IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 26 November 2023, accepted 12 December 2023, date of publication 15 December 2023,
date of current version 21 December 2023.

*Digital Object Identifier 10.1109/ACCESS.2023.3343523*

RESEARCH ARTICLE

# Beamforming-as-a-Service for Multicast and Broadcast Services in 5G Systems and Beyond

**NGUYEN HUU TRUNG AND NGUYEN THUY ANH**
School of Electrical and Electronic Engineering, Hanoi University of Science and Technology, Hanoi 100000, Vietnam
Corresponding author: Nguyen Huu Trung (trung.nguyenhuu@hust.edu.vn)

**ABSTRACT** Driven by the most advanced technologies, such as massive multiple-input multiple-output (MIMO), 3D beamforming, Software-Defined Networking, and network slicing, 5G/6G systems support radio connections and end-to-end network connectivity at ultrahigh speeds, low latency, high reliability, and massive connectivity. 5G/6G networks promise a faster, more streamlined future for broadcasters. It can handle rapid changes, along with improved quality standards in broadcasting. However, broadcasting services are becoming increasingly distributed and prevalent; thus, how to effectively provide broadcast services to end users has recently become a natural concern. In this paper, we propose a novel concept of Beamforming-as-a-Service (BFaaS) for delivering multicast and broadcast services in 5G/6G networks. We first comprehensively overview and study background standards and industrial activities through broadcast projects that have been carried over 5G platforms. Then, we shed light on the requirements of providing multicast and broadcast services to end users and the importance of beamforming in delivering multicast and broadcast services in 5G/6G networks. From this point, we define the BFaaS concept and vision and the benefits of the proposed BFaaS scheme. Next, we discuss the beam generation process, beamforming control and the channel model which is used in the proposed system model analysis. As a step further, with the proposed system model, we propose optimal algorithms for allocating beams to different service areas. The simulation results demonstrated the effectiveness of the proposed BFaaS scheme. Finally, we summarize the paper by identifying the potential areas of application of BFaaS for future research directions.

**INDEX TERMS** 5G networks, terrestrial broadcasting, broadband, multicast and broadcast services, new radio, as-a-service, beamforming.

## I. INTRODUCTION

Terrestrial Television (TV) broadcasting is becoming increasingly important along with other types of media. The coronavirus disease (COVID-19) pandemic has reinforced the importance of broadcasting and hybrid systems. TV broadcasting is currently undergoing drastic changes along with improved quality standards, such as High-Definition Television (HDTV), Ultra-High-Definition Television (UHDTV), high dynamic range television, high frame rate video, and Next-Generation Audio (NGA). On-demand services have changed the television and online video industries [1]. All-IP is the standard method for delivering television and video content. Artificial intelligence and big-data analytics have become key elements in intelligent content discovery. With the development of the future internet and next-generation mobile communication networks, traditional linear television services that are transmitted over IP and mobile communication networks and nonlinear services such as video-on-demand will coexist [2].

The Fifth Generation (5G) and Sixth Generation (6G) wireless communications systems provide a variety of services such as enhanced Mobile Broadband (eMBB), massive Machine-Type Communication (mMTC), Ultra-Reliable and Low-Latency Communication (URLLC) [3], and private networks that allow support of both a non-public network specific authentication mechanism for User Equipments (UEs) without a Universal Subscriber Identity Module (USIM) and an authentication and key agreement mechanism for UEs with a USIM [4].

The associate editor coordinating the review of this manuscript and approving it for publication was Yogendra Kumar Prajapati.

In previous literature, several works have been done on Multimedia Broadcast Multicast Services (MBMS), adapting existing broadcast systems such as Advanced Television Systems Committee (ATSC) 3.0, Digital Video Broadcasting - Second Generation Terrestrial (DVB-T2), and DVB - Next Generation Handheld (DVB-NGH) over 5G New Radio (NR) [5]. Gimenez et al. [6] presented a physical layer design for NR-MBMS, a system derived from 5G-NR specifications, with minor modifications and suitable for the transmission of linear TV and radio services in either single-cell or Single Frequency Network (SFN) operations. This design is based on a cyclic prefix Orthogonal Frequency Division Multiplexing (OFDM) solution, such as Long-Term Evolution (LTE), with a scalable numerology that enables radio resource allocation over different frequency bands. In this study, the NR-MBMS proposition was evaluated and compared to LTE-based Further evolved Multimedia Broadcast Multicast Service (FeMBMS) in terms of flexibility, performance, capacity, and coverage. The 5G NR using low-density parity-check and polar codes may have up to 7.2% higher bandwidth utilization than LTE-based FeMBMS.

The non-3GPP ATSC 3.0 broadcast Radio Access Technology (RAT) is aligned with 3GPP 5G NR unicast RAT in [7]. An innovative broadcast 5G convergence architecture was introduced, 5G NR frames and ATSC 3.0 frames were optimized, and ATSC 3.0 frames were time-aligned to 5G NR unicast frames and received using a dual simultaneously connected UE. In [8], a 5G NR mixed mode is presented for enabling the use of multicast/broadcast, which enables flexible, dynamic, and seamless switching between unicast and multicast or broadcast transmissions, and the multiplexing of traffic under the same radio structures.

An enhanced Next Generation - Radio Access Network (NG-RAN) architecture was introduced in [9] to support efficient, flexible, and dynamic selection between unicast and multicast/broadcast transmission modes and the delivery of terrestrial broadcast services. The NG-RAN is a cloud-RAN based on new concepts, such as the RAN broadcast/multicast areas, that allow a more flexible deployment in comparison to evolved Multimedia Broadcast and Multicast Services (eMBMS), using an enhanced NG-RAN architecture based on 3GPP Rel-15, which primarily focuses on broadcast/multicast capabilities to address requirements from multiple verticals.

Fallgren et al. [10] proposed an adaptive and robust beam-management algorithm at the air interface to improve the end-to-end architectural design of 5G networks, thereby enabling efficient broadcast and multicast transmissions for vehicle-to-anything services. Power-based Non-Orthogonal Multiplexing (P-NOM) technology has been proposed [11] in addition to the existing orthogonal time-division multiplexing scheme. By using P-NOM in a 5G-MBMS system, significant capacity gains can be achieved for delivering different types of broadcast services and delivering mixed unicast and broadcast services. A fast simplified multi-bit successive cancellation list decoding method for polar codes with FPGA implementation details for future 5G millimeter wave TV broadcasting services is introduced in [12].

As one of the emerging strategic information technologies, cloud computing is widely applied to deliver ''as-a-Service'' over 5G networks, including Software-as-a-Service (SaaS), Platform-as-a-Service (PaaS), Infrastructure-as-a-Service (IaaS), and Hardware-as-a-Service (HaaS). The well-known 5G SaaS [13], in which software applications are rented from a provider as opposed to purchasing them for enterprise installation and deployment. This allows the service providers to avoid the need to manage the hardware, infrastructure, operating system... All of which are managed and controlled by the network operator, to ensure that the application is always ready and working correctly.

The on-demand software development platform PaaS provides developers with a complete platform including an application, interface, database, and data storage. It also supports software development in the lifecycle cycle, such as code development, testing, application deployment, and services on cloud computing platforms [14].

5G IaaS is an on-demand computing infrastructure that provides remote access to physical resources such as servers, network devices, and storage drives. IaaS supports hardware, such as backhaul, backbone, and radio resource control units. In the IaaS framework, virtualization is a common method used to create on-demand delivery of network resources. IaaS is based on virtualization technology, which reduces operating costs owing to the efficient use of resources. Service providers do not need to maintain network hardware or systems. However, service providers must calculate the capacity, bandwidth, processing power, and storage [15].

More specifically, in [16], the concept of ANYthing-as-a-Service (ANYaaS) was proposed, wherein on-demand creation and orchestration of 5G services are specified and enabled by exploiting the benefits of both cloud computing and network function virtualization. The ANYaaS concept can create a single service instance operating individually or multiple correlated service instances, thereby ensuring efficient integration between them. This approach is based on the use of a Content Delivery Network (CDN).

Chang et al. [17] presented the concept of everything-as-a-service. From a wireless network virtualization perspective, instead of virtualizing computing resources in server virtualization, in wireless network virtualization, physical resources must be abstracted to isolate virtual resources from infrastructure service providers. Virtual resources can then be offered to different Network Service Providers (NSPs).

RAN-as-a-Service (RANaaS), which is based on generic datacenters, is presented in [18] to centrally execute part of the RAN functionalities, thus benefitting from centralization gains, which are fundamental in ultra-dense deployments. The approach is to achieve energy efficiency based on backhauling network power consumption with respect to the small-cell RF output power. RANaaS creates RAN sharing mechanisms to improve efficient usage of network resources. RAN resources are virtualized as a process in which physical wireless resources can be abstracted into

virtual resources, holding a subset of functionalities of the underlying physical counterpart, and shared by ensuring complete isolation from each other. Thus, RAN virtualization is the process of abstracting all elements of a RAN and slicing them into virtual elements that hold certain corresponding functionalities and are isolated from each other [19].

5G services such as eMBB, URLLC and mMTC are enabled through a key technology called Network Slicing (NS) [20] and are managed by End-to-End (E2E) service orchestration via Network Slice-as-a-Service (NSaaS) [21]. The NS creates virtual network segments for different services within the same 5G network. NS divides a physical network into independent logical subnets for different service types, each of which has a size and structure suitable for dedicated services [22]. Network-slicing solutions are based on 5G network functions across the RAN, core network, transport network, and orchestrator. The logical network concept of network slicing is based on a Software-Defined Network (SDN). Core network functions can be sliced to provide specific services to different users.

### A. MOTIVATIONS AND CONTRIBUTIONS

The motivation behind this work stems from the observation that on the convergence of broadcast and 5G broadband, the proposed solutions focus only on the improvement of RAT technology, allowing flexibility in the delivery of broadcast services via a High-Power High-Tower (HPHT) topology. The HPHT topology does not take advantage of MIMO transmissions [23]. Furthermore, in this architecture, each gNB can provide only a fixed service program, even without subscribers.

Today, we can already observe a variety of broadcasting services. The coverage areas for each service are distributed across many locations. However, in many situations, the subscribers are concentrated in one location. For example, a service provider must provide services to subscribers in a building, stadium, toll gates, and so on. Thus, to be more effective, different service companies provide different types of broadcasting services based on a shared network infrastructure with coverage areas corresponding to different groups of subscribers. Particularly for Internet of Things (IoT) services that are highly customizable, the quality of service also varies by service group.

Advanced antenna system deployment in 5G/6G networks enables state-of-the-art beamforming and spatial multiplexing techniques, which are powerful tools for capacity and coverage enhancement, and steering energy to a predetermined area adaptively and optimally [24]. This opens broadcast possibilities for providing diverse services to different customer groups and defines a novel area-based business model.

In this study, we seek to provide multicast and broadcast services to each specific service area while optimizing infrastructure resources through a set of quality assessment Key Performance Indicators (KPIs). From there, we propose beamforming-as-a-service (BFaaS) for Multicast and Broadcast Services in 5G systems and beyond, under the concept of service-oriented architecture. The idea is that network operators provide beams as services to NSPs. NSPs rent beams for different coverage areas that meet the requirements of KPIs regardless of the specific hardware architecture of the network infrastructure.

With the aim of providing Beamforming-as-a-Service for multicast and broadcast services in 5G systems, the key contributions of this study can be summarized as follows:

1) This paper introduces the background of standards of digital television terrestrial broadcasting and broadcast over 5G networks.

2) The study also provides a survey of broadcast projects that have been carried out over 5G platforms, and the projects are compared in terms of technological advantages and disadvantages.

3) We define the novel concept and system model of BFaaS for delivering multicast and broadcast services in 5G/6G networks. Our solution provides a means to effectively increase spectral efficiency.

4) Aiming to allocate beams to different service areas, we first propose one algorithm to gather the information from UEs for UE classification and beam assignment, and then the next three algorithms aim to determine the minimum number of beams to cover all coverage area. These algorithms are based on the greedy and dynamic programming algorithm. The fifth algorithm is used to maximize coverage area for a given number of beams. This algorithm is based on K-means clustering. The simulation results demonstrated the effectiveness of the proposed Beamforming-as-a-Service scheme.

5) Finally, because none of the proposals are ready for 5G NR systems, we propose using BFaaS for 5G NR technology.

### B. PAPER STRUCTURE AND ORGANIZATION

Table 1 lists the abbreviations used in this paper. The structure of the paper is illustrated in Fig. 1. The remainder of this paper is organized as follows. In Section II, we commence by providing an overview of the background standards and industrial activities through broadcast projects that have been carried over 5G platforms. We pay special attention to the implementation and use cases of the applied projects related to MBMS services and how to provide multicast and broadcast services to the end users.

Also, in this section we look at the pros and cons of the topologies used in standards and projects. As a step further, in Section III, we investigate the beamforming in delivering different multicast and broadcast services in 5G/6G systems.

For clarity, we pose research questions and define the novel BFaaS concept and vision and the benefits of the proposed BFaaS scheme. We present the beam generation process, beamforming control and the channel model in the proposed system model. We propose optimal algorithms for allocating beams to different service areas. The simulation results of the proposed optimal algorithms for allocating beams to different service areas are presented in Section IV.

**TABLE 1.** List of abbreviations.

| Abb. | Definition | Abb. | Definition | Abb. | Definition |
|---|---|---|---|---|---|
| 5G | Fifth Generation | FeMBMS | Further evolved Multimedia Broadcast and Multicast Services | NSaaS | Network Slice-as-a-Service |
| 6G | Sixth Generation | gNB | Next Generation NodeB | NSP | Network Service Provider |
| ANYaaS | ANYthing-as-a-Service | HaaS | Hardware-as-a-Service | PaaS | Platform-as-a-Service |
| AR | Augmented Reality | HDTV | High-Definition Television | PDP | Power Delay Profile |
| ATSC | Advanced Television Systems Committee | HPHT | High-Power High-Tower | PL | Path Loss |
| BF | Beamforming | IaaS | Infrastructure-as-a-Service | PMI | Precoding Matrix Indicator |
| BFaaS | Beamforming-as-a-Service | IoT | Internet of Things | P-NOM | Power-based Non-Orthogonal Multiplexing |
| BM-SC | Broadcast Multicast - Service Centre | KPI | Key Performance Indicator | QoE | Quality of Experience |
| BSA | Beam Service Area | KQI | Key Quality Indicator | QoS | Quality of Service |
| CDN | Content Delivery Network | LoS | Line of Sight | RAN | Radio Access Network |
| CIR | Complex Impulse Response | LTE | Long-Term Evolution | RANaaS | RAN-as-a-Service |
| CMAF | Common Media Application Format | MBMS | Multimedia Broadcast Multicast Services | RAT | Radio Access Technology |
| CPIX | Content Protection Interchange Format | MBMS-GW | MBMS - Gateway | RI | Rank Indicator |
| CSI-RS | Channel State Information - Reference Signal | MCE | Multi-Cell/Multicast Coordination Entity | RMS | Root Mean Square |
| DASH | Dynamic Adaptive Streaming over HTTP | MEC | Multiaccess Edge Computing | RSSI | Received Signal Strength Indicator |
| DRM | Digital Rights Management | MG-MIMO | Multi-Group MIMO | SaaS | Software-as-a-Service |
| DS | Delay Spread | MME | Mobility Management Entity | SDN | Software Defined Network |
| DTTB | Digital Television Terrestrial Broadcasting | mMTC | massive Machine Type Communications | SDoF | Spatial Degrees-of-Freedom |
| DVB-NGH | DVB - Next Generation Handheld | MooD | MBMS operation on-Demand | SE | Spectral Efficiency |
| DVB-T | Digital Video Broadcast - Terrestrial | NGA | Next Generation Audio | SlaaS | Slice-as-a-Service |
| E2E | End-to-End | NG-RAN | Next Generation - Radio Access Network | TSA | Total Service Area |
| eMBB | Enhanced Mobile Broadband | NLoS | Non-Line of Sight | TV | Television |
| eMBMS | evolved Multimedia Broadcast and Multicast Services | NR | New Radio | UHDTV | Ultra-High-Definition Television |
| EPC | Evolved Packet Core | NS | Network Slicing | USIM | Universal Subscriber Identity Module |

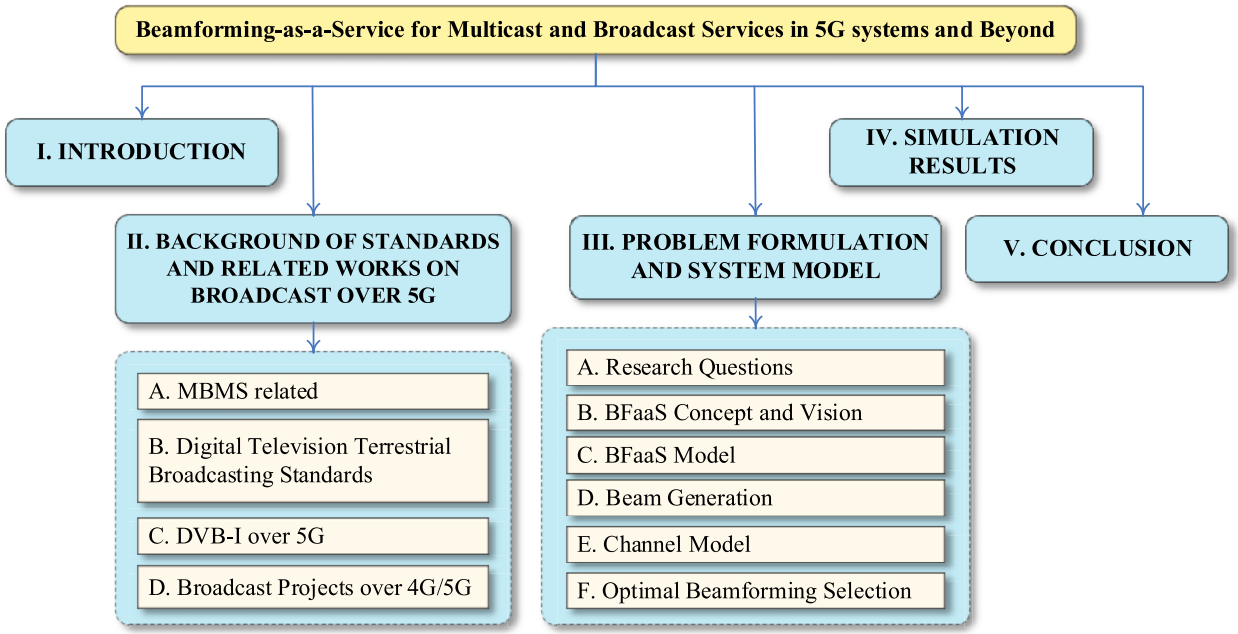


**FIGURE 1.** Structure of the paper.

Finally, we summarize the paper by identifying the potential areas of application of BFaaS for future research directions in Section V.

## II. BACKGROUND OF STANDARDS AND RELATED WORKS ON BROADCAST OVER 5G

In this section, we provide an overview on the background standards related to MBMS and Digital Television Terrestrial Broadcasting (DTTB). Then, we investigate the industrial activities through broadcast projects that have been carried over 5G platforms.

### A. MBMS RELATED

The MBMS is a point-to-multi-point specification for existing GSM/UMTS and CDMA2000 mobile networks. The MBMS standard first described in UMTS version 6 (Rel-6) introduces only minor changes to existing core and radio access network protocols. Therefore, MBMS can be implemented by upgrading software to an appropriate hardware platform. This reduces the cost of the terminal and core networks. MBMS allows broadcast on mobile technology platforms and is a relatively cheap technology compared to other broadcasting technologies, such as DVB-T2, which requires new receiver hardware and significant investment in network infrastructure [25].

The eMBMS, also known as the LTE Broadcast, of 3GPP was first described in version 9 and completed in release 13. eMBMS enables multiple mobile users to view the same content but uses only a fixed amount of network resources. With eMBMS, up to 60% of the network capacity can be allocated to broadcast services. Service providers have full control over the content to be broadcast within the broadcast bandwidth. eMBMS supports multicast and broadcast services. eMBMS was upgraded to FeMBMS in Release 14 [26].

FeMBMS is a further development of the LTE broadcast mode eMBMS in 3GPP Release 14 which was released in June 2017. FeMBMS is considered an added service for broadcasting from HPHT base stations, wider bandwidth, and no SIM card required. Instead of 60%, FeMBMS enables 100% of transmission capacity to be used for broadcasting services. This is a significant improvement over eMBMS, leading to growing interest among broadcasters regarding the potential of FeMBMS. The FeMBMS modulation mode is based on OFDM, similar to the DVB-T2 and ATSC 3.0. FeMBMS supports larger sites for single-frequency networks and allocates 100% of the resources to broadcast with independent signals for downlink transmission. The FeMBMS receive-only mode allows unregistered user devices to receive free-to-air signals without the need for uplink transmissions or SIM cards. The network architecture of FeMBMS in LTE/EPC networks with a HPHT is shown in Fig. 2. The FeMBMS architecture adds three new network elements to the existing LTE core network which includes Multi-Cell/Multicast Coordination Entity (MCE), MBMS Gateway (MBMS-GW), and Broadcast-Multicast Service Centre (BM-SC). The MCE manages the radio resources for the MBMS for all the radios that are part of the MBSFN service area. It coordinates the transmission of synchronized signals from different eNodeBs by using the M2 interface for the control plane. The MBMS-GW receives the broadcast data and forwards it to the relevant eNodeB in the network. The BM-SC is the core multicast/broadcast functionality that receives content through the xMB interface in unicast, converts it into multicast data, and sends it to the MBMS-GW. The Operation Support System/Business Support System (OSS/BSS) is used to support E2E telecommunication services, inventory, and service lifecycle. The MBMS operation on-Demand (MooD) is a multicast operation on demand. MooD enables automatic bandwidth partitioning between the unicast and broadcast modes. Because MBMS serves multiple mobile devices, there is no feedback such as hybrid automatic repeat request, and MIMO is not supported. Owing to the lack of MIMO, common-control pilot or reference signals are not delivered. Logically, however, the MBMS pilot signal or MBSFN reference signal is delivered. The PDN Gateway (P-GW) routes packets to and from external CDN. The Serving Gateway (S-GW) handles the user data traffic. The emulated Mobility Management Entity (MME) handles the control plane setup of the eMBMS session. The purpose is to bypass the MME when the operator's MME does not support the eMBMS [27].

### B. DIGITAL TELEVISION TERRESTRIAL BROADCASTING STANDARDS

Standards for DTTB have evolved worldwide, with different systems adopted in different regions. Currently, there are three leading DTTB systems: ATSC 3.0 system, DVB-T2 system, and the integrated service digital broadcasting terrestrial system. The ATSC 3.0 system has largely been adopted in North America, South America, Taiwan, and South Korea. This system adapts trellis coding and 8-level vestigial sideband modulation. The DVB-T2 system has now been widely adopted in Europe, the Middle East, Australia, and parts of Africa and Asia. The DVB-T2 system adapts coded OFDM.

The DVB-T2 system is an advanced DTTB system with high spectrum efficiency, robustness, and flexible services developed based on the DVB-T system. Although DVB-T2 can provide sufficient transmission speed for HDTV, the data rate remains insufficient for UHDTV services [28].

The mobile evolution of DVB-T2 is DVB-NGH, which was developed through a digital video broadcast terrestrial project. DVB-NGH deployment is motivated by the continuous growth of mobile multimedia services for handheld devices such as tablets and smartphones [6]. The main objective of DVB-NGH is to increase the network capacity and coverage area, outperforming existing mobile broadcasting standards, such as DVB-H (handheld) and DVB-SH (satellite services to handheld devices).

DVB-NGH uses a physical layer pipe, scalable video coding, time-frequency slicing, robust header compression, and MIMO. MIMO spatial multiplexing is specified in DVB-NGH as MIMO rate 2 codes, where the term "rate 2"

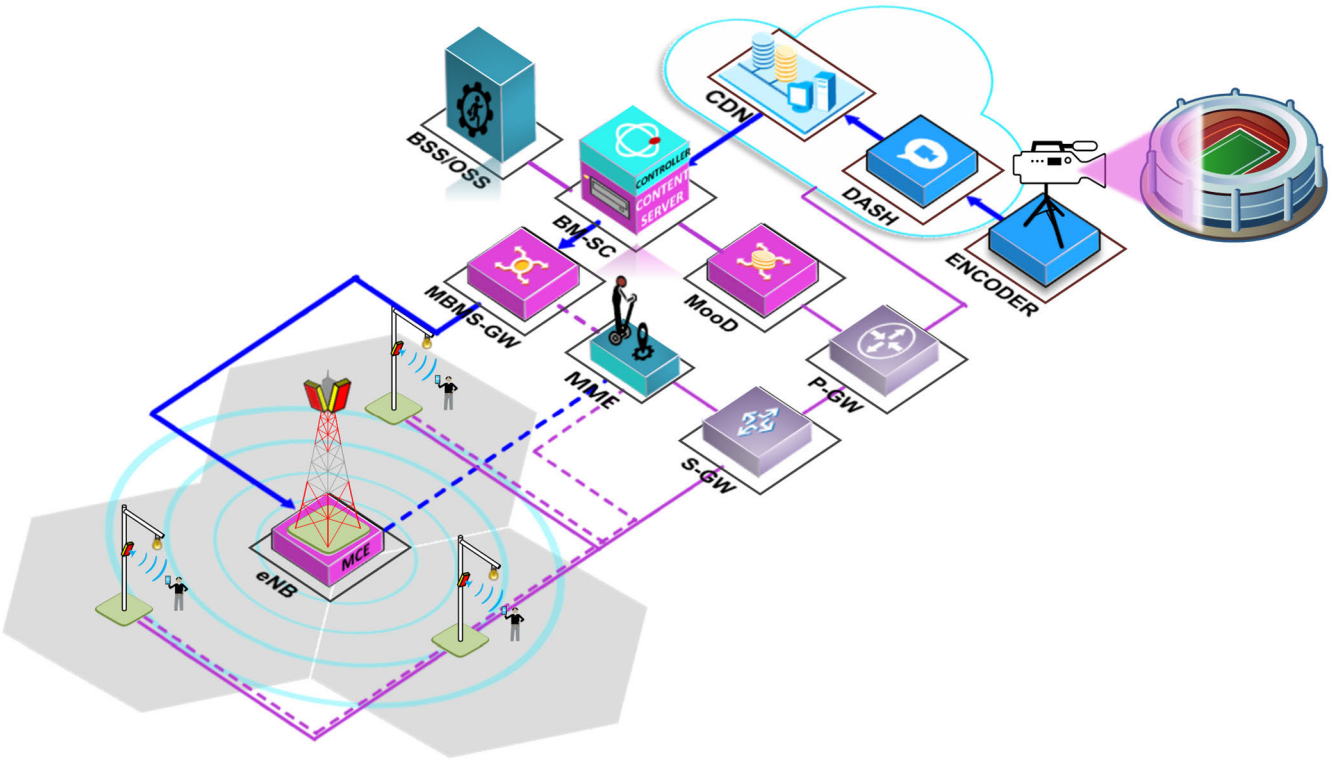


**FIGURE 2.** FeMBMS architecture in LTE/EPC networks with HPHT topology.

represents the transmission of two independent streams. MIMO rate 2 codes in DVB-NGH use a cross-polar antenna arrangement (antennas with orthogonal polarization) with two transmit and two receive antenna [29].

### C. DVB-I OVER 5G

The "I" in DVB-I stands for Internet. DVB-I provides an Internet-centric mechanism for delivering content services. DVB-I can be used in combination with DVB-T, C, and S. With DVB-I, we can watch TV on smartphones or tablets with internet access. We can select a program from a list of services and content, including DVB-I and broadcast services and we do not have to know or care whether a service arrives via broadcast or IP because DVB-I can offer stand alone or integrated with broadcast services. With these hybrid services, basic broadcast distribution is augmented with unicast for extended service coverage, lower distribution costs, improved quality, and additional user experience.

DVB-I concurrently delivers both unicast and multicast services to users and dynamically switches between these services. In DVB-I, the MBMS-SFN is broadcast using physical multicast channels. The physical multicast channel can be transmitted in QPSK, 16 QAM, or 64 QAM. No transmit diversity (i.e., MIMO) scheme is specified. Layer mapping and precoding shall be performed assuming a single antenna port, and the transmission uses antenna port 4 [30].

As mentioned before, we are interested in the transmission of multicast/broadcast services over 5G networks. Fig. 3 shows the DVB-I architecture for 5G systems to deliver multicast services.

Linear, on-demand TV services, ads, or objects are fed into the encoder, and the output is a Common Media Application Format (CMAF). CMAF is an emerging standard intended to simplify the delivery of HTTP-based streaming media. CMAF uses a common media format for video streams and reduces the costs, complexity, and latency.

The Content Protection Interchange Format (CPIX) is used to exchange key information between Digital Rights Management (DRM) and Dynamic Adaptive Streaming over HTTP (DASH) packagers. CPIX defines an XML schema for carrying content keys and encrypting the information. The DASH packager then puts each group of frames into a CMAF chunk and pushes it to the origin server of the CDN.

The transcaster server pulls source-adaptive bit rate streams and embeds them into multicast streams, which optimizes the network performance for streaming services while ensuring QoE to customers.

### D. BROADCAST PROJECTS OVER 4G/5G

This section presents the advantages and disadvantages of eMBMS/FeMBMS and 5G mobile TV pilot projects that have recently been conducted.

#### 1) 5G TODAY

The project 5G TODAY which is funded by the Bavarian Research Foundation, has been running since 2017. The project partners are IRT, Kathrein, Rohde & Schwarz, and

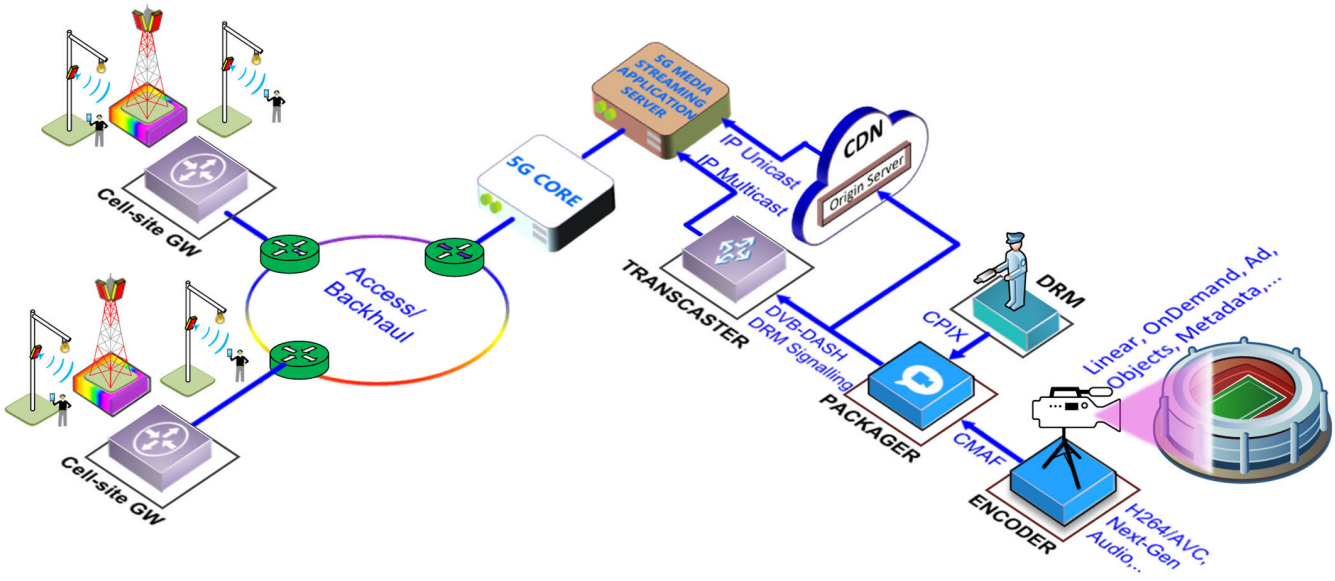


**FIGURE 3.** DVB-I over 5G E2E architecture with unicast/multicast dynamic switching.

associated partners. Bayerischer Rundfunk and Telefónica Deutschland jointly investigated the possibilities of a 5G-based broadcast solution. The main objective of the project is large-scale TV broadcasting in the higher Bavarian region in FeMBMS mode over 5G broadcast networks. The project investigated the implementation of FeMBMS transmitters and receivers and provided novel insights into network parameters, antenna design, and propagation models. The test field showed that 5G broadcasts offer fundamental advantages such as high video quality, low latency, and cost-effective distribution with a high coverage of up to 60 km [31].

### 2) 5G RURALFIRST

In the UK, the government funded the 5G RuralFirst project, which provides trials to harness the benefits of 5G networks in serving rural communities in agriculture, broadcasting, and public services. The 5G RuralFirst project aims to research and develop the following: 5G cloud core network, dynamic spectrum sharing, 5G RAT, agri-tech, broadcast, industrial IoT and community and infrastructure.

The 5G RuralFirst project is a new paradigm for spectrum sharing and deployment of disaggregated 5G lower-cost radios and RAN (at Orkney). Applications of the 5G RuralFirst project include radio broadcasts over 5G (Orkney), renewable energy/IoT/security (Orkney), Li-Fi (Light Fidelity) in rural settings in 5G (Orkney), salmon farming/safety/IoT monitoring (Orkney), agri-tech such as automomous farm vehicles, crop/soil condition mapping, and Augmented Reality (AR) veterinary information (Shropshire & Somerset). The following applications and E2E use cases were tested: IoT connectivity from fish farms to wind farms, connecting sensors, tourism with mobile connectivity, radio broadcast BBC, connecting drones for high-definition video for vehicles, and IoT monitoring of livestock. Business models are also investigated in this study.

The project used a mix of radio technologies, including pre-5G, such as 4G, 5G NR, SDR/custom-built, Li-Fi, LoRa, and citizens broadband radio service. The disaggregation of hardware and software plays an important role in designing a unified architecture that can operate with 2G, 3G, 4G, and 5G systems (all-G support, any-G solution) [32].

### 3) 5G-XCAST

The main objective of this project is to develop broadcast and multicast point to multi-point capabilities for 5G systems and dynamically adaptable 5G network architecture with layer-independent network interfaces to dynamically and seamlessly switch between unicast, multicast, and broadcast modes [33].

5G-Xcast provides efficient, scalable, and sustainable solutions for a large-scale distribution of media services fully consistent with the core 5G specifications, 4K UHDTV, and, in the future, 8K UHDTV and emerging new interactive services (e.g., augmented reality, virtual reality, and 360° visual media). Three main use cases are supported: hybrid broadcast services, object-based broadcast services, and public warning messages. Four different vertical markets are supported: media and entertainment (UHDTV, virtual reality, AR, $360^0$ video, content propositioning, push2talk), public warning (disasters, emergency alerts), automotive (autonomous driving, vehicle-to-everything broadcast service, infotainment, safety, signage information), and the IoT (massive software and firmware updates).

In 5G-Xcast, if there are two subscribers running on one stream (for example, a sports event) using unicast (two media flows), the system detects that it is a popular event, which may activate a broadcast session and steer that traffic onto the

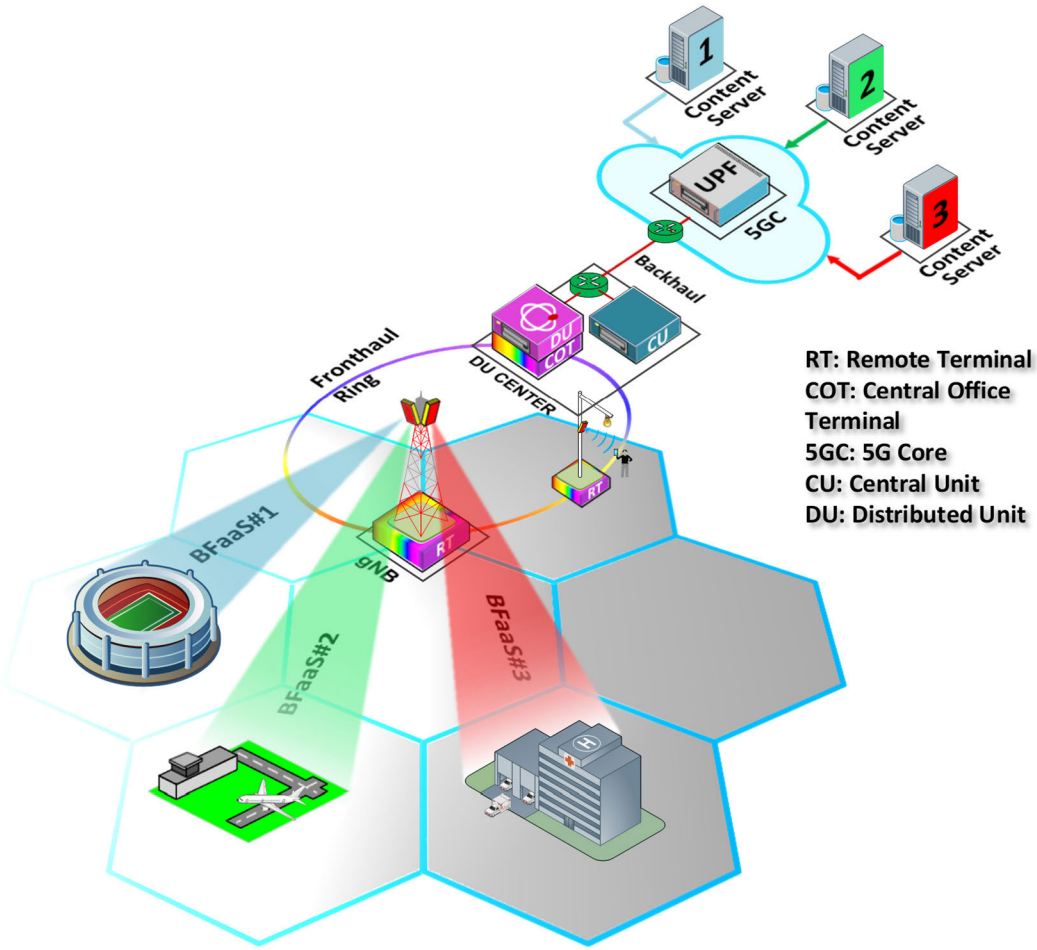


**FIGURE 4.** The proposed BFaaS concept and implementation in 5G NR for broadcast services.

broadcast. Thus, hundreds of unicasts can be replaced with a single broadcast stream. The unicast traffic is expected to decrease.

In addition, 5G-Xcast also has an object-based broadcasting mode, for example, an object-based weather forecast. This mode is implemented by dividing the content into objects, such as body language (sign language, assets, and background) rendered on the end device, depending on the situation and needs of the user to combine this information, and objects are delivered over either multicast or unicast. Objects for all devices are delivered over multicast, whereas personalized objects are delivered over unicast. This feature was implemented in the MEC.

#### 4) "TOWER OVERLAY" STAND-ALONE 4G/LTE BROADCAST NETWORK IN AOSTA VALLEY

RAI Research with the EBU and Technische Universitaet Braunschweig implemented a broadcast network using 4G/LTE towers in a SFN configuration. The test demonstrated the deployment of 4G/LTE technology in a conventional terrestrial broadcast network infrastructure for the distribution of public service media content and services [34].

#### 5) TRIAL OF LTE BROADCAST (LTE-B) IN RURAL NORWAY

Distribution of linear TV and big-screen TVs via 4G or 5G via LTE Broadcast (LTE-B) technology.

A summary of projects and their characteristics is presented in Table 2.

## III. PROBLEM FORMULATION AND SYSTEM MODEL

### A. RESEARCH QUESTIONS

This study aims to provide, manage, and monitor services to groups of subscribers, satisfying the corresponding set of KPIs for the quality assessment of services provided to these groups. The solution is to provide services in the form of beamforming-as-a-service based on optimizing system resources and ensuring the Multi-Group MIMO (MG-MIMO) criteria.

To achieve this aim, the following research questions were defined.

1) What is the importance of beamforming in delivering multicast/broadcast services in 5G/6G systems?

**TABLE 2.** Broadcast projects over 4G/5G.

| Project | Nation | Locations | Time | Technologies | Transmission power | Frequencies | Bandwidth | Speed, service | Tower |
|---|---|---|---|---|---|---|---|---|---|
| 5G TODAY | Germany | Main traffic routes between Munich and Salzburg | July-2017 to Oct-2019 | FeMBMS | 5 kW (100 kW ERP) | SFN, 750-760 MHz. | 5 MHz and 10 MHz. | Video and IP data | HPHT |
| 5G-Xcast | 5G-PPP, EU | Munich, Surrey and Turku | June-2017 to July-2019 | MooD for hybrid broadcast service | ca. 400W | SFN, 700 MHz (LTE band 28) | 2 x 10 MHz | 3.77 Mbps; 4K/8K UHDTV/ VR/AR/MR, 360° visual media and NGA | HPHT |
| 5G Rural First | UK | Orkney Islands (Scotland), Harper Adams and Somerset (England) | Mar-2018 to Sep-2019 | all-G O-RAN, HetNet Gateway, network slice, control and user plane separation | Coverage within the islands' 2000 square kilometers | Licence-exempt spectrum (shared use): TV whitespace, 2.4GHz/ 5GHz/ 60GHz | Broadband | 10 Mbps; spectrum sharing, and new applications & services | N/A |
| Tower Overlay | Italy and France | Eiffel Tower | Apri-2015 | LTE-Advanced eMBMS | 2.7 kW ERP | SFN, channel 53, 730 MHz | 8 MHz | 10 Mbps; LTE A+ / DVB-T2 | HPHT |
| Trial of LTE-B in rural Norway | Norway | Norwegian West Coast | 2017 to 2019 | LTE-Broadcast (LTE-B) | 1 x 40 W | 758 – 778 MHz downlink; 708 – 718 MHz uplink | 2 x 10 MHz | 4 TV channels in HD, 4÷5 Mbps per channel | HPHT |

2) How to generate beam(s) for a group of users for multicast/broadcast services?

3) Each or several (corresponding to the number of layers) BF will serve a certain number of UEs, so how does the BF management algorithm for each service area?

To address these issues, we identify the importance of BF in providing effective services to groups of subscribers.

### B. BFaaS CONCEPT AND VISION

The concept of BFaaS is to provide beams to user groups corresponding to different services so that the service provider is not concerned with the specific hardware architecture of the network, but rather only with the quantity and characteristics of the required beams.

BF is a limitation and an advantage of mMIMO. The use of BF increases the Signal-to-Interference-plus-Noise Ratio (SINR) and reduces interference, but also makes the management of beams more complex. Furthermore, broadcast systems currently only use the HPHT configuration for transmitting and broadcasting services. Therefore, the effectiveness of beamforming cannot be exploited fully. The proposed BFaaS solution considers user groups as the focal point. User groups connect to the network through beamforming based on the cost optimization and efficient energy usage of the base station. User groups are committed to the quality of service. Users who are not guaranteed quality of service will report to the network for appropriate adjustments based on the statistical performance of user usage.

The principle of BFaaS is based on the concept of service-oriented architecture. The service provider (e.g., multimedia and broadcast services) leases beams for each coverage area. The beam provider will have to design, adjust, and manage the beams that meet the service requirements specified in the contract between the service provider and the network operator. Payments vary depending on specific parameters such as QoS or bitrate. The BFaaS allows the establishment of new business models.

The effectiveness of BFaaS is that service providers do not need to be concerned with the physical details and specific architecture of the network but only need to focus on ensuring that the UE group they manage will receive the service with the required QoS and KPIs. Service providers do not need to have the know-how or in-house resources to maintain such radio solutions, thus avoiding downsides of ownership. However, the challenge is how many beams and protocols will be needed to meet the service requirements and how the BF control mechanism will be implemented in the gNB.

The principle of the proposed BFaaS scheme is illustrated in Fig. 4. The BFaaS principle is implemented in the 5G network architecture as follows: Suppose that we need to transmit three content data to three different service areas. For broadcast services, the content is distributed through Content Servers, which can be static or dynamic content that will change based on input data, personalized on each page, depending on user input data (as in the case of 5G-Xcast) entering the 5G core, through the UPF function, and through Backhaul to the DU Center. Hereafter, the content data of each group will be separated through BFaaS #1 to #3 to reach user groups #1 to #3. Because the data is layered, each data group can be transmitted to different layers depending on the speed and quality required.

The BFaaS architecture can be implemented using the SDN/NFV techniques. Here, NFV is used to migrate network functions, which are typically performed on network devices, to the cloud. When migrating the BFaaS functions to the cloud, an orchestrator must coordinate and schedule when to initiate and run the BFaaS service.

Table 3 presents a comparison of the proposed BFaaS solution with the other as-a-service solutions.

**TABLE 3.** Comparison of the proposed BFaaS solution with other as-a-Service solutions.

| | RANaaS | NSaaS | IaaS | Radio resource as a Service (RaaS) | Proposed BFaaS |
|---|---|---|---|---|---|
| **Concept** | Provide and supervise virtual RANs to meet the requirements of an E2E service. | Provide and supervise customized network slices as services to meet their requirements and the right level of services. | Provide and supervise on-demand computing infrastructure and remote access to physical resources such as servers, network devices, and storage drives. | Provide and supervise the radio resources as an online service, that is characterized by time, space, frequency, and physical resource block. | Provide and supervise customized massive MIMO based RF beams for groups of service. |
| **Principle** | RAN functionalities are virtualized and flexibly centralized on cloud and provided depending on the actual load and network characteristics. | Allows the operators to provision customized network slices to individual customers, and eventually enables these customers to gain access to some network slice management capabilities. | To provide virtual computing and offer networking and storage services to customers. With this service, the rented resources can be scaled up or down at any time if you want to integrate an additional server or reduce the computing power. | The radio resources can be pooled, abstracted, isolated, assigned and sliced properly according to the demands and requirements, and then are offered as a service | The Service Provider (for example, multimedia, broadcast) will lease beams for each different service area. |
| **Benefit** | Cost savings from COTS high processing density hardware. Efficient partitioning and Energy saving of network functions and resources to support the 5G use cases. | Permits automation of resource customization and capability exposure to the vertical business customers for managing their own slices. | Service providers do not need to care about maintaining network hardware devices, as well as the operation of the network system. | The available radio resources can be utilized more efficiently by permitting different parties to share the same spectrum. | The service provider does not need to concern themselves with the specific architecture of the network, but rather only with ensuring that the UEs they manage will enjoy the service with the required QoS and KPIs. |

### C. BFaaS MODEL

We define a Beam Service Area (BSA) as a service area that contains a set of UEs, with each BSA corresponding to a service that needs to be deployed on this set of UEs, that is, corresponding to eMBB content. In this regard, we must optimally determine a set of beams for each BSA. Each BSA has one or more beams, which can be combined using Type I or Type II precoding codebooks. Each beam corresponds to a codebook vector. The BFaaS model is illustrated in Fig. 7.

Next, we define the Total Service Area (TSA) as a coverage area where one or more BSAs can be formed, and $\boldsymbol{\mathcal{M}} = \{\mathcal{M}_1, \ldots, \mathcal{M}_n, \ldots, \mathcal{M}_M\}$ is a set of all BSAs in the Total Service Area (TSA) and $\boldsymbol{\mathcal{E}} = \{\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_M\}$ is the set of services in TSA. Each BSA corresponds to a subset of services. The total number of service layers of $n$-th BSA is $\varepsilon_n$. The QoS of each UE in a BSA can vary depending on the commitment to the UE of the Service Provider.

Denote $\boldsymbol{\mathcal{K}} = \{K_1, \ldots, K_n, \ldots, K_M\}$ is the set of beams. The subset $K_n$ beams serve the $n$-th BSA. The total number of beams is $K = \sum_{i=1}^{M} K_i$.

Let $\mathbb{U}$ be the set of all UEs. $\mathbb{U} = \{\mathcal{U}_1, \ldots, \mathcal{U}_n, \ldots, \mathcal{U}_M\}$, where $\mathcal{U}_n$ is the set of UEs belonging to the nth BSA. The total number of UEs in the TSA is $U = \sum_{i=1}^{M} \mathcal{U}_i$.

Assume we have to transmit $M$ broadcast streams, corresponding to $M$ service messages (the $m$-th service has $\varepsilon_m$ layers). Each broadcast stream is precoded by precoding code book $\mathbf{V}_{s,m}^{(i)}$, where $s$ is sub-band index. The precoding code book is described in beam generation part. The $m$-th broadcast message is transmitted by $i$-th beam with the transmit power of $P_m^{(i)}$.

### D. BEAM GENERATION

This section presents the precoding code book types and how to generate beams for the BSA service areas to provide service in the form of beams.

Without loss of generality, we consider the downlink of a BFaaS-based broadcasting service system model which is illustrated in Fig. 5. The system uses a massive MIMO system with Discrete Fourier Transform (DFT) based 3D beamforming, and the physical antenna configuration is a uniform planar array.

With a large number of antennas, beamforming creates beams both vertically and horizontally towards the UEs. In this study, we use 3D beamforming and a DFT based codebook. In 3D beamforming, it is necessary to control both the azimuth and elevation angles of the beams.

#### 1) ANTENNA CONFIGURATION

In DFT-based 3D beamforming, the question here is how many beams are managed and controlled. The 5G NR supports both a single panel and a uniform and non-uniform multipanel. In 5G NR, the logical antenna configuration is described by three parameters: $N_g$ is the number of panels, $N_1$ is the number of columns, and $N_2$ is the number of rows in a panel. The parameters are shown in Fig. 6.

In association with $N_1$ and $N_2$, $Q_1$ and $Q_2$ are oversampling factors to determine the sweeping steps of a beam during beam management (beam tracking). $Q_1$ determines the sweeping step in the horizontal direction and $Q_2$ determines the sweeping step in the vertical direction.

One of the parameters to be defined is the Channel State Information - Reference Signal (CSI-RS) antenna port.

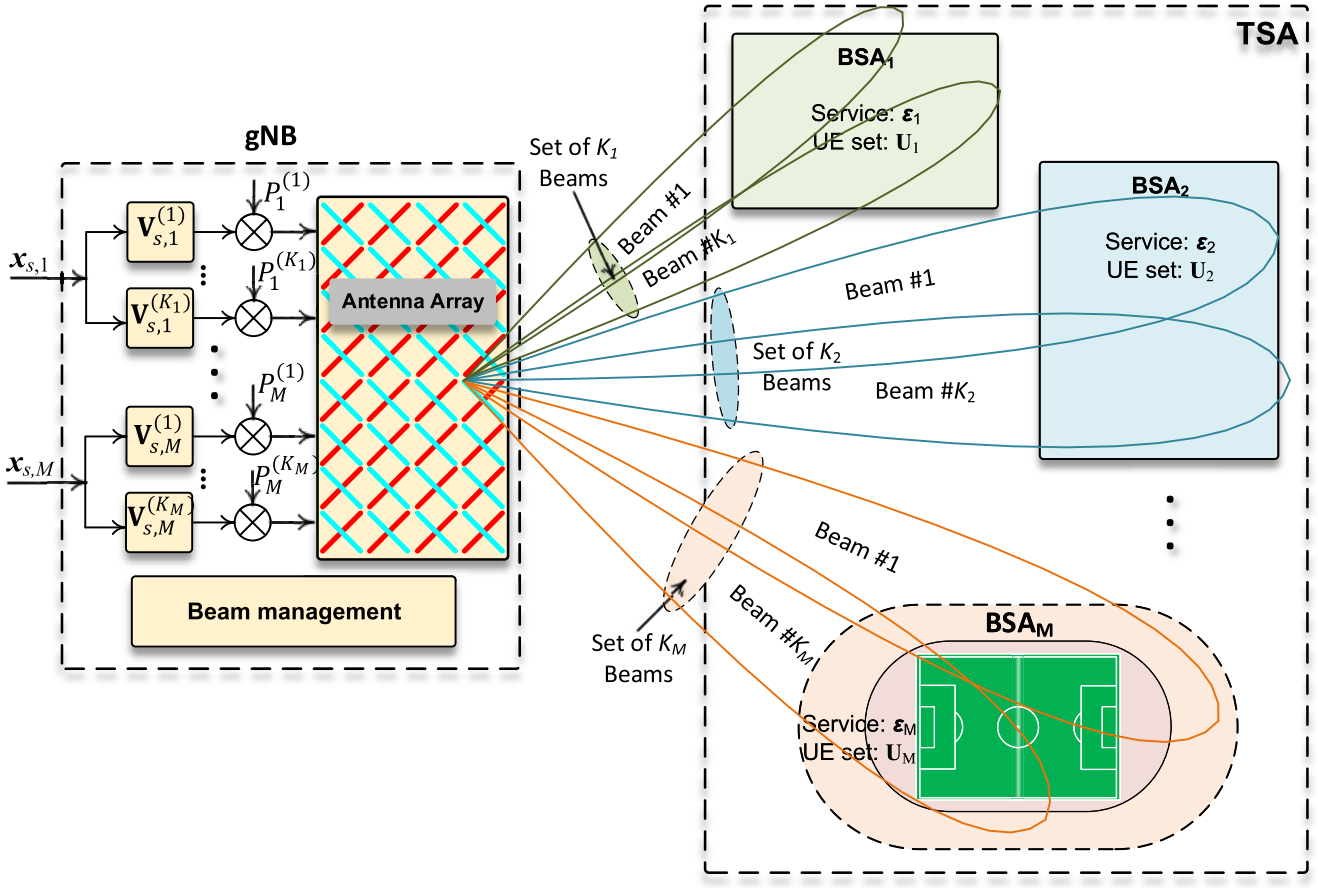


**FIGURE 5.** BFaaS model.

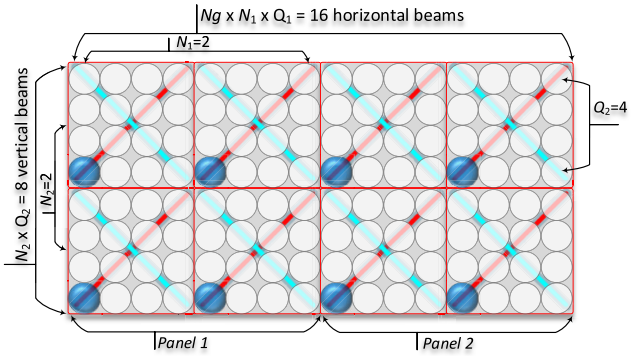


**FIGURE 6.** Multi panel antenna configuration.

Each antenna port is assigned a dedicated reference signal. Each antenna port represents a unique channel. The receiver can use a reference signal transmitted on the antenna port to estimate the channel model for this antenna port, which can subsequently be used to decode data transmitted on the same antenna port [35].

There are two types of beamforming control based on the codebook: Type I and Type II codebooks. Type I relies on creating only one specific beam, and controlling this beam based on phase control rather than amplitude. Type II is a linear combination of multiple beams with different phases and amplitudes. Therefore, controlling this beam is based on selecting a group of beams and controlling the phase and amplitude of each beam [36].

### 2) TYPE I CODEBOOK

Precoding Matrix Indicator (PMI) Type I codebook is defined as predesigned set of beamforming codevectors as

$$\mathbf{C} = \{\boldsymbol{\nu}_1, \boldsymbol{\nu}_2, \ldots, \boldsymbol{\nu}_N\}, \tag{1}$$

where a specific UE requests that the gNB use the best codevector or the best combination of a number of codevectors based on PMI and CQI information reported from that UE for Type I and Type II Codebooks, respectively.

The beamforming weights for $N_1$ dimension in horizontal direction is represented as

$$\mathbf{w}_{N_1}(n_l) = \frac{1}{\sqrt{N_1}} \left[ 1 \; e^{\frac{j2\pi n_l}{N_1 Q_1}} \; \ldots \; e^{\frac{j2\pi (N_1-1) n_l}{N_1 Q_1}} \right]^T, \quad n_l \in \{0, 1, \ldots, N_1 Q_1 - 1\}. \tag{2}$$

The beamforming weights for $N_2$ dimension in vertical direction is represented as

$$\mathbf{w}_{N_2}(m_l) = \frac{1}{\sqrt{N_2}} \left[ 1 \; e^{\frac{j2\pi m_l}{N_2 Q_2}} \; \ldots \; e^{\frac{j2\pi (N_2-1) m_l}{N_2 Q_2}} \right]^T, \quad m_l \in \{0, 1, \ldots, N_2 Q_2 - 1\}. \tag{3}$$

For ranks $\geq 2$, the precoder structure is a combination of the following two matrices

$$\boldsymbol{v}_s^{(l)} = \mathbf{W}_{\text{UPA}}^{(l)} \times \boldsymbol{\Phi}_s^{(l)}, \tag{4}$$

where $\mathbf{W}_{\text{UPA}}^{(l)}$ is a wideband component. The number of layers $N_l \leq \min(N_T, N_R)$, $N_T$ is the number of transmit antennas, and $N_R$ is the number of receiver antennas for the general case. $\mathbf{W}_{\text{UPA}}^{(l)}$ is used for beam selection in the vertical and horizontal directions and is of the form

$$\mathbf{W}_{\text{UPA}}^{(l)} = \begin{bmatrix} \mathbf{b}_{n_l m_l}^{(l)} & \mathbf{0} \\ \mathbf{0} & \mathbf{b}_{n_l m_l}^{(l)} \end{bmatrix}, \tag{5}$$

where

$$\mathbf{b}_{n_l m_l}^{(l)} = \mathbf{w}_{N_1}(n_l) \otimes \mathbf{w}_{N_2}(m_l) \tag{6}$$

is of 2D DFT structure, $\otimes$ is Kronecker product.

The sub-band component $\boldsymbol{\Phi}_s^{(l)}$ represents the NPSK co-phasing between the two polarizations for the $l$-th layer in the $s$-th sub-band. For rank $R = 1$, the subband component is of the form

$$\boldsymbol{\Phi}_s^{(1)} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ \varphi_s \end{bmatrix}, \quad \varphi_s = e^{j\pi s/2} \in \left\{0, \frac{\pi}{2}, \pi, \frac{3\pi}{2}\right\}. \tag{7}$$

and codebook for 1-layer CSI reporting is

$$\boldsymbol{v}_{n,m,s}^{(1)} = \frac{1}{\sqrt{2}} \begin{bmatrix} \mathbf{b}_{nm}^{(l)} \\ \varphi_s \mathbf{b}_{nm}^{(l)} \end{bmatrix}. \tag{8}$$

Codebook for 2-layer CSI reporting is

$$\boldsymbol{v}_{n_1,m_2,n_1,m_2,s}^{(2)} = \frac{1}{\sqrt{4}} \begin{bmatrix} \mathbf{b}_{n_1 m_1}^{(l)} & \mathbf{b}_{n_2 m_2}^{(l)} \\ \varphi_s \mathbf{b}_{n_1 m_1}^{(l)} & -\varphi_s \mathbf{b}_{n_2 m_2}^{(l)} \end{bmatrix}. \tag{9}$$

The general form of the PMI codebook $\mathbf{V}$ is as follows:

$$\mathbf{V}_{Paramerters}^{(No\ of\ Layers)} = \frac{1}{\sqrt{N_g 2R}} \begin{pmatrix} \mathbf{b}\ \mathbf{b} \cdots \mathbf{b} \\ \varphi\mathbf{b}\ \varphi\mathbf{b} \cdots \varphi\mathbf{b} \end{pmatrix}, \tag{10}$$

where the number of columns is equal to number of layers and number of rows is equal to number of CSI-RS ports, and $R$ is rank. For conclusion, we have:

- Number of polarizations = 2,
- Number of CSI-RS antenna ports $P_{CSI-RS} = 2 \times (N_1 \times N_2)$,
- Number of beams in a column = $N_1 \times Q_1$,
- Number of beams in a row = $N_2 \times Q_2$,
- Number of total 2D beams = $(N_1 \times Q_1) \times (N_2 \times Q_2)$.

### 3) TYPE II CODEBOOK

The difference between the Type I and Type II codebooks is that Type I allows the selection of the best beam for that connection, whereas Type II allows for the selection of more than one beam and the combination of these beams in the amplitude and phase. However, the Type II codebook supports only transmissions up to rank two to maintain a reasonable level of feedback overhead. Similar to Type I, the Type II codebook comprises two matrices $\mathbf{V} = \mathbf{W}_1 \mathbf{W}_2$. For rank-1, $\mathbf{V} = \begin{bmatrix} \boldsymbol{v}_0^{(1)} \\ \boldsymbol{v}_1^{(1)} \end{bmatrix}$, for rank-2, $\mathbf{V} = \begin{bmatrix} \boldsymbol{v}_0^{(1)} & \boldsymbol{v}_0^{(2)} \\ \boldsymbol{v}_1^{(1)} & \boldsymbol{v}_1^{(2)} \end{bmatrix}$.

The weighted combination of $K$ beams for polarization $p$, layer $l$ is given by

$$\boldsymbol{v}_p^{(l)} = \sum_{k=0}^{K-1} \mathbf{b}_{n_k m_k}^{(l)} a_{p,l,k} e^{j\varphi_{p,l,k}}, \tag{11}$$

where $a_{p,l,k}$ and $e^{j\varphi_{p,l,k}}$ represent the wideband beam amplitude scaling factor and the phase combiner for beam $k$ and on polarization $p$ and layer $l$.

The signal received at each UE in the specific BSA is transmitted to all beams belonging to the same service class, and the interference signal is transmitted to the other service layer. Therefore, the signal received at the single $k$-th UE in the $n$-th BSA on the $s$-th subcarrier can be written as

$$\begin{aligned} \mathbf{y}_{s,n}^{(k)} = & \underbrace{\left(\mathbf{w}_{s,n}^{(k)}\right)^H \mathbf{H}_{s,n}^{(k)} \sum_{i=1}^{K_n} \sqrt{P_n^{(i)}} \mathbf{V}_{s,n}^{(i)} \boldsymbol{x}_{s,n}}_{Desired\ signal} \\ & + \underbrace{\left(\mathbf{w}_{s,n}^{(k)}\right)^H \mathbf{H}_{s,n}^{(k)} \sum_{m=1, m\neq n}^{M} \sum_{i=1}^{K_m} \sqrt{P_m^{(i)}} \mathbf{V}_{s,m}^{(i)} \boldsymbol{x}_{s,m}}_{\text{Inter BSA interference}} \\ & + \underbrace{\left(\mathbf{w}_{s,n}^{(k)}\right)^H \mathbf{n}_s^{(k)}}_{\text{Gaussian noise}}, \end{aligned} \tag{12}$$

where $\mathbf{H}_{s,n}^{(k)} \in \mathbb{C}^{N_R \times P_{CSI-RS}}$ is the channel matrix between the gNB and $k$-th UE at the $s$-th subcarrier, which is constructed as in (5). $N_R$ is the number of receiving antennas. $\mathbf{V}_{s,n}^{(i)} \in \mathbb{C}^{P_{CSI-RS} \times \varepsilon_n}$ is the PMI codebook for $i$-th beam, $i = 1..K_n$ of size $(P_{CSI-RS} \times \varepsilon_n)$, where $P_{CSI-RS}$ is the number of CSI-RS ports. $\varepsilon_n$ is the number of layers and also the number of data streams spatially multiplexed to the UE, as previously described. $\boldsymbol{x}_{s,n} \in \mathbb{C}^{\varepsilon_n \times 1}$ denotes the transmission message for service $\varepsilon_n$ in the $n$-th BSA. $\boldsymbol{x}_{s,n}$ is the unit power, which implies that $\mathbb{E}\{\boldsymbol{x}_{s,n}(\boldsymbol{x}_{s,n})^H\} = \mathbf{I}_{\varepsilon_n}$. $P_n^{(i)}$ is transmit power of $i$-th beam in the $n$-th BSA. $\mathbf{w}_{s,n}^{(k)} \in \mathbb{C}^{N_R \times \varepsilon_n}$ is the receive beamforming vector for maximizing the received signal at the UE. $\mathbf{n}_s^{(k)} \in \mathbb{C}^{N_R \times 1}$ is the additive white Gaussian noise (AWGN) vector having the form of $\mathbf{n}_s^{(k)} \sim \mathcal{CN}(0, (\sigma_s^{(k)})^2 \mathbf{I}_{N_R})$, and $(.)^H$ stands for the conjugate transpose.

The total beamforming gain of the $i$-th beam at the $k$-th UE in the $n$-th BSA is

$$BG_{s,n}^{(k)} = \left\| \left(\mathbf{w}_{s,n}^{(k)}\right)^H \mathbf{H}_{s,n}^{(k)} \mathbf{V}_{s,n}^{(i)} \right\|^2. \tag{13}$$

The SINR for $k$-th UE is given by

$$\text{SINR}_{s,n}^{(k,l_n)} = \frac{\left| \left(\mathbf{w}_{s,n}^{(k,l_n)}\right)^H \mathbf{H}_{s,n}^{(k,l_n)} \sum_{i=1}^{K_n} \sqrt{P_n^{(i)}} \mathbf{V}_{s,n}^{(i,l_n)} \right|^2}{\left(\mathbf{w}_{s,n}^{(k,l_n)}\right)^H \mathbf{Q}_{s,n}^{(k,l_n)} \mathbf{w}_{s,n}^{(k,l_n)} + \left(\sigma_s^{(k)}\right)^2 \mathbf{I}_{N_R}}, \tag{14}$$

where $l_n$ is the layer index (data stream) $l_n = 1..\varepsilon_n$ in the $n$-th BSA, $n = 1, \ldots, M$. $\mathrm{Q}_{s,n}^{(k,l_n)} \in \mathbb{C}^{N_R \times N_R}$ is the interference covariance matrix, which is given by

$$\mathrm{Q}_{s,n}^{(k,l_n)} = \sum_{m=1,m\neq n}^{M} \sum_{i=1}^{K_m} \sqrt{P_m^{(i)}} \mathbf{H}_{s,n}^{(k)} \mathbf{V}_{s,m}^{(i)} \left(\mathbf{V}_{s,m}^{(i)}\right)^H \left(\mathbf{H}_{s,n}^{(k)}\right)^H. \tag{15}$$

The Received Signal Strength Indicator (RSSI) level received from $i$-th beam of $k$-th UE is

$$\mathrm{RSSI}_{s,n\ dB}^{(k,i)} = P_{n\ dB}^{(i)} + BG_{s,n\ dB}^{(k)} - PL\,(d)_{\,dB}, \tag{16}$$

where $PL$ is the path loss and $d$ is the distance from gNB to UE.

The Spectral Efficiency (SE) at the $s$-th subcarrier and $k$-th UE in the $n$-th BSA is computed as

$$R_{s,n}^{(k)} = \sum_{l_n=1}^{\varepsilon_n} \log_2\left(1 + \mathrm{SINR}_s^{(k,l_n)}\right). \tag{17}$$

The overall system throughput $\mathcal{C}$ is obtained as

$$\mathcal{C} = \sum_{k\in\mathbb{U}} \sum_{s=1}^{J} N_{s,n}^{PRB} R_{s,n}^{(k)}, \tag{18}$$

where $N_{s,n}^{PRB}$ and $J$ are the number of allocated physical resource blocks per BW per subcarrier and the number of aggregated subcarriers for $n$-th BSA, respectively.

### E. CHANNEL MODEL

This section describes the channel model which is used in the system model. In practice, there are two 5G channel models: the ring channel model and map-based channel model [37].

The one-ring MIMO channel model was developed using the geometry-based stochastic model method. The one-ring channel model is appropriate for describing environments in which the base station is elevated and unobstructed, whereas user equipment is surrounded by a large number of local scatterers. The two-ring and elliptical models are appropriate for environments in which both the base stations and users are surrounded by local scatterers.

The scatterers are assumed to be located on a ring with radius $r$, which is specified by the Root Mean Square (RMS) delay spread. The aim was to create a transmission delay fit for the measured Power Delay Profile (PDP) of the channel.

The map-based 5G channel model is a deterministic model of physical effects in which buildings and walls are modelled as rectangular objects [35]. Ray-tracing is a method of approximating the propagation of waves in an environment using discrete rays. The model is based on ray tracing, using a simplified 3D geometric description of the propagation environment [38].

In this study, we use map-based model. The Map-based model was chosen because it is suitable for massive MIMO and advanced beamforming techniques. It is also appropriate for the realistic modeling of path loss in the case of D2D and V2V. Therefore, this model is suitable for network planning, such as service group planning.

The main parameters of a channel model include PDP, Path Loss (PL) and Delay Spread (DS) as follows:

**Power delay profile**: PDP provides the intensity of a signal received through a multipath channel as a function of time delay. In this model, the received signals are scattering points in the form of a 5 × 5 grid, reflection and diffraction, Line of Sight (LoS) component, and ground reflected and scattered components.

Consider $N$ clusters, cluster index is $n = 1, \ldots, N$. $M_n$ is number of paths within cluster $n$. Denote $u$ is Rx antenna element and $s$ is Tx antenna element. The cluster powers are determined by

$$P'_n = \exp\left(-\tau_n \frac{r_\tau - 1}{r_\tau \mathrm{DS}}\right).10^{\frac{-Z_n}{10}}, \tag{19}$$

where $Z_n \sim \mathcal{N}(0, \zeta^2)$ is the per-cluster shadowing term in [dB], $\tau_n$ is the delay, DS is the delay spread, $r_\tau$ is the delay distribution proportionality factor, and cluster index $n$ =1, $\ldots, N$. $N$ is the number of clusters. Under the LoS condition, a single LoS ray with a power of

$$P_{1,\mathrm{LOS}} = \frac{K_R}{K_R + 1}, \tag{20}$$

and the power of cluster $n$ is

$$P_n = \frac{1}{K_R + 1} \frac{P'_n}{\sum_{n=1}^{N} P'_n} + \delta\,(n-1)\,P_{1,\mathrm{LOS}}, \tag{21}$$

where the Ricean K-factor $K_R$ is the ratio of the power in the direct path to that in the scattered paths [39]. $\delta(.)$ is Dirac's delta function. For a strong LoS signal, the K-factor was high, whereas its value was low in the presence of a strong multipath with a direct LoS signal. The K-factor increased as the LoS dominated.

The Complex Impulse Response (CIR) of the time-varying channel between RX antenna element $u$ and TX antenna elements $s$ for Non-Line of Sight (NLoS) of $m$-th ray within $n$-th cluster as

$$\begin{aligned} &h_{u,s,n,m}^{\mathrm{NLoS}}(t) \\ &= \sqrt{\frac{P_n}{M}} \begin{bmatrix} F_{Rx,u,\theta}\left(\theta_{n,m}^{ZoA}, \phi_{n,m}^{AoA}\right) \\ F_{Rx,u,\phi}\left(\theta_{n,m}^{ZoA}, \phi_{n,m}^{AoA}\right) \end{bmatrix}^T \\ &\times \begin{bmatrix} \exp\left(j\mathbf{\Phi}_{n,m}^{\theta\theta}\right) & \sqrt{(\kappa_{n,m})^{-1}} \exp\left(j\mathbf{\Phi}_{n,m}^{\theta\phi}\right) \\ \sqrt{(\kappa_{n,m})^{-1}} \exp\left(j\mathbf{\Phi}_{n,m}^{\phi\theta}\right) & \exp\left(j\mathbf{\Phi}_{n,m}^{\phi\phi}\right) \end{bmatrix} \\ &\times \begin{bmatrix} F_{Tx,s,\theta}\left(\theta_{n,m}^{ZoD}, \phi_{n,m}^{AoD}\right) \\ F_{Tx,s,\phi}\left(\theta_{n,m}^{ZoD}, \phi_{n,m}^{AoD}\right) \end{bmatrix} \exp\left(j2\pi \frac{(\mathbf{r}_{n,m}^{Rx})^T.\mathbf{d}_u^{Rx}}{\lambda_0}\right) \\ &\times \exp\left(j2\pi \frac{(\mathbf{r}_{n,m}^{Tx})^T.\mathbf{d}_s^{Tx}}{\lambda_0}\right) \exp\left(j2\pi \frac{(\mathbf{r}_{n,m}^{Rx})^T.\boldsymbol{v}^{UE}}{\lambda_0}\right). \end{aligned} \tag{22}$$

The line-of-sight channel coefficient is as follows:

$$\begin{aligned} & h_{u,s,1}^{\mathrm{LoS}}(t) \\ &= \begin{bmatrix} F_{Rx,u,\theta}\left(\theta_{LoS}^{ZoA}, \phi_{LoS}^{AoA}\right) \\ F_{Rx,u,\phi}\left(\theta_{LoS}^{ZoA}, \phi_{LoS}^{AoA}\right) \end{bmatrix}^T \\ &\quad \times \begin{bmatrix} \exp\left(j\boldsymbol{\Phi}_{LoS}\right) & 0 \\ 0 & -\exp\left(j\boldsymbol{\Phi}_{LoS}\right) \end{bmatrix} \\ &\quad \times \begin{bmatrix} F_{Tx,s,\theta}\left(\theta_{LoS}^{ZoD}, \phi_{LoS}^{AoD}\right) \\ F_{Tx,s,\phi}\left(\theta_{LoS}^{ZoD}, \phi_{LoS}^{AoD}\right) \end{bmatrix} \exp\left(j2\pi \frac{(\mathbf{r}_{LoS}^{Rx})^T.\mathbf{d}_u^{Rx}}{\lambda_0}\right) \\ &\quad \times \exp\left(j2\pi \frac{(\mathbf{r}_{LoS}^{Tx})^T.\mathbf{d}_s^{Tx}}{\lambda_0}\right) \exp\left(j2\pi \frac{(\mathbf{r}_{LoS}^{Rx})^T.\boldsymbol{v}^{UE}}{\lambda_0}\right), \end{aligned} \tag{23}$$

where $F_{Rx,u,\theta}$, $F_{Rx,u,\phi}$ and $F_{Tx,s,\theta}$, $F_{Tx,s,\phi}$ are the field patterns of the receive antenna element $u$ and transmit antenna element $s$ in the direction of the spherical basis vectors $\theta$ and $\phi$ respectively. The azimuth angle of arrival (AoA) is $\phi_{n,m}^{AoA}$, azimuth angle of departure (AoD) is $\phi_{n,m}^{AoD}$, zenith angle of arrival (ZoA) is $\theta_{n,m}^{ZoA}$, and zenith angle of departure (ZoD) is $\theta_{n,m}^{ZoD}$. $\kappa_{n,m}$ is the cross-polarization power ratio for $m$-th ray of $n$-th cluster.

$\{\boldsymbol{\Phi}_{n,m}^{\theta\theta}, \boldsymbol{\Phi}_{n,m}^{\theta\phi}, \boldsymbol{\Phi}_{n,m}^{\phi\theta}, \boldsymbol{\Phi}_{n,m}^{\phi\theta}\}$ are random initial phases for each ray $m$ of each cluster $n$ and for four different polarization combinations that are randomly generated with a uniform distribution $\mathcal{U}(-\pi, \pi)$. $\boldsymbol{\Phi}_{LoS}$ is the random initial phase of the LoS component.

$\mathbf{r}_{n,m}^{Rx}$ and $\mathbf{r}_{n,m}^{Tx}$ are the spherical unit vectors as

$$\mathbf{r}_{n,m}^{Rx} = \begin{bmatrix} \sin\left(\theta_{n,m}^{ZoA}\right)\cos\left(\phi_{n,m}^{AoA}\right) \\ \sin\left(\theta_{n,m}^{ZoA}\right)\sin\left(\phi_{n,m}^{AoA}\right) \\ \cos\left(\theta_{n,m}^{ZoA}\right) \end{bmatrix}, \tag{24}$$

$$\mathbf{r}_{n,m}^{Tx} = \begin{bmatrix} \sin\left(\theta_{n,m}^{ZoD}\right)\cos\left(\phi_{n,m}^{AoD}\right) \\ \sin\left(\theta_{n,m}^{ZoD}\right)\sin\left(\phi_{n,m}^{AoD}\right) \\ \cos\left(\theta_{n,m}^{ZoD}\right) \end{bmatrix}. \tag{25}$$

The UE velocity vector $\boldsymbol{v}^{UE}$ with speed $v$, travel azimuth angle $\phi_v$, and elevation angle $\theta_v$ is given by

$$\boldsymbol{v}^{UE} = \begin{bmatrix} \sin(\theta_v)\cos(\phi_v) \\ \sin(\theta_v)\sin(\phi_v) \\ \cos(\theta_v) \end{bmatrix}, \tag{26}$$

where $\mathbf{d}_s^{Tx}$ is the location vector of the transmit antenna element $s$, $\mathbf{d}_u^{Rx}$ is the location vector of the receive antenna element $u$, and $\lambda_0$ is the wavelength of carrier frequency.

The channel impulse response is obtained by adding the LoS channel coefficient to the NLoS channel impulse response and scaling both terms according to the desired K-factor $K_R$ as follows:

$$\begin{aligned} & h_{u,s}(\tau, t) \\ &= \sqrt{\frac{K_R}{K_R+1}} h_{u,s,1}^{\mathrm{LoS}}(t)\,\delta(\tau-\tau_1) \\ &\quad + \sqrt{\frac{1}{K_R+1}} \sum_{n=1}^{N}\sum_{m=1}^{M_n} h_{u,s,n,m}^{\mathrm{NLoS}}(\tau, t)\,\delta(\tau-\tau_n-\tau_{nm}), \end{aligned} \tag{27}$$

where $\tau_1$ is the delay of the LoS component and $\tau_n$ and $\tau_{nm}$ are the delays of the $n$-th cluster and $m$-th ray within the $n$-th cluster of the NLoS components, respectively.

The time-variant MIMO channel matrix between the gNB equipped with $N_t$ antennas and the UE equipped with $N_r$ antennas is obtained as

$$\mathbf{H}(\tau, t) = \begin{bmatrix} h_{11}(\tau, t) & \cdots & h_{1N_r}(\tau, t) \\ \vdots & \ddots & \vdots \\ h_{N_t1}(\tau, t) & \cdots & h_{N_tN_r}(\tau, t) \end{bmatrix}. \tag{28}$$

**The path loss model**: In a map-based channel model, the path loss is simplified using only three parameters, A, B, and C, as follows:

$$PL_{dB} = A\log_{10}\left(\frac{d}{1\mathrm{m}}\right) + B\log_{10}\left(\frac{f}{1\mathrm{GHz}}\right) + C. \tag{29}$$

For UMi Street Canyon, the LoS path loss model is

$$PL_{dB} = \begin{cases} PL_1, & 10\mathrm{m} \le d_{2D} \le d'_{BP}, \\ PL_2, & d'_{BP} \le d_{2D} \le 5\mathrm{km}, \end{cases} \tag{30}$$

where

$$PL_1 = 21\log_{10}\left(\frac{d_{3D}}{1\mathrm{m}}\right) + 20\log_{10}\left(\frac{f_c}{1\mathrm{GHz}}\right) + 32.4, \tag{31}$$

and

$$\begin{aligned} PL_2 = {} & 40\log_{10}\left(\frac{d_{3D}}{1\mathrm{m}}\right) + 20\log_{10}\left(\frac{f_c}{1\mathrm{GHz}}\right) + 32.4 \\ & - 9.5\log_{10}\left[\left(\frac{d'_{BP}}{1\mathrm{m}}\right)^2 + \left(\frac{h_{\mathrm{gNB}}-h_{\mathrm{UE}}}{1\mathrm{m}}\right)^2\right]. \end{aligned} \tag{32}$$

Breakpoint distance $d'_{BP}$ is given by [39]

$$d'_{BP} = \alpha_{BP}\frac{4h'_{gNB}h'_{UE}f_c}{c}, \tag{33}$$

where $f_c$ is the center frequency, $c$ is the speed of light, and $\alpha_{BP}$ is a breakpoint scaling factor, which is a function of the radio frequency and is introduced as

$$\alpha_{BP} = 0.87e^{-\frac{\log_{10}\left(\frac{f_c}{1\mathrm{GHz}}\right)}{0.65}}, \tag{34}$$

and $h'_{gNB}$ and $h'_{UE}$ are the effective antenna heights at gNB and UE, respectively.

The effective antenna heights $h'_{gNB} = h_{gNB} - h_E$ and $h'_{UE} = h_{UE} - h_E$, where $h_{gNB}$ and $h_{UE}$ are the actual antenna heights, and $h_E$ is the effective environmental height. For UMi $h_E = 1\mathrm{m}$.

**Delay Spread**: The frequency dependency of the DS is modelled as

$$DS_{[\log_{10}(s)]} = \mu_{\mathrm{DS}} + \gamma_{\mathrm{DS}}\log_{10}\left(\frac{f_c}{1\mathrm{GHz}}\right) + X(\sigma_{\mathrm{DS}}). \tag{35}$$

where $\mu_{\mathrm{DS}}$ represents the DS at 1 GHz, and $\gamma_{\mathrm{DS}}$ models the frequency dependency of the DS. $X(\sigma_{\mathrm{DS}})$ is a random variable with zero-mean and variance $\sigma_{DS}^2$. For UMi Street Canyon LoS model [38], $\mu_{\mathrm{DS}} = -7.1$, $\gamma_{\mathrm{DS}} = -0.75$, $\sigma_{\mathrm{DS}} = 0.38$.

An important aspect of the channel model is to understand the relationship between the UEs, based on which the beams are classified and designed to serve the UEs.

### F. OPTIMAL BEAMFORMING SELECTION

In this section, we present the proposed algorithms for the optimal beamforming selection.

Consider the proposed BFaaS model as illustrated in Fig. 5. Assume a service provider needs to provide services by BFaaS model. Let us remind that there are $\mathcal{M}$ set of BSAs in the Total Service Area (TSA) and $\mathbb{U}$ is the set of all UEs that the service provider has to serve. The total number of UEs is $U$. $\mathcal{K}$ is set of available beams.

Two problems are going to be solved in this study. The first problem is that, given a total number of UEs, find the minimum number of beams to ensure full coverage, and the overall system throughput is maximum. The constraints are that the total power is less than a given total power, and the number of UEs in any beam is greater than a given minimum number of UEs.

The problem is stated as follows:

$$\begin{aligned} \underset{\boldsymbol{\beta},\boldsymbol{P}}{\text{maximize}} \quad & \sum_{j\in\mathcal{K}}\sum_{i=1}^{U}\sum_{s=1}^{J}\beta_{i,j}N_{s,n}^{PRB}R_{s,n}^{(k)}, \\ \text{Subject to :} \quad & \mathbf{C1}: \sum_{j\in\mathcal{K}}\sum_{i=1}^{U}\beta_{i,j}P_n^{(j)} \le P_T, \\ & \mathbf{C2}: \min_j \sum_{i=1}^{K_j}\beta_{i,j} \ge N_{min}, \end{aligned} \tag{36}$$

where $\boldsymbol{\beta} = \{\beta_{i,j}\}, i = 1..U, j = 1..K$, where $\beta_{i,j} = 1$ if $i$-th UE is served by $j$-th beam and $\beta_{i,j} = 0$ otherwise, $U$ is the total number of UEs, $P_n^{(j)}$ is transmit power of $j$-th beam for $n$-th broadcast stream, $N_{s,n}^{PRB}$ is the number of radio resource blocks which are used for that transmission, $R_{s,n}^{(k)}$ is computed by (17), $P_T$ is total power, and $N_{min}$ is the limited number of UEs in a beam.

The second problem is that, given predetermined $M$ beams, select the appropriate beams in the set of available beams to maximize the overall system throughput. The constraints are similar to the first problem. The problem is stated as follows:

$$\begin{aligned} \underset{\boldsymbol{\beta},\mathbf{P}}{\text{maximize}} \quad & \sum_{m=1,j_m\in\mathcal{K}}^{M}\sum_{i=1}^{U}\sum_{s=1}^{J}\beta_{i,j_b}N_{s,n}^{PRB}R_{s,n}^{(k)}, \\ \text{Subject to:} \quad & \mathbf{C1}: \sum_{m=1,j_m\in\mathcal{K}}^{M}\sum_{i=1}^{U}\beta_{i,j_b}P_n^{(j_b)} \le P_T, \\ & \mathbf{C2}: \min_{j_m} \sum_{i=1}^{K_{j_m}}\beta_{i,j_m} \ge N_{min}. \end{aligned} \tag{37}$$

Next, we develop algorithms to solve the aforementioned problems.

#### 1) UE CLASSIFICATION AND BEAM ASSIGNMENT

The first algorithm is a pre-processing step for gathering information from the UEs. The next three algorithms (Algorithms 2, 3, and 4) solve the first problem to determine the minimum number of beams for that set of UEs. The last algorithm solves the second problem.

Multiple-stream spatial transmission is possible for mMIMO due to the available Spatial Degrees-of-Freedom (SDoF). UEs are classified based on the SDoF index. UEs in the line-of-sight (high K-factor) with a long transmission distance have a low SDoF index. UEs with a high SDoF index (high scattering) transmit multiple bit streams but receive low signal level [40].

The SDoF is defined as the number of singular values of the propagation channel matrix that exceeds the noise level. SDoF represents the number of bit streams that can be simultaneously transmitted, and it is equivalent to the number of orthogonal beams that must be generated simultaneously to transmit independent bit streams [41]. The gNB can simultaneously transmit the maximum of $r$ data streams to the target UE, thereby increasing channel throughput. In 5G NR, SDoF index is expressed as Rank Indicator (RI).

The UE classification and beam assignment algorithm is presented as follows:

**Algorithm 1** UE Classification and Beam Assignment

1:**Input**: The total user set $\mathbb{U} = \{\mathcal{U}_i\}$, the channel matrix of all users $\mathbf{H}$, the set of services $\boldsymbol{\mathcal{E}}$ and the total number of reference beams $K$.
2:**Initialization:** Initialization PMI codebook $\mathbf{V}$ for set of $K$ reference beams;
3: **for** $i = 1$ to $K$ **do**
4: Compute RSSI level of $k$-th UE in service $RSSI_k^{(i)}$;
5: **if** $(RSSI_k^{(i)} \ge threshold)$ **then**
6: Find service $\varepsilon_n$ of $k$-th UE by RI report;
7: Assign UE $k$ with service $\varepsilon_n$ into subset $\{\mathcal{U}_n\} \leftarrow UE_k^{\varepsilon_n}$;
8: Assign $UE_k^{\varepsilon_n}$ into $i$-th beam UE set: $\{\mathcal{U}^{(i)}\} \leftarrow UE_k^{\varepsilon_n}$;
9: Calculate total throughput of UEs in $i$-th beam: $R_i += R_k^{\varepsilon_n}$;
10: **end**
11: Find the UE with the highest RSSI level: $UE_{maxRSSI}^{(i)}$;
12: **end**
13: **Output**: UE subset of $i$-th beam $\{\mathcal{U}^{(i)}\}$, total throughput $R_i, i = 1..K$; Total number of UEs in service $U = \sum_{i=1}^{K}\mathcal{U}^{(i)}$ and $UE_{maxRSSI}^{(i)}$.

The input of **Algorithm 1** includes the set of users $\mathbb{U}$, the channel matrix of all users $\mathbf{H}$, the corresponding set of services $\boldsymbol{\epsilon}$ and the total number of reference beams $K$.

The algorithm works as follows. First, the system initializes precoding matrix $\mathbf{V}$ for the set of $K$ available beams. We call these beams reference beams.

Then, the system scans all UEs to determine the serving capability of each reference beam and the registered services of each UE using RSSI measurements and rank indicators which is calculated by (16). RI indicates the number of layers for multicast/broadcast streams transmission. Based on the measurements, the system determines the UE subset for each beam, total throughput of each beam, and total number of UEs in service.

The UE subset of $i$-th beam is denoted by $\{\mathcal{U}^{(i)}\}$. These subsets might overlap. It means that a specific UE can belong to some different beams.

The set of UEs with the highest RSSI level $UE^{(i)}_{maxRSSI}$ corresponding to each beam are considered as initial centroids for the fifth algorithm (K-means clustering).

We should also note that, in practice, it is difficult to report channel parameters to the gNB to estimate the centroids of the user groups [42]. In this study, the measured RSSI levels corresponding to the reference beams were used to determine the UE groups.

### 2) MINIMUM NUMBER OF BEAMS FINDING ALGORITHMS

Because system resources depend on the number of beams, minimizing the number of beams is of particular importance in optimizing system resources allocated to serving areas. In other words, the cost that service providers have to spend on network operators will be minimized for the total number of UEs. After all of UEs have been classified and all the reference beams are browsed. Next, we need to determine the minimum number of beams to serve a given set of UEs in order to optimize system resources.

Here, we consider three use cases:

- Select beams to serve the user set $\mathbb{U}$, and to ensure the service rates $\mathcal{E}$.
- Select beams based-on the user set $\mathbb{U}$ coverage.
- Select beams to serve the user set $\mathbb{U}$, and to ensure the service rates $\mathcal{E}$ and the power efficiency of the selected beams.

The algorithms 2,3,4 below solve these use cases, respectively.

The proposed **Algorithm 2** is an iterative greedy search algorithm. This algorithm is used to determine the minimum number of beams for service providers to hire from network operators.

In this algorithm, the weight of each beam is determined by the sum of the product of the number of UEs and its corresponding service rate. The weight of $i$-th beam is $b^{(i)} = u^{(i)} \times \varepsilon_i$, where $u^{(i)}$ is number of UEs in this beam after overlapping UEs have been removed. $\varepsilon_i$ is corresponding service rate.

The input of the algorithm is set of all UEs $\mathbb{U} = \{\mathcal{U}_i\}$, total power budget and number of available beams $K$.

The initial subset of UEs for each beam $\{\mathcal{U}^{(i)}\}$ is computed by the UE classification and beam assignment algorithm (**Algorithm 1**).

Start with the first iteration $i = 1$, $S := \emptyset$ which means no beam has been selected. $S$ is the set of selected beams.

The algorithm sorts the beams in descending order of weights and selects the beam with the highest weights if the remaining number of UEs is greater than the number of UEs in $i$-th beam until all UEs are selected.

Finding minimum number beams can be performed more efficiently using the dynamic programming algorithm.The next algorithm is as follows:

**Algorithm 2** Greedy Algorithm for Finding Minimum Number of Beams

**Input:** $\mathbb{U} = \{\mathcal{U}_i\}$, $\mathbf{H}$, $\boldsymbol{\mathcal{E}}$, $P_{total}$, $UE_{min}$,$K$;
**Initialization:** $S = \emptyset$; Total power $P_t = 0$;
Perform the UE classification and beam assignment by **Algorithm 1** to obtain UE subset of $i$-th beam $\{\mathcal{U}^{(i)}\}$ and total number of UEs in service $U$;
Assign a temporary variable $u^{(i)} = \mathcal{U}^{(i)}$, $i = 1..K$;
Calculate the beam weights $b^{(i)} = u^{(i)} \times \varepsilon_i$, $i = 1..K$;
Sort the beams in descending order of weights;
**for** $i = 1$ **to** $K$ **do**
  **if** $(U \geq u^{(i)})$ **then**
    $S \leftarrow S \cup i$;
    $P_t := P_t + P_i$;
    $U := U - u^{(i)}$;
    **for** $j = i$ **to** $K$ **do**
      Eliminate overlapping UEs with $i$-th beam in $j$-th beam, the remaining is $u^{(j)'}$;
      Recalculate beam weights: $b^{(j)} = u^{(j)'} \times \varepsilon_j$;
      **if** $(u^{(j)'} < UE_{\min})$ **then** $j := j + 1$;
    **end**
  **end**
  Sort the beams in descending order of weights;
  **if** $(U = 0)$ **then break**;
  **if** $(P_t \geq P_{total})$ **then break**;
**end**
**Output:** $S$;

**Algorithm 3** Dynamic Programming Algorithm for Finding Minimum Number of Beams

**Input:** $\mathbb{U} = \{\mathcal{U}_i\}$, $\mathbf{H}$, $P_{total}$, $UE_{min}$;$K$;
**Initialization:** $S[0] = 0$; $S[i] = \infty$,;$P_t = 0$;
Perform the UE classification and beam assignment by **Algorithm 1** to obtain UE subset of $i$-th beam $\{\mathcal{U}^{(i)}\}$ and total number of UEs in service $U$;
Assign a temporary variable $u^{(i)} = \mathcal{U}^{(i)}$, $i = 1..K$;
**for** $i = 1$ **to** $U$ **do**
  **for** $j = 1$ **to** $K$ **do**
    **if** $(u^{(j)} \leq i)$ **and** $(S[i - u^{(j)}] + 1 < S[i])$ **then**
      $S[i] := S[i - u^{(j)}] + 1$;
      Eliminate overlapping UEs with $j$-th beam in $k$-th beam, $k$=$j..K$;
      $P_t := P_t + P_j$;
      **if** $(P_t \geq P_{total})$ **then break**;
    **end**
  **end**
**end**
**Output:** $S$;

The proposed **Algorithm 3** uses dynamic programming methods to find minimum number beams. The input of this algorithm is similar to that of **Algorithm 2**.

In the case of regardless of the services rate, the algorithm is simply stated as follows. Given $K$ beams and the number of UEs that each beam can serve is $\mathcal{U}^{(i)}$, $i = 1..K$, find the

minimum number of beams so that the total UEs have been served.

Note that $\mathcal{U}^{(i)}$ is initial values of number of UEs for each beam, $u^{(i)}$ is number of UEs in this beam after overlapping UEs have been removed.

In this algorithm, an important variable is the state variable. State $S[i]$ is defined as the minimum number of beams required to serve $i$ UEs, where $i$ is less than the total number of UEs. $S[.]$ is actually an array with array index varies from 1 to the total number of UEs.

The algorithm works as follows. Initial state values are initialized to $S[0] := 0$ and $S[i] = \infty, \forall i$.

In order to find state $S[i]$, we need to find all the previous states $S[j]$ with $j < i$. Once the state $S[i]$ has been found, we can easily find the state of $S[i+1]$.

At every single beam of $j$ with $u^{(j)} \leq i$, find the minimum number of beams so that the total UEs is $i - u^{(j)}$. Assume this number of beams equals $m$. If $m + 1$ is smaller than the current number of beams, then we update state $S[i] = m + 1$.

After iteration from 1 to $U$, $S[U]$ is the minimum number of beams that we must look for.

In **Algorithm 3**, we do not care about the service rate nor the power efficiency of the beams. **Algorithm 4** is interested in the service rate as well as the transmit power of each beam.

**Algorithm 4** is based on the profit/weight ratio of each beam. We define the profit/weight ratio of a beam as the total throughput of the beam divided by its corresponding power as

$$pw^{(i)} = \left(u^{(i)} \times \varepsilon_i\right)/P_{i'}, \tag{38}$$

where $i$ is beam index, $P_i$ is transmit power of $i$-th beam.

Since the optimization problem is dealing with fractional amounts of profit/weight ratio, we propose to use fractional knapsack algorithm to maximize the profit of the selected beams. The input of the algorithm is set of all UEs $\mathbb{U} = \{\mathcal{U}_i\}$, the initial subset of UEs for each beam $\{\mathcal{U}^{(i)}\}$, total power budget and number of available beams $K$. The algorithm is as follows:

**Algorithm 4** sorts out the beams in the decreasing order of profit/weight and allocates them using a greedy approach.

**Algorithm 4** Fractional Knapsack Algorithm for Finding Minimum Number of Beams

**Input:** $\mathbb{U} = \{\mathcal{U}_i\}$, $\mathbf{H}$, $\boldsymbol{\mathcal{E}}$, $P_{total}$, $UE_{min}$; $K$;
**Initialization:** $S$:=Ø #number_UE = 0;
$P_t = 0$; UE_inservice:=Ø;
Perform the UE classification and beam assignment by **Algorithm 1** to obtain UE subset of $i$-th beam $\{\mathcal{U}^{(i)}\}$ and total number of UEs in service $U$;
Assign a temporary variable $u^{(i)} = \mathcal{U}^{(i)}$, $i = 1..K$;
**for** $i = 1$ **to** $K$ **do**
  $pw^{(i)} = (u^{(i)} \times \varepsilon_i)/P_i$;
**end**
Sort the beams in descending order of $pw^{(i)}$, $i = 1..K$;
$i := 1$;
**while** (#number_UE $< U$) **do**
  UE_inservice $\leftarrow$ UE_inservice $\cup\, \mathcal{U}^{(i)}$;
  #number_UE $+= u^{(i)}$;
  $S \leftarrow S \cup i$;
  **for** $j = i$ **to** $K$ **do**
    Eliminate overlapping UEs, remain $u^{(j)'}$;
    Recalculate $pw^{(j)} = (u^{(j)'} \times \varepsilon_j)/P_j$;
    **if** $(u^{(j)'} < UE_{min})$ **then** $j := j + 1$;//bypass
  **end**
  Resort the beams in descending order of $pw^{(j)}$, $j = 1..K$;
  $P_t := P_t + P_i$;
  **if** $(P_t \geq P_{total})$ **then break**;
  $i = i + 1$;
  **if** $(i > K)$ **then break**;
**end**
**Output:** $S$;

### 3) MAXIMIZATION OF COVERAGE AREA WITH PREDETERMINED NUMBER OF BEAMS

The problem now is that assume the service provides hires $M$ beams from network operators, select $M$ beams from $K$ available beams so that the coverage area is maximized.

This can be solved using K-means clustering method. The input of the algorithm is $M$, and this is the number of clusters to be separated. After separation, $M$ beams are allocated to the centroids of the clusters as required. The principle of K-means clustering includes the nearest-neighbor rule. According to the nearest-neighbor rule, each UE is associated with the closest centroid.

The vector $x_i = [x_{i1}, x_{i2}, \ldots, x_{iK}]$ represents the RSSI level of $i$-th UE that is received by $K$ reference beams.

These reference vectors form a sparse matrix with many zero values for orthogonal beams. The distance from the UE to the centroids is calculated based on this matrix space.

The RSSI levels of the total $U$ UEs are represented as follows:

$$\boldsymbol{\mathcal{X}} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_U] \in \mathbb{R}^{K \times U}. \tag{39}$$

Denote $\mathbf{y}_i = [\mathrm{y}_{i1}, \mathrm{y}_{i2}, \ldots, \mathrm{y}_{iM}]$ is a label vector. If $\boldsymbol{x}_i$ is classified into cluster $k$ then $y_{ik} = 1$, and $\mathrm{y}_{ij} = 0, \forall j \neq k$ elsewhere. This implies that exactly one element of vector $\mathbf{y}_i$ is equal to 1 (corresponding to the cluster of $\boldsymbol{x}_i$), and the remaining elements are equal to 0.

The constraint of $\mathbf{y}_i$ can be written as

$$\mathrm{y}_{ik} \in \{0, 1\}, \sum_{k=1}^{M} \mathrm{y}_{ik} = 1. \tag{40}$$

Select $M$ UEs with the highest RSSI levels for different reference beams as the initial centroids. The error for the entire dataset is

$$\boldsymbol{\mathcal{L}}(\mathbf{Y}, \mathbf{M}) = \sum_{i=1}^{U} \sum_{j=1}^{M} \mathrm{y}_{ij} \left\| \boldsymbol{x}_i - \mathbf{m}_j \right\|_2^2, \tag{41}$$

where $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_U]$, $\mathbf{M} = [\mathbf{m}_1, \mathbf{m}_2, \ldots, \mathbf{m}_M]$ are the matrices generated by the label vectors of each UEs and centroid, respectively.

The optimization problem is stated as follows:

$$\mathbf{Y}, \mathbf{M} = \arg\min_{\mathbf{Y},\mathbf{M}} \sum_{i=1}^{U}\sum_{j=1}^{M} \mathrm{y}_{ij} \left\|\boldsymbol{x}_i - \mathbf{m}_j\right\|_2^2,$$
$$\text{Subject to}: \quad \mathrm{y}_{ij} \in \{0, 1\}, \ \forall i, j; \ \sum_{j=1}^{M} \mathrm{y}_{ij} = 1, \forall i. \tag{42}$$

The pseudo code for the algorithm is as follows:

**Algorithm 5** ***K*-Means Clustering Algorithm for Maximization of Coverage Area**

1:**Input:** $\mathbb{U} = \{\mathcal{U}_i\}$, $\boldsymbol{\mathcal{X}}$, $\mathbf{M}$; $\mathbf{H}$;$K$;
2: **Initialization:**
$n = 0$; $\mathcal{L}_{tot}^{(n)} = 1$;
$n = 1$; $\mathcal{L}_{tot}^{(n)} = 0$; $\mathrm{S}_j^{(n)} = \{m_j\}, j = 1, 2, \ldots, M$; $\boldsymbol{\mathcal{V}} = \emptyset$;
3: Perform the UE classification and beam assignment by **Algorithm 1** to obtain UE subset of $i$-th beam $\{\mathcal{U}^{(i)}\}$ and total number of UEs in service $U$, and $UE_{maxRSSI}^{(i)}$
4: **while** $(|\mathcal{L}_{tot}^{(n)} - \mathcal{L}_{tot}^{(n-1)}| > \epsilon\, \mathcal{L}_{tot}^{(n-1)})$ **do**
5: **for** $i = 1$ **to** $U$ **do**
6: **for** $j = 1$ **to** $M$ **do**
7: Compute $\boldsymbol{\mathcal{L}}(\mathrm{x}_i, \mathbf{m}_j^{(n-1)}) = ||\mathrm{x}_i - \mathbf{m}_j^{(n-1)}||_2^2$
8: **end**
9: Find $j^* = \underset{j^*}{\operatorname{argmin}}\, \boldsymbol{\mathcal{L}}(\mathrm{x}_i, \mathbf{m}_{j^*}^{(n-1)})$;
10: $\mathrm{S}_{j^*}^{(n)} = \mathrm{S}_{j^*}^{(n)} \cup \{i\}$;
11: **end**
12: **for** $j = 1$ **to** $M$ **do**
13: $\mathbf{m}_j^{(n)} = \frac{1}{\left|S_j^{(n)}\right|} \sum_{\mathbf{k} \in S_j^{(n)}} \mathrm{x}_k$;
14: **end**
15: $\mathcal{L}_{tot}^{(n)} = \sum_{i=1}^{U} \sum_{j=1}^{M} \mathrm{y}_{ij} ||\mathrm{x}_i - \mathbf{m}_j^{(n)}||_2^2$;
16: $n = n + 1$;
17: **end//**while
18: **for** $i = 1$ **to** $M$ **do**
19: $\mathbf{m}_i = \mathbf{m}_i^{(n)}$; $S_i = S_i^{(n)}$;
20: **for** $j = 1$ **to** $M$ **do**
21: $j^* = \underset{j}{\operatorname{argmin}} \left\|\boldsymbol{b}_j - \boldsymbol{m}_i\right\|_2^2$;
22: $\boldsymbol{\mathcal{V}} = \boldsymbol{\mathcal{V}} \cup \{j\}$;
23: **end**
24: **end**
25: **Output:** $\{S_j\}$; $\boldsymbol{\mathcal{V}}$;

In this algorithm, the input is the set of RSSI levels of the UEs corresponding to reference beam $\boldsymbol{\mathcal{X}}$ and the set of $M$ UEs with the highest RSSI levels. Set $M$ UEs with maximum RSSI as the initial centroids.

The output of the algorithm is $M$ beams. The $i$-th beam is chosen such that the distance from the reference vector of the corresponding beam of the form

$$\mathbf{b}_i = [\underbrace{0, 0, \ldots, 1, \ldots, 0}_{=1 \text{at position } i}], \tag{43}$$

to the centroids is minimum as

$$j^* = \arg\min_j \left\|\mathbf{b}_j - \mathbf{m}_i\right\|_2^2. \tag{44}$$

#### 4) ALGORITHM SUMMARY AND COMPLEXITY ANALYSIS

In this subsection, we summarize the proposed algorithms includin UE classification and beam assignment algorithm, Greedy Algorithm (GA), Dynamic Programming algorithm (DP), Fractional knapsack algorithm (Knap) for finding minimum number of beams and K-means clustering algorithm (K-means) for maximization of coverage area and analyze the computational complexity of each algorithm because computational complexity is important to their implementation.

Start with the UE classification and beam assignment algorithm (**Algorithm 1**), the purpose of this algorithm is to gather information from the UEs and evaluate the serving capacity of each reference beam. This algorithm is a preprocessing step for the next algorithms. The computational complexity of **Algorithm 1** mainly depends on the RSSI calculation. The complexity of the calculation of the 2-norm of three matrices multiplication $\| \ (\mathbf{w}_{s,n}^{(k)})^H \mathbf{H}_{s,n}^{(k)} \mathbf{V}_{s,n}^{(i)} \ \|^2$ is $O(\varepsilon_n N_R N_{gNB} + \varepsilon_n^2 N_{gNB}) = O(N_{gNB})$ (it is applied the constants removing rule). The computational complexity of precoding matrix $\mathbf{V}$ calculation is $O(N_{gNB}^2)$, then the complexity of the RSSI calculation is $O(N_{gNB}^3)$. Repeating this for $U$ users, we have the complexity of $O(UN_{gNB}^3)$. The total complexity of **Algorithm 1** is $O(KUN_{gNB}^3)$, where $N_{gNB}$ is number of antennas at gNB.

For GA algorithm (**Algorithm 2**), the purpose of this algorithm is to find the minimum number of beams by greedy method.

At first, we run the UE classification and beam assignment algorithm. This part has the complexity of $O(KUN_{gNB}^3)$ as we calculated. The sorting beams codes at line 5 and line 18 have the same computational complexity of $O(K \log K)$.

The eliminating overlapping UEs (line 13 to line 15) has the computational complexity of $O(K \log K)$, the average complexity to browse all beams to select the optimal beams would be $O(K \log KK \log K) = O(K^2 \log^2 K)$. Therefore, the total average computational complexity of GA algorithm is

$$\begin{aligned} &O(KUN_{gNB}^3 + K \log K + K^2 \log^2 K + K \log K) \\ &\quad = O(KUN_{gNB}^3 + K^2 \log^2 K). \end{aligned} \tag{45}$$

Greedy algorithms are algorithms that search and select the local optimal solution at each step with the hope of finding a globally optimal solution. We hope to find a beam that serves as many UEs as possible at each iterative step.

Similar to GA algorithm, DP algorithm (**Algorithm 3**) also finds the minimum number of beams but using the dynamic programming method. In this method, the main problem is divided into subproblems and then it uses the

storage method (memoization) to remember the results of solved subproblems.

The total average computational complexity of DP algorithm is

$$O\left\{KUN_{gNB}^3 + UK(K\log K)\right\} = O(KUN_{gNB}^3 + UK^2\log K). \quad (46)$$

The complexity of the DP algorithm for finding minimum number of beams is greater than that of the algorithm using greedy method.

The purpose of Knap algorithm (**Algorithm 4**) is to find the minimum number of beams but using the profit/weight ratio of each beam from (38).

The principle of this algorithm is based on the greedy method. By calculating the computational complexity of each part of the algorithm, we have the total average computational complexity of Knap algorithm is

$$O\left\{KUN_{gNB}^3 + K\log K + \log U\left[K(K\log K) + K\log K\right]\right\} = O(KUN_{gNB}^3 + K^2\log U\log K). \quad (47)$$

For maximization of coverage area K-means algorithm (**Algorithm 5**), similar to the above algorithms, first, the algorithm runs the UE classification and beam assignment. This part has the complexity of $O(KUN_{gNB}^3)$.

The next part of the algorithm is based on K-means clustering approach. Typically, the time complexity or the running time of the K-means clustering using Lloyd's algorithm is $O(UMKi)$ where $U$ is the number of points (number of UEs in this case), $M$ is number of clusters, since each UE can belong to one of $K$ beams then the dimensionality of data is $K$, $i$ is the number of iterations. When $U$ is large, $i$ and $M$ are not too large the time complexity of this is approximately equal $O(UK)$.

The last part of the algorithm is to select the corresponding beams which has the computational complexity of $O(M^2K)$. The total average computational complexity of K-means algorithm is

$$O(KUN_{gNB}^3 + UK + M^2K). \quad (48)$$

The computational complexities of proposed algorithms are summarized in Table 4.

**TABLE 4. Comparison of computational complexity.**

| Algorithm | Computational complexity |
|---|---|
| GA (**Algorithm 2**) | $O(KUN_{gNB}^3 + K^2\log^2 K)$ |
| DP (**Algorithm 3**) | $O(KUN_{gNB}^3 + UK^2\log K)$ |
| Knap (**Algorithm 4**) | $O(KUN_{gNB}^3 + \log UK^2\log K)$ |
| K-means (**Algorithm 5**) | $O(KUN_{gNB}^3 + UK + M^2K)$ |

## IV. SIMULATION RESULTS

In this section, the numerical results are presented. MATLAB Monte Carlo simulation was used to evaluate the effectiveness of the proposed beamforming schemes under different simulation conditions. Each simulation was performed by averaging over 2000 random channel realizations.

**TABLE 5. System parameters in simulations.**

| Parameter | Value |
|---|---|
| Carrier frequency | FR1 |
| Bandwidth | 50 MHz |
| SNR | [-5 dB ÷25 dB] |
| Propagation environment | Dense urban |
| Path loss model | UMi |
| Delay spread ( $\mu_{DS}$ , $\gamma_{DS}$ , $\sigma_{DS}$ ) | (-7.1, -0.75, 0.38) |
| PDP K-factor ( $\mu_K$ , $\sigma_K$ ) | (9, 5) |
| Link topology | O2O/O2I |
| Number of antennas | 192 to 256 |
| Number of UEs | 48 to 884 |
| Monte-Carlo | 2000 |
| Antenna-element spacing | $\lambda/2$ |

The simulation parameters are listed in Table 5. The number of antennas at the gNB varies from 192 to 256. The resolution of the DFT codebook is set to be $K = N_{gNB}$. Each UE has one single antenna. For the massive MIMO antenna configuration at gNB, 2D planar array is used to allocate multiple beams to different service areas. With this configuration, it is able to control the width and height of a beam in horizontal and vertical directions, respectively. Also, in this study, we used zero forcing precoding for broadcasting service messages.

For channel model, we used map-based 5G channel model, which is described in section III, for the simulations. The PDP, PL, and DS parameters are calculated by cluster, where each cluster corresponds to the service area of each group. These parameters, together with the Ricean K-factor $K_R$ and path loss model, were computed step-by-step, as described in [39]. This is the real situation which is guided by following the guidelines of the 3GPP technical report. We assess the dense urban propagation environment for actual environment simulation.

The channel bandwidth is fixed to 50 MHz with link topology was Outdoor to Outdoor (O2O), Outdoor to Indoor (O2I). The O2I channel condition and low/high condition period updates are set equal to the LoS/NLoS channel condition update period. The O2I or O2O channel condition is based on the UE antenna height. O2I condition is considered only for UEs with random antenna heights. The path loss and delay spread model is calculated by (30) and (35), respectively.

In the simulation, we compare the proposed algorithms including GA, DP, Knap, K-means to existing Single User Random Request (Random) [42].

In the case of K-means clustering algorithm, due to the purpose of this algorithm is to maximize the coverage area with a given number of beams, however, in order to compare the performance of all algorithms, we run the algorithm with every single number of beams to cover all the total number of UEs. In all mentioned algorithms, the UE classification and beam assignment algorithm is run first.

In the simulation for the first algorithm (**Algorithm 1**), UEs are uniformly distributed throughout the entire coverage area of a gNB and single layered services ($\varepsilon = 1$). For the following algorithms, the simulation process is conducted sequentially as the steps in the algorithm. The effectiveness of the proposed optimal beamforming selection algorithms for BFaaS scheme is evaluated by the achieved average SE and the total sum rate.

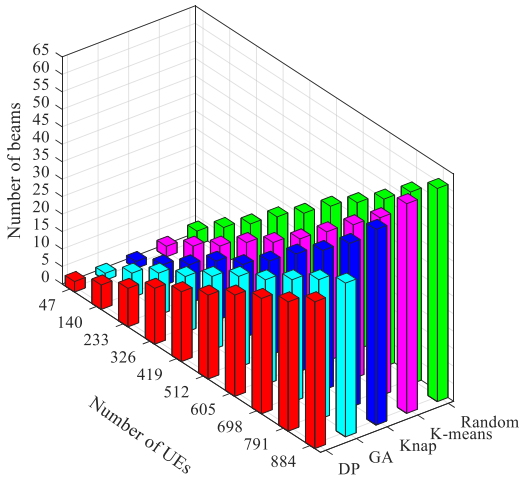


**FIGURE 7.** Minimum number of beams comparison.

Fig. 7 shows the minimum number of beams comparison among the 4 algorithms against the number of UEs. In this simulation, the number of UEs varies from 48 to 884. We can observe that the DP algorithm achieves the best performance and the more UEs the system manages, the more benefit the system will get with the DP algorithm. At low number of UEs regime, the performance of all the algorithms is very similar, but it diverges as the number of UEs increases.

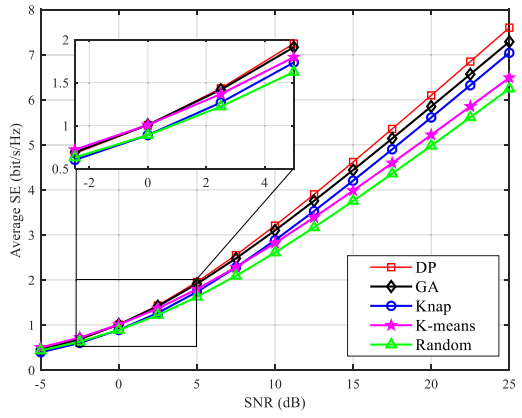


**FIGURE 8.** Average spectral efficiency comparison.

Fig. 8 shows a comparison of the average SE of the proposed algorithms for the Beamforming-as-a-Service for Multicast and Broadcast services. The number of antennas at the gNB in the simulation is set to 256 and the number of UEs is set to 512.

It can be observed from the plot that the performance of the proposed algorithms is better than that of the existing scheme. This is because the optimal beamforming selection algorithms eliminate the overlapping UEs from different beams to minimize number of beams. From there, the allocation of beams and the power utilization of the system are optimized. Thus, it can improve the sum rate and the average SE of the system. The Fig. 8 also showed that the DP algorithm provides the best performance in term of the average SE. At high SNR regime (25 dB), the average SE of the DP algorithm is better than that of the GA, Knap, K-means algorithms and the existing scheme from 5%, 7%, 14% to 17%, respectively. In low SNR regime (<5 dB), the DP and GA algorithms have similar performance and the DP algorithm outperform the existing scheme.

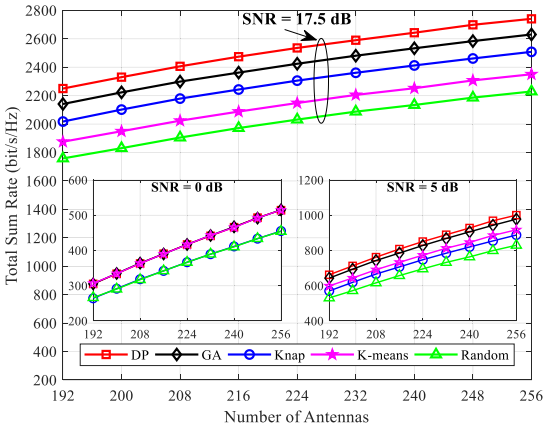


**FIGURE 9.** Total sum rate vs. number of antennas.

Next, in Fig. 9, we consider the total sum rate of the system when the number of UEs is fixed at 512 and the number of antennas varies from 192 to 256. We notice that the total sum rate of the system almost linearly increases as the number of antennas increases. Consistent with the results in the previous simulation, the performance of the DP and GA algorithms is similar in the low SNR regime.

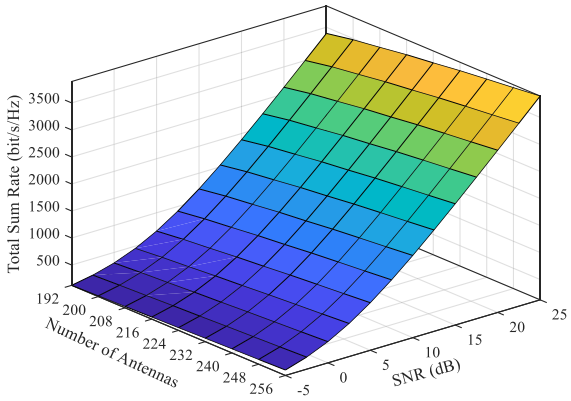


**FIGURE 10.** Total sum rate vs. number of antennas and SNR.

Finally, interested in the case where the best algorithm is DP, in Fig. 10 we plot the total sum rate against the number of antennas and SNR in 3D view. It is clear that the total sum rate

also increases slightly as the number of antennas from 192 to 256 elements.

The proposed BFaaS scheme not only brings efficiency in spectral usage but also creates a service that provides beams on demand.

## V. CONCLUSION

In this paper, we provided our concept and vision of Beamforming-as-a-service scheme for delivering multicast and broadcast Services in 5G systems and beyond. To begin with, we overviewed the background standards and industrial activities through broadcast projects that have been carried over 5G platforms. Focusing on the multicast and broadcast services delivering, we presented, discussed the system model and architectures of standards, industrial projects with pros and cons. The overview is shedding light on the requirements of providing multicast and broadcast services to end users. The end users can watch videos on demand at anytime, anywhere, in any form. This is suitable to the criteria in which everything is sensed, everything is connected, and everything is intelligent in the evolution towards 6G. The results of the investigation clearly demonstrate that the convergence of telecommunication, television and broadcast networks will enable the provision of multimedia services based on the existing 5G platform.

We proposed a Beamforming-as-a-service scheme for multicast and broadcast services in 5G systems and beyond, which allows service providers to optimize the use of beams for each coverage area that meets the requirements regardless of the specific hardware architecture of the network infrastructure. We have shown the benefits achieved by using BFaaS scheme in comparison to other methods.

The value of this work is that BFaaS scheme enables highly efficient network transmission by transmitting multicast/broadcast streams through different beams to different service areas instead of multiple unicasts. Therefore, managing broadcast streams is much easier. This opens up a new business model for Service Providers. However, this scheme addressed challenges to real-world deployment for overlapping service areas. This can be solved by designing a PMI matrix for the overlapping part and allocating appropriate data layers.

In another aspect, we developed algorithms to optimize the beam selection based on greedy methods, dynamic programming, and K-means clustering. To evaluate the system performance, we used a map-based channel model in the simulations to demonstrate the effectiveness of the proposed optimal beam selection algorithms.

In the next research direction, we will address the proposed BFaaS in open RAN architecture, beam management, and resource allocation for high-quality broadcast multimedia transmissions.

## REFERENCES

[1] J. G. Oh, Y. J. Won, J. S. Lee, Y.-H. Kim, J. H. Paik, and J. T. Kim, "A study of development of transmission systems for next-generation terrestrial 4 K UHD and HD convergence broadcasting," *EURASIP J. Wireless Commun. Netw.*, vol. 2015, no. 1, May 2015, Art. no. 128, doi: 10.1186/s13638-015-0362-x.

[2] P. Yu, F. Zhou, X. Zhang, X. Qiu, M. Kadoch, and M. Cheriet, "Deep learning-based resource allocation for 5G broadband TV service," *IEEE Trans. Broadcast.*, vol. 66, no. 4, pp. 800–813, Dec. 2020, doi: 10.1109/TBC.2020.2968730.

[3] J. Kaur and M. A. Khan, "Sixth generation (6G) wireless technology: An overview, vision, challenges and use cases," in *Proc. IEEE Region 10 Symp. (TENSYMP)*, Jul. 2022, pp. 1–6, doi: 10.1109/TENSYMP54529.2022.9864388.

[4] D. Mi, J. Eyles, T. Jokela, S. Petersen, R. Odarchenko, E. Öztürk, D.-K. Chau, T. Tran, R. Turnbull, H. Kokkinen, B. Altman, M. Bot, D. Ratkaj, O. Renner, D. Gomez-Barquero, and J. J. Gimenez, "Demonstrating immersive media delivery on 5G broadcast and multicast testing networks," *IEEE Trans. Broadcast.*, vol. 66, no. 2, pp. 555–570, Jun. 2020, doi: 10.1109/TBC.2020.2977546.

[5] G. George, S. Roy, S. Raghunandan, C. Rohde, and T. Heyn, "5G new radio in nonlinear satellite downlink: A physical layer comparison with DVB-S2X," in *Proc. IEEE 4th 5G World Forum (5GWF)*, Oct. 2021, pp. 499–504, doi: 10.1109/5GWF52925.2021.00094.

[6] J. J. Gimenez, J. L. Carcel, M. Fuentes, E. Garro, S. Elliott, D. Vargas, C. Menzel, and D. Gomez-Barquero, "5G new radio for terrestrial broadcast: A forward-looking approach for NR-MBMS," *IEEE Trans. Broadcast.*, vol. 65, no. 2, pp. 356–368, Jun. 2019, doi: 10.1109/TBC.2019.2912117.

[7] M. Simon, E. Kofi, L. Libin, and M. Aitken, "ATSC 3.0 broadcast 5G unicast heterogeneous network converged services starting release 16," *IEEE Trans. Broadcast.*, vol. 66, no. 2, pp. 449–458, Jun. 2020, doi: 10.1109/TBC.2020.2985575.

[8] E. Garro, M. Fuentes, J. L. Carcel, H. Chen, D. Mi, F. Tesema, J. J. Gimenez, and D. Gomez-Barquero, "5G mixed mode: NR multicast-broadcast services," *IEEE Trans. Broadcast.*, vol. 66, no. 2, pp. 390–403, Jun. 2020, doi: 10.1109/TBC.2020.2977538.

[9] M. Säily, C. B. Estevan, J. J. Gimenez, F. Tesema, W. Guo, D. Gomez-Barquero, and D. Mi, "5G radio access network architecture for terrestrial broadcast services," *IEEE Trans. Broadcast.*, vol. 66, no. 2, pp. 404–415, Jun. 2020, doi: 10.1109/TBC.2020.2985906.

[10] M. Fallgren, T. Abbas, S. Allio, J. Alonso-Zarate, G. Fodor, L. Gallo, A. Kousaridas, Y. Li, Z. Li, Z. Li, J. Luo, T. Mahmoodi, T. Svensson, and G. Vivier, "Multicast and broadcast enablers for high-performing cellular V2X systems," *IEEE Trans. Broadcast.*, vol. 65, no. 2, pp. 454–463, Jun. 2019, doi: 10.1109/TBC.2019.2912619.

[11] L. Zhang, W. Li, Y. Wu, Y. Xue, E. Sousa, S.-I. Park, J.-Y. Lee, N. Hur, and H.-M. Kim, "Using non-orthogonal multiplexing in 5G-MBMS to achieve broadband-broadcast convergence with high spectral efficiency," *IEEE Trans. Broadcast.*, vol. 66, no. 2, pp. 490–502, Jun. 2020, doi: 10.1109/TBC.2020.2983563.

[12] T. Hong, T. Tang, X. Dong, R. Liu, and W. Zhao, "Future 5G mmWave TV service with fast list decoding of polar codes," *IEEE Trans. Broadcast.*, vol. 66, no. 2, pp. 525–533, Jun. 2020, doi: 10.1109/TBC.2020.2977561.

[13] E. Ahvar, S. Ahvar, S. M. Raza, J. M. S. Vilchez, and G. M. Lee, "Next generation of SDN in cloud-fog for 5G and beyond-enabled applications: Opportunities and challenges," *Network*, vol. 1, no. 1, pp. 28–49, Jun. 2021, doi: 10.3390/network1010004.

[14] L. Qing, "A 5G PaaS collaborative management and control platform technology based on cloud edge collaboration based on particle swarm optimization algorithm," in *Proc. IEEE Asia–Pacific Conf. Image Process., Electron. Comput. (IPEC)*, Apr. 2021, pp. 144–147, doi: 10.1109/IPEC51340.2021.9421164.

[15] R. Bruschi, F. Davoli, F. D. Bravo, C. Lombardo, S. Mangialardi, and J. F. Pajo, "Validation of IaaS-based technologies for 5G-ready applications deployment," in *Proc. Eur. Conf. Netw. Commun. (EuCNC)*, Jun. 2020, pp. 46–51, doi: 10.1109/EuCNC48522.2020.9200926.

[16] T. Taleb, A. Ksentini, and R. Jantti, "'Anything as a Service' for 5G mobile systems," *IEEE Netw.*, vol. 30, no. 6, pp. 84–91, Nov. 2016, doi: 10.1109/MNET.2016.1500244RP.

[17] Z. Chang, Z. Zhou, S. Zhou, T. Chen, and T. Ristaniemi, "Towards service-oriented 5G: Virtualizing the networks for everything-as-a-service," *IEEE Access*, vol. 6, pp. 1480–1489, 2018, doi: 10.1109/ACCESS.2017.2779170.

[18] D. Sabella, A. de Domenico, E. Katranaras, M. A. Imran, M. di Girolamo, U. Salim, M. Lalam, K. Samdanis, and A. Maeder, "Energy efficiency benefits of RAN-as-a-service concept for a cloud-based 5G mobile network infrastructure," *IEEE Access*, vol. 2, pp. 1586–1597, 2014, doi: 10.1109/ACCESS.2014.2381215.

[19] M. Kist, J. F. Santos, D. Collins, J. Rochol, L. A. DaSilva, and C. B. Both, "AIRTIME: End-to-end virtualization layer for RAN-as-a-service in future multi-service mobile networks," *IEEE Trans. Mobile Comput.*, vol. 21, no. 8, pp. 2701–2717, Aug. 2022, doi: 10.1109/TMC.2020.3046535.

[20] A. A. Barakabitze, A. Ahmad, R. Mijumbi, and A. Hines, "5G network slicing using SDN and NFV: A survey of taxonomy, architectures and future challenges," *Comput. Netw.*, vol. 167, Feb. 2020, Art. no. 106984, doi: 10.1016/j.comnet.2019.106984.

[21] V. Sciancalepore, F. Cirillo, and X. Costa-Perez, "Slice as a service (SlaaS) optimal IoT slice resources orchestration," in *Proc. IEEE Global Commun. Conf.*, Dec. 2017, pp. 1–7, doi: 10.1109/GLOCOM.2017.8254529.

[22] J. F. Santos, M. Kist, J. Rochol, and L. A. DaSilva, "Virtual radios, real services: Enabling RANaaS through radio virtualisation," *IEEE Trans. Netw. Service Manag.*, vol. 17, no. 4, pp. 2610–2619, Dec. 2020, doi: 10.1109/TNSM.2020.3009863.

[23] J. Dobruna, E. Spahiu, M. Pogacnik, and M. Volk, "5G streaming: IP-based vs. high-power high-tower broadcast," in *Proc. 45th Jubilee Int. Conv. Inf., Commun. Electron. Technol. (MIPRO)*, May 2022, pp. 1507–1511, doi: 10.23919/MIPRO55190.2022.9803623.

[24] F. A. Pereira de Figueiredo, "An overview of massive MIMO for 5G and 6G," *IEEE Latin Amer. Trans.*, vol. 20, no. 6, pp. 931–940, Jun. 2022, doi: 10.1109/TLA.2022.9757375.

[25] F. Hartung, U. Horn, J. Huschke, M. Kampmann, and T. Lohmar, "MBMS—IP multicast/broadcast in 3G networks," *Int. J. Digit. Multimedia Broadcast.*, vol. 2009, pp. 1–25, Apr. 2009, doi: 10.1155/2009/597848.

[26] E. Öztürk, W. Zia, V. Pauli, and E. Steinbach, "Performance evaluation of ATSC 3.0 DASH over LTE eMBMS," in *Proc. IEEE Int. Symp. Broadband Multimedia Syst. Broadcast. (BMSB)*, Jun. 2018, pp. 1–6, doi: 10.1109/BMSB.2018.8436708.

[27] Z. Xia, B. Xu, X. Meng, and Y. Zhang, "Comparative study on KPIs between FeMBMS and DTMB," in *Proc. Int. Wireless Commun. Mobile Comput. (IWCMC)*, Jun. 2020, pp. 520–524, doi: 10.1109/IWCMC48107.2020.9148226.

[28] *Digital Video Broadcasting (DVB), Next Generation Broadcasting System to Handheld, Physical Layer Specification (DVB-NGH)*, document A160, DVB, 2012.

[29] H. T. Nguyen, T. T. Le, and T. H. Nguyen, "Capacity improvement for DVB-NGH with dual-polarized MIMO spatial multiplexing and hybrid beamforming," *Int. J. Digit. Multimedia Broadcast.*, vol. 2020, pp. 1–11, Aug. 2020, doi: 10.1155/2020/9578521.

[30] *Digital Video Broadcasting (DVB); Service Discovery and Programme Metadata for DVB-I*, document TS 103.770, Version 1.1.1, Nov. 2020.

[31] M. Iordache, O. Badita, B. Rusti, A. Bonea, G. Suciu, E. Giannopoulou, G. Landi, and N. Slamnik-Krijestorac, "Future 5G network implementation and open testbeds deployment for real 5G experiments," in *Proc. IEEE Future Netw. World Forum (FNWF)*, Oct. 2022, pp. 355–360, doi: 10.1109/FNWF55208.2022.00069.

[32] 5G RuralFirst. (2019). *Project Conclusion Report*. [Online]. Available: https://www.5gruralfirst.org/project-conclusion-report/

[33] 5G PPP. (2019). *5G-Xcast Broadcast and Multicast Communication Enablers for the Fifth Generation of Wireless System*. [Online]. Available: http://5g-xcast.eu/wp-content/uploads/2018/09/5G_Xcast_Brochure_Designer_version.pdf

[34] S. Ilsen, F. Juretzek, L. Richter, D. Rother, and P. Brétillon, "Tower overlay over LTE-advanced+(TOoL+): Results of a field trial in Paris," in *Proc. IEEE Int. Symp. Broadband Multimedia Syst. Broadcast. (BMSB)*, Jun. 2016, pp. 1–6, doi: 10.1109/BMSB.2016.7521952.

[35] M. Beccaria, A. Massaccesi, P. Pirinoli, N. K. Kiem, N. H. Trung, and L. H. Manh, "Innovative MIMO antennas for 5G communication systems," in *Proc. IEEE Int. Conf. Environ. Electr. Eng. IEEE Ind. Commercial Power Syst. Eur.*, Jun. 2018, pp. 1–4, doi: 10.1109/EEEIC.2018.8493747.

[36] N. H. Trung, "Multiplexing techniques for applications based-on 5G system," in *Multiplexing—Recent Advances and Novel Applications*, S. Mohammady, Ed. London, U.K.: IntechOpen, 2022, ch. 6, pp. 101–125, doi: 10.5772/intechopen.101780.

[37] Z. Dong and Y. Zeng, "Near-field spatial correlation for extremely large-scale array communications," *IEEE Commun. Lett.*, vol. 26, no. 7, pp. 1534–1538, Jul. 2022, doi: 10.1109/LCOMM.2022.3170735.

[38] V. Nurmela, "METIS channel models," METIS, Tech. Rep., ICT-317669-METIS/D1.4, Jul. 2015.

[39] *5G; Study on Channel Model for Frequencies From 0.5 to 100 GHz*, document TR 138.901, Version 16.1.0, Release 16, 3GPP, 2020.

[40] T.-T. Le, T.-H. Nguyen, and H.-T. Nguyen, "User grouping for massive MIMO terrestrial broadcasting networks," in *Proc. IEEE 8th Int. Conf. Commun. Electron. (ICCE)*, Jan. 2021, pp. 467–471, doi: 10.1109/ICCE48956.2021.9352134.

[41] N. H. Trung, N. T. Anh, M. Duc, D. T. Binh, and L. T. Tan, "System theory based multiple beamforming," *Vietnam J. Sci. Technol.*, vol. 55, no. 5, pp. 653–665, Oct. 2017, doi: 10.15625/2525-2518/55/5/9149.

[42] R. Tian, Y. Liang, X. Tan, and T. Li, "Overlapping user grouping in IoT oriented massive MIMO systems," *IEEE Access*, vol. 5, pp. 14177–14186, 2017, doi: 10.1109/ACCESS.2017.2729878.

**NGUYEN HUU TRUNG** received the Dipl.Eng., M.E., and Ph.D. degrees from the Department of Communications Engineering, School of Electrical and Electronic Engineering, Hanoi University of Science and Technology, in 1996, 1998, and 2004, respectively. Since 1998, he has been a Lecturer with the School of Electrical and Electronic Engineering, Hanoi University of Science and Technology. He is currently tenured as an Associate Professor, a Senior Lecturer, and the Head of the Aerospace Electronic Laboratory. His research interests include advanced modulation techniques based on wavelet theory, GNSS, SDR, massive MIMO, MC-CDMA, 5G/6G broadcasts, satellites, and UAV.

**NGUYEN THUY ANH** received the Dipl.Eng., M.E., and Ph.D. degrees from the Department of Circuits and Signal Processing, School of Electrical and Electronic Engineering, Hanoi University of Science and Technology, in 1996, 1998, and 2008, respectively. Since 1999, she has been a Lecturer with the School of Electrical and Electronic Engineering, Hanoi University of Science and Technology. She is currently an Associate Professor with the School of Electrical and Electronic Engineering. Her research interests include digital signal processing for communication systems, information theory, massive MIMO, multicarrier communication systems, and state optimization control theory.