DP-16QAM superchannel transmission over 640km EDFA-only SSMF and two 280GHz WSSs," Proc. ECOC 2012, Amsterdam, Netherland, paper Th.2.C.1, Sep. 2012.

[8] M.V. Eyuboglu and S.U.H. Qureshi, "Reduced-state sequence estimation with set partitioning and decision feedback," IEEE Transactions on Communications, vol. 36, no. 1, pp. 13-20, Jan. 1988.

[9] Y. Yu, Y. Che, T. Bo, D. Kim, and H. Kim, "Reduced-state MLSE for an IM/DD system using PAM modulation," Optics Express, Vol. 28, No. 26 / 21 December 2020.

6G ISAC-THz Opens New Possibilities for Wireless Communication Systems

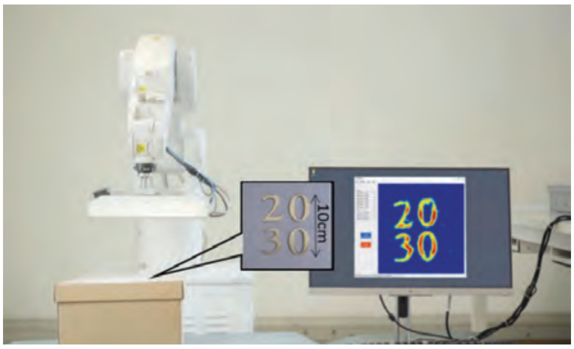
High data rate communications and high-precision sensing are emerging new capabilities for delivering future wireless services. The Huawei 6G Research Team has developed and demonstrated an integrated sensing and communications with a Tera-Hertz (ISAC-THz) prototype. Using wireless electromagnetic waves, the prototype can sense and produces images of blocked objects with millimeter-level resolution and communicates at an ultra-high rate of 240 Gbit/s, opening up new service possibilities for 6G systems.

In his speech at China's first 6G Symposium on September 16, 2021, Huawei Wireless CTO Dr. Wen Tong said, "6G is no longer just a platform that connects everything. It is an intelligent platform that offers both integrated sensing and communications (ISAC) and integrated computing and communications. This platform will provide intelligent services and applications for industries to create greater social value. Bit transmission is not the only function of the 6G network. We will reconstruct and represent the physical world using the propagation properties of radio waves such as reflection, scattering, refraction, and multipath. The 6G network will serve as a sensing network and 6G terminals will serve as sensing terminals. With network sensing and terminal sensing working in tandem, we can model the physical world covered by the entire network based on 6G. This will create two new features — sensing-assisted communications and network-wide crowdfunding for AI big data."

The sensing data extracted from the 6G network will not just be used for modeling the physical world, it will also serve as a big data source and entry for AI learning. Network sensing enables a new usage scenario beyond communications, covering a series of use cases, including device-based or device-free target positioning, imaging, environment reconstruction and monitoring, and gesture and action recognition. Such use cases will be widely applied to industries, including human-machine coordination, environment reconstruction for smart cities, climate sensing, healthcare, and security detection. More application examples are given in the book that Huawei 6G research team published 6G: The Next Horizon.

1 ISAC-THz Prototype Verification

THz lies between the mmWave and infrared frequencies, and is considered an important alternative solution for achieving Tbit/s communication rates thanks to the ultra-large communication bandwidth. Due to its high frequency, THz has a millimeter-level and even sub-millimeter-level wavelength, so it can be applied to relatively small handheld or wearable devices to achieve functions like high-precision positioning, high-res 3D imaging, and mass spectrometry analysis for materials. Unlike optical cameras, THz can penetrate certain obstacles, which achieves high-precision imaging and all-weather sensing in invisible conditions with enhanced privacy protection. As such, THz bands can be used in many daily life and production scenarios, including non-invasive health monitoring, checking food safety, finding defects in high-precision manufacturing, monitoring pollution, and supporting machine vision. As a result, THz is one of the most important techniques for ISAC.



Huawei's ISAC-THz prototype achieves millimeter-level imaging resolution for objects in a closed box

The Huawei 6G Research Team has built a unified platform for integrated sensing and communications in the terahertz (ISAC-THz) band, which is applicable to the 100–300 GHz bands. The team has also verified the technical feasibility and prototype in two challenging scenarios: high-precision sensing and imaging on terminals, and outdoor medium-distance ultra-high-speed transmission.

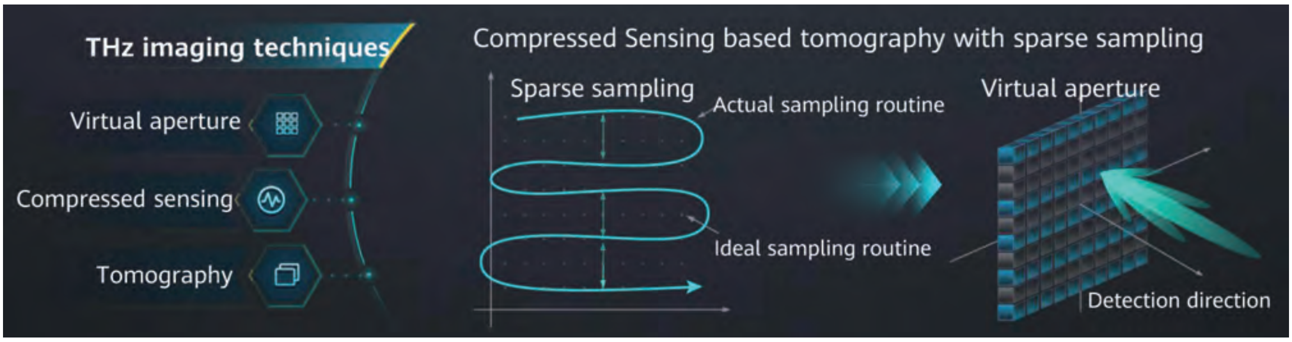
2 Millimeter-Level, High-Precision Sensing and Imaging

In the THz imaging prototype, a robotic arm simulates a person holding a THz terminal to scan and produce an image of an object in a closed box. The prototype uses 140 GHz carrier frequency, 8 GHz bandwidth, and a 4TX16R MIMO array. The THz wave is sent from the antenna array, penetrates the box, and is reflected back to the antenna by the object inside the box. After sampling and processing in real time using an algorithm, the imaging result is generated and displayed.

To achieve millimeter-level imaging resolution, the research team proposes a virtual aperture MIMO array architecture. A limited number of physical antenna elements on the terminal forms a small array. By moving the handheld terminal and scanning the target object, users can form a virtual aperture with greater degrees of freedom, which is equivalent to a real physical aperture formed by thousands of antenna elements without increasing the terminal size. Because the scanning traces are generally sparse and irregular, the prototype uses compressed sensing, tomography, and sparse aperture algorithms to process the sparsely sampled signal waveforms to obtain millimeter-level high-resolution images.

3 Outdoor Medium-Distance Ultrahigh-Speed Transmission

The THz communications prototype is tested in an urban scenario. The transmitter simulates a typical base station installed on the roof of a building and the terminal receiver on the ground level of a city street. The roof-to-ground distance is about 500 meters, and a line-of-sight (LOS) link between the base station and terminal is available. This prototype operates at 220 GHz with a bandwidth of 13.5 GHz. It combines a 2x2 polarized MIMO architecture



Sparse sampling enables high-precision imaging with a virtual aperture

Huawei's ISAC-THz outdoor communications prototype achieves the industry's highest transmission data rate of 240 Gbit/s at a 500-m distance and ultra-wideband, low-bit quantized digital baseband processing technology to perform channel estimation, equalization, non-linear compensation, demodulation, and decoding.

This is the first prototype to achieve 240 Gbps outdoor transmission over medium-to long-distance LOS air interfaces, proving that it is technically feasible to use THz for outdoor ultra-high-speed communications.

4 Outlook

The Huawei 6G Research Team will continue researching and verifying ISAC-THz technology together with channel measurement and modeling in different frequency bands and scenarios, and exploring areas such as miniaturization implementation, 3D stereoscopic imaging, THz mass spectrometry, THz networking, and mobility.

The ISAC concept will not be limited to THz bands. Instead, it will be integrated into the full spectrum to meet different sensing range and precision requirements. We look forward to cooperating with more partners in this field to make "Intelligence of Everything" a reality.

Ultra-Low Power and High-Data Rate Short-Range Wireless Enables Fully Immersive 6G

The Huawei 6G Research Team has developed a prototype featuring ultra-low power consumption, ultra-high throughput, and ultra-low latency for short-range communications using the 70 GHz mmWave spectrum.

With the large-scale commercialization of 5G, the wireless industry, including Huawei, has started researching 6G, the next generation of mobile communications technology. Through extreme connectivity, 6G will provide an all-wireless and immersive experience for human-centric communications and enable the era of "connected intelligence."

Short-range communications typically operates in high frequency bands such as millimeter wave (mmWave) or even Tera-Hertz (THz), and is expected to provide wireless connections as integrated side-links over the "last meters" with very high throughput, very low latency, and very low power consumption. Such extreme performance for short-range communications aims to replace wired connections with free movement and a truly immersive experience. Examples include immersive interactions based on extended reality (XR), holographic communications, and novel interfaces for the metaverse.

The Huawei 6G Research Team has developed a prototype featuring ultra-low power consumption, ultra-high throughput, and ultra-low latency for short-range communications using the mmWave band 70 GHz. This enables devices to communicate at a throughput higher than 10 Gbit/s (Gigabits per second) with sub-millisecond latency. The transmission rate is several times higher than that of wired USBs, and the power consumption of the entire system is less than 560 mW.

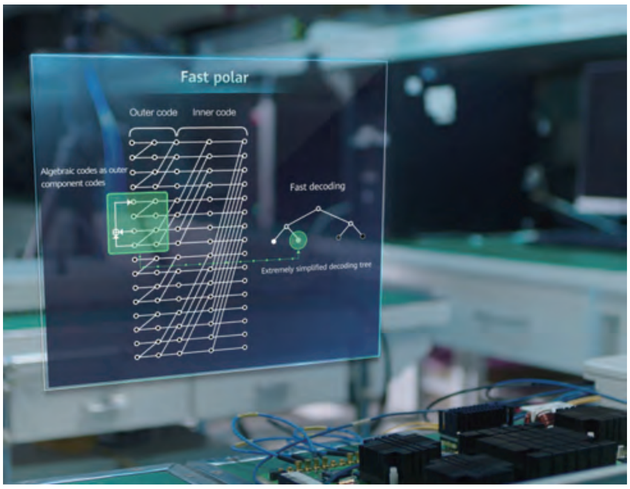
The prototype uses a host of cutting-edge technologies, including:

- **Tbit/s throughput with low-power polar encoding/decoding.** Based on algebraic codes, this technology replaces moderate-rate outer codes to simplify the SC decoding process, boost the decoding throughput, and reduce the chip area by 80%, compared with traditional short-range coding schemes.
- **Low-power 1-bit ADC.** With a limited number of ADC bits, the power consumption of RF chains is significantly reduced. This technology uses the zero-crossing modulation with oversampling at the receiver side to enhance system spectral efficiency.

- **Adaptive beam sweeping with a high-speed short-range phased array.** This technology uses novel dual-polarized phased arrays to independently transmit dual-stream data at a high speed. When combined with the adaptive beam sweeping mechanism powered by AI-based prediction algorithms, this technology accurately adjusts beam directions with an ultra-low beam scanning overhead, even in scenarios with mobility.

- **High-efficiency SiGe large-scale antenna array with antenna in package (AIP).** The large-scale, high-gain irregular array architecture can be integrated in a LTCC (Low Temperature Co-Fired Ceramic) packaging module with a real smartphone form factor, so that the miniature AIP can also be implemented in wearables.

Huawei is committed to providing cost-effective extreme connectivity for customers. The Huawei 6G Research Team is dedicated to enhancing the communication experience of next-generation communications systems and providing immersive human-centric services for the "connected intelligence" era. To achieve that, extreme-performance short-range communications is an important enabling technology.


Fast polar
Outer code
Inner code
Algebraic codes as outer component codes
Fast decoding
Extremely simplified decoding tree

6G ISAC-OW Extends the Frontier of Spectrum for Wireless Communication Systems

To meet the high communication rate and high-precision sensing requirements in EMF-free scenarios such as healthcare and industry automation, the Huawei 6G Research Team has proposed integrated sensing and communications with optical wireless (ISAC-OW).

As a next-generation wireless communications technology, 6G will integrate sensing and communications functions. It will continue to use multiple frequency bands, spanning low-band, mid-band, and high-band spectrum. Among the high frequency bands and in addition to millimeter wave (mmWave), 6G will extend its reach to terahertz (THz) and even optical spectrum in certain home and industrial scenarios.

ISAC-OW technology can be naturally integrated into existing lighting and light display systems, making every lamp and every screen part of the 6G ISAC-OW system. Given the ultra-high communication bandwidth, ISAC-OW is also a potential candidate to achieve ultra-high throughput. Because of the gap between optical spectrum and conventional electromagnetic spectrum, there will be no electromagnetic interference in conventional radio frequency bands. ISAC-OW is especially suitable for electromagnetic radiation-sensitive environments, like smart healthcare and industrial manufacturing.

The sub-millimeter wavelength of the optical spectrum can achieve high-precision positioning and high-resolution imaging, which, when combined with the response of substances to the characteristics of light waves, will enable more precise and accurate health sensing and monitoring.

The Huawei 6G Research Team made a key technology breakthrough with the research and prototype verification of the ISAC-OW system. The team's first ISAC-OW prototype implements integrated communication, positioning, and sensing in both architecture and capability. The prototype simulates a medical environment where robots are accurately sensed and localized through optical wireless links (such as visible light and infrared spectrum) and can be remotely commanded to pick up and carry objects. The optical links in the prototype also wirelessly transmit the real-time videos between robots and the controller at a high speed, achieving integrated sensing and communications. By detecting subtle facial color changes or abdominal fluctuations, the ISAC-OW prototype is able to contactlessly monitor a person's heartbeat and breathing status in real time, with an accuracy equivalent to commercial smart watches.

Prototype

During the prototype development, the team made breakthroughs in key technologies such as distributed optical antenna for joint detection and transmission, integrated sensing and communication architecture, and high-precision Time of Flight (ToF) modeling analysis. For example:

- For simultaneous communications and localization, unified waveforms, hardware architecture, and signal processing algorithms jointly enable high-speed video transmission and centimeter-level precision of indoor localization. During localization, devices use enhanced reflecting surfaces to reflect optical signals from base stations without generating interference for other base stations. As synchronization is not required, base stations can measure phase differences to provide high-precision localization.
- For contactless health monitoring and considering the impact of the heartbeat on blood vessels in the face, the team has combined ToF modeling analysis and deep learning technologies to accurately detect subtle changes in the light intensity reflected by faces, so as to measure the heartbeat and breathing frequency. The latter can also be measured by abdominal fluctuations during breathing. Results show that this approach is comparable to commercial smart watches.

Outlook

Capitalizing upon the advantages of the optical spectrum, which include high bandwidth, line-of-sight transmission, no RF interference, and high energy efficiency, the ISAC-OW system creates a larger imagination space and a solid technical foundation for 6G applications.

The Huawei 6G Research Team will continue researching ISAC-OW technology, with the aim of providing the most advanced core technologies and solutions for future 6G scenarios such as holographic hospitals and industrial automation.

HUAWEI TECHNOLOGIES CO., LTD.
Huawei Industrial Base
Bantian, Longgang
Shenzhen 518129, PRC
Tel: +86-755-28780808

General Disclaimer

The information in this journal may contain predictive statement including, without limitation, statements regarding the future financial and operating results, future product portfolios, and new technologies. There are a number of factors that could cause actual results and developments to differ materially from those expressed or implied in the predictive statements. Therefore, such information is provided for reference purpose only and constitutes neither an offer nor an acceptance. Huawei may change the information at any time without notice.