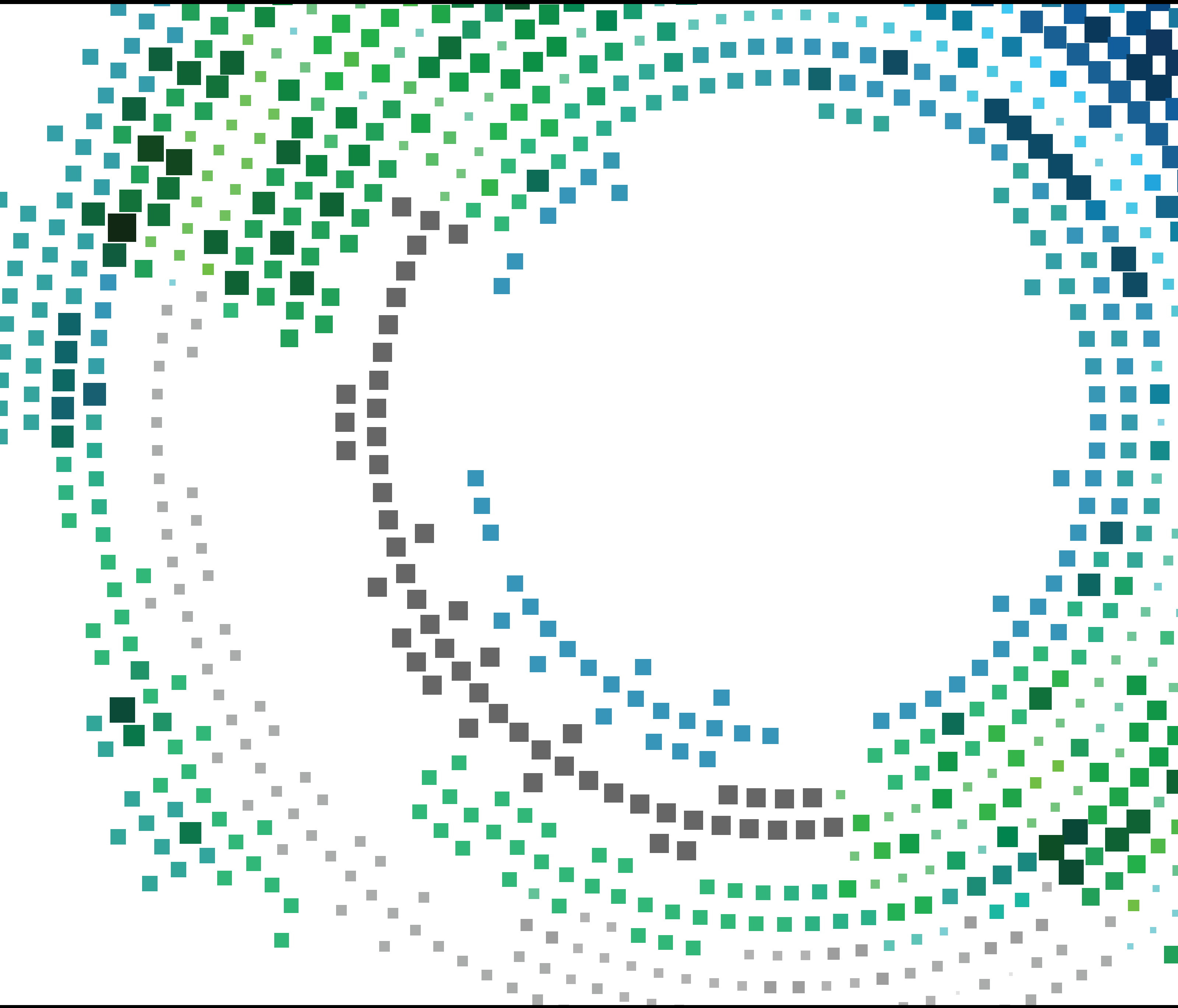
Mobile Information Systems

End-to-End Automation of 5G Networks

Lead Guest Editor: Tara A. Yahiya
Guest Editors: André-Luc Beylot and Pinar Kirci



End-to-End Automation of 5G Networks

Mobile Information Systems

End-to-End Automation of 5G Networks

Lead Guest Editor: Tara A. Yahiya
Guest Editors: André-Luc Beylot and Pinar Kirci

Editorial Board

Contents

Hindawi
Mobile Information Systems
Volume 2019, Article ID 6425419, 2 pages
https://doi.org/10.1155/2019/6425419

Editorial

End-to-End Automation of 5G Networks

Tara Ali Yahiya,[1] **Pinar Kirci**,[2] **and André-Luc Beylot**[3]

[1] *Department of Computer Science and Engineering, University of Kurdistan-Hewler, Erbil, KRG, Iraq*
[2] *Department of Engineering Sciences, Faculty of Engineering, Istanbul University-Cerrahpasa, Istanbul, Turkey*
[3] *Department of Telecommunications and Networks, ENSEEIHT, University of Toulouse, Toulouse, France*

Correspondence should be addressed to Tara Ali Yahiya; t.ibrahim1@ukh.edu.krd

Received 28 April 2019; Accepted 28 April 2019; Published 17 June 2019

The fifth-generation (5G) network will be the next big revolution in the area of mobile and wireless communication. This is not due only to the high data rate that it is offering but also due to its openness to different types of applications like the ultralow latency applications, Internet of things (IoT), and massive device-to-device communications. These kinds of applications cannot be supported without the big changes witnessed by the core network of the 5G through the virtualization and softwarization techniques that revolutionize not only the core but also the radio network, bringing higher performance, low latency, and quality of service guarantee to these types of applications.

The revolution in 5G networks is represented by their automation and adaptability to fulfill different requirements of users. To do so, the service and the network infrastructure are configured through techniques like software-defined networking (SDN) and network function virtualization (NFV) to enable the dynamicity of both radio and core networks instead of having a fixed and static architecture. However, this automation cannot be achieved only through the use of tools but can also be achieved by using intelligent mechanisms and predictive analytics in order to make them responsive to different customers' requirements.

Providing techniques of automation in both radio and core networks will provide reliability through collaborative intelligent resource allocation management. Thus, this will result in providing agile services and new on-demand services in order to meet the requirements of the users as well as their quality of service (QoS), especially for real-time services.

The article "Mobile Edge Assisted Live Streaming System for Omnidirectional Video" by X. Hu et al. proposes a mobile edge assisted live streaming system for omnidirectional video (MELiveOV); the MELiveOV can intelligently offload the processing tasks to the edge computing-enabled 5G base stations. The MELiveOV consists of an omnidirectional video generation module, a streaming module, and a viewpoint prediction module. A prototype system of MELiveOV is implemented to prove its complete end-to-end OV live streaming service. The evaluation result demonstrates that compared with the traditional solution, the MELiveOV can reduce the network bandwidth requirement by about 50% and the transmission delay by more than 70% while ensuring the quality of the user's experience.

The article "An End-to-End Automation Framework for Mobile Network Testbeds" by A. D. Zayas et al. describes the end-to-end automation framework developed as part of the TRIANGLE project. The TRIANGLE project is devoted to the benchmarking of apps and devices in mobile networks. For that purpose, it is needed to ensure the repeatability in the behaviour of all the components of the mobile network during the execution of the same test. This is why one of the main objectives of the project was to develop an end-to-end automation framework to provide repeatable testing. The paper describes in detail the design and the implementation of the framework.

The article "NSAF: An Approach for Ensuring Application-Aware Routing Based on Network QoS of Applications in SDN" by J. Park et al. proposes the network situation-aware framework (NSAF) to more efficiently handle application routing based on the QoS requirements and changing network status. A mechanism for supporting the NSAF consisting of application registration, network status monitoring, violation detection, and routing control is also presented. The applicability and feasibility of the proposed NSAF are verified by implementing a prototype. The NSAF may be used as a reference model for efficiently managing SDN-based networks to ensure application QoS.

The article "Expanding GÉANT Testbeds Service to Support Pan-European 5G Network Slices for Research in the EuWireless Project" by Á. Rios et al. presents the design options for creating a Pan-European mobile network for research in the context of the European Horizon 2020 EuWireless project. The most likely direction is a platform that makes it easier to create network slices for research. In this context, we identify one promising technology to implement network slicing in 5G networks: the framework GÉANT Testbeds Service (GTS). GTS is currently a production service by GÉANT that offers remote construction and use of virtual testbeds for wired networks mapped to the real GÉANT infrastructure. These GTS virtualized testbed environments conform to software-defined network (SDN) principles and offer computing, storing, and switching resources, at scale and with line rate performance. In this paper, we explain how the current (wire oriented) GTS can be extended with the 5G components, such as radio access nodes (gNBs), transport networks, and user devices, in order to implement 5G network slices. Our first conclusion is that using GTS for EuWireless implementation is feasible, dramatically increasing the potential impact of this service in the research community.

In the article "Independence and Fairness Analysis of 5G mmWave Operators Utilizing Spectrum Sharing Approach" by M. L. Attiah et al., an analytical framework involving a flexible hybrid mmWave SSA is presented to assess the effectiveness of SSA and investigate its influence on network functionality in terms of independence and fairness among operators. Two mmWave frequencies (28 GHz and 73 GHz) are used with different spectrum bandwidths. Various access models have been presented for adoption by four independent mobile network operators that incorporate three types of spectrum allocation (exclusive, semipooled, and fully pooled access). Furthermore, an adaptive multistate mmWave cell selection scheme is proposed to associate typical users with the tagged mmWave base stations that provide a great signal-to-interference-plus-noise ratio, thereby maintaining reliable connections and enriching user experience. Numerical results show that the proposed strategy achieves considerable improvement in terms of fairness and independence among operators, which paves the way for further research activities that would provide better insight and encourage mobile network operators to rely on SSA.

Conflicts of Interest

The editors declare that there are no conflicts of interest regarding the publication of this special issue.

Acknowledgments

The editors would like to thank the authors for their contributions and the reviewers of the papers for their help in bringing this issue to its current form.

Tara Ali Yahiya
Pinar Kirci
André-Luc Beylot

Hindawi
Mobile Information Systems
Volume 2019, Article ID 8487372, 15 pages
https://doi.org/10.1155/2019/8487372

Research Article

Mobile Edge Assisted Live Streaming System for Omnidirectional Video

Xinjue Hu,[1,2] Wei Quan,[1,2] Tao Guo,[1,2] Yu Liu,[1,2] and Lin Zhang[1,2]

[1] *School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing 100876, China*
[2] *Research Center of Networks and Communications, Peng Cheng Laboratory, Shenzhen 518055, China*

Correspondence should be addressed to Lin Zhang; zhanglin@bupt.edu.cn

Received 28 January 2019; Accepted 14 April 2019; Published 21 May 2019

Guest Editor: Tara A. Yahiya

As a popular form of virtual reality (VR) media, omnidirectional video (OV) has been continuously developed in recent years. OV contains the view of the scene in every direction, which will ask for around 120 Mbps with 8k resolution and 25 fps (frames per second). Although there has been a lot of work to optimize the transmission for on-demand of OV, the research on the live streaming of OV is still very lacking. Another big challenge for the OV live streaming system is the huge demand for computing resources. The existing terminal devices are difficult to completely carry tasks such as stitching, encoding, and rendering. This paper proposes a mobile edge assisted live streaming system for omnidirectional video (MELiveOV); the MELiveOV can intelligently offload the processing tasks to the edge computing enabled 5G base stations. The MELiveOV consists of an omnidirectional video generation module, a streaming module, and a viewpoint prediction module. A prototype system of MELiveOV is implemented to prove its complete end-to-end OV live streaming service. Evaluation result demonstrates that compared with the traditional solution, MELiveOV can reduce the network bandwidth requirement by about 50% and the transmission delay of more than 70% while ensuring the quality of the user's experience.

1. Introduction

According to the report [1] commissioned by Intel and conducted by Ovum, VR and AR applications will account for 90 percent of 5G data use over the next decade. Omnidirectional video (OV) is one of the most mature forms of VR, and it is expected to become the killer application for the future of 5G networks [2]. Driven by more powerful network performance, the 5G-powered OV not only focuses on realistic visual effects but also emphasizes the user's interactive experience. The basis of the interactive OV application is the ability to implement a complete end-to-end live streaming service system, which is the core problem that this work wants to solve. The popularity of OV technology brings viewers a novel immersive multimedia experience, but this new experience is supported by video contents with very high resolution (usually 4k or 8k) multiplied by 360-degree panoramic viewpoint. The transmission of OV usually consumes 4~6x the bandwidth of a regular video with the same viewable resolution, which means a huge challenge to traditional video streaming architecture. The more amount of data and more complex computing tasks are the two major challenges that the OV live streaming system needs to address.

On the one hand, OV uses head mount displays (HMDs) with stereoscopic capabilities to provide the immersive experience. When the omnidirectional content is viewed by users, only a subset of the entire video frame will be displayed on the HMD's screen. To reduce the waste of network bandwidth caused by the redundancy of OV data, various improved solutions have been proposed both in academic and industrial communities. Some research works [3–5] have designed the tile-based coding scheme, which can effectively optimize the OV transmission. Many relevant Standards Development Organizations (SDOs) also have started work on the scope of OV [6]. But most of these works are carried out around the application for on-demand of

OV. Now more and more users are paying attention to the live streaming experience, which is also the development trend of digital multimedia technology in the future. Therefore, there is an urgent need for a new feasible solution that can minimize the bandwidth requirement of OV while simultaneously maximizing the user's experience.

On the other hand, in addition to the high bandwidth consumption during transmission, the huge demand for computing resources is another big challenge for the design of the OV live streaming system. The acquisition and generation of OV content require extensive stitching and encoding work. Especially when performing OV live streaming service, these computing works need to be completed in real time, which puts extremely high demands on the performance of the processing platform. When the OV streaming is viewed, the system can customize the process of rendering according to different fields of view (FOV) for multiple users. Accurate prediction of the user's viewpoint can bring great benefits to the optimization of the OV live streaming system. By predicting the FOV areas that users may be watching in the near future, the transfer of data in those useless areas can be avoided. And the OV live streaming system is able to use limited bandwidth to maximize the image quality of the FOV area. Viewpoint prediction relies on deep learning neural network algorithm, which also has a high demand for computational power. Offloading computationally intensive tasks to resourceful cloud/fog servers is necessary to reduce the pressure of the users' devices while saving the cost of the OV devices. Compared with the traditional central cloud server, the mobile edge computing (MEC) architecture can bring computing resources closer to users, thus greatly reducing the response delay of users requesting services.

A mobile edge assisted live streaming system for omnidirectional video (MELiveOV) is presented in this work to address the above challenges. As shown in Figure 1, it consists of the omnidirectional video generation module, the streaming module, and the viewpoint prediction module. Through the collaborative work of each module, MELiveOV achieves the complete end-to-end live streaming service of the omnidirectional video. Meanwhile, its edge computing architecture closely matches the needs of the 5G network and has a very broad application prospect. We implement the prototype system of MELiveOV and evaluate various performance metrics for it. The evaluation result shows that MELiveOV can effectively reduce the network bandwidth requirement and the transmission delay during the OV live streaming.

The contributions of this paper are summarized as follows:

(i) We build an end-to-end mobile edge assisted live streaming system for omnidirectional video (MELiveOV). With the help of the MEC architecture, MELiveOV is able to perform well in both service latency and bandwidth requirements.

(ii) In order to speed up the real-time generation of the omnidirectional video after the acquisition, we design an improved stitching algorithm based on the overall mapping table.

(iii) A tile-based omnidirectional video transmission scheme is introduced to MELiveOV to reduce the pressure on network bandwidth during OV live streaming.

(iv) In order to enhance the user's quality of experience and reduce the service delay, we design a user's viewpoint prediction algorithm, which enables MELiveOV to provide proactive service for users.

The rest of this paper is organized as follows. Section 2 discusses related work. Section 3 introduces the system architecture of MELiveOV. Section 4 presents the design of omnidirectional video generation module based on the overall mapping stitching table. Section 5 presents the structure of tile-based streaming module. Section 6 introduces the architecture of the viewpoint prediction module using deep learning. Section 7 describes our implementation and evaluation. Section 8 concludes the paper and discusses the future work.

2. Related Works

Live streaming of events has been traditionally done by using broadcast TVs. DASH can also be applied to live streaming over the Internet [7, 8], despite the tighter latency constraints compared to on-demand video services. The challenge of live streaming is to minimize the end-to-end delay between the content generation (at the server) and presentation (at the client). The main research in video streaming areas focuses on optimization of different aspects like higher resolution (e.g., omnidirectional video and virtual reality streaming), lower latency, higher compression ratio, and better quality of experience (QoE). Many focus on adaptive streaming to fit as many network situations as possible and make full usage of possible bandwidth. In [9], subjective studies that cover QoE aspects of adaptation dimensions and strategies are revisited. As a result, QoE influence factors of HAS and corresponding QoE models are identified, and open issues and conflicting results are also discussed. The tiled video source is a sounding way to obtain adaptive streaming [10]; they describe how spatial access can be performed in an adaptive HTTP streaming context, using MPEG-DASH and its SRD extensions. They describe a configurable implementation of these technologies, within the GPAC open-source player, allowing experimentations of different adaptation policies for tiled video content. New scenarios enabled by the development of technology like virtual reality (VR) have attracted great attention. Ozcinar et al. [5] proposed an end-to-end streaming system implementation that contains tiling, a novel extension of the MPD, and DASH bitrate level selection in a viewport-aware manner, which can bring significant quality enhancements compared with the traditional streaming approach. In [11], a novel wireless video transmission method is developed, where the authors jointly investigate how to conquer the problem of source video's huge size, how to efficiently satisfy a user's view switch request, and how to handle packet loss. In [12], they develop a prototype of VR live architecture that combines RTP and DASH to deliver 360° VR content to a

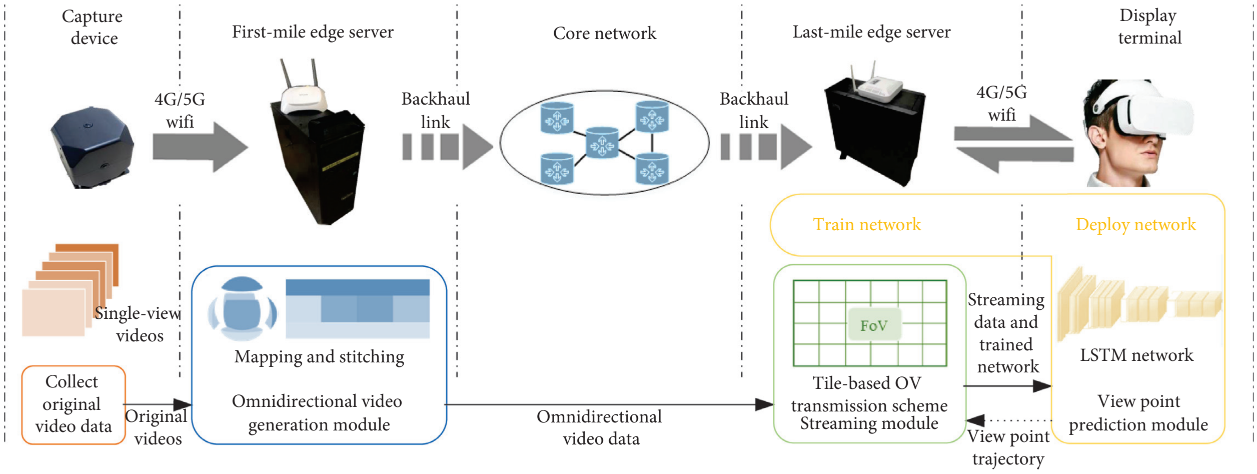


Figure 1: Overview of mobile edge assisted live streaming system for omnidirectional video (MELiveOV).

Huawei set-top-box and a Samsung Galaxy S7. The system multiplexes a single HEVC hardware decoder to provide faster quality switching than at the traditional group of pictures (GOP) boundaries.

As for the QoE aspect, in [13], it is believed that cellular operators and content providers can tremendously improve video QoE by predicting available bandwidth and sharing it through APIs. To be more specific, when combined with rate stabilization functions, prediction outperforms existing video streaming algorithms and reduces the gap with optimal to 4%. Besides, in [14], a layered framework for migrating active service applications that are encapsulated either in virtual machines (VMs) or containers is presented. This layering approach allows a substantial reduction in service downtime. The framework is easy to implement using readily available technologies, and one of its key advantages is that it supports containers, which is a promising emerging technology that offers benefits over VMs. Reducing delay is an attracting field as well. Machen et al. [15] presented a layered framework for migrating active service to MEC. This layering approach allows a substantial reduction in service downtime [16]. ENA develops a novel transmission scheduling framework dubbed AdaPtive HFR vIdeo Streaming (APHIS). It is proved by intensive experiments that APHIS framework is able to appropriately filter video frames and adjust data protection levels to optimize the quality of HFR video streaming. Sanchez et al. [17] presented a video coding and slicing scheme for OV streaming. In a delay-constrained circumstance, their scheme significantly reduces the transmission cost and enhances the quality of the reconstructed video sequences compared with the nonadaptive transmission scheme.

Omnidirectional video (OV) enables direct surround immersive viewing of a scene by warping the original image into the correct perspective given a viewing direction. A live streaming system for OV has been achieved in [18]. They design periodic and adaptive optimization frameworks to adapt to the bandwidth variations and FoV prediction errors in real time. OV can offer immersive visual experience when a user equipped with HMD, but transmit OV with high bitrate will bring a heavy burden to transmission system especially in a real-time scenario. So, how to compress a video without affecting the user experience is very important. Chen et al. [19] reviewed the recent advances in the pipeline of omnidirectional video processing including projection and evaluation. An efficient way was achieved to facilitate motion-constrained HEVC tiles. Sreedhar et al. and Skupin et al. [20, 21] investigated various viewpoint dependent projection schemes, and they developed a methodology for comparing the rate-distortion performance of these projections. Yu et al. and Lee et al. [22, 23] considered the problem of evaluating the coding efficiency in the context of viewing with a HMD. They compared the original and coded videos on the viewpoint after sphere-to-plane mappings. It is observed that the equal-area mapping yields around 8.3% bitrate savings relative to the commonly used equirectangular mapping. Ghaznavi-Youvalari et al. and Curcio et al. [24, 25] adopted subjective assessment results of experiments using a tile-based streaming system for OV. This work reduces streaming bitrates by an average of 44% under a subjective DMOS value of 4.5. Yu et al. [26] showed a computationally efficient solution using Lagrangian optimization by separating the sampling and bit allocation constraints and got coding gains over standard representations. Graf et al. [27] described the usage of tiles in HEVC/H.265 and VP9, enabling bandwidth efficient adaptive streaming of omnidirectional video over HTTP. Various streaming strategies have been defined, which can effectively improve the quality of OV streaming service. Li et al. [28] proposed a tile-based omnidirectional video segmentation scheme which can save up to 28% of the pixel area and 20% of BD-rate averagely compared to the traditional equirectangular projection-based approach. Gudumasu et al. [29] showed a viewing orientation tracking and real-time viewpoint extraction platform. Generally, the user can only view a restricted field of view of the content. This means that a significant part of the bandwidth is wasted by transmitting quality video in regions that are not being visualized. So, there appears tile-based transmit method along with user's viewpoint prediction. Ozcinar and Smolic [30] created a new visual attention user dataset for OV, investigated the behavior of viewers when consuming the content, and analyzed the prediction performance of state-of-the-art visual attention models. Ninan and Atluru [31] generated a second reconstructed image with view direction of the viewer when the user watched the first reconstructed image. Ghaznavi-Youvalari and Aminlou [32] proposed a geometry-based motion vector scaling method in order to compress the motion information of omnidirectional content efficiently. The result shows a 2.2% bitrate reduction with Versatile Video Coding (H.266/VVC) standard. Ghaznavi-Youvalari and Aminlou [33] divided the image into tiles and set different priorities with FOV information. The high-priority tiles will encode with a high bitrate. To measure the objective quality of omnidirectional video in observation space more accurately, a weighted-to-spherically-uniform quality evaluation method has been proposed in [34].

Many Standards Development Organizations in the field of multimedia and communications have also begun working on OV [6]. In R15 [35], 3GPP has started to consider the application of virtual reality (VR) media services in the next generation of mobile network. The Digital Video Broadcasting (DVB) Project established a VR-related commercial module to follow up on this area [36]. The Video Coding Experts Group (VCEG, ITU-T Q6/16) and the Moving Picture Experts Group (MPEG, ISO/IEC JTC 1/SC 29/WG 11) begun the standardization process for OV, respectively, which started from the research of OV coding and transmission technology, and are expected to guide the development of the entire OV application ecosystem. There have been some joint groups to carry out some works in the field of OV, for example, the Joint Collaborative Team on Video Coding (JCT-VC), responsible for developing the High-Efficiency Video Coding (HEVC) standard and its extensions [37] and the Joint Video Exploration Team (JVET) that investigates new video coding approaches for coding efficiency beyond HEVC [38]. The coding of the omnidirectional video has attracted enough attention and has gradually become the focus of multimedia technology development. Compared with the VCEG, MPEG concentrated more on the technologies of delivery and display. MPEG established the subgroup of Omnidirectional MediA Format (OMAF) [2], which is envisioned to become Part 2 of the emerging ISO/IEC 23090 MPEG-I.

3. System Architecture of MELiveOV

As an end-to-end service system, MELiveOV covers the entire service chain from acquisition to playback. As shown in Figure 1, we have designed the corresponding modified functional module at every stage of the OV live streaming service. In the conventional scheme, after the raw data are collected by the omnidirectional camera, the OV is generated directly locally. Constrained by the limited computing capability of the capture device, the omnidirectional video generation process will be extremely time consuming, which greatly affects the real-time performance of the live streaming service. In MELiveOV, we offload the computational tasks required for the omnidirectional video generation process to the first-mile edge server. It is usually deployed at the access point of the first hop in the mobile communication network to provide the most timely service to the capture device. The mapping and stitching operations are then performed by the omnidirectional video generation module to obtain the OV data containing omnidirectional scene information.

Similarly, we deployed the last-mile edge server at the access point of the 5G network closest to the viewer. Before the last hop, the streaming module on the last-mile edge server will optimize the transmission in real-time based on the viewer's viewpoint trajectory fed back by the display terminal, which can effectively reduce the bandwidth requirement during the OV download process. The streaming module mainly applies a tile-based OV transmission scheme. By dividing the complete video into multiple tiles by spatial region, we can control the quality of different tiles to optimize transmission. We use a high bitrate for the tiles in the user's field of view (FOV) and a lower bitrate for the area outside the user's FOV. Through the interaction with the display terminal, the streaming module can minimize the transmission bandwidth requirement of the OV without sacrificing the quality of the user viewing area.

This can save the power consumption of VR devices and the costs of traffic for the users. On the display terminal of MELiveOV, we introduced the architecture of the proactive service. In the traditional reactive service architecture, the server can only respond and process after waiting for the user's request to arrive. For example, when our system is without the proactive service architecture, the optimization process performed by the streaming module can only rely on the user's past viewpoint data. This lag in user information can result in reduced performance of the streaming module. Therefore, we deploy the viewpoint prediction module to proactively predict the user's possible viewpoint location in the future. This can further improve the QoE (quality of experience) of MELiveOV's users. The viewpoint prediction module is designed based on the LSTM (long short-term memory) network. The LSTM is a model often used in deep learning to process time series predictive data problems. Our prediction model can not only learn the information of the user's personalized viewing habits but also can perceive the statistical distribution of video saliency only through the multiuser viewpoint data.

4. Omnidirectional Video Generation Module

The omnidirectional camera is generally composed of multiple cameras so that the image data of the scene can be collected from various directions. The most representative camera is a 6-lens omnidirectional camera that can capture up, down, left, right, front, and back six channels of video. These raw data need to be mapped and stitched to generate OV content. There are a variety of omnidirectional image unfolding methods, such as equirectangular projection (ERP, the upper part of Figure 2), which is the most familiar rendering method for the average user, achieving unfolding by transforming the spherical image into rectangular space according to longitude and latitude; cube maps (the lower part of Figure 2) transform the sphere into cubes and then expand the six faces of the cube; Equi-Angular Cubemap (EAC) is an optimization of the traditional cube expansion, correcting the deformation of the cube expansion by keeping the pixels evenly sampled.

Usually, the omnidirectional camera needs to be processed offline for several hours after the acquisition to finally generate the omnidirectional video. This is obviously unacceptable for OV's live streaming service system, so we need to develop a dedicated fast real-time stitching algorithm for MELiveOV. Next, we will introduce the functional design of the omnidirectional video generation module in detail.

4.1. Overview Structure of the Module. Traditional omnidirectional image stitching requires dynamic estimation of the input of each camera at each moment. Firstly, feature point matching is needed to estimate the intrinsic parameters and extrinsic parameters of the camera, and then the overall white balance is performed on each image to facilitate deriving the best stitching mask between images. Finally, with best stitching masks between every two pictures found, all original pictures can be merged into the same coordinate system to form the omnidirectional frame. Apparently, the stitching procedure is considerably time consuming. However, real-time processing of omnidirectional images requires both high resolution and image quality and low latency simultaneously. Due to the computational power and the high complexity of the algorithm itself, quality and efficiency are a pair of mutually exclusive indicators. Limited by this situation, the traditional stitching method can hardly be implemented, resulting in the current fisheye image real-time stitching technology's lack of variety. In our scheme, the stitching mapping table, which describes the projection from pixel coordinate in each unit lens' image to the pixel coordinate in the final omnidirectional frame, is firstly decided, and then the mapping table is embedded in the image processing algorithm to achieve omnidirectional image stitching in real time each frame by each frame.

The procedure for obtaining the parameter mapping table in our scheme is as follows:

(1) Input fisheye images and separately estimate the camera model to obtain the mapping of points on two-dimensional fisheye images to corrected three-dimensional points on hemispheres.

(2) Scale three-dimensional corrected images, with the equirectangular projection (ERP) method, to unfolding pattern to prepare for subsequent processions.

(3) Extract feature points and estimate to find the best math. Then, accordingly calculate the intrinsic parameters and the extrinsic parameters to register the spatial positional relationship between the images.

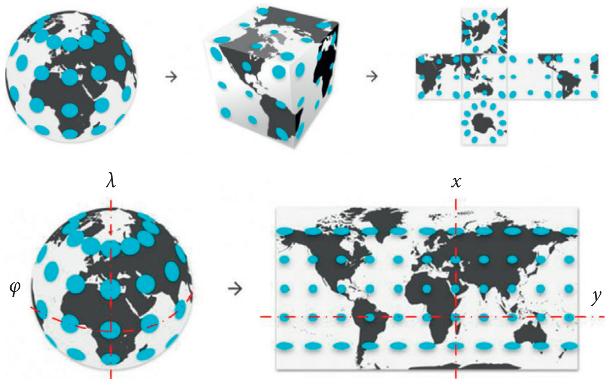


Figure 2: Cube maps and equirectangular projection.

(4) According to the registration result, adjust the spatial positional relationship between the five hemispherical planes, superimpose the five-way corrected hemisphere in the world coordinate system, fuse image pixels on the overlap part, and then convert the three-dimensional image into an omnidirectional frame with ERP.

(5) Extract and save the overall homography matrix of the coordinate on the fisheye image to the coordinate on the final omnidirectional frame for subsequent real-time processing.

4.2. Camera Calibration and Camera Model Estimation. The process of calibrating a camera model is actually a transformation estimate of a two-dimensional vector p in the original fisheye plane to a three-dimensional vector P in the world coordinates. Namely, to accomplish this process, the intrinsic and extrinsic parameters of cameras and the distortion parameters of lenses need to be estimated. The most commonly used technique for lens distortion parameter correction is polynomial fitting, and the pose estimation parameters of the unit lens are 3×3 dimensional matrices. The relationship between the parameters of these two parts is a composite function. This composite optimization technique has a strong dependence on the initial value of the parameters, and the mutual interference is obvious as well, leading to difficulty in achieving global optimization. Our scheme treats the camera parameters and lens distortion as a combined system and estimates the transformation process as a whole. Map two-dimensional points on the fisheye plan to three-dimensional vectors and then convert them to points on the surface of a unit sphere to give a coordinate.

The camera model we use is presented in Figure 3. Let p be a pixel point in the original fisheye image, (u, v) is the pixel coordinate in term of the image center point as the origin, let P be its corresponding three-dimensional vector emanated from the single effective viewpoint, and (x, y, z) is the unit point in term of the optical axis as the origin. Since the plane coordinate transformation is an affine transformation, the relationship between p and P can be expressed as

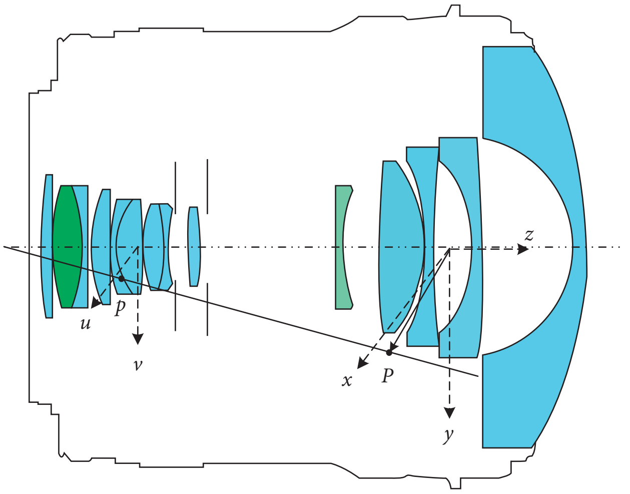


FIGURE 3: Camera model of fisheye camera.

$$\begin{bmatrix} x \\ y \end{bmatrix} = \alpha \begin{bmatrix} u \\ v \end{bmatrix}. \tag{1}$$

Then, the overall mapping of the two-dimensional plane coordinate to the three-dimensional vector can be written as

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} \alpha \cdot u \\ \alpha \cdot v \\ f(u, v) \end{bmatrix} = \begin{bmatrix} u \\ v \\ f'(u, v) \end{bmatrix} = \begin{bmatrix} u \\ v \\ f'(\rho) \end{bmatrix}, \tag{2}$$

$f(u, v)$ is a function of two-dimensional coordinates, $\rho = \sqrt{u^2 + v^2}$. $f'(\rho)$ is the polynomial function that needs fitting:

$$f' = \alpha_0 + \alpha_1 x + \alpha_2 x^2 + \alpha_3 x^3 + \alpha_4 x^4 + \ldots. \tag{3}$$

The polynomial fitting process is assisted by the Matlab toolbox ocam_calib. After a lot of experiments, the number of polynomial terms to be fitted is not as good as possible. To be more specific, the phenomenon of fitting degradation will occur for the sake of too many polynomial terms. Finally, it is determined that the four-term polynomial is used for fitting.

4.3. Unfolding of the Spherical Image. After obtaining the corrected three-dimensional hemisphere image from the fisheye image, we need to perform spherical unfolding for subsequent processing. In our scheme, the most widely used ERP is implemented to achieve unfolding.

As can be seen in the lower part of Figure 2, λ is the longitude of the location to project; φ is the latitude of the location to project; φ_1 are the standard parallels (north and south of the equator) where the scale of the projection is true; λ_1 is the central meridian of the map; x is the horizontal coordinate of the projected location on the map; y is the vertical coordinate of the projected location on the map. It can be concluded that

forward mapping:

$$x = (\lambda - \lambda_0)\cos(\varphi_0), y = (\varphi - \varphi_0), \tag{4}$$

reverse mapping:

$$\lambda = \frac{x}{\cos(\varphi_0)} + \lambda_0, \varphi = y + \varphi_0. \tag{5}$$

4.4. Spatial Registration. For physically setting the five cameras to be mutually orthogonal, theoretically, using the central camera's coordinate system as the world coordinate system and, respectively, rotating the corresponding coordinate system of other cameras by 90°, namely, multiplying the original three-dimensional coordinate matrix by the corresponding rotation matrix can guarantee a strict registered system, achieving three-dimensional space registration. However, considering the physical placement and the camera lens may introduce errors, and the center estimated by the fisheye correction process is not sufficiently the center of the original image; the edge may be misaligned when stitching, so the corrected three-dimensional spherical image needs to be registered again.

To perform registration, we first need to match and filter the feature points of the two images, select the best matching points, calculate the homography matrix, and then calculate the rotation matrix between adjacent two corrected pictures according to the homography matrix.

According to the principle of pinhole imaging, points in the camera coordinate system can be mapped to the world coordinate system via rotation and translation. The translation can be written as follows:

$$\begin{gathered} \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix} = \begin{bmatrix} R & t \\ 0^T & 1 \end{bmatrix} \begin{bmatrix} x_w \\ y_w \\ z_w \\ 1 \end{bmatrix}, \\ R = \begin{bmatrix} 0 & \theta_z & \theta_y \\ \theta_z & 0 & \theta_x \\ \theta_x & \theta_y & 0 \end{bmatrix}, \\ t = \begin{bmatrix} t_x \\ t_y \\ t_z \end{bmatrix}, \end{gathered} \tag{6}$$

where R is the rotation matrix and θ_x, θ_y, and θ_z represent the angles at which the camera rotates around three coordinate axes. t is the translation vector, and t_X, t_y, and t_z are the translation distances of the camera along three coordinate axes.

Through calibrating, the extrinsic parameters matrix $[R|t]$ of each camera relative to the center camera can be obtained, thereby completing the spatial registration.

4.5. Generating the Overall Mapping Table. Finally, the results of previous parts are combined, and overlapped pixels on the spherical surface are fused to produce a mapping table. The table describes how the source coordinate on the overall fisheye map will be transformed to the destination coordinate on the omnidirectional frame. With the mapping table fixed, by preparing multiple threads simultaneously and executing pixel mapping operations on different regions of the panoramic frame, real-time stitching is available. The function display diagram of the omnidirectional video generation module is shown in Figure 4. Our module can complete the stitching process of an OV frame within 20 ms.

5. Streaming Module

High resolution and low transmission delay are the key points in the OV live streaming system. When the transmission delay reaches up to 13 ms or the bitrate is too low, users will feel tired and dizzy [39]. To ensure a good watching experience, the best way is to transmit omnidirectional video to display terminal, but this transmission method does not consider that viewer only watches a small portion of the entire full image. In fact, if OV player offers a 90 rectangle-view when the user looks at a certain direction, only one of the six spheres appear in the user's version and other parts will be out of sight. Transmitting non-FOV with a high bitrate will cause a huge waste of the network bandwidth. Therefore, we adopt a two-layer tile-based transmission mechanism to reduce the heavy burden on the transmission system.

5.1. Projection of the OV Content. After the omnidirectional video generation process, we get a spherical omnidirectional video which we cannot encode with existing coding standards such as H.264/AVC, H.265/HEVC. Since the encoder can only encode rectangular pictures, the omnidirectional video must be mapped into rectangular user view. A common method called equirectangular projection is to map the 3D sphere image to a 2D rectangular plane with longitude as a reference. However, the different visual angle on a panoramic sphere will result in different map areas. The closer to the two poles of the sphere, the more serious the image distortion. As Figure 5 shows, when the user looks at the equator position of the sphere, the projection area corresponding to the 2D plane is 11.9% of the entire panoramic frame. And the area gets the maximum value of 26.7% when view at the poles and is very distorted [40].

5.2. Two-Layer Streaming Scheme for OV. One method is to intercept the FOV area and transmit the FOV image with high bitrate to the client alone. Though it does not consider that the real-time OV system is latency sensitive, if user's head movement is too fast or the image cannot reach to display terminal in time, the display terminal will not have enough time to match the image properly. Users may see a blank area in this view, and it will seriously reduce the user's QoE.

So, we adopt a two-layer tile-based transmission mechanism, and Figure 6 shows the detailed process. First, after the equirectangular projection, the panoramic frame will be encoded by H.265/HEVC. And a low bitrate layer which we called basic layer (BL) will be generated. The BL represents the omnidirectional view at a low bitrate. At the same time, the panoramic frame is divided into 4 × 6 tiles, and the tiles in the FOV area will be extracted by the encoder and encoded with high bitrate as tile enhanced layer (TEL). The FOV area information such as the coordinate of the screen center collected by video client will return to the encoder side. BL and TEL will be transmitted to the client, and these two layers are superimposed on the client side to display.

In this two-layer tile-based transmission mechanism, encoder needs to encode the FOV tiles according to information which returned from display terminal to ensure system performance of MELiveOV. However, due to the random nature of viewer motion, it is very difficult to predict a long-term movement of the user's head. The accuracy will drop from 92% down to 71% when the time for prediction increases from 1 second to 2 seconds [41]. So, the prediction time we set is 1 second. According to the visual movement trajectory of the user in the first few seconds, the prediction algorithm shows the position of the user's viewpoint in the next second.

If the TEL has not arrived in time or TEL matches error, the client can display the BL to ensure a basic view experience rather than generate a blank area. Although this method will cause huge computation, transmission bandwidth is a more valuable resource. With the two-layer tile-based method, we solve the problem of unexpected head movement and network flow. The client can obtain a panoramic frame with the high bitrate of the FOV area and the low bitrate of the non-FOV area, which saves about 55% of the bandwidth consumption without affecting the user's QoE.

5.3. Adaptive FOV Size Selection. In the above two-layer transmission mechanism, a fixed FOV area is used. If the size of the FOV area can be dynamically selected according to different network condition, the system will be more adaptive. When the network is in a good condition, a larger area of high bitrate omnidirectional video can be obtained by the display terminal so that the user can get a better QoE.

Therefore, we adopt an adaptive strategy which allows the encoder to choose different FOV region size based on the network condition. Focusing on the user's viewpoint, we set the FOV area to 90° and 120-degree FOV areas, respectively. When the network is in a bad condition, the encoder selects the FOV with 90°, a smaller panoramic frame area is encoded with the high bitrate. When the network condition is ideal, the encoder selects the FOV with 120° so that the larger panoramic frame area will be encoded with a relatively high bitrate.

The actual function of our two-layer OV transmission scheme is shown in Figure 7. It can be easily observed from the panoramic frame of OV that there is a significant difference in the video quality between FOV and non-FOV.

6. Viewpoint Prediction Module

Accurately predicting the viewer's future viewpoint trajectory can help MELiveOV to better enhance the user experience. Thus, we designed a special prediction model, which can provide users with effective viewpoint prediction at long intervals by using the local historical data and global multiuser information.

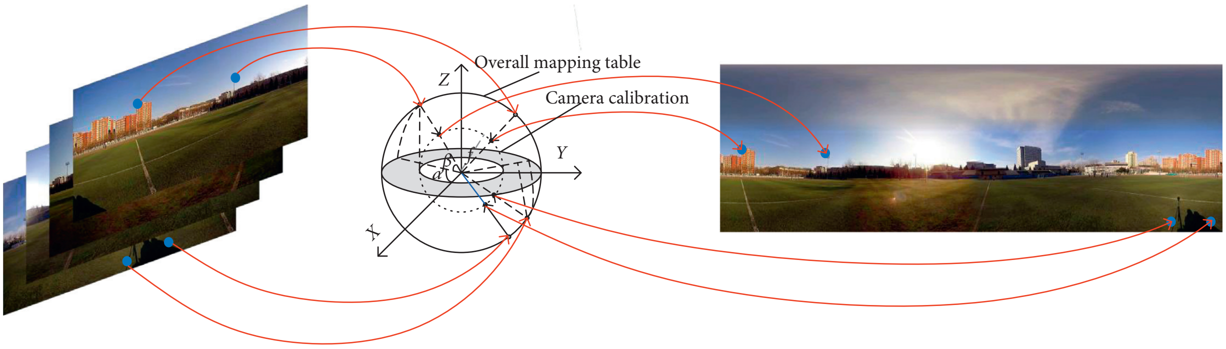


Figure 4: Function display of omnidirectional video generation module.

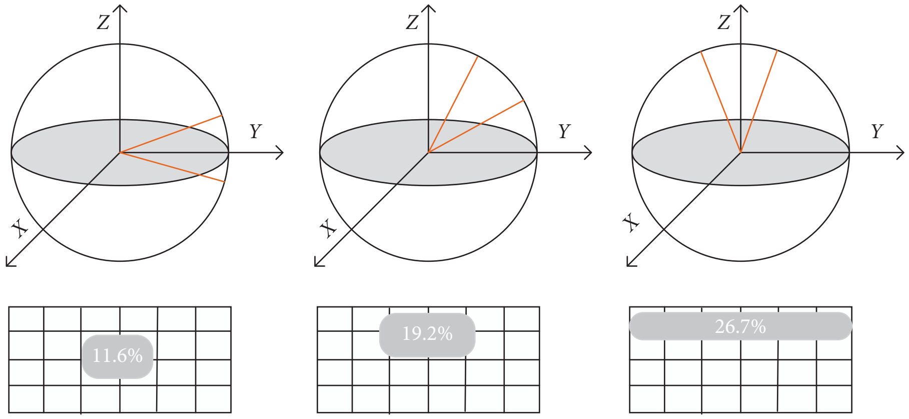


Figure 5: Projection from different angles.

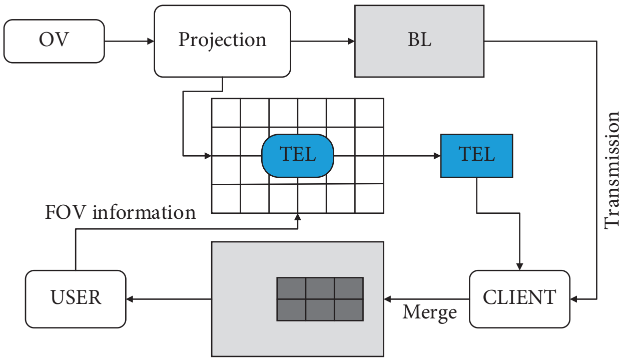


Figure 6: Two-layer streaming process.

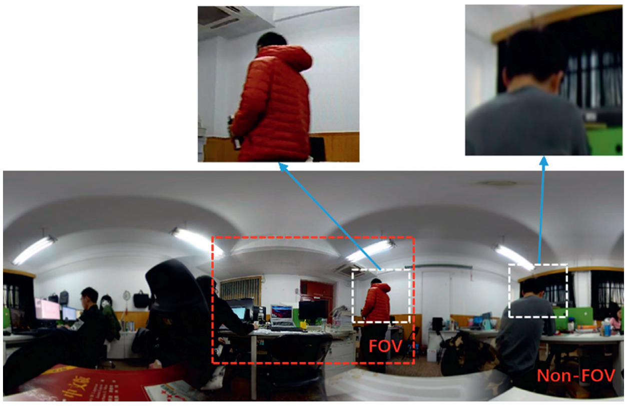


Figure 7: Function display of two-layer OV transmission.

6.1. Overview of the Module. The problem of viewpoint prediction is considered from two perspectives in the viewpoint prediction module. On the one hand, most users are not watching the OV for the first time. Therefore, the historical viewpoint data of the OVs they have seen may contain some information about the user's viewing habits. For example, some users may prefer to move their viewpoints slowly and smoothly, while other users prefer faster viewpoint movements. This customized information allows our module to be adaptable to different users. On the other hand, the OV content provider may already have collected the viewpoint trajectory data from multiple users for the same OV source. Through the analysis of the dataset, it can be found that when different users watch the same OV, their viewpoint trajectory will have a similar movement pattern. This is because some frames of the OV have the content that can arouse most users' interest. When viewing these frames, different users tend to focus on the same region of interest, so the viewpoint trajectory will have a similar movement pattern. In this way, these existing models will help provide more accurate viewpoint prediction services as new users begin to watch.

The overview flow of viewpoint prediction module is shown in Figure 8. In the proposed method, the viewpoint prediction system includes two independent channels, one

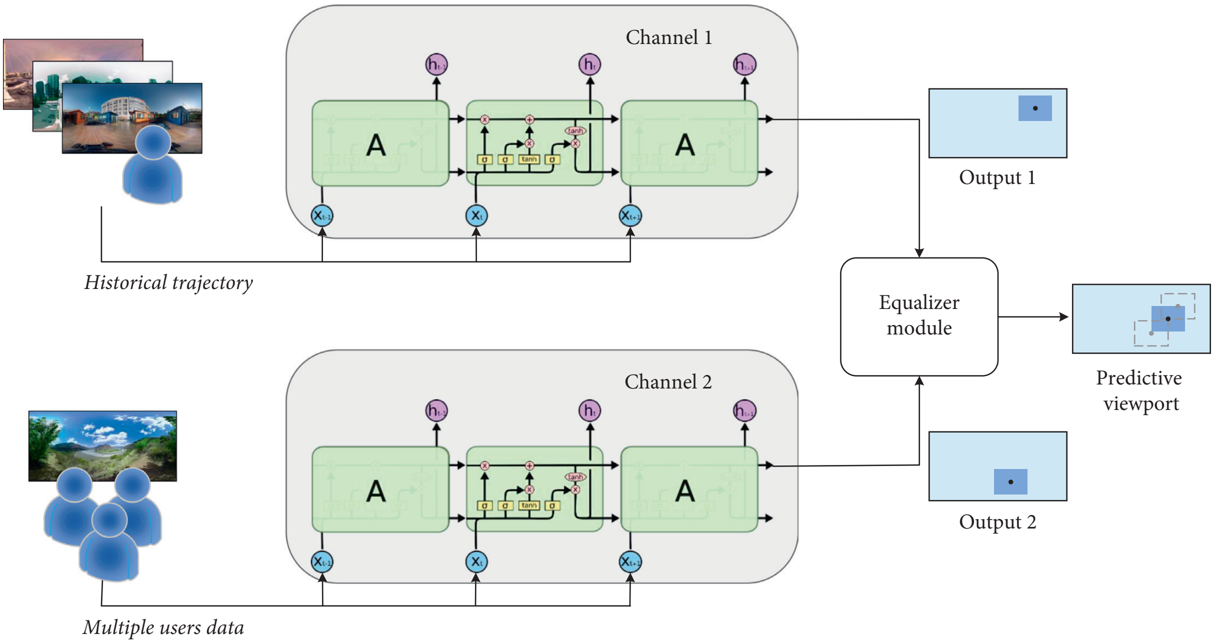


Figure 8: System flow of CPVp-LSTM.

of which makes the prediction based on historical viewpoint data of the single user. And the second channel will use the trajectory data of other people from the same OV content to predict the viewpoint. After the output of both channels passes through the equalizer model, the final prediction result can be obtained.

As shown in Figure 8, both channels of viewpoint prediction module implement prediction functions through the LSTM (long short-term memory) network. The LSTM network is often used to implement the prediction of time series data in deep learning. It is a good way to detect and fit the deep rules of the data. Based on these advantages, the LSTM network is well suited as the basic predictor for the proposed module.

6.2. Basic Predictor Based on LSTM. As shown in Figure 8, both channels of CPVp-LSTM implement prediction functions through the LSTM network. The LSTM network is often used in the prediction of time series signals. It can well detect and fit to the in-depth features of the dataset. Based on these advantages, the LSTM network is well suited as the basic predictor for the proposed algorithm. Suppose that the time series of the user's viewport can be expressed by $\left\{\overrightarrow{V_{t_1}}, \overrightarrow{V_{t_2}}, \cdots, \overrightarrow{V_{t_{n-1}}}, \overrightarrow{V_{t_n}}\right\}$. $\overrightarrow{V_{t_i}}$ represents the viewport coordinates of the user at time t_i. The core function of the basic predictor is to calculate $\overrightarrow{V_{t_{i+M}}}$ from $\overrightarrow{V_{t_{i-N}:t_i}}$ with LSTM networks, N is the length of the input sequence, and M is the length of the predicted interval. The historical viewport coordinate sequence from time t_{i-N} to time t_i is used to predict the position of the viewport at the time t_{i+M} in the future.

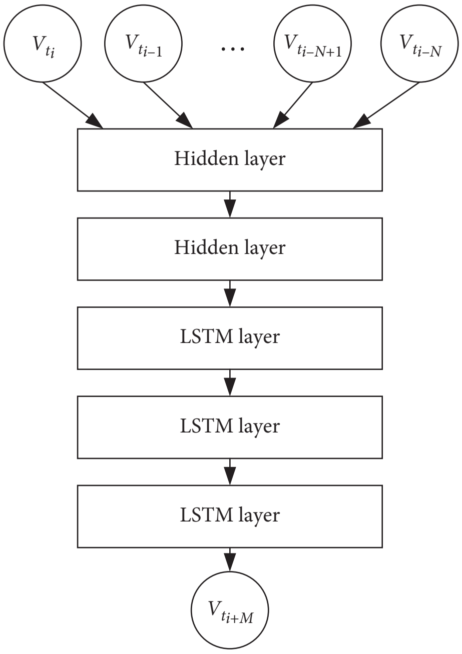


Figure 9: LSTM network of basic predictor.

The proposed basic predictor contains two hidden layers and three LSTM layers, as shown in Figure 9. The rectified linear unit (ReLU) activation function is used after the hidden layer to enhance nonlinearity. The LSTM layer is composed of N LSTM units. Each unit generates two values simultaneously; one is the output of the current unit, and the other one is the collection of memory information from all previous units. Both of these two output values will be sent into the next unit as the input, so the LSTM layer can be memorable. The loss function is modified based on cross entropy, which is used to update various parameters of the network during each iteration of the training. The user's viewport position $\overrightarrow{V_t}$ can be described by its Euler angle coordinates, which includes 3 degrees of freedom, pitch, yaw, and roll (i.e., X, Y, and Z angles). X and Y angles are within −90° ~ 90° and −180° ~ 180°, respectively. In 90% of time, Z angles are within −11° ~ 9°. Based on this special range of values for viewport coordinates, we define an improved cross entropy loss function L. Its definition of the Y component is shown as equations (7) and (8). T_{h} is a threshold used to determine whether an out-of-bounds condition has occurred, which is generally set to a default value of 10. Y is the predicted output and $\widehat{Y}$ is the actual value:

$$\acute{Y} = \begin{cases} Y - 360 & \text{where } 0 < 360 - (Y - \widehat{Y}) < T_{\mathrm{h}}, \\ Y + 360 & \text{where } 0 < 360 + (Y - \widehat{Y}) < T_{\mathrm{h}}, \\ Y & \text{otherwise.} \end{cases} \tag{7}$$

After normalizing $\acute{Y}$ and $\widehat{Y}$,

$$L = -[\acute{Y} \log(\widehat{Y}) + (1 - \acute{Y})\log(1 - \widehat{Y})]. \tag{8}$$

The cross entropy definition of the X component is similar to the Y component. Due to the small distribution range of the Z component, there is no out-of-bounds condition in most cases, so the cross entropy of the Z component does not change.

In CPVp-LSTM, the predictors used in the two channels are similar in structure, but the size and some parameters of each layer are adjusted according to the difference between the input sequence.

6.3. Prediction Model Based on User Viewing Habits.

The difference in viewing habits between different users is enormous, which needs to be fully considered when making viewpoint predictions based on personal historical data. We use the user's ID as an index to create a separate viewpoint trajectory database for each user. The database will contain historical viewpoint data for all OVs that the user has viewed. Since the user's behavioral habit information is mainly included in the relative movement of the user's viewpoint (slow or fast) and is not closely related to the absolute position of the user's viewpoint, we extract the differential data of the user's viewpoint trajectory and send them to the LSTM network for training.

At time t_i, its difference value can be obtained by the following formula:

$$D_{t_i} = V_{t_i} - V_{t_i - 1}, \tag{9}$$

where V_{t_i} is the current viewpoint coordinate at time t_i and $V_{t_i - 1}$ is the last coordinate at time $t_i - 1$. The LSTM network finally obtains the predicted value D_M of the viewpoint coordinate change amount, and the final output result of channel 1 is $V_{t_i + M} = V_{t_i} + D_M$.

6.4. Prediction Model Based on ROI Perception of OV Content.

Inspired by some existing viewpoint prediction schemes, they are able to improve the accuracy of prediction by acquiring regions of interest (ROI) in OV frames. This type of method first locates the ROI by performing image feature extraction on each frame that is predecoded and then simultaneously sends the ROI coordinates into the prediction model along with the viewpoint coordinates acquired by the sensor of the display terminal. The ROI information of every frame can effectively improve the accuracy of the prediction model, but this operation of predecoding and extracting features is very expensive in terms of resource consumption of most display devices.

In this paper, we consider that the information of this ROI should also be included in the time series of viewpoint coordinates. When a frame of the OV has an ROI that attracts the attention of most users, the user's viewpoint position should tend to converge at this moment. In order to get that ROI information, we cluster the set of viewpoint coordinates of each frame in one OV. These viewpoint data are collected from all users when they independently watch this OV. Because the number of ROIs contained in one frame cannot be predetermined, the DBSCAN (density-based spatial clustering of applications with noise) algorithm is used for clustering. DBSCAN can automatically determine the number of clusters by specifying the distance between members and the maximum boundary of the cluster.

Figure 10 shows the analysis results of two typical frames. The left side of Figure 10(a) is the picture of the OV frame, and the right side is the clustering result of the frame viewpoint coordinates of this frame. It can be seen that most of the points are clustered to cluster-1, which are colored yellow. The remaining isolated points are shown in blue and their number is too small to be grouped together. The area indicated by the yellow box in the OV frame on the left corresponds to cluster-1 in the clustering result. It can be clearly observed that the concentration of the viewpoint at this time is due to the presence of the diver in the area of the yellow box. Similarly, the cluster-1 of the clustering results in Figure 10(b) is caused by the diver in the yellow box of the OV frame, and the cluster-2 is caused by the underwater wreckage of the green box.

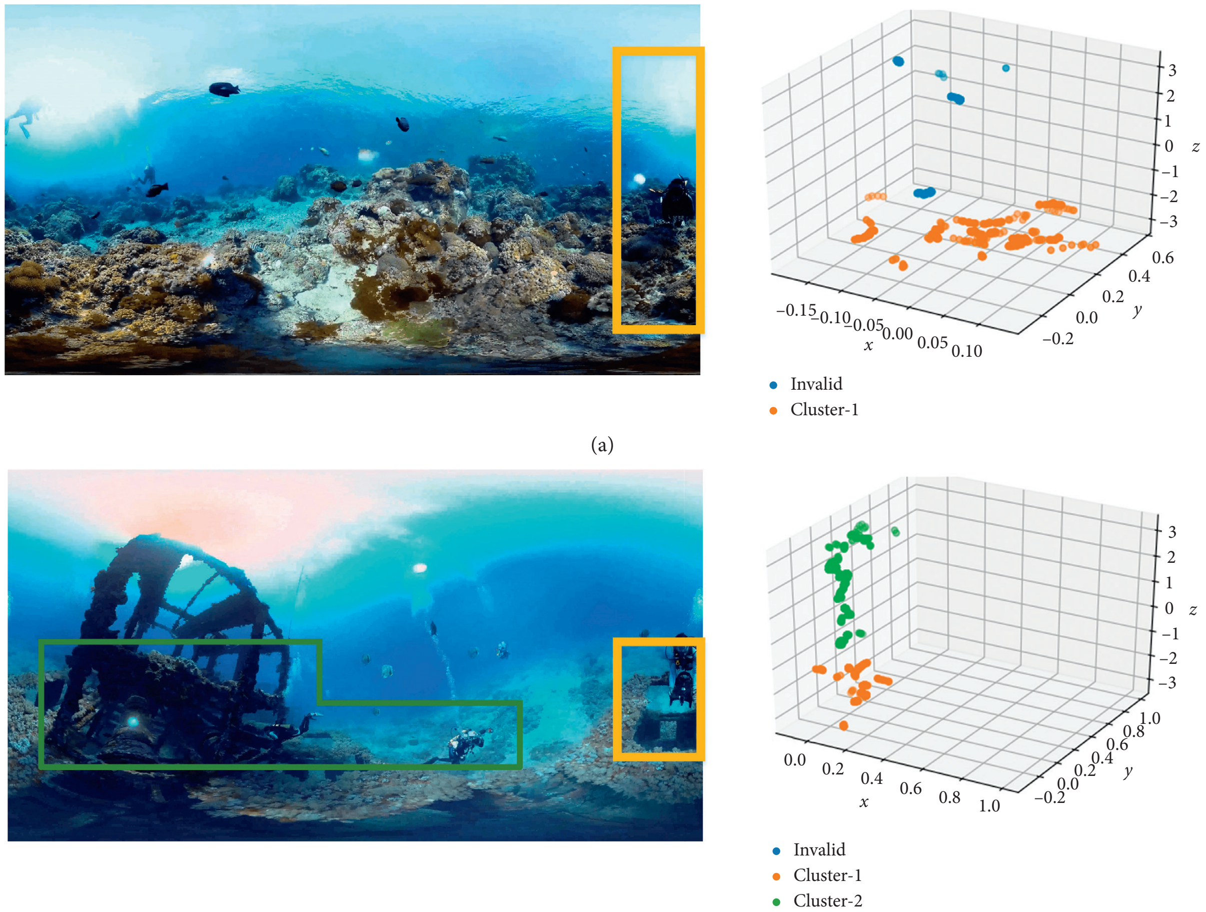
Because channel 2 mainly refers to the information of the absolute coordinates of the user's viewpoint, the V_t sequence is directly used as the input of the predictor. At the same time, we introduce the clustering results of each frame into the prediction model to improve accuracy. In actual deployment, after the viewpoint prediction module collects the viewpoint data from different users according to the OV ID, the clustering operation can be completed with only a small amount of resources. Channel 2 will directly output the predicted viewpoint coordinates.

7. Implementation and Evaluation

In this section, we will show the implementation of the MELiveOV prototype system and discuss the performance of it.

7.1. Experimental Prototype System.

Figure 11 shows the capture device of the prototype system. It consists of a



(a)

(b)

Figure 10: Cluster result of OV diving.

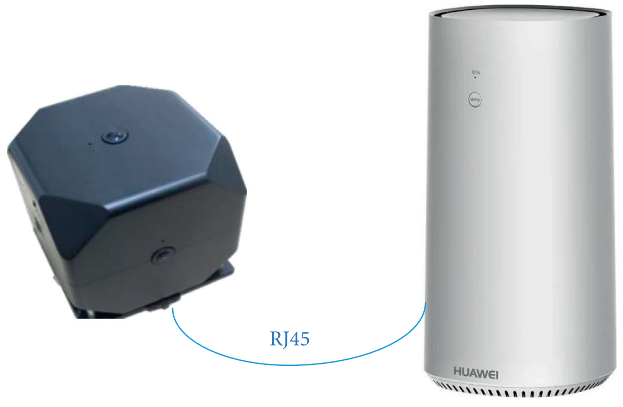


Figure 11: Omnidirectional capture device.

customized omnidirectional camera with 6 lenses that can simultaneously capture video data in 6 directions (up, down, left, right, front, and back) and a 5G CPE. They communicate through the RJ45 network ports. The structure of the customized camera is shown in Figure 12. We use HiSilicon's Hi3559AV100 as the control board, which is responsible for collecting all the original lens data and generating standardized video sequences. Data are transmitted between lens and control board through MIPI interface.

Our prototype system also includes two edge servers, as shown in Figure 13. The edge server consists of a 5G small cell and a regular server. The regular server has an Intel(R) Xeon(R) CPUE5-2630 v4 and six GTX 1080TI 11G; the size of the server is 32G. We modified the forwarding strategy of the 5G small cell so that after the data arrives, it will be processed by the server before forwarding. There are two sets of such edge servers, one as the first-mile edge server and the other as the last-mile edge server. Communication between them is achieved through a virtual core network inside the lab.

On the display terminal, the prototype system supports access to multiple heterogeneous playback devices. Such as Android phones, PC, and HMD. We have designed dedicated player software on each platform to implement the functionality of the viewpoint prediction module. All player software can collect the user's viewpoint data with the sampling frequency of 30 Hz.

As presented in Figure 14, the prototype system of MELiveOV implements the end-to-end live streaming service of OV. The left part of Figure 14 is the picture inside the FOV, which can be seen by the user on the display terminal

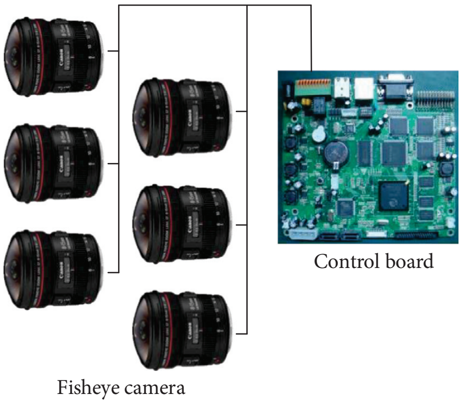


Figure 12: Structure of customized camera.

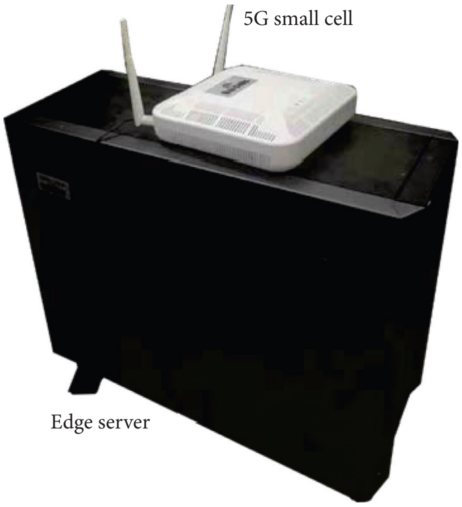


Figure 13: 5G edge server.

through the screen of the device. The upper right part of the figure shows the actual situation of the user watching the OV live streaming through the Android phone. The lower right part of the figure shows the working scene of the capture device of MELiveOV. As shown in the figure, we placed the omnidirectional camera on a handcart with power supply, and the camera communicates with the 5G small cell of the edge server over the wireless network.

7.2. Experimental and Evaluation Results. In this subsection, we tested the MELiveOV prototype system in different scenarios and analyzed the system performance. As shown in Figure 15, we conducted four experiments of OV live streaming in the *Playground, Road, Office*, and *Night* scenes. We collected data about the video quality and network bandwidth consumption of MELiveOV in four sets of experiments.

The overall resolution of OV in all four scenarios is around 4k (the resolution of the OV panoramic frame is not fixed due to the two-layer transmission scheme) and the frame rate is 25 fps. Besides, we used FFMPEG as our coding tool and H.264/AVC as our coding standard. The PSNR of the OV picture is shown in Figure 11 during the live streaming. In Figure 16, we used PSNR (peak signal to noise ratio) to evaluate the picture quality during OV live streaming. The red column represents the quality of the video picture within the user's FOV, and the yellow column represents the quality of the non-FOV area. In the *Night*, the quality of the OV is relatively high because the picture content is relatively simple (mainly black) and the camera is fixed. In the *Road*, the camera is moving and there are too many objects (buildings and trees) in the scene, so the PSNR is the worst. The results of *Playground* and *Office* are more common. MELiveOV can guarantee that the PSNR of the user's FOV in OV live streaming is about 50 dB. At the same time, we can also ensure that the PSNR of non-FOV areas is maintained above 30 dB. When the user's viewpoint trajectory is predicted to be wrong, MELiveOV can still avoid image incompleteness in the user's field of view.

Figure 17 is analyzed with SSIM (structural similarity index) as a quality evaluation indicator. The results show that MELiveOV can also achieve better performance on SSIM, the quality of the FOV region is maintained above 0.98, and the non-FOV region is around 0.9.

We have verified the reliability of the picture quality of MELiveOV during live streaming. Next, we will show the network bandwidth situation of the MELiveOV. We set up a comparison system that puts the omnidirectional video generation task on the central cloud server (which is a cloud server leased on the public network). The comparison system does not include the streaming module of the last-mile edge server and the prediction module of the display terminal. It can only implement the most basic OV live streaming function. The results of the network bandwidth consumption experiment are shown in Table 1. We can see that in all scenarios, MELiveOV can save about 50% of the bandwidth demand, which can effectively reduce the transmission pressure of the network.

In terms of transmission delay, we also compared the two sets of schemes. The results are shown in Table 2. We can see that the service request during the OV live streaming can be responded in time due to the introduction of the MEC architecture. MELiveOV's average transmission delay can be reduced by 70% to 80%, which greatly enhances the real-time performance of OV live streaming. It can also be seen from the table that in the case of indoor scenes and fixed cameras, the transmission delay of the system is small. When the camera is outdoors and moving, the overall system latency rises significantly. We believe this is mainly due to the limited transmit power of the 5G small cell we used in the experiment. By the way, we noticed that the comparison system also achieved good latency performance in night scenes. This is mainly because of the fewer network users at night. And the network condition is better, so the transmission delay is significantly improved.

8. Conclusion and Future Work

In order to meet the needs of omnidirectional video (OV) live streaming services, this paper proposes a mobile edge assisted live streaming system for omnidirectional video

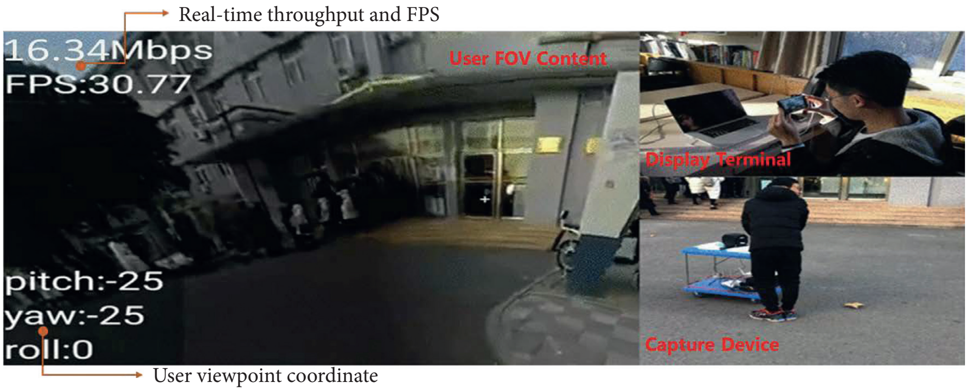


Figure 14: Presentation of MELiveOV prototype system.

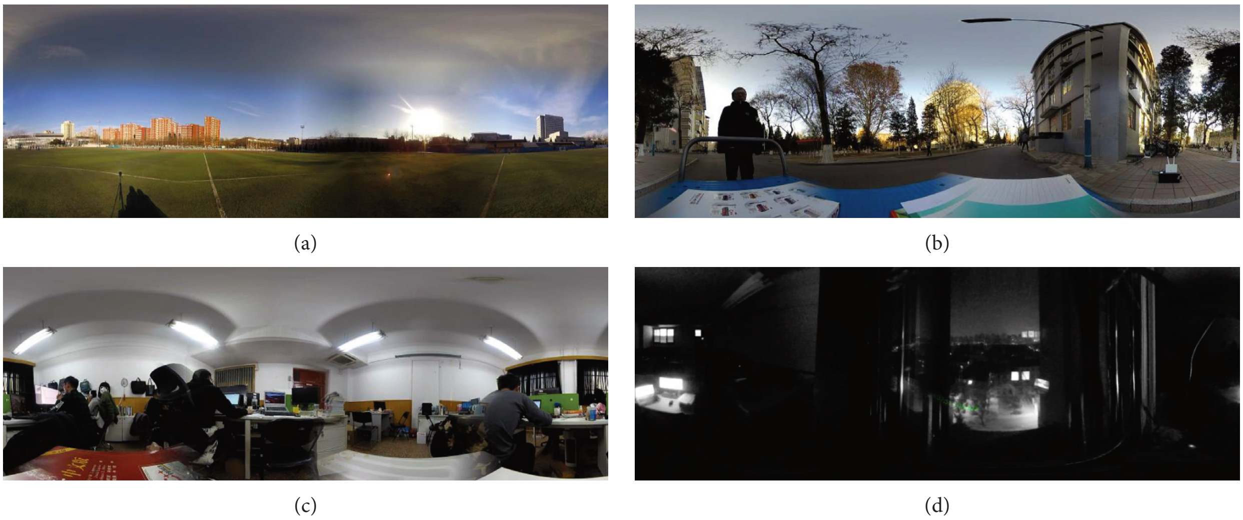

(a)
(b)
(c)
(d)

Figure 15: MELiveOV in different scenarios: (a) playground; (b) road; (c) office; (d) night.

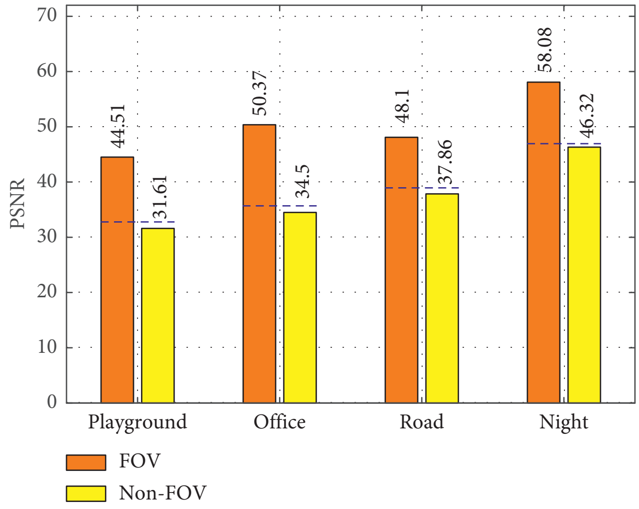


Figure 16: PSNR of OV during live streaming.

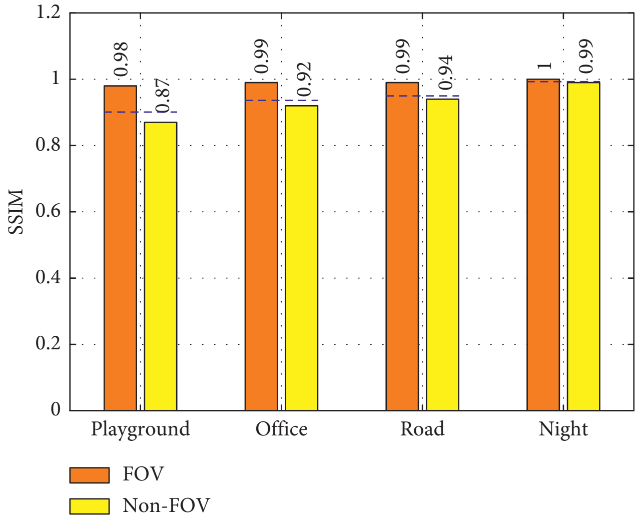


Figure 17: SSIM of OV during live streaming.

(MELiveOV). Enabled by the 5G edge servers with abundant computing resources, MELiveOV can offload the computational OV stitching tasks to the edge and introduce more complex prediction algorithms to optimize live streaming performance. An end-to-end prototype system was built, and a complete service chain from capture to display for OV live streaming was implemented. The results of the

Table 1: Average bandwidth consumption of OV live streaming system.

Scenarios	Unoptimized streaming scheme (Mbps)	MELiveOV (Mbps)	Decrease ratio (%)
Playground	35.52	15.10	57.5
Road	38.88	16.64	57.2
Office	29.20	14.48	50.4
Night	22.48	11.52	48.8

Table 2: Average transmission delay of OV live streaming system.

Scenarios	Unoptimized streaming scheme (seconds)	MELiveOV (seconds)	Decrease ratio (%)
Playground	22.7	6.7	70.5
Road	26.4	5.2	80.3
Office	13.5	2.8	79.3
Night	8.7	2.4	72.4

evaluation experiment show that MELiveOV can reduce the network bandwidth requirement by about 50% and the transmission delay of more than 70% under the premise of ensuring the picture quality of viewers.

There are still many problems to be solved in the research of OV live streaming. For example, cameras may switch between multiple 5G base stations during long-distance movement. It is very important to design reliable mechanisms to ensure seamless migration of computational tasks between different edge servers. And how to achieve resource scheduling and data fusion in multiuser scenarios is also one of our future research directions. To conclude, 5G MEC is a promising solution and can well meet the needs of high-resolution OV live streaming services.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work was supported by the BUPT Excellent Ph.D. Students Foundation (CX2019102). This work was also supported by the project “Stereoscopic Coverage Communication Network Verification Platform for China Sea” (PCL2018KP002).

References

[1] Intel Ovum, *5G Economics of Entertainment Report*, Technical Report, 2018.

[2] S. Oh and S. Hwang, “OMAF: generalized signaling of region-wise packing for omnidirectional video,” in *Proceedings of the 118th MPEG Meeting of ISO/IEC JTC1/SC29/WG11, MPEG2017/m40423*, 2017.

[3] F. Qian, B. Han, Q. Xiao, and V. Gopalakrishnan, “Flare: practical viewport-adaptive 360-degree video streaming for mobile devices,” in *Proceedings of the 24th Annual International Conference on Mobile Computing and Networking*, pp. 99–114, ACM, New Delhi, India, October 2018.

[4] M. Xiao, C. Zhou, Y. Liu, and S. Chen, “OpTile: toward optimal tiling in 360-degree video streaming,” in *Proceedings of the 2017 ACM on Multimedia Conference*, pp. 708–716, ACM, Mountain View, CA, USA, October 2017.

[5] C. Ozcinar, A. De Abreu, and Aljosa Smolic, “Viewport-aware adaptive 360 video streaming using tiles for virtual reality,” in *Proceedings of the 2017 IEEE International Conference on Image Processing (ICIP)*, pp. 2174–2178, IEEE, Beijing, China, September 2017.

[6] R. Skupin, Y. Sanchez, Y.-K. Wang, M. M. Hannuksela, J. Boyce, and W. Mathias, “Standardization status of 360 degree video coding and delivery,” in *Proceedings of the 2017 IEEE Visual Communications and Image Processing (VCIP)*, pp. 1–4, IEEE, St. Petersburg, FL, USA, December 2017.

[7] D. V. Nguyen, H. T. T. Tran, A. T. Pham, and T. Cong Thang, “An optimal tile-based approach for viewport-adaptive 360-degree video streaming,” *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 9, no. 1, pp. 29–42, 2019.

[8] H. Ahmadi, O. Eltobgy, and H. Mohamed, “Adaptive multicast streaming of virtual reality content to mobile users,” in *Proceedings of the on Thematic Workshops of ACM Multimedia*, pp. 170–178, ACM, Mountain View, CA, USA, October 2017.

[9] M. Seufert, S. Egger, M. Slanina, T. Zinner, T. Hobfeld, and P. Tran-Gia, “A survey on quality of experience of http adaptive streaming,” *IEEE Communications Surveys & Tutorials*, vol. 17, no. 1, pp. 469–492, 2015.

[10] J. Le Feuvre and C. Concolato, “Tiled-based adaptive streaming using MPEG-DASH,” in *Proceedings of the 7th International Conference on Multimedia Systems, MMSys ’16*, pp. 41:1–41:3, New York, NY, USA, 2016.

[11] Z. Liu, S. Ishihara, Y. Cui, Y. Ji, and Y. Tanaka, “Jet: joint source and channel coding for error resilient virtual reality video wireless transmission,” *Signal Processing*, vol. 147, pp. 154–162, 2018.

[12] C. Griwodz, M. Jeppsson, H. . Espeland et al., “Efficient live and on-demand tiled hevc 360 VR video streaming,” in *Proceedings of the 2018 IEEE International Symposium on Multimedia (ISM)*, pp. 81–88, IEEE, Taichung, Taiwan, 2018.

[13] X. Kelvin Zou, J. Erman, V. Gopalakrishnan et al., “Can accurate predictions improve video streaming in cellular networks,” in *Proceedings of the 16th International Workshop on Mobile Computing Systems and Applications*, pp. 57–62, ACM, Santa Fe, NM, USA, February 2015.

[14] C. Li, L. Toni, J. Zou, H. Xiong, and P. Frossard, “QoE-driven mobile edge caching placement for adaptive video streaming,” *IEEE Transactions on Multimedia*, vol. 20, no. 4, pp. 965–984, 2018.

[15] A. Machen, S. Wang, K. K. Leung, B. J. Ko, and T. Salonidis, “Live service migration in mobile edge clouds,” *IEEE Wireless Communications*, vol. 25, no. 1, pp. 140–147, 2018.

[16] J. Wu, C. Yuen, N.-M. Cheung, J. Chen, and C. W. Chen, “Enabling adaptive high-frame-rate video streaming in mobile cloud gaming applications,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 25, no. 12, pp. 1988–2001, 2015.

[17] Y. Sanchez, G. Singh Bhullar, R. Skupin, C. Hellge, and T. Schierl, “Delay impact on mpeg OMAF’s tile-based

viewport-dependent 360° video streaming," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 9, no. 1, pp. 18–28, 2019.

[18] L. Sun, F. Duanmu, Y. Liu et al., "A two-tier system for on-demand streaming of 360 degree video over dynamic networks," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 9, no. 1, pp. 43–57, 2019.

[19] Z. Chen, Y. Li, and Y. Zhang, "Recent advances in omnidirectional video coding for virtual reality: projection and evaluation," *Signal Processing*, vol. 146, pp. 66–78, 2018.

[20] K. K. Sreedhar, A. Aminlou, M. M. Hannuksela, and M. Gabbouj, "Viewport-adaptive encoding and streaming of 360-degree video for virtual reality applications," in *Proceedings of the 2016 IEEE International Symposium on Multimedia (ISM)*, pp. 583–586, IEEE, San Jose, CA, USA, December 2016.

[21] R. Skupin, Y. Sanchez, C. Hellge, and T. Schierl, "Tile based HEVC video for head mounted displays," in *Proceedings of the 2016 IEEE International Symposium on Multimedia (ISM)*, pp. 399-400, IEEE, San Jose, CA, USA, December 2016.

[22] M. Yu, H. Lakshman, and B. Girod, "A framework to evaluate omnidirectional video coding schemes," in *Proceedings of the 2015 IEEE International Symposium on Mixed and Augmented Reality*, pp. 31–36, IEEE, Fukuoka, Japan, October 2015.

[23] S.-H. Lee, S.-T. Kim, E. Yip, B.-D. Choi, J. Song, and S.-J. Ko, "Omnidirectional video coding using latitude adaptive downsampling and pixel rearrangement," *Electronics Letters*, vol. 53, no. 10, pp. 655–657, 2017.

[24] R. Ghaznavi-Youvalari, A. Zare, H. Fang et al., "Comparison of HEVC coding schemes for tile-based viewport-adaptive streaming of omnidirectional video," in *Proceedings of the 2017 IEEE 19th International Workshop on Multimedia Signal Processing (MMSP)*, pp. 1–6, IEEE, London, UK, October 2017.

[25] I. D. D. Curcio, H. Toukomaa, and D. Naik, "Bandwidth reduction of omnidirectional viewport-dependent video streaming via subjective quality assessment," in *Proceedings of the 2nd International Workshop on Multimedia Alternate Realities*, pp. 9–14, ACM, CA, USA, August 2017.

[26] M. Yu, H. Lakshman, and B. Girod, "Content adaptive representations of omnidirectional videos for cinematic virtual reality," in *Proceedings of the 3rd International Workshop on Immersive Media Experiences*, pp. 1–6, ACM, Brisbane, Australia, October 2015.

[27] M. Graf, C. Timmerer, and C. Mueller, "Towards bandwidth efficient adaptive streaming of omnidirectional video over http: design, implementation, and evaluation," in *Proceedings of the 8th ACM on Multimedia Systems Conference*, pp. 261–271, ACM, Taipei, Taiwan, June 2017.

[28] J. Li, Z. Wen, S. Li, Y. Zhao, B. Guo, and J. Wen, "Novel tile segmentation scheme for omnidirectional video," in *Proceedings of the 2016 IEEE International Conference on Image Processing (ICIP)*, pp. 370–374, IEEE, Phoenix, AZ, USA, 2016.

[29] S. Gudumasu, H. Ahmed, Y. He, and Y. Ye, "A sub-picture-based omnidirectional video live streaming platform," in *Proceedings of the Applications of Digital Image Processing XLI*, vol. 10752, p. 1075234, International Society for Optics and Photonics, San Diego, CA, USA, August 2018.

[30] C. Ozcinar and A. Smolic, "Visual attention in omnidirectional video for virtual reality applications," in *Proceedings of the 2018 Tenth International Conference on Quality of Multimedia Experience (QoMEX)*, pp. 1–6, IEEE, Berlin, Germany, September 2018.

[31] A. Ninan and C. Atluru, "View direction based multilevel low bandwidth techniques to support individual user experiences of omnidirectional video," US Patent App. 15/842,703, 2018.

[32] R. Ghaznavi-Youvalari and A. Aminlou, "Geometry-based motion vector scaling for omnidirectional video coding," in *Proceedings of the 2018 IEEE International Symposium on Multimedia (ISM)*, pp. 127–130, IEEE, Taichung, Taiwan, 2018.

[33] J. Le Feuvre and C. Concolato, "Tiled-based adaptive streaming using MPEG-DASH," in *Proceedings of the 7th International Conference on Multimedia Systems*, vol. 41, ACM, Wörthersee, Austria, May 2016.

[34] Y. Sun, A. Lu, and Y. Lu, "Weighted-to-spherically-uniform quality evaluation for omnidirectional video," *IEEE signal processing letters*, vol. 24, no. 9, pp. 1408–1412, 2017.

[35] 3GPP, Virtual Reality (VR) media services over 3GPP (release 15), Technical Specification (TS) 26.918, 3rd Generation Partnership Project (3GPP), 06 2017, Version 1.0.0.

[36] DVB, "DVB study mission on virtual reality (CM1706)," *Commercial Module, Digital Video Broadcasting*, DVB Project, 2016.

[37] Y. Ye, E. Alshina, and J. Boyce, "JVET-E1003: Algorithm descriptions of projection format conversion and video quality metrics in 360 Lib," *Joint Video Exploration Team ITU-T SG 16 WP3 ISO/IEC JTC 1/SC 29/WG 11 5th Meet*, Joint Video Exploration Team, Geneva, Switzerland, 2017.

[38] J. Boyce, E. Alshina, A. Abbas, and Y. Ye, "JVET common test conditions and evaluation procedures for 360 video," in *Joint Video Exploration Team (JVET) of ITU-T SG16 WP3 and ISO/IEC JTC1/SC29/WG11, JVET-E1030*, pp. 1–6, 2017.

[39] M. C. Potter, B. Wyble, C. E. Hagmann, and E. S. McCourt, "Detecting meaning in RSVP at 13 ms per picture," *Attention, Perception, & Psychophysics*, vol. 76, no. 2, pp. 270–279, 2014.

[40] G. He, J. Hu, H. Jiang, and Y. Li, "Scalable video coding based on user's view for real-time virtual reality applications," *IEEE Communications Letters*, vol. 22, no. 1, pp. 25–28, 2018.

[41] F. Duanmu, E. Kurdoglu, Y. Liu, and Y. Wang, "View direction and bandwidth adaptive 360 degree video streaming using a two-tier system," in *Proceedings of the 2017 IEEE International Symposium on Circuits and Systems (ISCAS)*, pp. 1–4, IEEE, Baltimore, MD, USA, May 2017.

Hindawi
Mobile Information Systems
Volume 2019, Article ID 3971598, 16 pages
https://doi.org/10.1155/2019/3971598

Research Article

NSAF: An Approach for Ensuring Application-Aware Routing Based on Network QoS of Applications in SDN

Joonseok Park,[1] **Jeseung Hwang**,[2] **and Keunhyuk Yeom**[3]

[1]*Research Institute of Logistics Innovation and Networking, Pusan National University, Busandaehak-ro 63beon-gil, Geumjeong-gu, Busan 46241, Republic of Korea*
[2]*Busan Grand ICT R&D Center Associate Research, Pusan National University, Busandaehak-ro 63beon-gil, Geumjeong-gu, Busan 46241, Republic of Korea*
[3]*Department of Electrical and Computer Engineering, Pusan National University, Busandaehak-ro 63beon-gil, Geumjeong-gu, Busan 46241, Republic of Korea*

Correspondence should be addressed to Keunhyuk Yeom; yeom@pusan.ac.kr

Received 17 January 2019; Accepted 31 March 2019; Published 23 April 2019

Guest Editor: Pinar Kirci

The advent of software-defined networking (SDN) has led to the paradigm of programmable network environments. Conceptually, the structure of SDN is divided into three layers: application, control, and infrastructure. The SDN controller in the control layer can configure and execute the routing of applications to the infrastructure layer consisting of network devices, including hosts and switches. Current studies on SDN have predominantly focused on control layer aspects, such as controller development, performance aspects of the controller, and interaction among different controllers and between controllers and network devices. However, to provide seamless network services and efficiently manage network environments, application-aware routing is essential because applications may have different quality of service (QoS) requirements, such as maximum bandwidth and minimum delay. This study proposes the Network Situation-Aware Framework (NSAF) to more efficiently handle application routing based on the QoS requirements and changing network status. A mechanism for supporting the NSAF consisting of application registration, network status monitoring, violation detection, and routing control is also presented. The applicability and feasibility of the proposed NSAF are verified by implementing a prototype. The NSAF may be used as a reference model for efficiently managing SDN-based networks to ensure application QoS.

1. Introduction

Developments in network technologies have resulted in new programmable network environments based on software-defined networking (SDN) [1–5], in which the control plane is separated from the data plane. As shown in Figure 1, SDN can generally be divided into three layers: application, control, and infrastructure. The controller in the control layer can order the routing of business applications to the infrastructure layer.

Following the advent of SDN, relevant studies have primarily focused on the control layer aspects [6, 7]. Various SDN open-source projects, such as Floodlight [8], OpenDaylight [9], ONOS [10], and Ryu [11], are currently active. Research into protocols, including OpenFlow [12] and OpFlex [13], which support communication between the control and infrastructure layers, is also underway. To further spread the use of the SDN technology and maximize its network management benefits, research on the application layer aspect is needed (e.g., a method to control the network from the application point of view).

In SDN, business applications can have different network quality of service (QoS) requirements. A particular application may need maximum bandwidth, whereas another may require minimum delay. In addition, the network condition changes according to network traffic and failure. Therefore, application-aware routing [14–16] is essential. This study proposes a framework called the Network Situation-Aware Framework (NSAF), and its mechanism satisfies this need and ensures a stable execution of SDN applications. The NSAF analyzes each application's network QoS requirements, monitors current network status, detects violation of network QoS, and finds the most appropriate network routing paths.

The remainder of this paper is organized as follows: Section 2 outlines the basic concepts; Section 3 presents the NSAF and its design and architecture; Section 4 presents a case study; Section 5 discusses related work and evaluates the proposed mechanism; and Section 6 concludes this paper and outlines future work.

2. Basic Concepts

2.1. DiffServ. Differentiated services (DiffServ) specify a simple and scalable mechanism for classifying and managing network traffic and providing QoS on modern IP networks [17]. DiffServ streamlines flow and simplify complicated packet processing in the network by clustering traffic into traffic classes according to the predefined QoS. This feature makes DiffServ lightweight and easy to implement.

Applications with similar traffic characteristics and performance requirements are mapped into DiffServ classes based on the end-to-end behavior requirements of the applications according to RFC 5127. Table 1 shows the service classes of DiffServ. Service class is used herein as an application type.

2.2. Ontology and Semantic Web Rule Language (SWRL). Ontology is a formal and explicit specification of shared conceptualization [18, 19]. The ontology is a core element of the semantic web that increases the quality of information search on the web by enabling machines to decipher and understand the data existing on the web without human involvement by assigning semantics to the data. The key elements of the ontology are the following:

(1) Classes: sets, collections, concepts, or types of objects

(2) Individuals: instances of objects

(3) Relations: ways in which classes and individuals can be related to one another

(4) Attributes: aspects, properties, features, characteristics, or parameters that can be associated with objects

SWRL [20, 21] is used to express rules in the semantic web. An SWRL rule consists of an inference relationship between antecedent and consequent.

2.3. Genetic Algorithms. Genetic algorithms (GAs) [22–24] are a class of global optimization algorithms first introduced in John Holland's book "Adaptation on Natural and Artificial Systems." A typical GA has the following execution process:

(1) Randomly initialize population (t)

(2) Determine the fitness of population (t)

(3) Loop

  (a) Select parents from population (t)

  (b) Perform crossover and mutation on parents to create offspring population

  (c) Determine the fitness of the combined population $(t+1)$

(4) Stop when the best individual is good enough

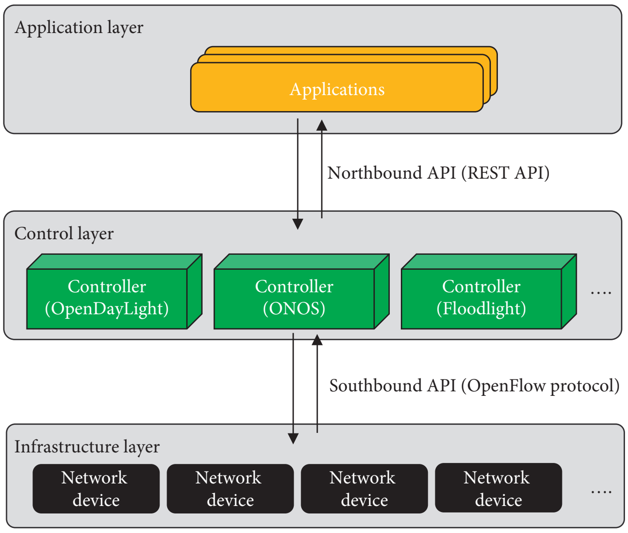


Figure 1: SDN layers.

Table 1: DiffServ classes [17].

Service class (application type)	QoS tolerance Packet loss	 Delay	 Jitter
Network control	Low	Low	Yes
Telephony	Very low	Very low	Very low
Signaling	Low	Low	Yes
Multimedia conferencing	Low-medium	Very low	Low
Real-time interactive	Low	Very low	Low
Broadcast video	Very low	Medium	Low
Multimedia streaming	Low-medium	Medium	Yes
Low-latency data	Low	Low-medium	Yes
OAM	Low	Medium	Yes
High-throughput data	Low	Medium-high	Yes
Standard	Not specified		
Low-priority data	High	High	Yes

Holland [22] suggested the notion of schemata for the convergence analysis of genetic algorithms. Schemata are bit patterns, which function as representatives of a set of binary strings. The population consists of a set of N binary strings of length L at time t, where N is the number of strings in the population, which contains the bit pattern, such as 100000, 110011, and 010010. Furthermore, H is called $o(H,t)$. Let us assume that function f has to be maximized. f is defined as overall binary strings of length L, and it is called fitness of the strings. Two parent strings from the current population are always selected for the creation of a new string. The probability that a parent string H_j will be selected from N strings H_1, H_2,. . ., H_N is shown in the following equation [23]:

$$p\left(H_j\right) = \frac{f\left(H_j\right)}{\sum_{j=1}^{N} f\left(H_j\right)}. \quad (1)$$

Strings with greater fitness are more likely to be selected than those with lesser fitness. Let $f\mu$ be the average fitness of all strings in the population, as shown in the following equation [23]:

$$f\mu = \frac{1}{N}\sum_{i=1}^{N} f\left(H_i\right). \quad (2)$$

The probability $p(H_j)$ can be written as $(f(H_j))/(Nf\mu)$.

The fitness function is used to assess the excellence of the individuals in the population. That is, it evaluates whether each individual should survive into the next generation. A surviving individual becomes a parent and creates a new generation through crossover or mutation. Crossover is used to create offspring by copying positions in two parents that do not overlap with each other, while mutation creates offspring by changing the information of an individual. For the new generation, the fitness function repeatedly evaluates whether the optimization goal has been reached to solve the problem. The optimal solution is derived by repeatedly evaluating the new generation with the fitness function to determine whether the optimization goal of solving the problem has been achieved.

3. Network Situation-Aware Framework (NSAF)

The main techniques of the approach for embodying the conceptual NSAF architecture and realizing the NSAF are presented in this section. The NSAF is located between the application and control layers (Figure 2). It acts as an intermediator that manages the QoS requirement of applications and controls the SDN controller. Controller dependencies are mitigated by separating the application and the controller.

In a conventional structure, in which the application is directly connected to a controller, the application has to configure an additional module for interconnection for each controller. However, in a separated architecture, the application only needs to consider the interface of the NSAF, and the connection to each controller has the advantage of being abstracted and managed as a common NSAF interface. The advantage of monitoring and controlling the SDN controller through the NSAF API is that one does not need to know the usage and function of various kinds of SDN controllers.

3.1. Requirements and Architecture of the NSAF. Table 2 presents the identified key requirements of the NSAF.

Figure 3 shows the NSAF architecture in terms of its key requirements. The architecture consists of five modules: interface manager, application-aware service-level agreement (SLA) manager, application manager, flow rule manager, and topology manager. It has resources, including application profile, network service-level agreement (NSLA), topology information, and route information.

The interface manager utilizes access points to control the SDN controller through the NSAF by abstracting and mapping the REST API provided by the SDN controller. The application manager collects and manages NSLA information, which represents the application's detailed profile information and the network SLA of the application. The application-aware SLA manager determines whether or not the network QoS of the path used by the current application violates the NSLA and calculates new paths. The flow rule manager creates and manages rules that control routing paths. The topology manager collects, monitors, and manages network status and traffic information.

3.2. NSAF Execution. The NSAF execution covers application QoS assurance, including application profile and QoS registration, network status monitoring, identification of violations, and routing changes (Figure 4). It is divided into four phases: application registration, network status monitoring, violation detection, and routing control.

The service type of the application, basic information, and NSLA, which is the network QoS requirement information of the application, is registered in the application registration phase. Network status monitoring requests the API provided by the SDN controller to collect traffic and resource information for the connected network topology. The network status is analyzed and visualized based on the collected information. The violation detection process determines whether or not the route used by the current application violates the NSLA of the application based on the collected traffic and resource information and the application of NSLA information. Routing control is conducted if the NSLA is determined to be violated (i.e., the route currently being used by the application does not guarantee QoS). The routing control process constructs a path that satisfies the application QoS.

3.3. Application Registration. The RFC 5127 "Aggregation of DiffServ Service Class" standardized by the Internet Engineering Task Force [17] presents 12 application types: network control, telephony, signaling, multimedia conferencing, real-time interactive, multimedia streaming, broadcast video, low-latency data, OAM, high-throughput data, standard, and low-priority data. We used DiffServ's QoS tolerance as a condition for satisfying the application performance. We also analyzed the QoS factors defined in DiffServ, and the QoS metrics measurable by the SDN controller [25]. Based on this, we defined packet loss, bandwidth, delay, and jitter as the application quality factors.

As shown in Figure 5, the application profile is registered in the application registration phase. This profile consists of the application type information and the NSLA, which describes the application's network QoS requirements.

3.4. Network Monitoring. Collecting information on the current network situation is necessary in ensuring the

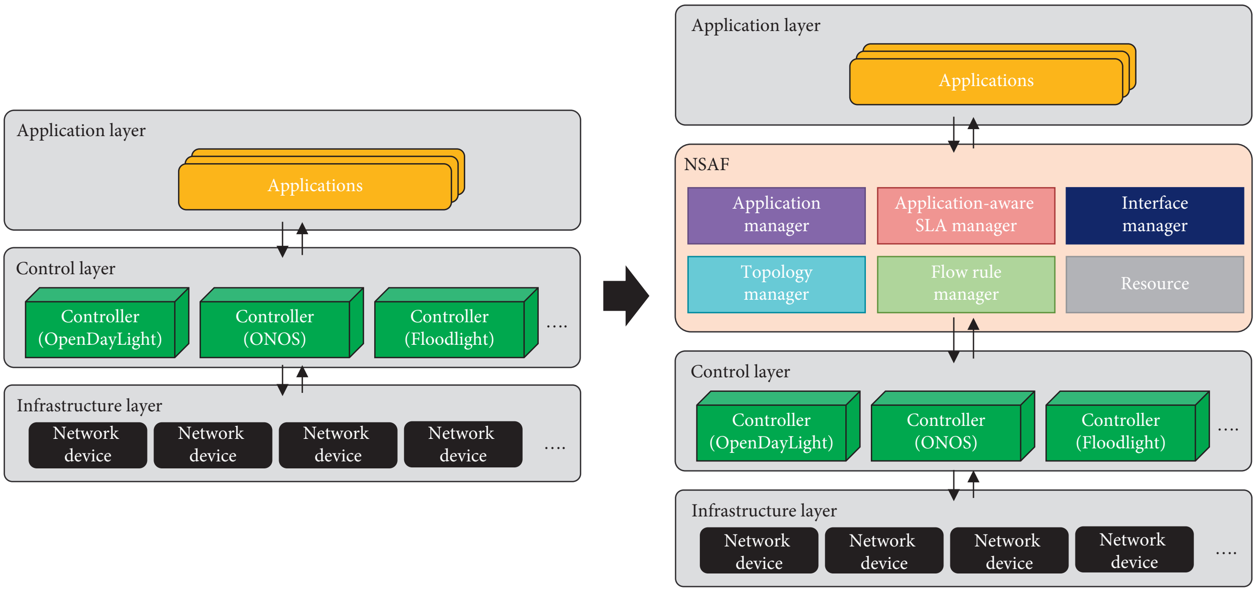


Figure 2: SDN vs extended SDN with the NSAF.

Table 2: Key requirements of the NSAF.

ID	Requirements	Description
R-01	Application profile management	Collects and analyzes the application type and application requirements
R-02	NSLA management	Manages network service-level agreement (SLA) for applications
R-03	Network status monitoring	Monitors node and link information in the network topology
R-04	Application-specific situation awareness	Ascertains QoS violation of the network route using application profile and NSLA
R-05	Application-specific path calculation	Calculates and determines path according to the network status

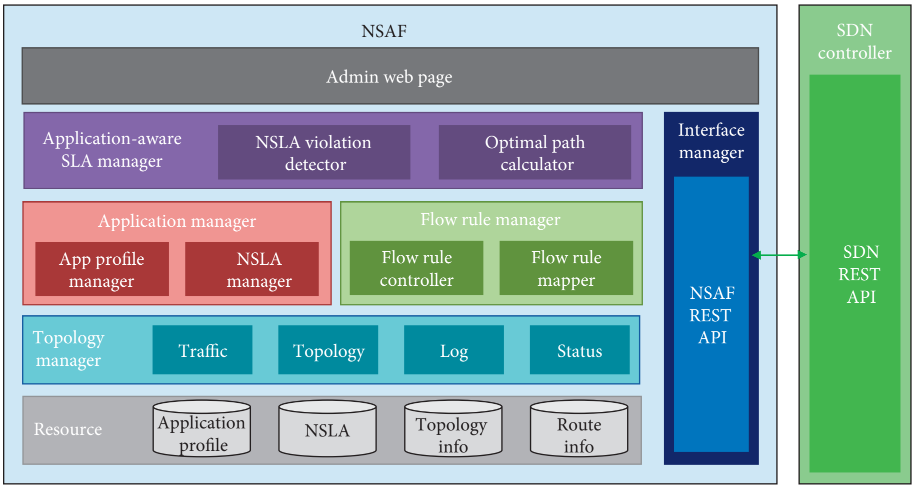


Figure 3: NSAF architecture.

application QoS. The proposed NSAF uses the REST API to determine the status of the SDN controller and collect and update the network status information in real time.

Figure 6 shows the topology model defined for collecting and managing the network status information in the NSAF.

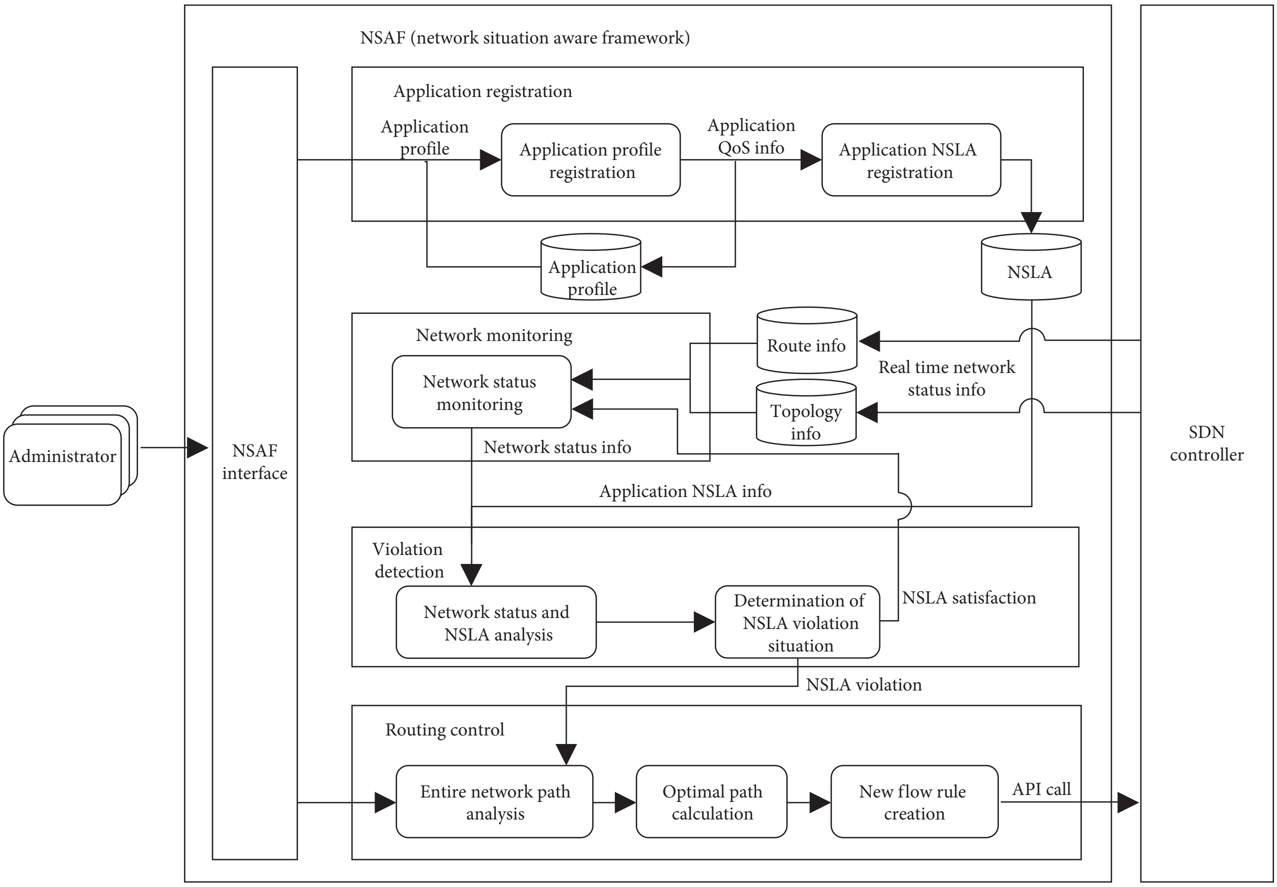


Figure 4: Execution of the NSAF.

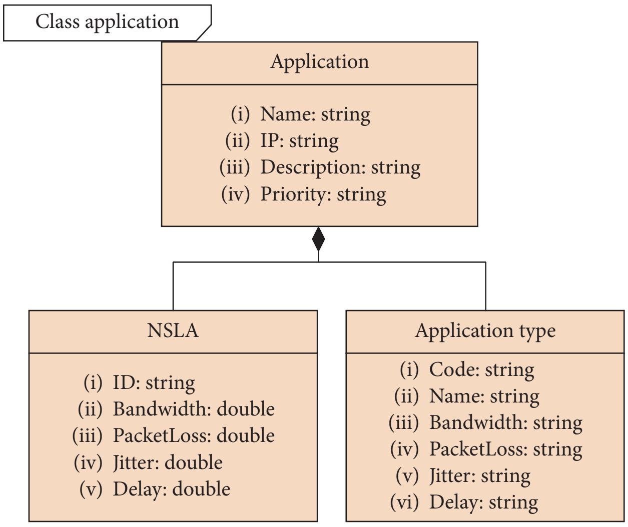


Figure 5: Application profile.

The topology for collecting the network status information models consists of at least one node and a link. Nodes are used to collect information on network devices, such as switches and routers. A link is defined to manage the connection between nodes for routing path configuration. The flow table is defined to manage information for network routing control.

3.5. Violation Detection. Two types of violations are defined to determine the application QoS violation situation. A hard fail occurs if the link between two nodes is broken. In contrast, a soft fail occurs if the connection between the nodes is not a problem, but the NSLA of the application is not satisfied. Table 3 shows the types of violations and details of the violations by type.

The ontology is applied to detect violations. Figure 7 shows the internal structure used to determine whether or not a violation has occurred. As illustrated in Figure 7, the application profile information and the network status information are sent to the NSLA Violation Controller. Context information (e.g., network QoS) and traffic information that can be measured for a link connecting nodes (e.g., switches and routers constituting the entire network topology) are extracted based on the collected information. The ontology manager reflects the extracted context information on the ontology. The ontology is loaded into the reasoning engine, and the violation situation is identified. The violation detect manager collects the violation and sends it to the monitoring manager through the NSLA violation controller to notify the NSLA of the violation of the current application.

We constructed the ontology model by abstracting the entities of the elements derived from the application registration and network state. We then defined the relationships between them. Table 4 shows an example of the attributes for the identified class, defined attribute, and data

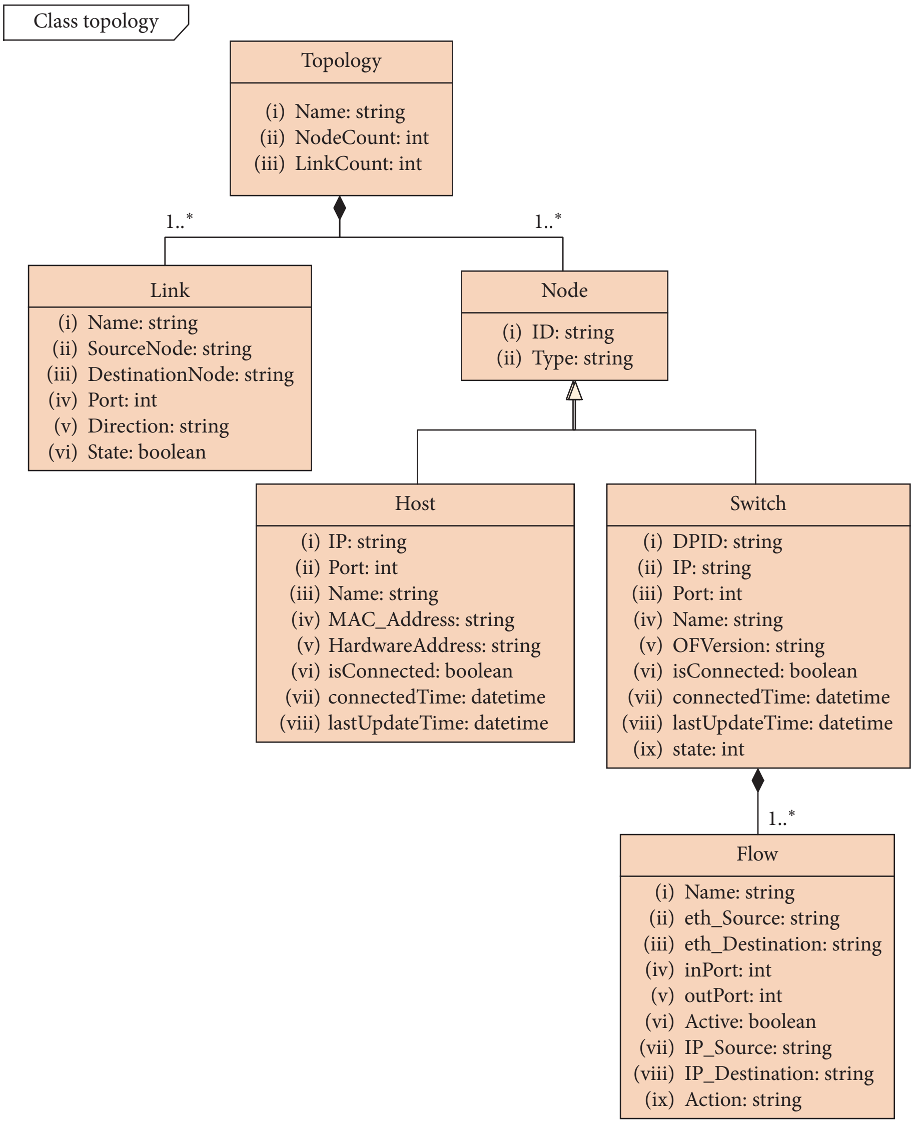


Figure 6: Topology model for the network status monitoring.

Table 3: Defined violation types.

Violation type	Details of violation type	Description
Hard fail	Node violation	The topology node is disconnected
	Link violation	The topology link is disconnected
Soft fail	Bandwidth violation	The application does not have the required bandwidth to execute
	Packet loss violation	The maximum packet loss for the application is exceeded
	Delay violation	The maximum delay for the application is exceeded
	Jitter violation	The application’s jitter is violated

property elements. Figure 8 presents the ontology metamodel constructed using the elements analyzed in Table 4.

3.6. Routing Control. The network QoS of the application was ensured by applying two path algorithms to derive different paths when the defined NSLA is violated. First, the route algorithm was used to calculate the digest algorithm, which is the most used in the path algorithm, by using the QoS classified herein as the cost. In addition, a number of alternative paths can be derived by applying a GA to find optimal solutions for multiobjective problems.

The path calculation process using Dijkstra is elaborated below:

(5) Calculate the T-score using the average and standard deviation values for four kinds of application QoS. The T-score is used to set the reference value because

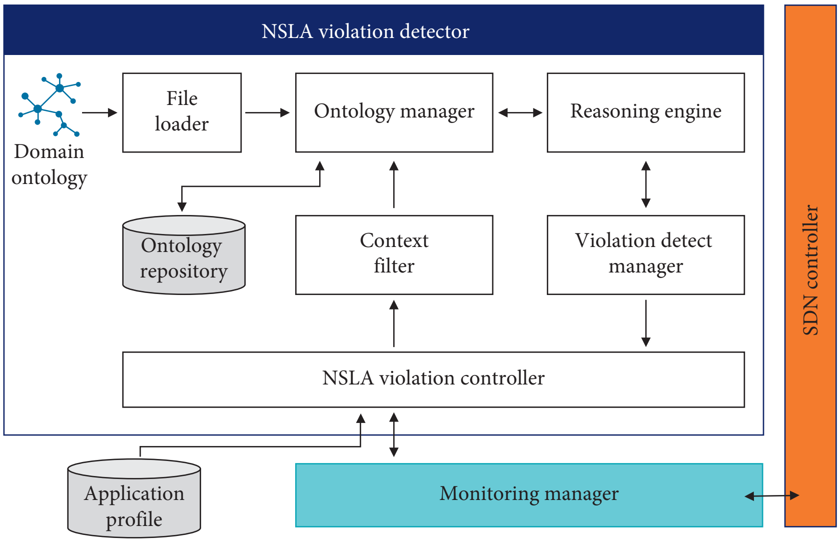


Figure 7: Internal structure used for the violation detection.

Table 4: Identified class, attribute, and data property information.

Class	Attribute	Data property	Range	Description
Application	ID, name, admin, IP address (v4, v6), application type	hasName hasAdmin hasIPAddress hasDescription	String String String String	Application name Application administrator Application IP address Application description
NSLA	Application ID, bandwidth, packet loss, delay, jitter	hasApplicationID hasBandwidth hasPacketloss hasDelay hasJitter	String Int Double Double Double	Application ID Bandwidth Packet loss Delay time Variation of delay time
Violation	Name, occurred time, target application, reason	hasName hasOccurredTime hastargetapplicaton hasreason	String Datetime String String	Violation name Time of occurrence Target application Reason for occurrence

each QoS value unit is different. Equation (3) is the T-score formula:

$$\mathrm{T-score}\,(\mathrm{QoS}) = \left(\frac{(\mathrm{QoS}) - \mathrm{average}\,(\mathrm{QoS})}{\mathrm{standard\ deviation}\,(\mathrm{QoS})} \times 20\right) + 100. \tag{3}$$

(6) Calculate each QoS cost by multiplying the calculated T-score by the weight according to the application type. The final cost is calculated by summing each QoS cost of the computed node, as presented in the following equation:

$$\begin{gathered}\mathrm{node\ cost} = \mathrm{T-score} \times \mathrm{weight}_{\mathrm{QoS}},\\ \mathrm{final\ cost} = \sum_{i=1}^{4} \mathrm{node\ cost}.\end{gathered} \tag{4}$$

(7) The final good-quality path is calculated as the minimum cost.

The GA performs the following application process:

(1) The topology of each of the four QoS choices is composed of adjacency matrices [26]. The path from the source to the destination is randomly selected and expressed as the initial population (1: selected, 0: not selected). Let T be the topology and S be the topology nodes. An initial population is formed by a bit pattern re2presented by 0 and 1 according to Equation 5:

$$\forall S_i \in T = \begin{cases} 1, & S_i \in \text{routing path}, \\ 0, & \text{else}. \end{cases} \tag{5}$$

(2) A fitness assessment is conducted for the initial population. The criterion of the fitness evaluation is the measurement of whether or not the value of the node QoS element of each path exceeds the NSLA of the application. The fitness score increases by 1 point if the QoS value of the node satisfies the NSLA. Equation (6) shows the fitness evaluation case of the NSLA bandwidth:

$$\forall \text{ pair of node } (s_i, s_j), \text{ where } i \neq j \text{ and directly connected,}$$

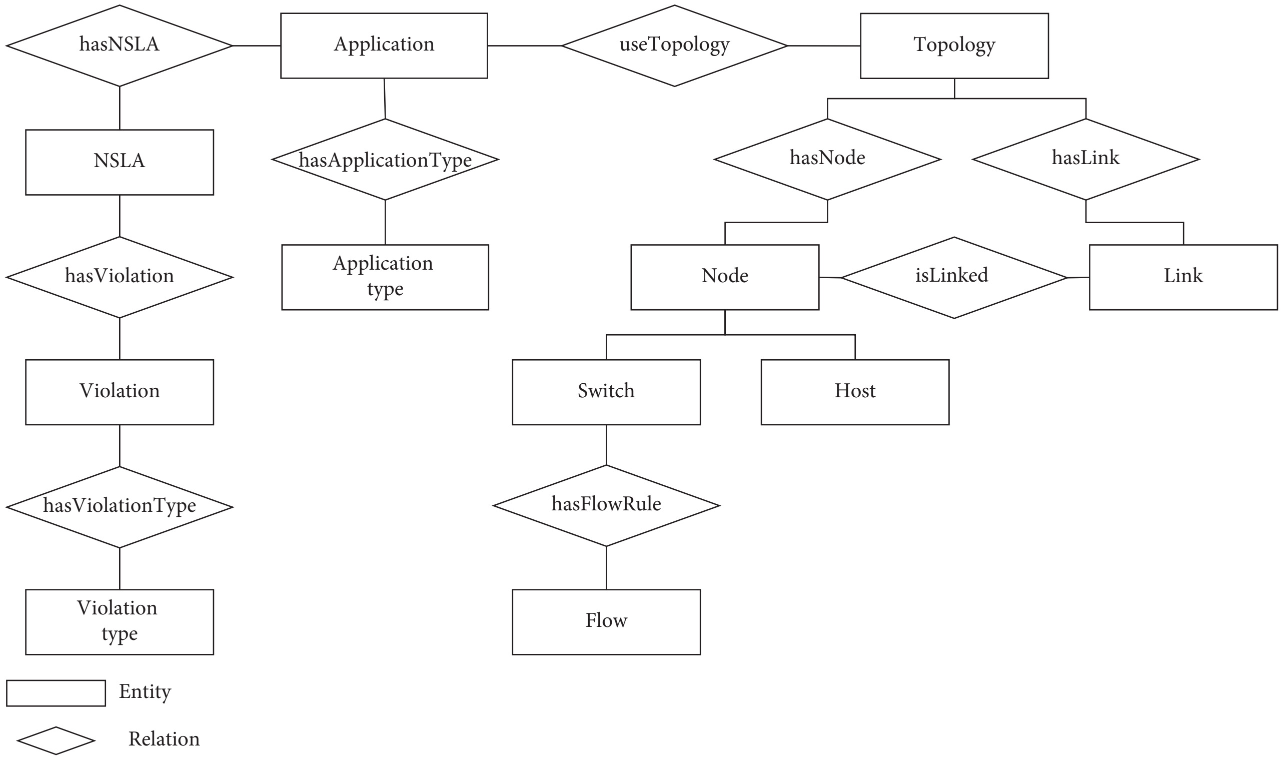


Figure 8: Ontology metamodel.

$$\begin{aligned} \text{then bandwidth QoS}_{ij} &= \begin{cases} 1, & \text{if value of bandwidth QoS}_{ij} \\ & \geq \text{value of NSLA (bandwidth)}, \\ -1, & \text{else,} \end{cases} \\ \text{fitness evauation (bandwidth QoS)} &= \sum \text{bandwidth QoS}_{ij}. \end{aligned} \tag{6}$$

Figure 9 depicts an example in which the fitness evaluation formula is applied to determine if the bandwidth corresponding to the initial generation of the solution meets the application's NSLA.

(3) The scores from the evaluation of the fitness of the path selected as the initial generation are then divided by the number of hops and calculated as the final fitness score. Equation (7) is the fitness evaluation formula. Figure 10 shows an example of the final fitness score obtained using

$$\text{QoS fitness} = \frac{\sum_{h=1}^{\text{Hop Count}} \text{path (QoS)}}{\text{hop count}}. \tag{7}$$

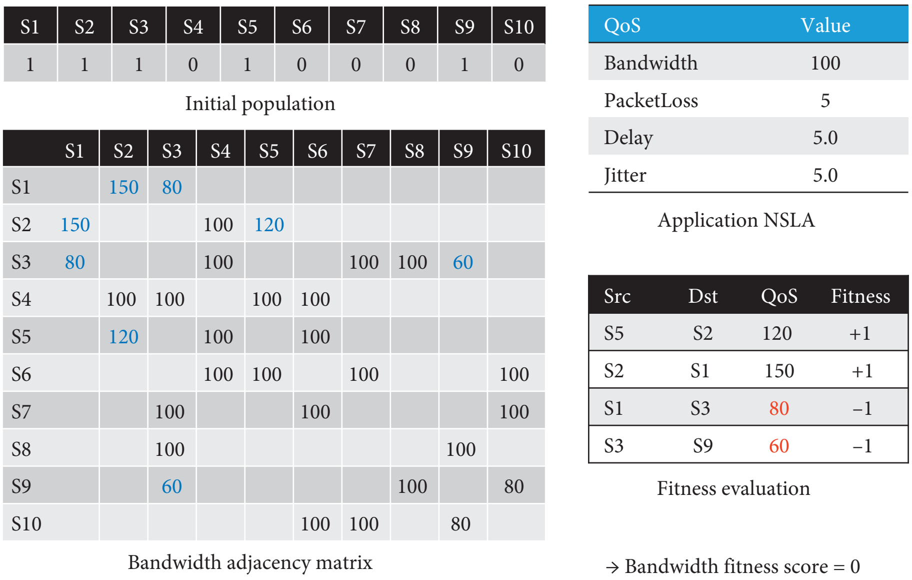
(4) The top 20% of the early generations with the best fitness evaluation is judged as the dominant gene and mated or mutated to the next generation. We determined herein whether or not crossing is possible. If the gene cannot be crossed, it generates a child gene through mutation. Figure 11 shows the concept of crossover and mutation.

(5) Steps 2 to 4 are repeated on the new genes until a path with four points of target fitness score satisfying the NSLA satisfies all the four QoS factors and derives the paths.

4. Case Study of the NSAF

A prototype was implemented to evaluate the proposed NSAF and its process using Floodlight controller and Mininet [27]. Figure 12 shows the application registration. Figure 13 illustrates the network status monitoring of the NSAF that reflects the model and element described in Sections 3.3 and 3.4.

Figure 12 shows the application registration interface. The data to be input on the interface comprise the application manager member information (ID, password, manager name, and number to be contacted) and the application profile information (application name, IP address, application type, and application description). The NSLA's criteria will be set depending on the application type. The QoS criteria for each application type are based on the DiffServ standard.



S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
1	1	1	0	1	0	0	0	1	0

Initial population

	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
S1		150	80							
S2	150			100	120					
S3	80			100			100	100	60	
S4		100	100		100	100				
S5		120		100		100				
S6				100	100		100			100
S7			100			100				100
S8			100						100	
S9			60					100		80
S10						100	100		80	

Bandwidth adjacency matrix

QoS	Value
Bandwidth	100
PacketLoss	5
Delay	5.0
Jitter	5.0

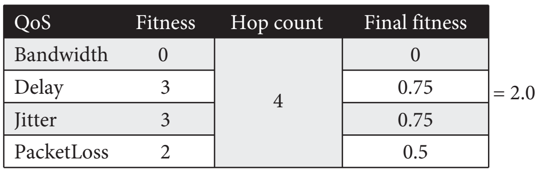
Application NSLA

Src	Dst	QoS	Fitness
S5	S2	120	+1
S2	S1	150	+1
S1	S3	80	−1
S3	S9	60	−1

Fitness evaluation

→ Bandwidth fitness score = 0

Figure 9: Examples of initial population fitness evaluations.



QoS	Fitness	Hop count	Final fitness
Bandwidth	0	4	0
Delay	3		0.75
Jitter	3		0.75
PacketLoss	2		0.5

= 2.0

Figure 10: Final fitness score.

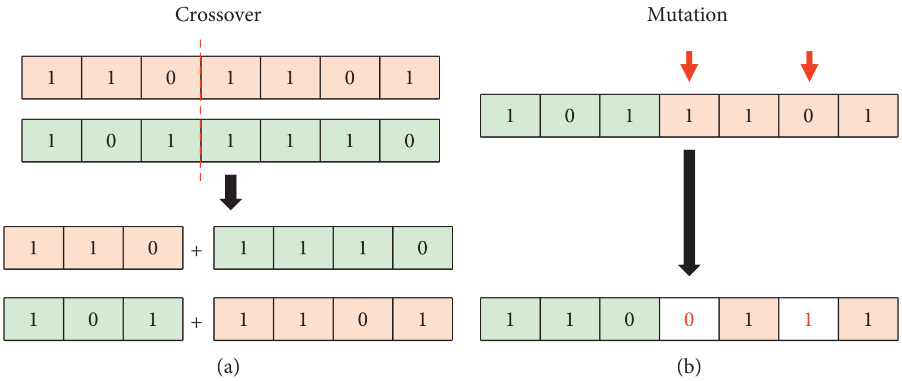
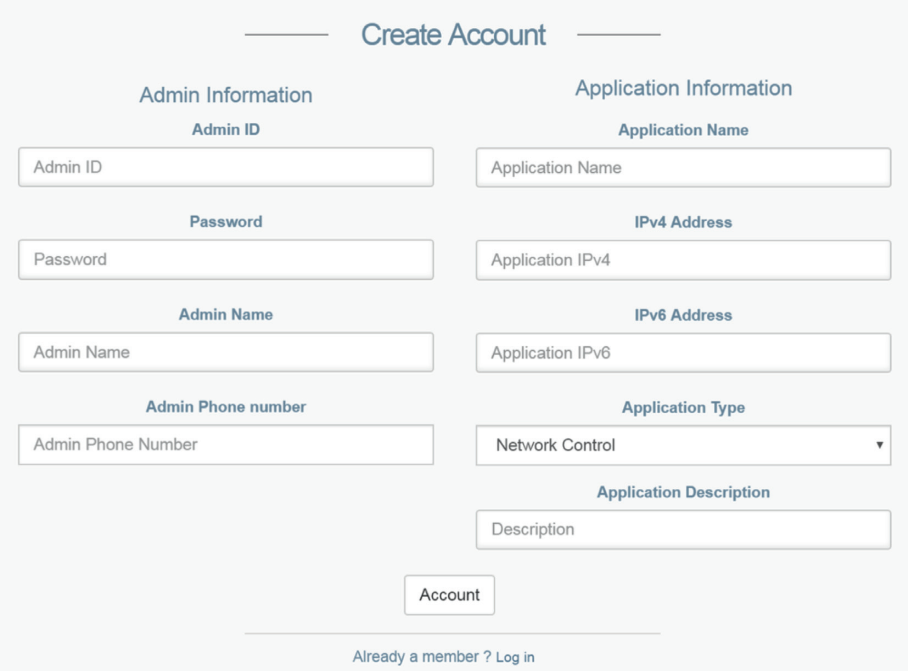


Figure 11: (a) Crossover and (b) mutation.

Figure 13 depicts that the NSAF Network Status Monitoring Interface connects and calls the Floodlight controller's data. It shows the structure of the switches and hosts that make up the topology at a glance and displays simple information on the top right by selecting each switch and host. The network monitoring provided by the NSAF supports not only topology monitoring but also information about the switches and the hosts connected to the SDN controller and dashboard monitoring that visually shows real-time network traffic.

An ontology model for detecting the NSLA violation was constructed using Protégé 5.0. Figure 14 shows the implemented ontology.

Based on the NSAF ontology, the SWRL rule shown in Table 5 was created to test the NSLA violation status defined in Section 3.5. Figure 15 shows the test application multimedia streaming application called PNU Tube. The NSLA information for the application is as follows: a bandwidth of 100 Mbps, a packet loss of 20 bytes, a delay under 1.5 ms, and a jitter under 0.5 ms.



Create Account

Admin Information

Admin ID

Admin ID

Password

Password

Admin Name

Admin Name

Admin Phone number

Admin Phone Number

Application Information

Application Name

Application Name

IPv4 Address

Application IPv4

IPv6 Address

Application IPv6

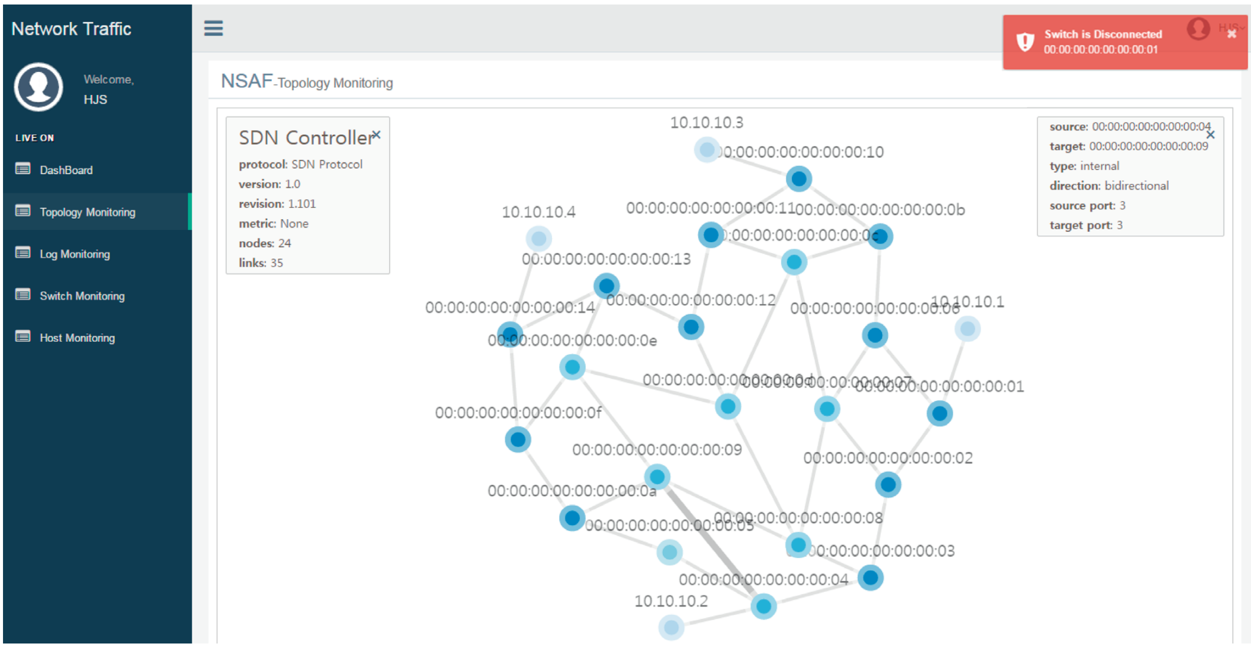
Application Type

Network Control ▾

Application Description

Description

Figure 12: NSAF application registration interface.



Switch is Disconnected
00:00:00:00:00:00:00:01

SDN Controller

protocol: SDN Protocol
version: 1.0
revision: 1.101
metric: None
nodes: 24
links: 35

source: 00:00:00:00:00:00:00:04
target: 00:00:00:00:00:00:00:09
type: internal
direction: bidirectional
source port: 3
target port: 3

Figure 13: NSAF Network Status Monitoring Interface.

The virtual topology structure was composed of 10 switches such that the NSLA violation status can be determined. The path that the application used to send packets from H1 to destination H2 was S5 ⟶ S6 ⟶ S7 ⟶ S3 ⟶ S9. Using the application profile information and the topology QoS value, we confirmed that a soft fail situation occurred by applying the SWRL rules, as shown in the example in Table 5. When we checked the result of discrimination of the NSLA violation, three soft fail-type violations were found to have occurred. The QoS value for the S3 bandwidth, delay, and packet loss did not satisfy the NSLA; hence, bandwidth, delay, and packet violations occurred.

The following is an example of the bandwidth violation among the SWRL rules defined in Table 5: [Application (?A)] denotes the presence of an application called “a.” [hasNSLA (?a, ?n)] indicates that “a” has an NSLA of “n.” [hasBandwidth (?n, ?nb)] has a bandwidth value of “nb.” [useTopology (?a, ?t)] uses a topology named “t,” while hasNode (?t, ?s)] has a node named “s.” [hasBandwidth (?s, ?sb)] indicates that node “s” has a bandwidth value of “sb.” [swrl: greaterThan (?nb, ?

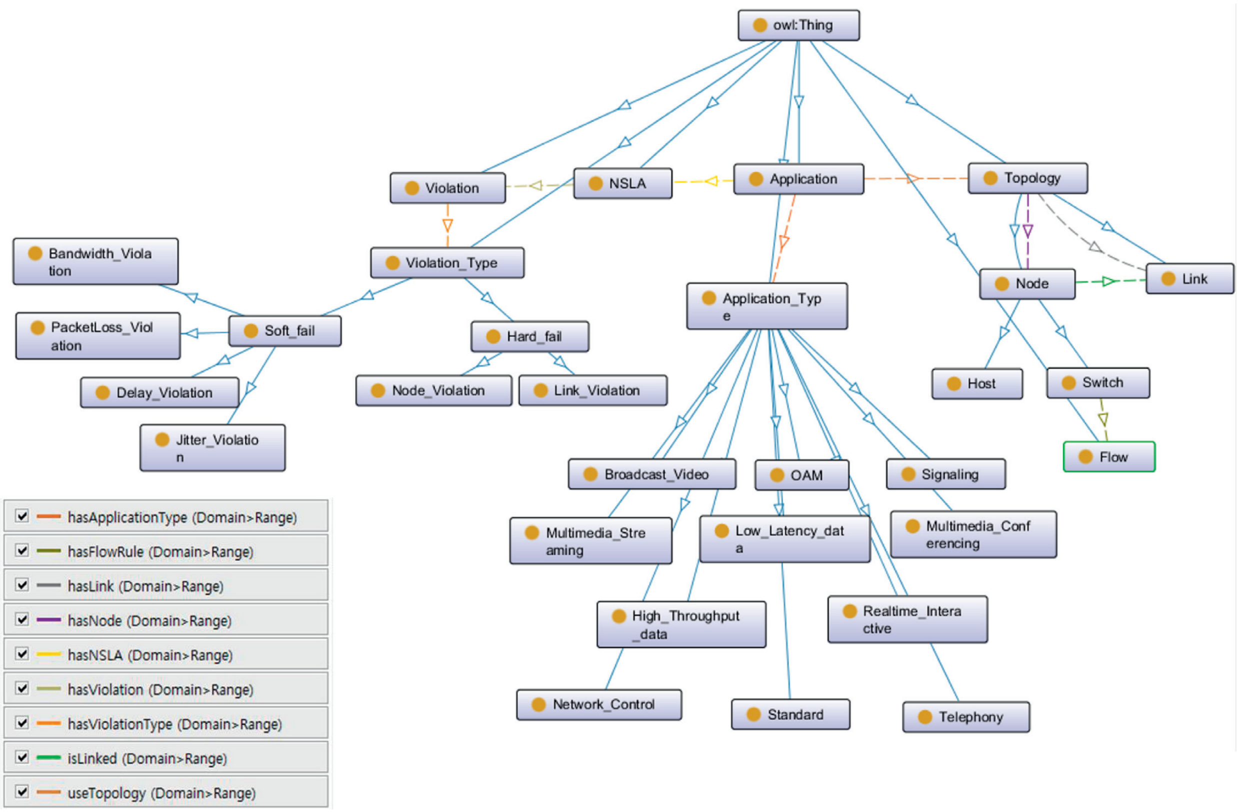


Figure 14: Implemented NSAF ontology.

Table 5: Example SWRL rule for the violation detection.

Violation type	Details of the violation type	SWRL rule
Hard fail	Node violation	Application (?a)ˆuseTopology (?a, ?t)ˆhasNode (?t, ?n)ˆhasState (?n, false) -> Node_Violation (?a)
Soft fail	Bandwidth violation	Application (?a)ˆhasNSLA (?a, ?n)ˆhasBandwidth (?n, ?nb)ˆuseTopology (?a, ?t)ˆhasNode (?t, ?s)ˆhasBandwidth (?s, ?sb)ˆswrlb:greaterThan (?nb, ?sb) -> Bandwidth_Violation (?a)

sb)] is a syntax that compares the NSLA bandwidth value "nb" with the bandwidth value of the topology node. Therefore, it is deduced that [->Bandwidth_Violation (?a)], that is, application "a" occurred as a bandwidth violation.

As an example, we tested the path that guarantees the network QoS by setting S2 as the starting node and S28 as the destination node in the virtual environment composed of 30 nodes (Figure 16) and executing Dijkstra's algorithm and the GA presented in Section 3.6.

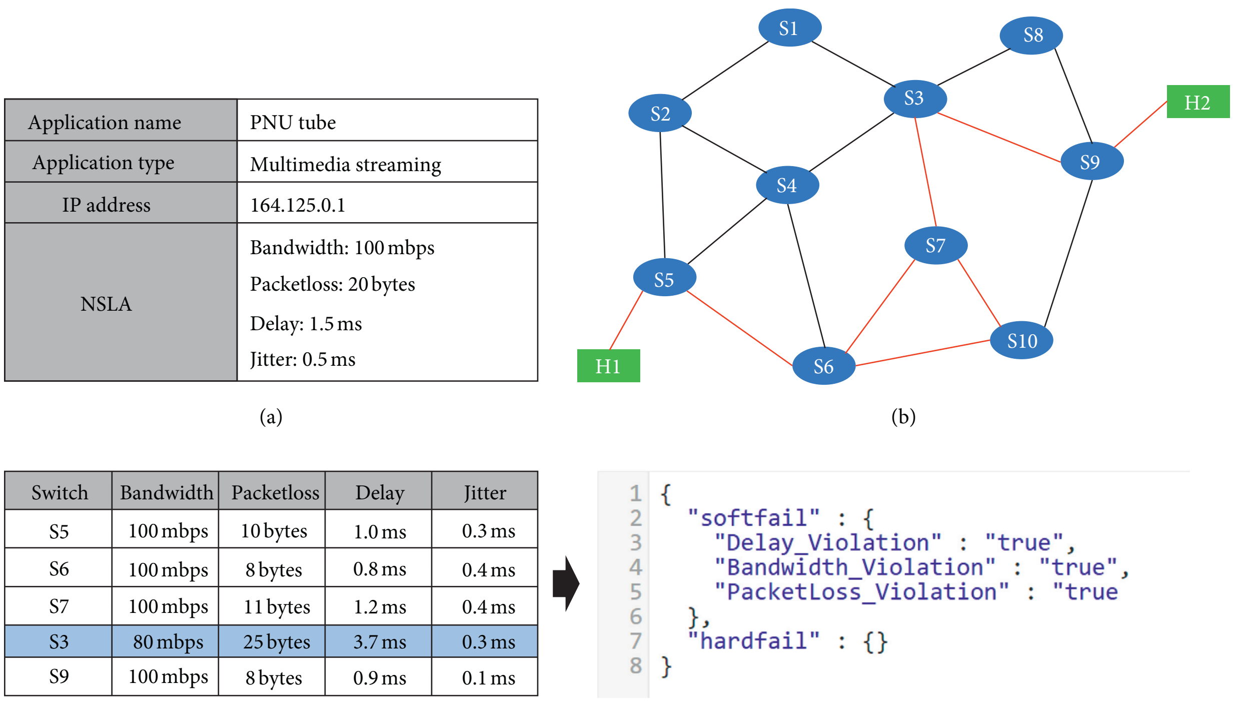
Figure 17 presents an example of the logarithm of the values calculated in the network configuration environment of Figure 18 on the NSAF according to Dijkstra's algorithm, which was presented in Section 3.6.

As shown in the log in Figure 17, the individual T-scores for each QoS were calculated, and the node with the lowest cost was selected by summing the node cost. Finally, the path should be {S2, S7, S12, S13, S18, S23, S28}.

Figure 19 shows the result of the GA application. Figure 18 illustrates the calculated path (i.e., paths 1, 2, and 3). As shown in the results of the execution log of the NSAF on the left side, we performed crossover and mutation on the selected initial solution and confirmed that a number of alternative paths with a fitness score of 4 were derived.

5. Related Work and Discussion

Various studies have been conducted as regards support monitoring in SDN environments. Isolani et al. [28] proposed an SDN interactive manager to monitor, visualize, and configure SDN with the administrator in the management loop. Their proposed system includes a management plane that is responsible for managing elements in other SDN layers, such as monitoring device status and allocating resources. Their SDN interactive manager sits in this proposed



Application name	PNU tube
Application type	Multimedia streaming
IP address	164.125.0.1
NSLA	Bandwidth: 100 mbps Packetloss: 20 bytes Delay: 1.5 ms Jitter: 0.5 ms

(a)

(b)

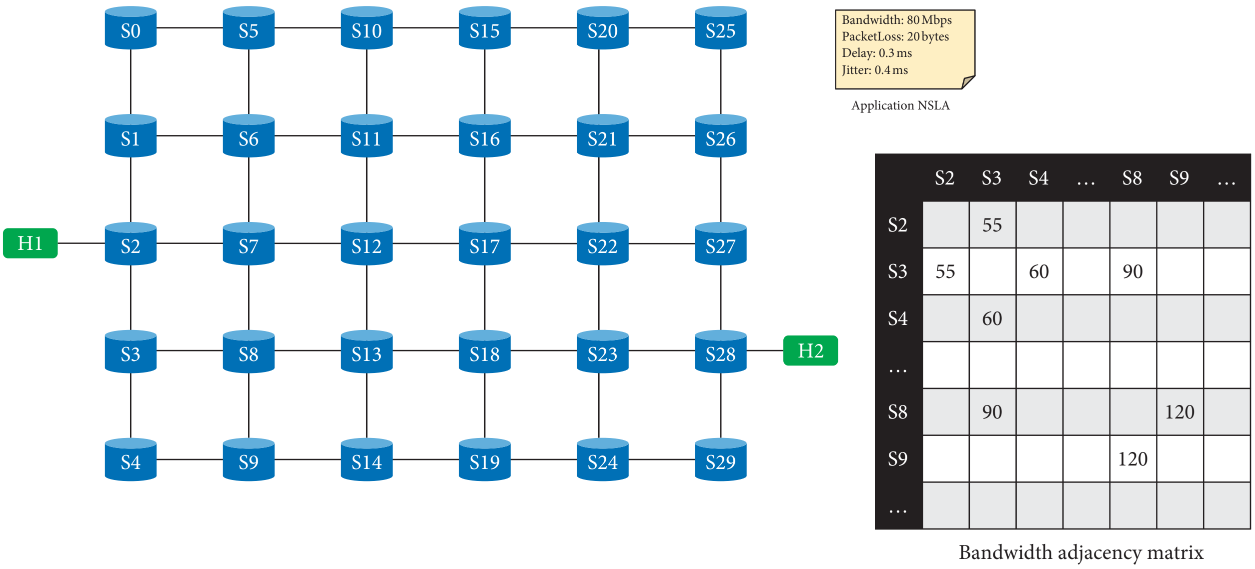
Switch	Bandwidth	Packetloss	Delay	Jitter
S5	100 mbps	10 bytes	1.0 ms	0.3 ms
S6	100 mbps	8 bytes	0.8 ms	0.4 ms
S7	100 mbps	11 bytes	1.2 ms	0.4 ms
S3	80 mbps	25 bytes	3.7 ms	0.3 ms
S9	100 mbps	8 bytes	0.9 ms	0.1 ms

(c)

```
{
  "softfail" : {
    "Delay_Violation" : "true",
    "Bandwidth_Violation" : "true",
    "PacketLoss_Violation" : "true
  },
  "hardfail" : {}
}
```

(d)

Figure 15: Violation detection: (a) application information; (b) routing path in use (red line); (c) routing table information; (d) violation detection result.



Bandwidth: 80 Mbps
PacketLoss: 20 bytes
Delay: 0.3 ms
Jitter: 0.4 ms

Application NSLA

	S2	S3	S4	…	S8	S9	…
S2		55					
S3	55		60		90		
S4		60					
…							
S8		90				120	
S9					120		
…							

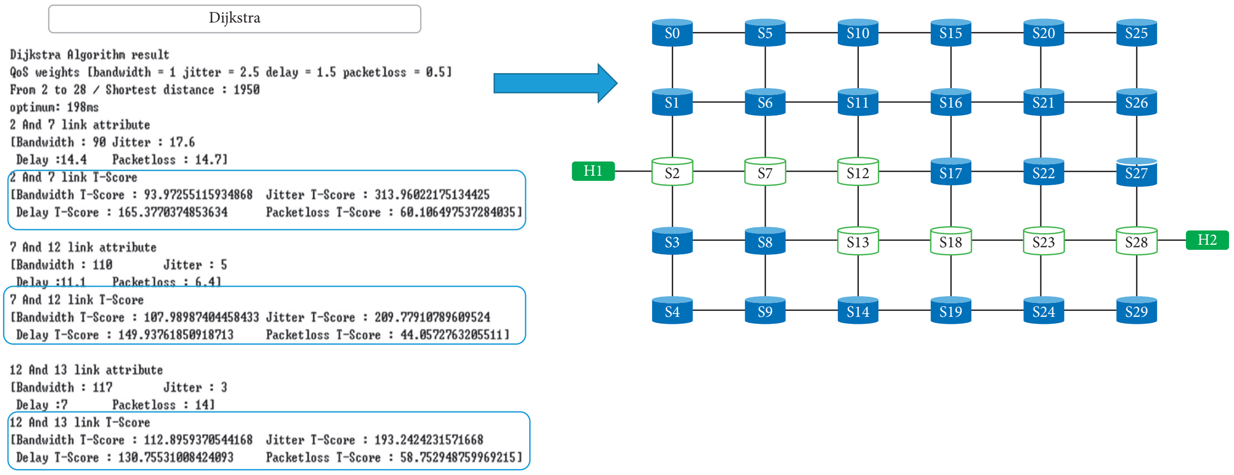
Bandwidth adjacency matrix

Figure 16: Network environment and application NSLA.

plane and comprises three main components, namely, monitoring manager, visualization manager, and configuration manager.

Raumer et al. [29] proposed a flow sampling application that receives information from the analyzer module, called MonSamp, for QoS monitoring. In their proposed system, the monitor is directly connected to SDN and receives part of the network traffic information. MonSamp can reactively decide to reduce the number of flows using the received information. It is an SDN application for extraction and direct sampling of the network traffic.

Chowdhury et al. [30] proposed a network statistics collection framework called PayLess. The proposed framework operates on the top of the control layer and uses the northbound API of the controller. It consists of a request interpreter, a scheduler, a switch selector, and an aggregator and data store. The request interpreter translates high-level primitives expressed by the application to flow-level

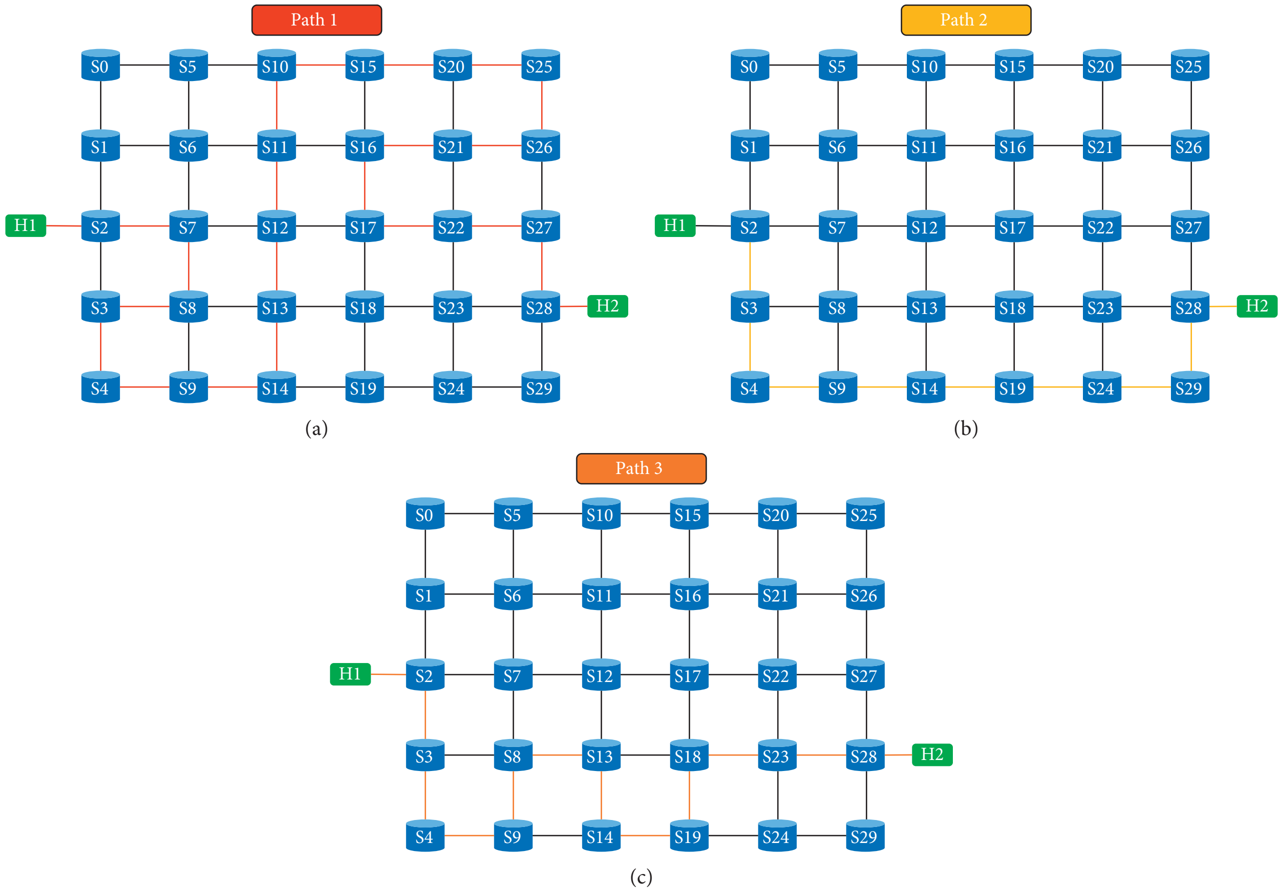


```
Dijkstra Algorithm result
QoS weights [bandwidth = 1 jitter = 2.5 delay = 1.5 packetloss = 0.5]
From 2 to 28 / Shortest distance : 1950
optimum: 198ms
2 And 7 link attribute
[Bandwidth : 90 Jitter : 17.6
 Delay :14.4    Packetloss : 14.7]
2 And 7 link T-Score
[Bandwidth T-Score : 93.97255115934868  Jitter T-Score : 313.96022175134425
 Delay T-Score : 165.3770374853634      Packetloss T-Score : 60.106497537284035]

7 And 12 link attribute
[Bandwidth : 110        Jitter : 5
 Delay :11.1    Packetloss : 6.4]
7 And 12 link T-Score
[Bandwidth T-Score : 107.98987404458433 Jitter T-Score : 209.77910789609524
 Delay T-Score : 149.93761850918713     Packetloss T-Score : 44.0572763205511]

12 And 13 link attribute
[Bandwidth : 117        Jitter : 3
 Delay :7       Packetloss : 14]
12 And 13 link T-Score
[Bandwidth T-Score : 112.8959370544168  Jitter T-Score : 193.2424231571668
 Delay T-Score : 130.75531008424093     Packetloss T-Score : 58.752948759969215]
```

Figure 17: Dijkstra's path calculation results.



(a)

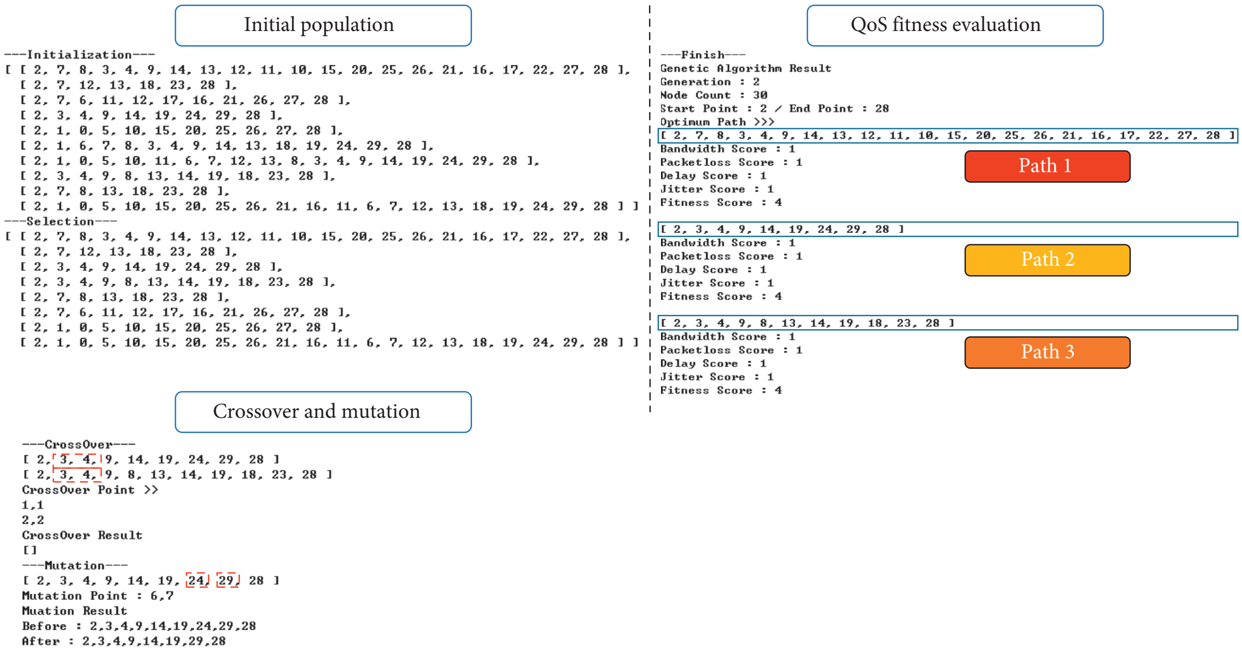
(b)

(c)

Figure 18: Path results.

primitives. The scheduler schedules polling of switches. The switch selector selects switches for the statistics collection. The aggregator and data store collects raw data from the selected switches and stores these raw data in the data store.

As outlined earlier, an element can be used to monitor the application QoS, but the focus of this element is on monitoring the network resources in the SDN environment, such as visualizing the current network status and providing the traffic statistics information, rather than the application QoS information. To the best of our knowledge, no approach that guarantees the QoS of the overall application from application QoS registration to the calculation of the QoS guaranteed path to guarantee the QoS of the application has yet been proposed.



```
---Initialization---
[ [ 2, 7, 8, 3, 4, 9, 14, 13, 12, 11, 10, 15, 20, 25, 26, 21, 16, 17, 22, 27, 28 ],
  [ 2, 7, 12, 13, 18, 23, 28 ],
  [ 2, 7, 6, 11, 12, 17, 16, 21, 26, 27, 28 ],
  [ 2, 3, 4, 9, 14, 19, 24, 29, 28 ],
  [ 2, 1, 0, 5, 10, 15, 20, 25, 26, 27, 28 ],
  [ 2, 1, 6, 7, 8, 3, 4, 9, 14, 13, 18, 19, 24, 29, 28 ],
  [ 2, 1, 0, 5, 10, 11, 6, 7, 12, 13, 8, 3, 4, 9, 14, 19, 24, 29, 28 ],
  [ 2, 3, 4, 9, 8, 13, 14, 19, 18, 23, 28 ],
  [ 2, 7, 8, 13, 18, 23, 28 ],
  [ 2, 1, 0, 5, 10, 15, 20, 25, 26, 21, 16, 11, 6, 7, 12, 13, 18, 19, 24, 29, 28 ] ]
---Selection---
[ [ 2, 7, 8, 3, 4, 9, 14, 13, 12, 11, 10, 15, 20, 25, 26, 21, 16, 17, 22, 27, 28 ],
  [ 2, 7, 12, 13, 18, 23, 28 ],
  [ 2, 3, 4, 9, 14, 19, 24, 29, 28 ],
  [ 2, 3, 4, 9, 8, 13, 14, 19, 18, 23, 28 ],
  [ 2, 7, 8, 13, 18, 23, 28 ],
  [ 2, 7, 6, 11, 12, 17, 16, 21, 26, 27, 28 ],
  [ 2, 1, 0, 5, 10, 15, 20, 25, 26, 27, 28 ],
  [ 2, 1, 0, 5, 10, 15, 20, 25, 26, 21, 16, 11, 6, 7, 12, 13, 18, 19, 24, 29, 28 ] ]
```

```
---CrossOver---
[ 2, 3, 4, 9, 14, 19, 24, 29, 28 ]
[ 2, 3, 4, 9, 8, 13, 14, 19, 18, 23, 28 ]
CrossOver Point >>
1,1
2,2
CrossOver Result
[]
---Mutation---
[ 2, 3, 4, 9, 14, 19, 24, 29, 28 ]
Mutation Point : 6,7
Muation Result
Before : 2,3,4,9,14,19,24,29,28
After : 2,3,4,9,14,19,29,28
```

```
---Finish---
Genetic Algorithm Result
Generation : 2
Node Count : 30
Start Point : 2 / End Point : 28
Optimum Path >>>
[ 2, 7, 8, 3, 4, 9, 14, 13, 12, 11, 10, 15, 20, 25, 26, 21, 16, 17, 22, 27, 28 ]
Bandwidth Score : 1
Packetloss Score : 1
Delay Score : 1
Jitter Score : 1
Fitness Score : 4

[ 2, 3, 4, 9, 14, 19, 24, 29, 28 ]
Bandwidth Score : 1
Packetloss Score : 1
Delay Score : 1
Jitter Score : 1
Fitness Score : 4

[ 2, 3, 4, 9, 8, 13, 14, 19, 18, 23, 28 ]
Bandwidth Score : 1
Packetloss Score : 1
Delay Score : 1
Jitter Score : 1
Fitness Score : 4
```

Figure 19: Genetic path calculation results.

Table 6: Execution time of the NSAF.

Framework element	Execution time
Collecting network status information	Collect network status information from the SDN controller every 10 s
Determining application QoS violations	Perform within 0.4 s to derive results
Route discovery to meet application QoS	Dijkstra's algorithm: calculate one path within 0.3 s; genetic algorithm: derive various paths satisfying the QoS within 5 s

The SDN controller herein was controlled using the NSAF without the need for expert knowledge of SDN. Moreover, application-aware routing was supported to improve usability. The NSAF operated between the control and application layers; hence, reusability can be improved by reusing the architecture even if an SDN controller integration environment is introduced in the future. The violation was identified through the ontology model; hence, in the case of another new violation situation being added, it can be simply discriminated by defining an associated SWRL rule. Extensibility can, thus, be increased.

We implemented the NSAF prototype that guarantees the application QoS and verified the feasibility of the proposed method. Table 6 shows the execution time of the proposed NSAF. The path derivation time to guarantee the QoS of the application was composed of 20, 30, 40, and 50 nodes of the virtual topology. The average time to route derivation was measured by executing the algorithm for five times.

The NSAF utilized Dijkstra's algorithm to apply T-score and GA for optimization and guarantee the application QoS. Dijkstra's algorithm is fast; hence, the route must be calculated using Dijkstra's algorithm first in case of a QoS violation. However, if a problem occurs in the path calculated using Dijkstra's algorithm, multiple paths satisfying the QoS can be found as backups using GA because Dijkstra's algorithm computes only one path satisfying the minimum cost. The NSAF applied a mechanism that adjusts the application to run continuously by changing the path to ensure the application QoS.

In addition, we conducted a qualitative comparison evaluation of the framework proposed herein and monitoring research in SDN environments. The evaluation criteria were compared with those of the monitored aspects and those of the application's QoS monitored aspects. We compared application QoS monitoring and QoS monitoring methods and investigated whether or not path control considering QoS is supported. Table 7 presents the related studies and comparative assessments.

6. Conclusion

This study proposed a framework, called NSAF, that guarantees the application QoS by recognizing the QoS requirements, ascertaining QoS violation status, and discovering network paths that satisfy the application QoS to support a stable application operation in SDN environments.

The NSAF made it possible to register 12 types of DiffServ applications and set the application QoS on measurable

Table 7: Comparison with the related studies.

Criteria		Interactive monitoring [28]	MonSamp [29]	PayLess [30]	NSAF
Monitored targets		SDN controller and device	Traffic pattern	Application QoS	Application QoS and network path that satisfy application QoS
Application QoS monitoring aspects	How to monitor QoS	Registration of the user traffic profile (configuration parameter setting)	Flow sampling application (switch flow rule)	Monitoring request object creation (specification)	NSLA based on the ontology (binding value at the ontology)
	Route control with QoS	Not supported	Feedback from the flow sampling application	Not supported (only monitoring)	Route control using NSAF
	How to ensure QoS	Not supported (visualization support about the monitoring result)	Not presented	Not presented (statistical result about monitoring)	(i) QoS violation detection using the ontology (ii) Network routing with Dijkstra's algorithm and GA considering QoS

quality factors, such as bandwidth, packet loss, delay, and jitter. The NSAF was supported by a proposed application profile model. Furthermore, a topology model was used in the proposed framework to monitor the network status in terms of the application QoS violation. We also classified the application violations as either hard fail or soft fail and identify the application QoS violations based on the ontology.

Dijkstra's algorithm was applied to compute the path cost of four QoS in the network path search satisfying the application QoS, and the alternative optimized paths were determined using GA.

The proposed framework can be used as a reference structure to change the network paths of applications according to their QoS requirements and varying network status. In the future work, we plan to expand the proposed NSAF to predict the changes in the network and control paths when applications are executed.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (nos. NRF-2014R1A1A2007061, NRF-2016R1D1A1B03935865, and NRF-2017R1D1A1B03030243).

References

[1] M. Jarschel, T. Zinner, T. Hossfeld, P. Tran-Gia, and W. Kellerer, "Interfaces, attributes, and use cases: a compass for SDN," *IEEE Communications Magazine*, vol. 52, no. 6, pp. 210–217, 2014.

[2] W. Xia, Y. Wen, C. H. Foh, D. Niyato, and H. Xie, "A survey on software-defined networking," *IEEE Communications Surveys & Tutorials*, vol. 17, no. 1, pp. 27–51, 2015.

[3] A. Mediola, J. Astorga, E. Jacob, and M. Higuero, "A survey on the contributions of software-defined networking to traffic engineering," *IEEE Communications Surveys & Tutorials*, vol. 19, no. 2, pp. 918–953, 2017.

[4] S. Sezer, S. Scott-Hayward, P. Chouhan et al., "Are we ready for SDN? implementation challenges for software-defined networks," *IEEE Communications Magazine*, vol. 51, no. 7, pp. 36–43, 2013.

[5] K. M. Modieginyane, B. B. Letswamotse, R. Malekian, and A. M. Abu-Mahfouz, "Software defined wireless sensor networks application opportunities for efficient network management: a survey," *Computers & Electrical Engineering*, vol. 66, pp. 274–287, 2018.

[6] M. Jammal, T. Singh, A. Shami, R. Asal, and Y. Li, "Software defined networking: state of the art and research challenges," *Computer Networks*, vol. 72, pp. 74–98, 2014.

[7] Y. E. Oktian, S. Lee, H. Lee, and J. Lam, "Distributed SDN controller system: a survey on design choice," *Computer Networks*, vol. 121, pp. 100–111, 2017.

[8] Project Floodlight, 2018, http://www.projectfloodlight.org.

[9] OpenDaylight, 2018, https://www.opendaylight.org.

[10] ONOS, 2018, http://onosproject.org.

[11] Ryu, 2018, https://osrg.github.io/ryu.

[12] OPenFlow, 2018, https://www.opennetworking.org/technical-communities/areas/specification/open-datapath.

[13] Cisco Systems:OpFlex:An Open Policy Protocol White Paper, 2018, http://www.cisco.com/c/en/us/solutions/collateral/data-center-virtualization/application-centric-infrastructure/white-paper-c11-731302.html.

[14] T. Tsai, K. Wang, and T. Y. Chao, "Dynamic flow aggregation in SDNs for application-aware routing," in *Proceedings of the 2016 10th International Symposium on Communication Systems, Networks and Digital Signal Processing (CSNDSP)*, Prague, Czech Republic, July 2016.

[15] Y. Desmouceaux, P. Pfister, J. Tollet, M. Townsley, and T. Clausen, "6LB: scalable and application-aware load

balancing with segment routing," *IEEE/ACM Transactions on Networking*, vol. 26, no. 2, pp. 819–834, 2018.

[16] L. Cheng, K. Wang, and Y. Hsu, "Application-aware routing scheme for SDN-based cloud datacenters," in *Proceedings of the 2015 Seventh International Conference on Ubiquitous and Future Networks*, pp. 820–825, Sapporo, Japan, July 2015.

[17] J. Chan, J. Babiarz, and F. Notel Baker, *Cisco Systems: Aggregation of Diffserv Service Classes*, Network Working Group, Internet Engineering Task Force (IETF), Fremont, CA, USA, 2018, https://tools.ietf.org/html/rfc5127.

[18] OWL (web ontology language), 2018, http://www.w3.org/2004/OWL.

[19] I. Horrocks, P. F. Patel-Schneider, and F. van Harmelen, "From SHIQ and RDF to OWL: the making of a web ontology language," *Journal of Web Semantics*, vol. 1, no. 1, pp. 7–26, 2003.

[20] SWRL: A Semantic Web Rule Language Combining OWL and RuleML, 2018, https://www.w3.org/Submission/SWRL.

[21] T. M. de Farias, A. Roxin, and C. Nicolle, "SWRL rule-selection methodology for ontology interoperability," *Data & Knowledge Engineering*, vol. 105, pp. 53–72, 2016.

[22] H. J. Holland, *Adaptation in Natural and Artificial Systems*, The MIT Press, Cambridge, MA, USA, 1992.

[23] R. Rojas, *Neural Networks—A Systematic Introduction*, Springer-Verlag, Berlin, Germany, 1996.

[24] M. Gen and R. Cheng, *Genetic Algorithms and Engineering Optimization*, John Wiley & Sons, Hoboken, NJ, USA, 2000.

[25] M. Karakus and A. Durresi, "Quality of service (QoS) in software defined networking (SDN): a survey," *Journal of Network and Computer Applications*, vol. 80, no. 15, pp. 200–218, 2017.

[26] R. B. Bapat, "Adjacency matrix," in *Graph and Matrices*, Springer, Berlin, Germany, 2010.

[27] Mininet, 2018, http://mininet.org.

[28] P. H. Isolani, J. A. Wickboldt, C. B. Both, J. Rochol, and L. Z. Granville, "Interactive monitoring, visualization, and configuration of OpenFlow-based SDN," in *Proceedings of the 2015 IFIP/IEEE International Symposium on Integrated Network Management (IM)*, pp. 207–215, Ottawa, Canada, May 2015.

[29] D. Raumer, L. Schwaighofer, and G. Carle, "MonSamp: a distributed SDN application for QoS monitoring," in *Proceedings of the 2014 Federated Conference on Computer Science and Information Systems*, pp. 961–968, Łódź, Poland, September 2014.

[30] S. Chowdhury, M. F. Bari, R. Ahmed, and R. Boutaba, "PayLess: a low cost network monitoring framework for Software Defined Networks," in *Proceedings of the 2014 IEEE Network Operations and Management Symposium (NOMS)*, pp. 1–9, Kraków, Poland, May 2014.

Hindawi
Mobile Information Systems
Volume 2019, Article ID 6249247, 13 pages
https://doi.org/10.1155/2019/6249247

Research Article

Expanding GÉANT Testbeds Service to Support Pan-European 5G Network Slices for Research in the EuWireless Project

Álvaro Rios,[1] **Barbara Valera-Muros**,[1] **Pedro Merino-Gomez**,[1] **and Jerry Sobieski**[2]

[1] *Universidad de Málaga, Andalucia Tech, Edificio de Investigacion Ada Byron, Málaga 29071, Spain*
[2] *NORDUnet, Nordic Gateway for Research and Education NORDUnet, Copenhagen, Denmark*

Correspondence should be addressed to Pedro Merino-Gomez; pedro@lcc.uma.es

Received 21 December 2018; Accepted 4 April 2019; Published 23 April 2019

Guest Editor: Pinar Kirci

This paper presents the design options for creating a Pan-European mobile network for research in the context of the European Horizon 2020 EuWireless project. The most likely direction is a platform that makes it easier to create network slices for research. In this context, we identify one promising technology to implement network slicing in 5G networks: the framework GÉANT Testbeds Service (GTS). GTS is currently a production service by GÉANT that offers remote construction and use of virtual testbeds for wired networks mapped to the real GÉANT infrastructure. These GTS-virtualized testbed environments conform to Software Define Networks (SDNs) principles and offer compute, storage, and switching resources, at scale and with line rate performance. In this paper, we explain how the current (wired oriented) GTS can be extended with the 5G components, such as radio access nodes (gNBs), transport networks, user devices, etc., in order to implement 5G network slices. Our first conclusion is that using GTS for EuWireless implementation is feasible, dramatically increasing the potential impact of this service in the research community.

1. Introduction

Traditionally, experimentation in computer networks has a low barrier entry. The equipment needed to set up complex architectures is inexpensive, and the algorithm to control communication between peers is usually open and flexible to test new communication protocols and processes. Moreover, with the adoption of virtualization and Software Defined Networks (SDNs) techniques, researchers can create realistic large-scale networks with hosts behaving identically to physical ones [1]. When we move to wireless, Wi-Fi access technology devices offer the same Application Programming Interface (API) to the upper layers of the network as the wired ones; so the interface that controls access to the lower layers can also be easily modified to test new configurations, maintaining this barrier entry at its lower level [2].

In contrast, the barrier entry in experimentation with cellular mobile networks is high, mainly due to two reasons. Firstly, obtaining the required equipment and software to set up a mobile network for experimentation is arduous because of the costs of acquaintance and the hesitance of manufacturers to sell the components to anyone other than commercial operators, as it can disrupt their business model. Secondly, and more important, research in mobile networks requires access to state-regulated radio spectrum in order to recreate realistic environments and to use commercial equipment as part of the new developments. Governments worldwide use public auction schemes for the frequencies needed to operate, and commercial operators are normally the only ones capable of bidding for their use. Even when the experimenters have access to the right equipment, they can only set up a complete network by isolating it in radio-frequency shields to prevent radiation outside the boundaries of the laboratory, greatly diminishing the value of the tests [3].

Regarding the lack of access to the basic infrastructure needed to perform experiments, it was recently lessened by the appearance of the Software Defined Radio (SDR) technology. SDR allows recreating base stations and other radio elements of the network with commercial off-the-shelf

(COTS) components placed inside a computer. In addition, several open-source projects [4, 5] were created to provide the elements of the core network by implementing the published standards directly in software. Nevertheless, SDR technology is not able to achieve the radio performance of the real equipment, and the different software solutions are still unstable or do not provide all the characteristic of the commercial ones.

Recognizing that mobile technology research is a key component of scientific prowess, different initiatives try to provide researchers with access to commercial equipment outside the umbrella of telecommunications operators. The European Commission, as part of research programs like FIRE, has set up different technology laboratories, i.e., NITOS in Greece or PerformNetworks in Spain [6], equipped with state-of-the-art components that are offered to researchers and businesses interested in this kind of experimentation. Although these projects allow basic research in the mobile networks field, they still have a major caveat. The experiments are linked to fixed laboratories, and the mobility part of wireless technology research is limited or has to be simulated by other means [7, 8].

In this context, the EuWireless project [9] is envisioned with the purpose of helping researchers and experimenters to test new developments in a live, realistic mobile network across Europe. The main objective of the project is to create a research environment capable of providing access to the usually hidden internal mechanism of the networks and also coexisting with commercial operators, without interfering with commercial exploitation. In that direction, this paper addresses how to implement 5G network slices for research at a large scale reusing a mature technology for automatic creation of network testbeds, namely, the GÉANT Testbed Service (GTS) [10, 11]. GTS is already being used by the GÉANT research community to seamlessly create virtual networks between different locations across Europe, and, with some modifications, it could be able to integrate radio resources, spectrum sharing, and mobile network entities into a single homogeneous user interface. In the paper, we introduce the design options for EuWireless and analyze how to map EuWireless concepts to GTS technology. This analysis confirms the technical feasibility of implementing EuWireless 5G slices in the midterm.

The rest of the paper is organized as follows. Section 2 lists the design principles and goals that the EuWireless architecture must fulfil to be differentiated from other projects. Section 3 compares different high-level architectures to be used as the foundation of the EuWireless infrastructure, briefly summarizing pros and cons of each one. A technology capable of coordinating all the components to be used in the initial architecture proposed, the GÉANT Testbeds Service (GTS), is detailed in Section 4. In Section 5, we present the most relevant related work that is currently ongoing, and, finally, conclusions are addressed in Section 6.

2. EuWireless Design Principles

The main goal of the EuWireless project is to design and build a comprehensive architecture capable of providing the researcher community, experimenters, and end-users access to low-level 5G resources, network capabilities, and configuration options by combining a pool of heterogeneous resources into a homogeneous interface, without the constraints of a laboratory setting or the rigidity of commercial environments. This section introduces the design principles of EuWireless and the main concepts to be considered as inputs to design the architecture, as discussed in the next section.

2.1. Design Principles. EuWireless has identified the following design principles that the final architecture must meet to differentiate itself from other existing projects:

(i) Support for concurrent researchers in their own isolated network: The framework that controls the infrastructure must be able to accommodate several experiments in parallel, sharing resources from the physical network, but without interfering with each other.

(ii) Cross-country deployment: Although the EU has invested in the spread of the European research infrastructure, many sites are concentrated around certain universities or technological areas. EuWireless aims to be able to offer experimentation capabilities near researchers across Europe.

(iii) Scalable Point of Deployment as the main core object: Instead of relying on a centralized control of all the infrastructures deployed in Europe, it is desirable to build a distributed architecture that provides local high-performance and great scalability.

(iv) Adaptable to integrate new mobile technologies: the design must be flexible enough to accommodate new standards and paradigms in the future, without relying too much on current, fixed implementations.

(v) Fully automated: researchers must be able to define, reserve, and use the resources needed for any given experiment in an autonomous way, whereas EuWireless must maintain the separation between experiments and a fair user policy to prevent resource exhaustion from greedy experiments.

2.2. Network Slicing. The key enabler to meet the project objectives and design principles is the ongoing convergence between traditional wired networks and mobile ones, which is already a goal in the recently published Fifth Generation (5G) standard. This convergence is known as Network Slicing (NS) and proposes the adoption of virtualized network paradigms and new schemas of resource sharing between the different stakeholders that provide network services to the end user [12]. NS technology combines a set of diverse resources such as radio spectrum, network capacity, or processing power and offers them for a specific purpose or service [13]. This concept can go further, allowing the aggregation of virtual resources with physical ones in a transparent way from the point of view of the service being offered [14]. Thus, a telecommunications operator is able to create core network instances and reserve them for the exclusive use of a group of users or kind of traffic, guaranteeing certain Key Performance Indicators (KPIs) regardless of the network load. These KPIs are defined for each use case based on the needs of the different vertical industries, as analyzed by the GSMA in [15].

The key concept in NS is the complete separation of resources to be used by different services, in the understanding that this separation is a logical one and the physical resources are used concurrently by all the users but with different priorities. An example of this kind of resource sharing is depicted in Figure 1.

Thanks to the slicing concept, EuWireless can offer each experimenter a network composed of real resources, by using an abstraction layer that distributes those resources in a differentiated way, that is, completely isolated from networks used in other experiments. A remarkable advantage of network slicing is that the same principles can be applied to resources that are not owned by the same company or operator providing the service. This offers the network operator the opportunity of expanding the coverage area and the range of services that could be included in the networks provided for experimentation by entering into agreements with commercial operators with key resources, such as spectrum, link capacity, or raw deployments in a wide area.

2.3. 5G Slices for Research as a Service. Expanding on the design principles stated previously, the prime goal is to deliver a platform that would provide the research community with an end-to-end solution to perform experiments in a realistic way, with commercial equipment but without the constraints of maintaining a highly reliable commercial operation.

In order to do this, the idea is to allow the experimenters to reserve the specific resources needed, run the experiment and free the resources afterwards. In each experiment, the researchers would see a complete 5G operator, from the radio hardware communicating with the UE to the network level, including the core network and any Multiaccess Edge Computing (MEC) services they may need to perform all kinds of experiments [16]. The project should be able to create a private network for each of the experiments in a way that would not interfere with one another. From the low-level radio access to the components that route the user's traffic through the operator network to other users and networks, each experimenter will perceive a complete "telecommunication infrastructure" at his or her disposal, regardless of the networks deployed by other experimenters. Figure 2 shows the entities and logical relations of a complete 5G network, from the User Equipment (i.e., the smartphone or modem), to authentication and network access.

A traditional approach would involve replication of these components for each researcher and compartmentalization to ensure isolation. However, network slicing enables this isolation in a seamless way without the need of duplicating the resources for each experiment. Moreover, the combination of the network components in a virtual environment and the proficient allocation of the resources needed to perform each task of the experiment will result in the possibility of creating an ample number of "virtual" networks, always maintaining a strict separation between them (Figure 3).

2.4. Distributed Deployment of EuWireless. Another goal of the project is to have it distributed, instead of being physically concentrated in a laboratory or a testbed, with the intention of providing equal opportunities for access and performance to researchers located across Europe. If the resources are located in a single Point of Presence (PoP), this goal would be impossible due to unavoidable physical constraints that will decrease data flow performance. Thus, the main components of EuWireless must be decentralized and decoupled in nature, making it possible to expand the coverage of the service just by deploying one set of hardware and software at each desired location. Additionally, it is desirable for these services to be easily connectable with already deployed installations, so as to compose a mesh network of nodes capable of providing the same kind of features to the experimenters.

This leads to the concept of PoP as the core object in the EuWireless infrastructure. As mentioned above, there is a need to provide services geographically close to where the experiment is running. A PoP would consist in the set of hardware and software that is capable of configuring and managing the slices created for each of the experiments, with the capacity of running as a single node or as a part of a network of PoPs. PoPs must be able to provide services isolated in a certain geographical or logical area, but also to interconnect in a seamless and decentralized way. All the software running in the PoP must be able to interact with the entities present in the current deployment, hardware or software, as well as with the ones present on other PoPs. All of this will naturally provide scalability, since an overloaded deployment can be scaled up just by adding processing power or network links, or even installing new PoPs in areas with a great demand of services to increase performance. As a result, there will be a homogeneous interface with all the resources available to be used, even if the PoP located close to the researcher does not have the resources needed by the experimenter. Furthermore, the project can grow organically from the initial pilots to large deployments running an ample range of experiments concurrently by just connecting new PoPs, as shown in Figure 4, avoiding the reconfiguration of the underlying infrastructure.

The EuWireless project aims to build a future-proof framework where new technologies are tested in conjunction with the research on the current one. Any architecture chosen to provide the services of a project like EuWireless should be sufficiently abstract or expandable to accommodate new paradigms to be tested without fixating in the rigid structure of a particular generation of mobile networks. The following section discusses some design options for this architecture.

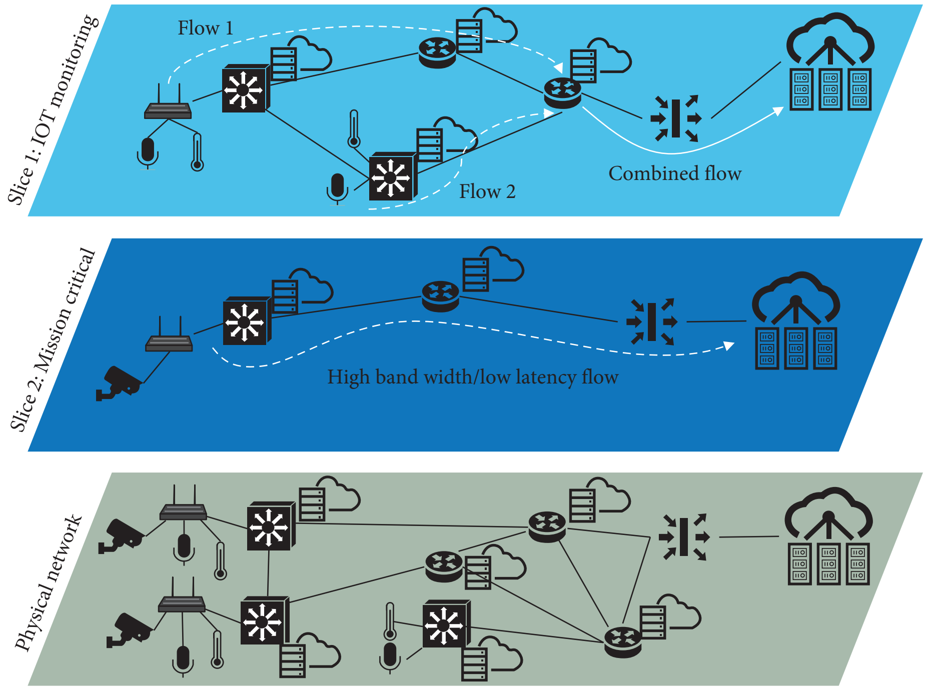


Figure 1: Different slices over the same physical topology.

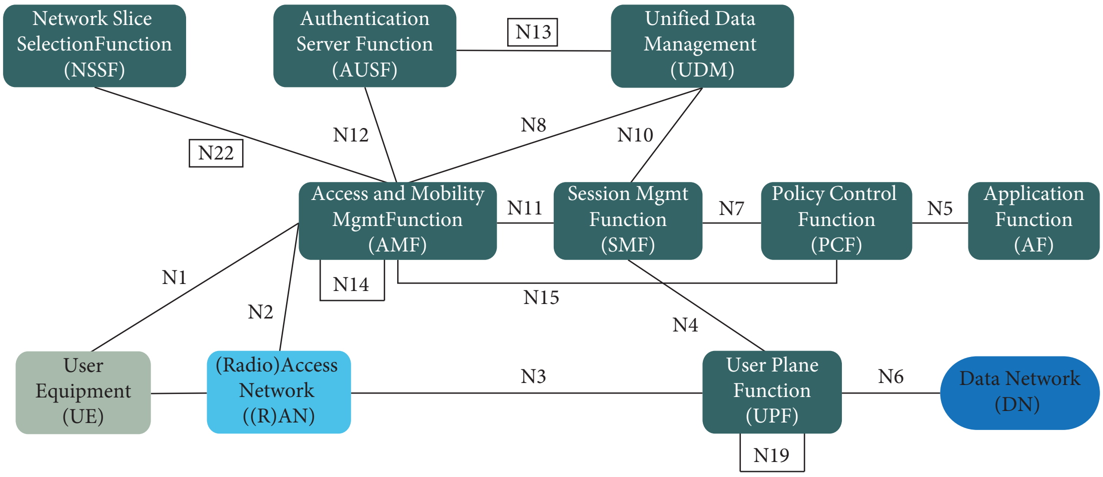

Network Slice SelectionFunction (NSSF)
Authentication Server Function (AUSF)
N13
Unified Data Management (UDM)
N22
N12
N8
N10
Access and Mobility MgmtFunction (AMF)
N11
Session Mgmt Function (SMF)
N7
Policy Control Function (PCF)
N5
Application Function (AF)
N1
N14
N15
N2
N4
User Equipment (UE)
(Radio)Access Network ((R)AN)
N3
User Plane Function (UPF)
N6
Data Network (DN)
N19

Figure 2: Entities in a 5G network.

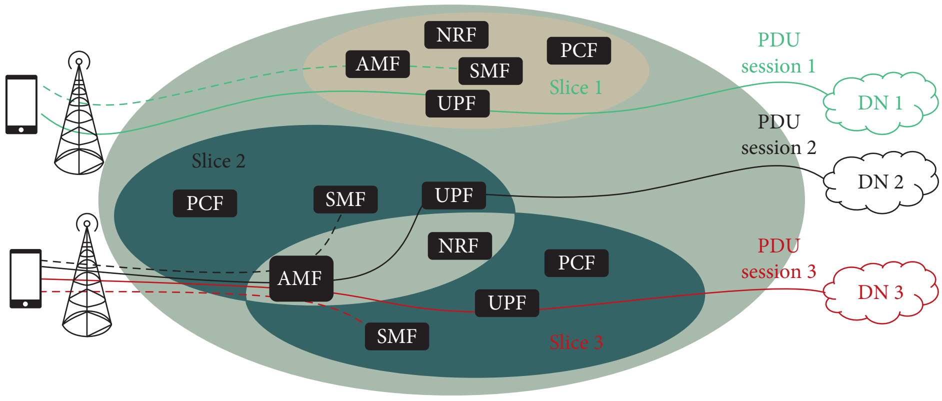


Figure 3: Network slicing in 5G networks.

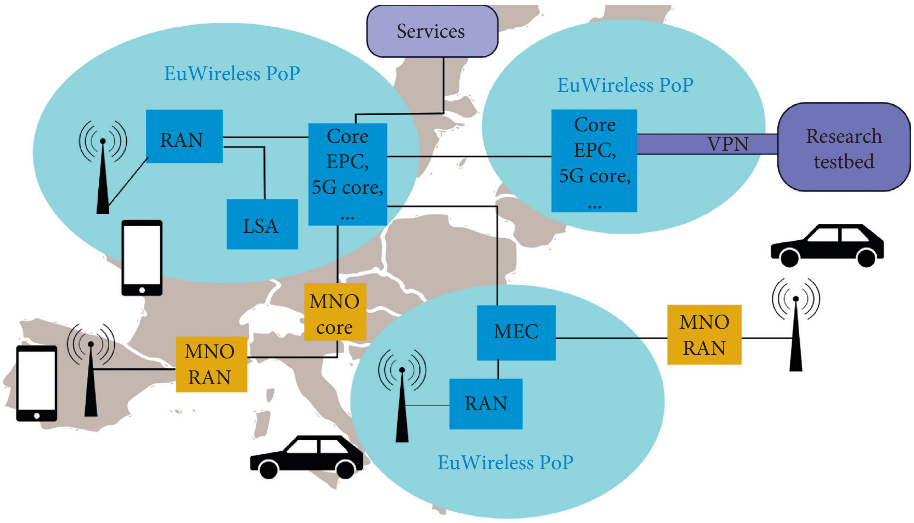


Figure 4: EuWireless high-level deployment overview [9].

3. Design Options for EuWireless

Several reference architectures are considered in order to implement the functionality to be offered by EuWireless, combining private and public infrastructures, which will also drive the kind of access and resources to be offered to the researchers.

3.1. Creation of a Full 5G Operator. The first option is to create a full 5G operator to provide the kind of network for experimentation with all the resources owned by EuWireless. This would give the project complete freedom in the configuration and selection of the elements that compose a particular slice. Any kind of end-to-end experiment can be created focusing on any layer of the stack, and the researchers will have direct access to monitor the internals of the architecture to gather the raw performance data from within the network.

The main drawback of this approach is the cost, both in economic and in human resources, needed to complete the deployment of such an operator across Europe. Moreover, additional regulatory constraints might hamper completeness of the service provided: to be able to operate in a multicountry environment EuWireless would need to comply with local legal requirements in each territory in which it operates [17], as well with European regulations regarding the security and the privacy of the data passing through its network, even when these data are generated automatically by other systems [18]. In addition, even if all the infrastructures are owned by the project, there are still limitations on the kind of access that can be provided to low-level network capabilities in order to maintain the operational needs of the whole network.

A related, more realistic approach to this architecture would be one in which EuWireless owns the core of the network where experiments take place but relies on external operators to ensure network capacity and last-mile access in areas where it cannot provide radio coverage. Figure 5 presents the schematic architecture for this kind of infrastructure.

3.2. Management of External Resources. The second approach is the opposite of the first one. In this case, EuWireless would not own any physical infrastructure apart from the minimum required to provision, configure, and manage the slices and could be considered a Virtual Mobile Network Operator (vMNO) that provides services on top of the physical resources of a single commercial operator, or several of them. EuWireless would lease/rent 5G slices of the commercial operators and offer them to the researchers to test new services. By entering into an agreement with operators of different countries, it would be possible to extend EuWireless coverage as far as desired, using PoPs near the point of connection with the carriers to provide access to the EuWireless network and the slice-monitoring facilities.

The main problem of this architecture is that it completely depends on the resources of external operators, which might not be available for the entire duration of the experiment or for which the Quality of Service (QoS) offered may suddenly change depending on the commercial operation of the physical infrastructure. Another constraint would be, most probably, that slices would be limited to 5G resources, as it is the only technology that supports the concept of slicing, and research in the interconnection with other technologies would require external capabilities outside the scope of the project.

Nevertheless, the openness of the configuration is attractive since it enables creating proprietary slices with advanced features and minimal costs, as the network elements are already being used in a commercial operation, even when this flexibility depends on the agreements with the commercial provider, which is ultimately responsible for limitations on the usage of its network, and may or may not be eager to offer the concrete resources experimenters would like to use in some deployments. Figure 6 shows the

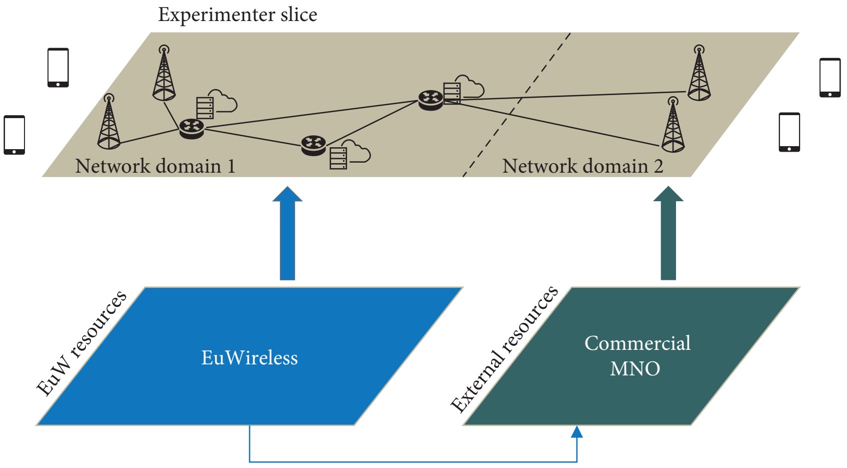


Figure 5: First architecture considered for the project. EuWireless has resources with extended coverage provided by commercial MNOs.

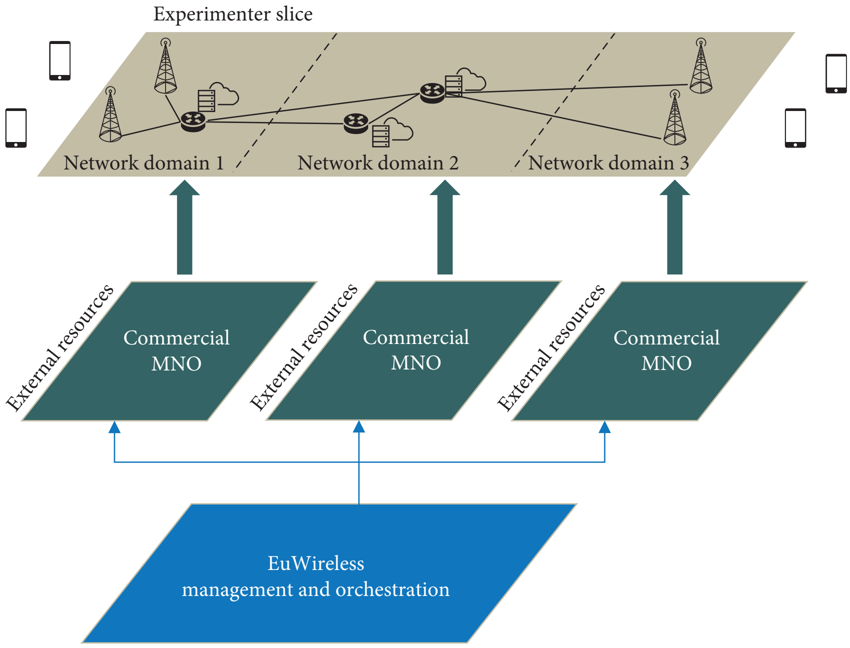


Figure 6: Second architecture considered for the project. EuWireless manages resources of commercial MNOs.

only component required, Management and Orchestration, coordinating resources of different operators to create a slice for a particular experiment.

3.3. Network Slices Virtualization and Management. The previous architectures have the issue of only being able to provide 5G slices and, even in the case when the infrastructure is completely owned and controlled by EuWireless, the inability to access low-level components of the network. One solution would be to use the slices created as the foundation of a complete network. Ultimately, a network slice defines the links between entities with different characteristics and the performance expected when they are interconnected. However, the state of the art of real networks is moving towards virtualization, and these entities could be perceived as virtual objects; so, in fact, any software can be deployed and any network infrastructure can be defined in the slices to interconnect them. By adding a new level of abstraction to the resources, it is possible to present different elements aggregated as “raw slices” on top of which new services can be deployed, treating different resources of different operators in a homogeneous way.

Moreover, with this concept of raw connectivity slices, researchers are not strictly bound to experiment in 5G technology because additional network elements, even entities from previous generations, can be instantiated as Virtualized Network Functions (VNF) [19]. With this architecture (Figure 7), some of the constraints of the two previous proposals are reduced: for example, the limitation of only experimenting in a 5G environment is lowered because the researcher can deploy any kind of “virtualized” environment, previous generation entities can be deployed as virtual machines using the computer processing power of

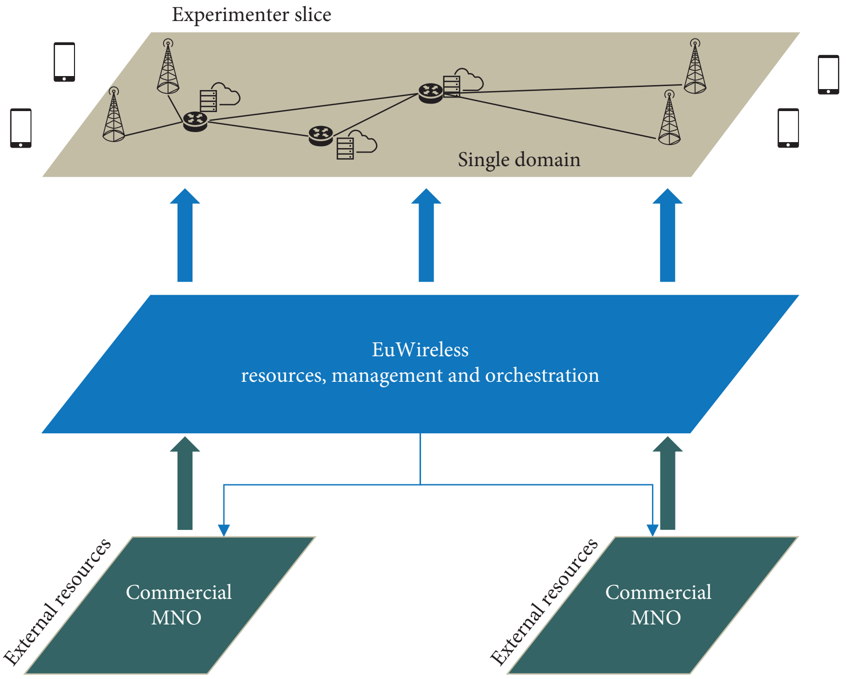


Figure 7: Third architecture considered for the project. EuWireless is a "virtual slice" provider.

the slice, and connected with 5G entities, Wi-Fi access points can be attached in any link of the network as an IP device [20], and so on. In addition, the architecture "seen" by the experimenter will not depend directly on the network status of the external operators, as the EuWireless coordination layer will manage the resources during the experiment lifecycle, being capable, for example, of selecting different links if the conditions of the underlying physical network worsen.

In this approach, the only thing outside the scope of the experiment will be the infrastructure that allows creating the slice, such as the MANO, or the slice manager used to coordinate the resources provided by the different actors. But even in this case, if the researcher needs those components for a specific experiment, they can actually be installed as virtual appliances to control another virtual layer on top of the raw one, although this kind of virtual-on-virtual slices imposes a nonzero performance penalty that must be analyzed before undertaking the complex task of deploying a scenario of this type.

Another benefit of this approach is hinted in Figure 7: in the two previous architectures, the traffic had to pass through the different resource providers' network domains, whereas in this one, a single domain can be created for each slice and experiment.

4. Role of GTS in the Implementation of 5G Slicing

The third architecture, chosen as the basis of the EuWireless infrastructure, allows great flexibility in the selection and configuration of the network elements, but a key component is still missing: the coordination layer that manages and orchestrates all the different resources into a single unified slice to be offered to the experimenter, ensuring link quality and network isolation between different experiments. During the evaluation of available standards and technologies for this purpose, a new architecture developed by the European research network GÉANT emerged, namely, the GÉANT Testbeds Service (GTS) [10, 11]. This section explains how to use GTS to implement the selected architecture.

4.1. GÉANT Testbed Service. The GÉANT Testbeds Service (GTS) is an innovative network virtualization architecture that offers experimenters access to wide area network infrastructure integrated within the GÉANT network footprint. Through GTS, the research community can deploy and refine novel and/or experimental networking concepts "at scale," using real network components, with real users, under real-world conditions with the addition of Software Defined Networks (SDNs) paradigms and network-based computing, storage, and switching of resources for research networking and distributed applications. This notion of allocating network infrastructure components to particular projects or applications, under the direct control of the research user, is quite similar to the "Network Slicing" in 5G but applied only to wired networks. From the user perspective, GTS offers virtual testbeds, as represented in Figure 8.

In the original GÉANT implementation, these various components were allocated and stitched together by human-mediated operational processes or using different applications and protocols on a project-by-project basis. GTS offers enhancements to this original work with dynamic and automated slice provisioning and employs advanced abstraction techniques to enable a broad range of new network

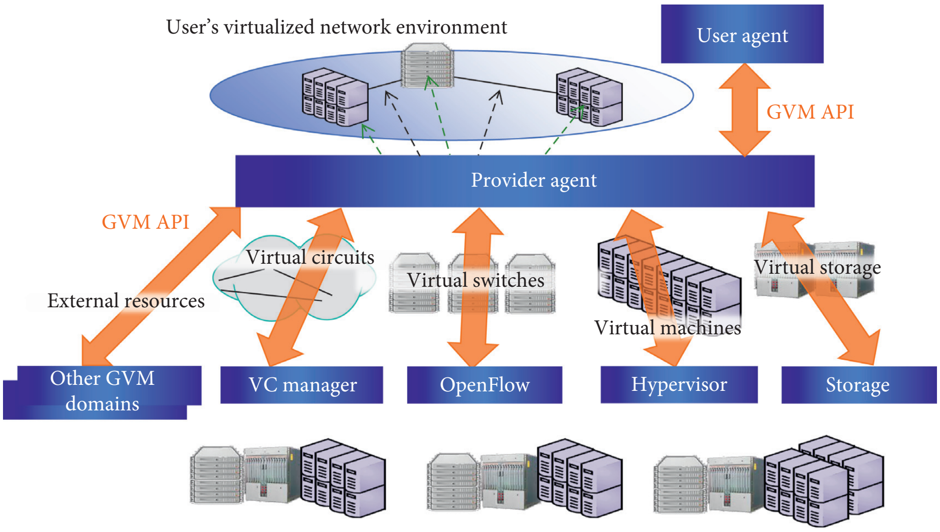


Figure 8: Vision of a virtualized network environment provided by GÉANT.

resources to be offered as fully virtualized service objects—such as VMs, virtual circuits, virtual switching/forwarding elements, storage, instruments, sensors, wireless/cellular resources, and other functional network objects.

4.2. Generic Virtualization Model architecture. Since its initial implementation in 2013, GTS has evolved into a sophisticated open-source Software Suite and a Generic Virtualization Model (GVM) architecture. The service is able to dynamically allocate a wide range of SDN-capable network infrastructure elements and provide these resources as insulated and isolated networks, according to the user's direction and under user control. GTS is also able to dynamically allocate both virtual machines (VMs), in a range of increasingly powerful configurations, and dedicated hardware platforms, or bare metal servers, where the user has access to the entire physical server interconnected with a fully virtualized SDN. To connect all of these virtual "nodal" objects, there are virtual circuits provisioned with hard QoS guarantees. All of these resources can be scheduled or reserved in the time domain and placed spatially within the geographic footprint of the GTS service reachability.

The GVM is conceptually simple: it defines an opaque service domain that offers virtualized resources at user request, meaning that the underlying infrastructure and processes that produce the resources are not exposed (and are not relevant) to the user. Within the GVM architecture, a resource can be anything, and resources that share common characteristics can be grouped into types, or classes. Each resource class defines the attributes that govern the behavior of the object, which users can tune to their needs when requesting such an object. A key aspect of defining resources is specifying how the resource communicates with other resources. This is done by specifying I/O ports and networking capabilities. Ports enumerate data paths from/to each resource instance. These ports will be connected to other ports to create the network slice topology. Using the GTS approach, when requesting a slice, the user must specify two aspects: the set of resources the experiment requires and the interconnection of those resources relative to one another by specifying port adjacencies. This set of links defines the topology of the network slice, or the data flow graph when the resources are functional objects rather than hardware infrastructure analogs. In this model, all objects that make up the virtual network slice are realized as resources—including virtual circuits. Thus, the network slice is reduced to a set of individual resource objects derived from the original network diagram, and a set of adjacencies that indicate which ports are interfaced to or flow into other ports. Figure 9 shows the real architecture of the virtualized network using the GVM and hints what is the work to be done in EuWireless to expand GTS.

4.3. GTS Experiment Lifecycle. The user interacts with the service through a set of web service primitives that move the resources through a basic lifecycle. The user experiences the resources as virtually the same as the real objects they are intended to model. By hiding these internal allocation and provisioning processes, the service providers are accorded maximum flexibility to use whatever products, external resources, or technologies they prefer. This opaque principle is also the basis of the Network Service Interface (NSI) [21]. This technology-agnostic service model allows the service architecture to scale by allowing providers to retain control of their local infrastructure engineering and design; as long as the service provider delivers the "virtualized" service objects as defined, they are not otherwise constrained to use particular products or technologies.

Each virtual object class requires a means of realizing instances of that object. This "runtime" environment, which supervises the sharing of the underlying infrastructure, is essential for each GVM resource class. In addition to these class-specific runtime services, there are processes that manage the lifecycle of the service object and that each virtual object requires in order to become real. Since the

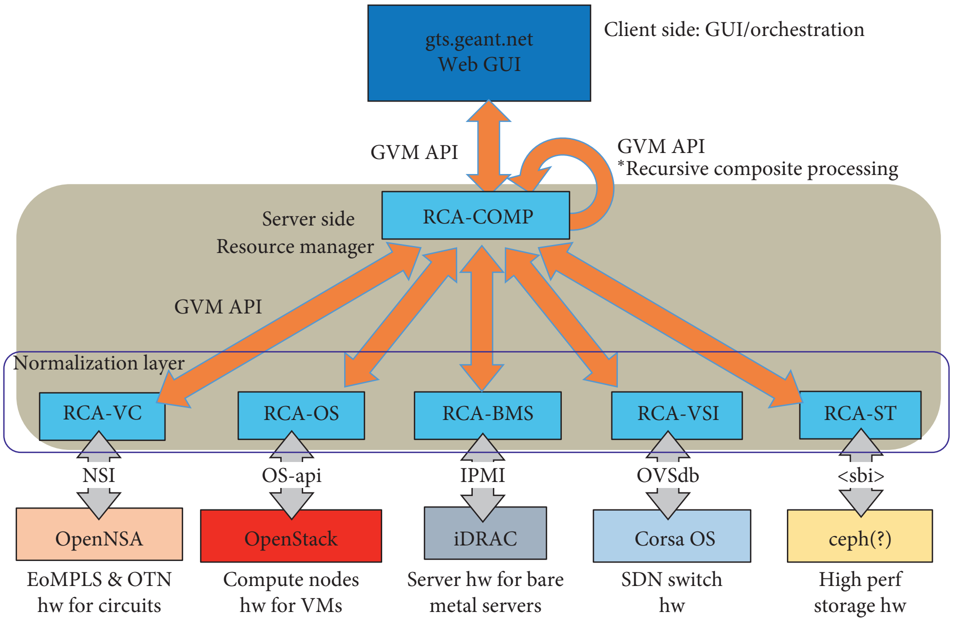


Figure 9: GVM.

internal mechanisms required to find and reserve infrastructure for each resource class may be quite different, GTS defines Resource Control Agents (RCAs) as the software modules that implement these processes for each class.

Every resource class must implement the following five basic control primitives, which move the resource, whether physical or virtual, through the states of its lifecycle, shown in Figure 10:

Reserve(): Instantiate the resource, allocate physical infrastructure components

Activate(): Configure the hardware to realize the resource, configure the VNF with the required information, move from RESERVED state to ACTIVE state

Deactivate(): Move from ACTIVE state to RESERVED state, deconfigure VNFs

Release(): Destroy the instance, shutdown the VNF

Query(): Get information about the actual state of the resource

4.4. Expanding GTS with 5G components. The EuWireless effort will be focused on the definition and implementation of the lifecycle of 5G network components to be used within the GTS environment. Once their GVM template is complete, we can leverage the already existing traditional network models to create slices with 5G access capabilities, as the additional core network entities of 5G can be deployed as virtualized functions. Figure 11 reflects this virtualization proposed to address network research in the context of GTS and EuWireless. Compared with Figure 9, the radio access node (gNB) and the AMF (one of the key components of the 5G core) are now virtualized to be offered to the researchers in their testbeds.

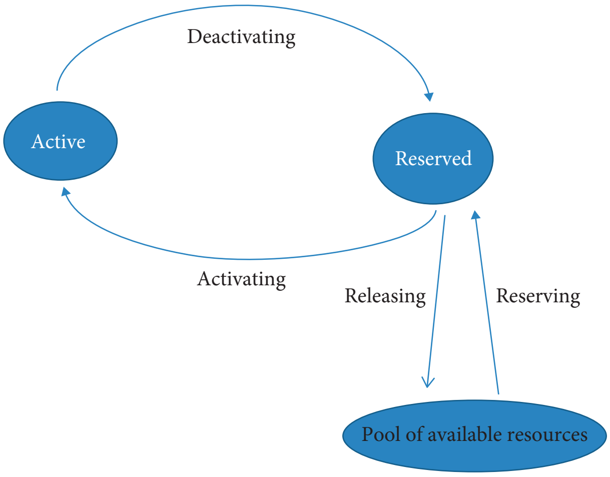


Figure 10: GTS objects lifecycle.

Each 5G component will be considered a virtual object in the GTS architecture. As mentioned above, virtual object instances interact with each other by exchanging information across their virtual I/O ports. Thus, the definition of the ports required for each 5G entity is fundamental in the implementation and corresponds with the interfaces of each component. The definition of a virtual object includes the name of the object, the description of its behavior, the ports required, and the set of attributes, where the latter are the parameters the user can tweak in defining the object instance. These attributes characterize the reservation of resources to be taken into account by the system.

Following the previous example, in the case of the virtual object that abstracts the Access and Mobility Management Function (AMF) entity, it will have, at least, eight ports that represent the logical relations with the entities of the 5G network, i.e., N1 with the User Equipment (UE), N2 with the

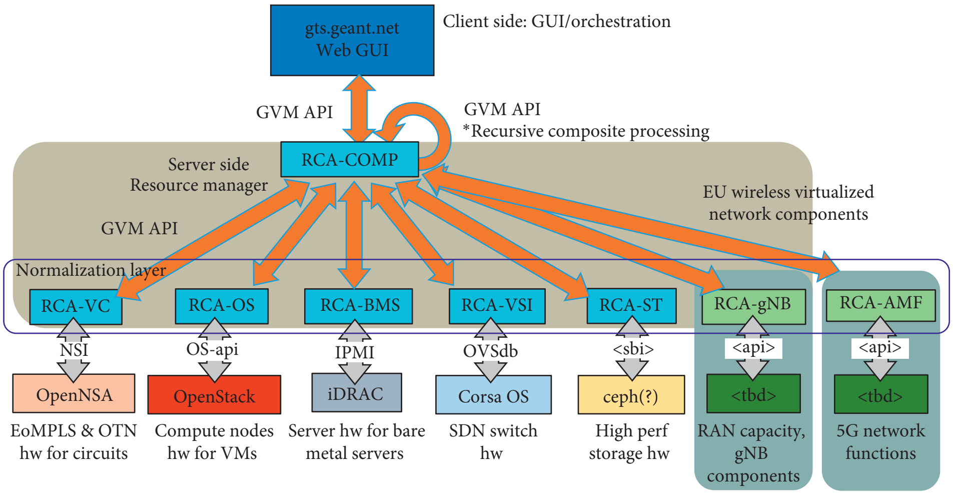


Figure 11: GVM incorporating EuWireless RCAs.

(Radio) Access Network ((R)AN), N8 with the Unified Data Management (UDM), N11 with the Session Management Function (SMF), N12 with the Authentication Server Function (AUSF), N14 with other AMF entities in the network, N15 with the Policy Control Function (PCF), and N22 with the Network Slice Selection Function (NSSF). Ports in GTS are linked with virtual circuits that might or might not be distributed and will establish the connection between the virtual object, in this case the AMF entity, and the destination entity or external component. These virtual circuits are able to transport any protocol the entities might use, such as Stream Control Transmission Protocol (SCTP) or User Datagram Protocol (UDP), that are the most used between mobile network components. In addition, there might be optional ports available such as the port connecting the AMF with the control plane through the N_{amf} in case the architecture is based on services, as shown in Figure 12. In this case, only three ports are required in the AMF, two if the UE connects via 3GPP technologies since the N1 will be in this case a logical interface through the physical N2 interface.

In accordance with the control primitives of the object lifecycle introduced in the previous section, after the reservation of resources and the initialization of the element, each entity is contextualized by the software that will configure the element as a network function in the 5G slice. In the case of the RCA that manages the virtual object representing a gNB, the reservation implies to find a compute node sufficient to run the virtual gNB software stack, with port capacity and spectrum capacity. Thus, steps will be reserve the port capacity towards IP network, then ensure that spectrum capacity is available, then reserve compute resources such as memory, number of cores, disk capacity, and finally any other facilities to/from/among physical gNB agents. The activation will spin up the VMs and establish bridges to the phy-gNB interfaces, that is, spin up the v-gNB VM(s) with context, set bridges to physical gNB ports, and configure them for spectrum sharing.

The virtual objects in GTS might be considered atomic, which hide all lower layer constructs, or composite, that enable the specification of complex virtual objects as a grouping of others. The latter may be parameterized during reservation, contain other composites allowing the existence of hierarchically constructed object instances, have external ports that can be set adjacent to ports of internal children, and be used to further parameterize children object instances. Following this reasoning, the network slice is understood as the root level composite object.

Table 1 shows a summary of the actions to be performed by the main components of the 5G stack in response to the execution of each GVM primitive.

5. Related Work

Experimentation in 5G networks is the main scope of the EuWireless project. However, there are several projects that already offer experimenters different state-of-the-art components to allow basic research in mobile networks. One of them is the European Commission's FIRE program. This program provides technology laboratories equipped with components to perform mobile networks research. However, the experiments must be executed in the laboratory, and thus wireless technology is usually simulated by other means. Similarly, in the US, the GENI program offers large-scale Internet testbeds and in 2013 funded the project SciWiNet [22] to support wireless networking systems research based on the vMNO model. SciWiNet offers cellular data services via Sprint's 3G, WiMAX, and LTE networks. Thus, a wireless testbed for the academic research community is provided, enabling research in areas such as eHealth, intelligent transportation systems, smart buildings and structures, homeland security, and Internet of Things, although the underlying infrastructure is not under investigation.

Recently, the European 5G PPP Initiative launched three projects as part of Horizon 2020 that will address the challenge of creating 5G end-to-end facilities. One of these projects, 5GENESIS [23], will deploy five large-scale platforms to compose a pan-European test platform addressing multiple vertical use cases. Thus, the architecture will be owned by each project, with an approach completely opposed to the SciWiNet vMNO.

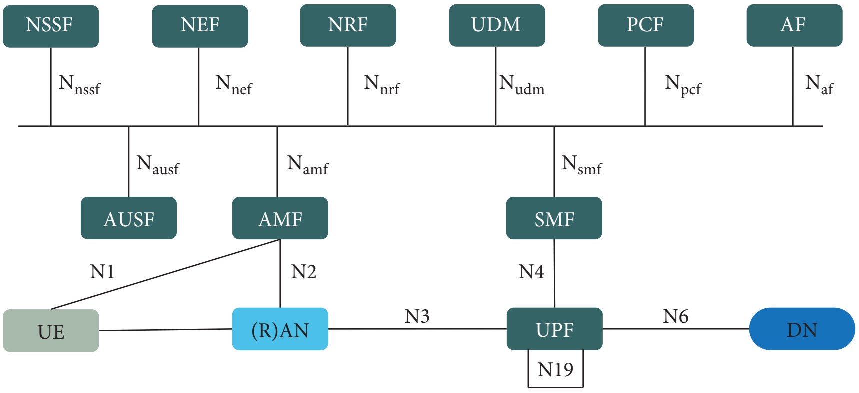


FIGURE 12: Entities in a 5G network service-based architecture.

TABLE 1: Actions in response to primitives.

GTS primitive/5G component	User equipment (UE)	gNB	AMF	UPF	AUSF/UDM	SMF
Reserve	Locate one specific UE and connect	Ensure that the gNB is available on the testbed	Instantiate AMF VNF	Instantiate UPF VNF	Instantiate AUSF VNF	Instantiate SMF VNF
Activate	Attach to 5G slice	Activate N2, N3 interfaces to the correct AMF, UPF and start PLMN broadcast	Connect interfaces N8, N11, N12 with the rest of the components	Connect interfaces N4, N6 with SMF and DN	Register user credentials and keys, connect to AMF	Configure UE IP address pool, traffic steering, connect to AMF/UPF
Deactivate	Detach from 5G slice	Close logical association with AMF/UPF	Disconnect from other entities	Disconnect from other entities	Disconnect from other entities	Disconnect from other entities
Release	Free from testbed	Return resource to the pool	Shutdown VNF	Shutdown VNF	Shutdown VNFs	Shutdown VNF
Query	Get connectivity status and measurements	Get logical association status, number of UE connected, measurements	Get connection status	Get connection status	Get connection status, registered users	DL data notification, traffic routing

In comparison with the above-mentioned projects, EuWireless proposes an architecture that combines both solutions by using virtualized network slices that will reserve the resources required to perform the different experiments from a pool of external and internal resources available, which will be transparent to the researchers due to the implementation of an abstraction layer.

There are also less theoretical projects in which slices are not only described but also created. In this context, the main goal of the MATILDA project [24] is to provide clear interfaces for the multisite management of cloud/edge computing and Internet of Things (IoT) resources. For this purpose, the project proposes a three-layer architecture: the development environment layer, the 5G application orchestrator, and the programmable 5G infrastructure. Thus, MATILDA includes the design and development of 5G-ready applications on top of network slices that are application-aware and can lead to optimal application execution [25].

The 5Gtango project aims at improving the flexibility and programmability of 5G networks. Thus, 5Gtango proposes the creation of a platform with validation and verification mechanisms for VNFs and Network Services, which is vendor-independent and owns an orchestrator compatible with existing Virtual Infrastructure Managers (VIMs) and SDN controllers [26].

The SLICENET project [4] focuses on the deployment of real end-to-end slicing in virtualised multidomain, multi-tenant 5G networks. The project targets three main use cases: the smart grid oriented to the energy vertical, the eHealth with connected ambulances, and the smart city use case. In [5], SLICENET presents results for a disaggregated Ultra-Dense Network deployment with slicing capabilities based on the RAN runtime.

The 5G-EmPOWER open platform [27] targets an open ecosystem where new 5G services can be tested in realistic conditions. Currently, the platform supports both Wi-Fi and LTE RAN, and developers are able to fully visualize the state of the network and deploy and orchestrate network services dynamically.

6. Conclusions

This paper presents different infrastructure choices that can be used in the final design of the research network proposed by the EuWireless project. The goals of EuWireless seek a long reach but they are realistic and achievable, and architectures discussed represent the state of the art in the deployment of mobile networks worldwide. Moreover, they represent a further step from the simple commercial exploitation of radio resources and computational power, putting them at the service of the researchers and experimenters that cannot access the required infrastructure any other way. Several architectures with different levels of resource ownership have been considered, ranging from a completely EuWireless-owned infrastructure to a "virtual service provider" that leverages on external resources and only acts as a coordinator of those resources to compose slices to be offered to the research community. The architecture chosen lies in between these two and is a compromise that combines both the reliability of resources owned by the project and the extensibility of using external capacity when needed.

The decision of always presenting a homogeneous "virtual slice" to the researcher, by combining different resources as a single network, ensures the uniformity of the experimentation process and diffuses the physical separation between components from different providers. This choice complicates the design with the introduction of an abstraction layer needed to coordinate and manage resources owned by EuWireless and those owned by external operators. However, by choosing a mature technology like GTS as the base for physical and virtual resource management, which has already proved its usefulness in academic and research environments, the work to be done in the design of the final architecture is reduced and the chances for success are increased by minimizing the development and debug time needed to test different approaches and configurations. Our planned future work is to complete a first implementation of relevant 5G components to offer 5G virtual slices as testbeds to the GTS and to the general research communities.

Data Availability

The research and academic data used to support the findings of this study are included within the article as references to journals, research papers, and other academic and public material.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

The EuWireless project is funded by the EU Commission under the H2020 research and innovation program under grant agreement no. 777517 and coordinated by the University of Malaga (UMA). Barbara Valera has a research grant from University of Malaga (UMAJI94).

References

[1] T. Bakhshi, "State of the art and recent research advances in software defined networking," *Wireless Communications and Mobile Computing*, vol. 2017, Article ID 7191647, 35 pages, 2017.

[2] H. A. Omar, K. Abboud, N. Cheng, K. R. Malekshan, A. T. Gamage, and W. Zhuang, "A survey on high efficiency wireless local area networks: next generation WiFi," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 4, pp. 2315–2344, 2016.

[3] K. De Moor, I. Ketyko, W. Joseph et al., "Proposed framework for evaluating quality of experience in a mobile, testbed-oriented living lab setting," *Mobile Networks and Applications*, vol. 15, no. 3, pp. 378–391, 2010.

[4] C.-Y. Chang and N. Nikaein, "RAN runtime slicing system for flexible and dynamic service execution environment," *IEEE Access*, vol. 6, pp. 34018–34042, 2018.

[5] C. Chang, N. Nikaein, O. Arouk et al., "Slice orchestration for multi-service disaggregated ultra-dense RANs," *IEEE Communications Magazine*, vol. 56, no. 8, pp. 70–77, 2018.

[6] P. Demeester, P. Van Daele, T. Wauters, and H. Hrasnica, *Fed4FIRE—The Largest Federation of Testbeds in Europe, Building the Future Internet through FIRE*, River Publishers, Gistrup, Denmark, 2017.

[7] C. A. García-Pérez and P. Merino, "Experimental evaluation of fog computing techniques to reduce latency in LTE networks," *Transactions on Emerging Telecommunications Technologies*, vol. 29, no. 4, article e3201, 2018.

[8] C. Garcia-Perez, A. Diaz-Zayas, A. Rios et al., "Improving the efficiency and reliability of wearable based mobile eHealth applications," *Pervasive and Mobile Computing*, vol. 40, pp. 674–691, 2017.

[9] P. Merino, L. Panizo, and A. Diaz Zayas, "EuWireless: Design of a pan-European Mobile Network Operator for Research," in *Proceedings of the 2018 European Conference on Networks and Communications (EuCNC)*, Ljubljana, Slovenia, June 2018.

[10] F. Farina, P. Szegedi, and J. Sobieski, "GÉANT world testbed facility: federated and distributed testbeds as a service facility of GÉANT," in *Proceedings of the 26th International Teletraffic Congress*, Karlskrona, Sweden, September 2014.

[11] J. Sobieski, F. Farina, S. Naegele-Jackson, K. Kramaric, B. Pietrzak, and M. Hazlinsky, "GÉANT testbed service external domain ports: a demo on multiple domain connectivity," in *Proceedings of the 2015 Fourth European Workshop on Software Defined Networks (EWSDN)*, pp. 113-114, Bilbao, Spain, September 2015.

[12] I. Afolabi, T. Taleb, K. Samdanis, A. Ksentini, and H. Flinck, "Network slicing and softwarization: a survey on principles, enabling technologies, and solutions," *IEEE Communications Surveys & Tutorials*, vol. 20, no. 3, pp. 2429–2453, 2018.

[13] A. Kaloxylos, "A survey and an analysis of network slicing in 5G networks," *IEEE Communications Standards Magazine*, vol. 2, no. 1, pp. 60–65, 2018.

[14] J. Ordonez-Lucena, P. Ameigeiras, D. Lopez, J. J. Ramos-Munoz, J. Lorca, and J. Folgueira, "Network slicing for 5G with SDN/NFV: concepts, architectures, and challenges," *IEEE Communications Magazine*, vol. 55, no. 5, pp. 80–87, 2017.

[15] GSM Association, *Network Slicing Use Case Requirements*, GSM Association, London, UK, 2018.

[16] I. Sarrigiannis, E. Kartsakli, K. Ramantas, A. Antonopoulos, and C. Verikoukis, "Application and network VNF migration in a MEC-enabled 5G architecture," in *Proceedings of the 2018 IEEE 23rd International Workshop on Computer Aided Modeling and Design of Communication Links and Networks (CAMAD)*, Barcelona, Spain, September 2018.

[17] B. Custers, F. Dechesne, A. M. Sears, T. Tani, and S. van der Hof, “A comparison of data protection legislation and policies across the EU,” *Computer Law & Security Review*, vol. 34, no. 2, pp. 234–243, 2018.

[18] S. Wachter, “Normative challenges of identification in the internet of things: privacy, profiling, discrimination, and the GDPR,” *Computer Law & Security Review*, vol. 34, no. 3, pp. 436–449, 2018.

[19] ETSI, *Network Functions Virtualisation, An Introduction, Benefits, Enablers, Challenges & Call for Action*, ETSI, Sophia Antipolis, France, 2012.

[20] M. Richart, J. Baliosian, J. Serrat, J.-L. Gorricho, and R. Agüero, “Slicing in WiFi networks through airtime-based resource allocation,” *Journal of Network and Systems Management*, 2018.

[21] T. Kudoh, G. Roberts, and I. Monga, “Network Services Interface: an interface for requesting dynamic inter-datacenter networks,” in *Proceedings of the Optical Fiber Communication Conference and Exposition and the National Fiber Optic Engineers Conference (OFC/NFOEC)*, Anaheim, CA, USA, March 2013.

[22] M. Berman, J. S. Chase, L. Landweber et al., “GENI: a federated testbed for innovative network experiments,” *Computer Networks*, vol. 61, pp. 5–23, 2014.

[23] H. Koumaras, D. Tsolkas, G. Gardikis et al., “5GENESIS: the genesis of a flexible 5G facility,” in *Proceedings of the 2018 IEEE 23rd International Workshop on Computer Aided Modeling and Design of Communication Links and Networks (CAMAD)*, pp. 1–6, Barcelona, Spain, September 2018.

[24] P. Gouvas, A. Zafeiropoulos, C. Vassilakis et al., “Design, development and orchestration of 5g-ready applications over sliced programmable infrastructure,” in *Proceedings of the First International Workshop on Softwarized Infrastructures for 5G and Fog Computing (Soft5 2017), Co-Located with the 2017 29th International Teletraffic Congress (ITC 29)*, pp. 13–18, Genoa, Italy, September 2017.

[25] P. Gouvas, A. Zafeiropoulos, E. Fotopoulou et al., “Separation of concerns among application and network services orchestration in a 5G ecosystem,” in *Proceedings of the 2018 European Conference on Networks and Communications (EuCNC) Workshop “From Cloud Ready to Cloud Native Transformation: What It Means and Why It Matters”*, Ljubljana, Slovenia, June 2018.

[26] C. Parada, J. Bonnet, E. Fotopoulou et al., “5GTANGO: A beyond-MANO service platform,” in *Proceedings of the IEEE European Conference on Networks and Communications (EuCNC)*, Ljubljana, Slovenia, June 2018.

[27] R. Riggio, M. K. Marina, J. Schulz-Zander, S. Kuklinski, and T. Rasheed, “Programming abstractions for software-defined wireless networks,” *IEEE Transactions on Network and Service Management*, vol. 12, no. 2, pp. 146–162, 2015.

Hindawi
Mobile Information Systems
Volume 2019, Article ID 2563917, 8 pages
https://doi.org/10.1155/2019/2563917

Research Article

An End-to-End Automation Framework for Mobile Network Testbeds

Almudena Díaz Zayas, Bruno García García, and Pedro Merino

University of Malaga, Málaga, Spain

Correspondence should be addressed to Almudena Díaz Zayas; adz@uma.es

Received 18 January 2019; Accepted 6 March 2019; Published 1 April 2019

Guest Editor: Tara A. Yahiya

This paper describes the end-to-end automation framework developed as part of the TRIANGLE project. The TRIANGLE project is devoted to the benchmarking of apps and devices in mobile networks. For that purpose, it is needed to ensure the repeatability in the behaviour of all the components of the mobile network during the execution of the same test. This is why one of the main objectives of the project was to develop an end-to-end automation framework to provide repeatable testing. This paper describes in detail the design and the implementation of the framework.

1. Introduction

Conformance to standards [1] at the radiofrequency level is a legal requirement to allow the deployment of new mobile devices in the (regulated) mobile networks. Since the origins of 2G networks, Europe has led the creation and exploitation of conformance testing procedures and tools for mobile devices. The major vendors of smartphones apply testing methods to certify that their devices comply with the radio and signalling standards and are authorized to operate in the regulated spectrum.

However, the certification of radio and protocol signalling related features does not guarantee a good behaviour of the mobile devices when running applications, and this could be a huge problem considering the critical role apps will play in the 5G era. Underperformance of apps or unexpected failures due to the interaction of different apps in the same device could be detected with proper testing and qualification methods. The main objective of the TRIANGLE project is to develop the TRIANGLE testbed [2] for the testing of applications and the performance of devices when running apps.

In this context, the project focused on the development of a framework for testing applications in a set of repeatable scenarios on reference handsets. This means a fully controlled environment (from the network up to the handset) for which the elements behave in a reproducible manner which allows identifying differences generated by or due to the application under test. The controlled environment will allow the emulation of different radiofrequency scenarios and backend network impairments. The outcome of the test is a benchmark score comparing the application to a set of reference values and defined KPIs (key performance indicators) [3].

Current solutions for testing apps with different networking conditions are based just on software tools to emulate the provision of specific network qualities in terms of latencies or error rate (some examples are Facebook's Augmented Traffic Control [4] or Dymmynet [5]). Such software emulation means an advance with respect to pure simulation [6, 7] of the whole system in a computer because they support the connection of the real devices as part of the end-to-end path. However, the network itself is just an illusion. Emulating 5G network effects like seamless handover or changes in the quality of a bearer dynamically can hardly exhibit the same behaviour than real network equipment. Furthermore, the behaviour of the mobile device running the application should require a lot of (not always feasible) modifications to be connected to such software emulated network.

The main objective of the TRIANGLE testbed is to provide realistic 5G behaviours to unmodified commercial mobile devices. For this purpose, the testbed combines actual industrial 4G/5G testers, commercial EPCs, commercial mobile user equipment, etc. to offer emergent networking scenarios in a very flexible and programmable environment that ensures repeatability of experiments. With this approach, commercial devices can be directly connected to the platform to run the apps in the selected networking scenarios.

The main building blocks of the TRIANGLE testbed architecture are the testing framework and the testbed infrastructure, composed by hardware and software components. The testing framework covers all the software, coordination/sequencing that automates the control and management of the testbed infrastructure. It is in charge of handling and converting the test requests into actionable steps within the software and hardware portion of the testbed. This paper focuses on the description of the TRIANGLE testing framework, whose main target is the full automation of the underlying software and hardware infrastructure of the testbed.

The most relevant contributions of the paper are the description of the methodology used to implement the TRIANGLE testing framework and the introduction of the Test Automation Framework (TAP) from Keysight, a modern Microsoft.NET-based application that can be used for the development of advanced automation software for testing systems. The methodology is inspired by the testing methodology used in the telecommunication domain that is based on the specification of a set of test cases which contains the description and configuration of the testing environment and the actions to be executed during each test. This methodology ensures that the testing is repeatable, regardless of the equipment and the entity performing the certification. That means that the same test cases are adopted and implemented by a different testbed, and the results provided by both of them should be comparable. The TAP tool enables to translate the test cases into executable TAP test plans. The TAP test plans allow automating the configuration, control, and execution of the tests.

Section 2 describes the testbed infrastructure. Section 3 introduces the TAP software and additional components used to build the so-called TRIANGLE testing framework. The interaction between the testing framework and the testbed infrastructure is explained in Section 4; this section details the control interfaces implemented by the testing framework to manage each one of the components of the testbeds. Finally, Section 5 remarks the main outcomes of the TRIANGLE testbed.

2. Testbed Infrastructure

In order to provide a better understanding on how the testing framework performs, this section provides an overview of the different components which integrate the testbed.

The TRIANGLE testbed is composed of several interconnected hardware units, computers, and virtual machines. All these components work together to provide the means to execute tests over applications or devices and to provide test reports. On top of the infrastructure layer, operates the testing framework which fully automates the configuration, control, and execution of the test cases specified in the project [8] for the testing of applications and devices.

Figure 1 shows an overview of the physical interconnections between the testbed components.

The commercial mobile devices are connected to the UXM Wireless Test Set from Keysight [9]. This equipment is used traditionally in the conformance testing of mobile devices. The UXM plays the role of RAN (radio access network) in the TRIANGLE testbed. Some of its key features of the UXM are flexible intercell interference coordination (eICIC) schemes, WLAN (wireless local area network) offloading, and IMS (IP multimedia subsystem)/end-to-end VoLTE (voice over LTE) communications between multiple devices.

The radio signal is not radiated (over-the-air); instead, it is conducted through calibrated RF cables to the UE antenna connector. For testing purposes, most UEs typically contain small antenna connectors which are hidden from the user. The UXM supports the interconnection of two UEs. To connect more devices to the same UXM, the testbed uses RF switches, controlled by a switch driver. The switches are placed in the RF connection between the UEs and the UXM, and the switch driver will select which RF connections (RF paths) to be routed to the UXM.

The UXM emulates all the network signalling and physical signals, including MIMO (multiple input multiple output) configurations. All the protocol layers in the emulated network operate realistically as defined in the 3GPP (3rd Generation Partnership Project) test specifications and can be configured. Moreover, for testing purposes, the UXM instrument provides additional useful capabilities, such as a downlink channel emulator to emulate the effect of actual radiated propagation, detailed logging, and friendly control.

The battery pins of the UE (user equipment) are connected to the power analyzer N6705B. This allows both the control of the input voltage into the phone and to measure the instantaneous current drawn by the device. The N6705B power analyzer supports up to four devices connected at the same time.

In addition, the mobile device should provide some control and automation interface that can be used from the testbed orchestration tools. For instance, in the case of Android, this means a USB connection to a testbed computer to support connection through the ADB (Android Debug Bridge) tool. To support several mobile devices, the testbed use a DUT USB hub.

The following elements are connected via Ethernet in a local network: UXM, N6705B, switch driver, management server, core network, and transport. The testbed also includes a virtualized infrastructure based on OpenStack to support the local deployment of services.

The core network is a commercial EPC (evolved packet core) from Polaris Networks, which includes the main elements of a standard core network: MME (mobility management entity), SGW (serving gateway), PGW (packet data network gateway), HSS (home subscriber server), and PCRF (policy and charging rules function), which have been

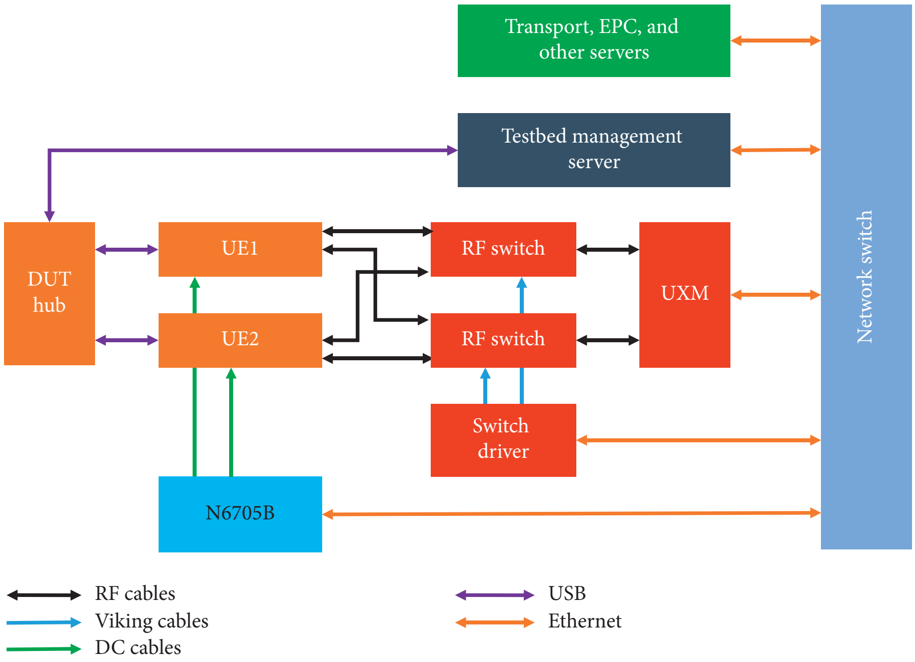


Figure 1: TRIANGLE testbed infrastructure.

integrated. In addition, this EPC includes the ePDG (evolved packet data gateway) and ANDSF (access network discovery and selection function) components for dual connectivity scenarios.

To emulate the transport network between the eNodeB and the EPC, TRIANGLE uses an SDN (software defined networking) deployment that provides features such as traffic prioritization, separation of data and control plane traffic, and transparent mirroring of selected traffic flows. Moreover, the testbed offers the possibility of integrating artificial impairments in the interfaces of the core network and the application servers.

Two software probes are deployed at the UE, the DEKRA TACS4 agents [10], and the TestelDroid tool [11] from the University of Malaga. TRIANGLE also provides an instrumentation library for application developers, in order to provide additional measurements that cannot be extracted by other means. There are additional software probes running at the core network and at the backhaul. Hardware probes include a power analyzer connected to the UE to measure power consumption and the probes available in the radio access emulator. All the probes are also under the control of the testing framework.

Finally, in the management servers, there is deployed a software tool called Quamotion WebDriver [12]. This tool is based on the Selenium technology used for web testing [13]. Quamotion WebDriver has adapted this technology for the testing of Android applications. This tool is also under the control of the testing framework to run the tests.

3. Triangle Testing Framework

The TRIANGLE testing framework has been developed in the scope of the TRIANGLE project. A key enabler for the automation implemented in the testing framework is TAP. TAP provides flexible and extensible test sequence and test plan creation. TAP is a Microsoft.NET-based application that can be used stand-alone or in combination with higher-level test executive software environments: Leveraging C# and Microsoft Visual Studio; TAP is a platform upon which it is possible to build test solutions.

Particularly, TAP has been used for the implementation of the test cases defined in the context of the TRIANGLE project [8] and the implementation of the interfaces to control the components of the testbed infrastructure. This section focuses on the implementation of the test cases, while Section 4 focuses on the description of the plugins developed to automate the control of the testbed infrastructure.

TAP allows the definition of test plans. TAP test plans consist of a sequence of test steps, which specify the set of actions that are performed on different TAP instruments (the logical entities that control the physical equipment on the platform) or control the execution flow of the test plan (for instance, repeating certain steps or executing different actions depending on the results of a previous test step). TAP defines a set of basic test steps that are suited for most user requirements, and it is possible to develop custom steps in C# for complex or very specific needs.

TAP instruments, and the functionally equivalent DUTs (devices under test), define the logic for interacting with other entities in the facility. TAP includes a generic instrument for controlling SCPI (standard commands for programmable instruments) compatible equipment, while custom TAP instruments can be developed using C#. Each TAP instrument encapsulates the configuration and management logic of the equipment, as well as defining the actions that can be performed on it.

TAP plays a key role in the implementation of the testing framework as it allows controlling and automating all the components present in the testbed. In the context of TRIANGLE, a test plan is composed by the sequence of steps specified by the test cases. In each test step, there is a set of configurable parameters. The value of the configurable parameters can be made available externally in the test plan; hence, it is possible to modify these values before the execution of a test plan, making possible to orchestrate the execution of TAP using an external entity. In this way, it is possible to use TAP for controlling the fine-grained interaction with the different equipment of the platform, while keeping a separate, upper layer for general test case orchestration.

Python has been used to implement the logic on top of TAP that decides the specific values to be configured in the test plan. The decision is based on the information stored in the YAML configuration files. The YAML files contain the values of the parameters specified in the test cases.

Python is a modern programming language with a large community of developers that continues to improve with the inclusion of new built-in modules and features on each subsequent release. Additionally, there is a large amount of ready to use, open source, frameworks, and libraries suited for common needs. This can facilitate the development of the components of the platform, for example, by using Python's standard libraries on the definition of the REST API (representational state transfer application programming interface) that supports the communication between different components.

Since Python is an interpreted language, it is also possible to modify the code rapidly, with no need for a separate compilation stage on a development machine, which can drastically reduce the complexity of testing and the time needed for the resolution of simple issues.

Figure 2 shows a more detailed view of the main functional blocks that make up the TRIANGLE testing framework. The architecture can be divided into several subsystems, whose role will be introduced briefly in this section.

End users access the TRIANGLE testbed through the Portal [14]. In this Portal, they can upload the applications under test. In addition, they will have to declare the features or capabilities of their applications. These features will define what can be tested and which test case specifications will be applicable during the testing of the application. In addition to the application features, end users will have to provide additional information; for example, they have to choose the device on which the application will be executed during the test and the scenario (network and propagation conditions, for example, pedestrian or vehicular scenarios).

In order to improve the utilization of the TRIANGLE testbed, it is possible to queue several campaign executions at a time. These campaigns will be executed successively by the testbed, avoiding idle periods of time while waiting for a user to start the execution of another campaign. This has been achieved by means of a scheduler layer between the TRIANGLE portal and the Orcomposutor.

The ORchestrator-COMPOSer-execUTOR (Orcomposutor), implemented in Python, is a server with a REST API that runs on the same Windows machine as TAP.

Once all the required information has been entered in the Portal, the TRIANGLE testbed end user can proceed with the test. The first step would be to take the collected information and turn it into executable TAP test plans. This is the task of the test plan Orcomposutor. According to the features introduced in the Portal for the application under test, the Orcompusutor will generate the applicable test plans.

To create the required TAP test plans, the Orcomposutor uses predefined TAP test plan templates. The TAP test plan plays the same role that the TTCN3 implementations of the test cases defined by the 3GPP for the certification of mobile protocols implementation [15]. When possible, the Orcomposutor will take advantage of two TAP features: the ability to expose parameters of a test step to external callers and a test step that allows the execution of another test plan. Figure 3 shows an example of test plan. The test plan is composed by a test plan reference which contains the network configuration of the test. The rest of test steps are used to configure and control the components of the testbed. In the right side of Figure 3, the configurable parameters exposed by the test steps are shown.

For instance, many test plans will start by setting up the network scenario and configuring the required parameters in the testbed equipment. This setup is the same, regardless of the body of the test plan. Thus, the Orcomposutor reuses existing TAP test plans that configure particular network scenarios.

For an application test, the body of the test plan typically includes replaying the user actions contained in an application user flow provided by the application developer. The Orcomposutor gets the application user flow from the Portal and sets the corresponding external parameter of the WebDriver replay test step. WebDriver is the tool used to interact automatically with the application under test (click on buttons, scroll).

The Orcomposutor is also aware of which KPIs are going to be measured with each of the generated TAP test plans. If necessary, the test plan should provide explicit support for performing the measurements required for the KPIs. For instance, if a test plan will contribute to a KPI on power consumption, the power analyzer must be configured and used in the test plan.

Each of the TAP test plans created by the Orcomposutor can then be executed in the Testbed using TAP. The TAP test plan contains all the information required to execute a test automatically.

During the execution of the TAP test plan, the measurement tools will gather measurements. The measurement tools that are fully integrated with TAP will publish them as usual. In this case, the results will be handled by a TAP result listener that sends them to a central OML server [16]. This OML server uses a PostgreSQL database server to store the measurements. Some tools may include OML support, and thus send their measurements directly to the OML server.

The main functions of the Orcomposutor can be summarized as follows:

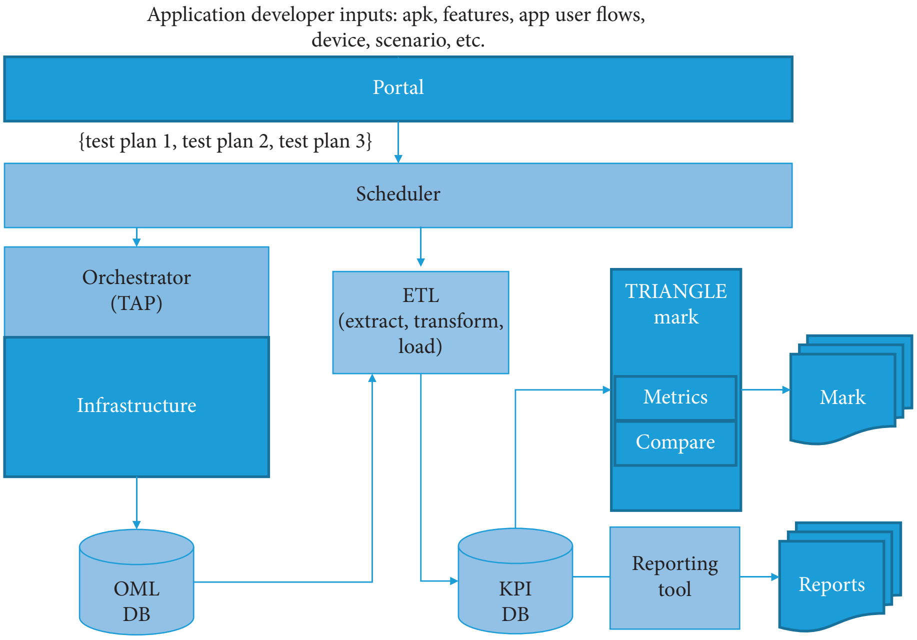
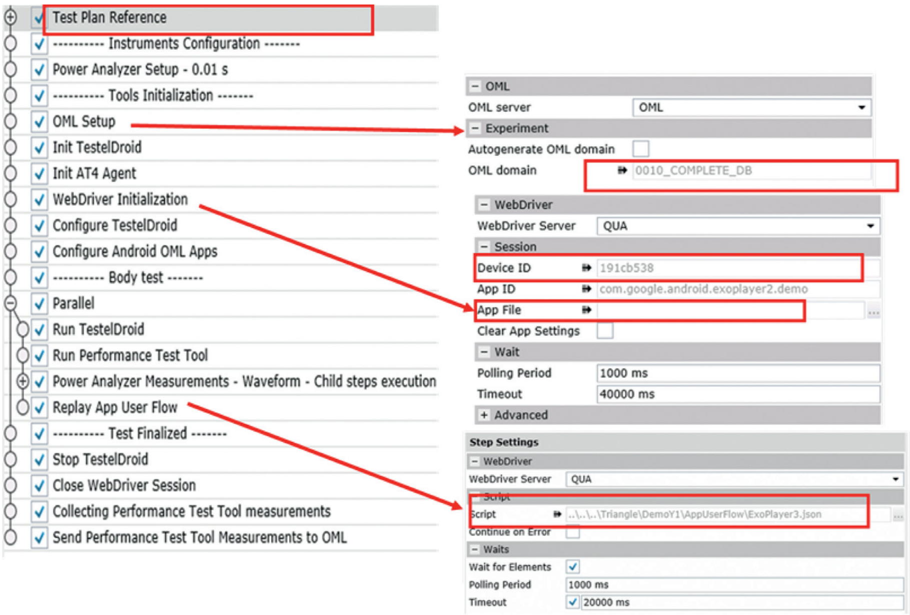

Application developer inputs: apk, features, app user flows, device, scenario, etc.
Portal
{test plan 1, test plan 2, test plan 3}
Scheduler
Orchestrator (TAP)
Infrastructure
ETL (extract, transform, load)
TRIANGLE mark
Metrics
Compare
Mark
OML DB
KPI DB
Reporting tool
Reports

Figure 2: TRIANGLE testing framework architecture (light blue boxes).


☑ Test Plan Reference
☑ ---------- Instruments Configuration -------
☑ Power Analyzer Setup - 0.01 s
☑ ---------- Tools Initialization -------
☑ OML Setup
☑ Init TestelDroid
☑ Init AT4 Agent
☑ WebDriver Initialization
☑ Configure TestelDroid
☑ Configure Android OML Apps
☑ ---------- Body test -------
☑ Parallel
☑ Run TestelDroid
☑ Run Performance Test Tool
☑ Power Analyzer Measurements - Waveform - Child steps execution
☑ Replay App User Flow
☑ ---------- Test Finalized -------
☑ Stop TestelDroid
☑ Close WebDriver Session
☑ Collecting Performance Test Tool measurements
☑ Send Performance Test Tool Measurements to OML

− OML
OML server: OML
− Experiment
Autogenerate OML domain ☐
OML domain: 0010_COMPLETE_DB
− WebDriver
WebDriver Server: QUA
− Session
Device ID: 191cb538
App ID: com.google.android.exoplayer2.demo
App File:
Clear App Settings ☐
− Wait
Polling Period: 1000 ms
Timeout: 40000 ms
+ Advanced

Step Settings
− WebDriver
WebDriver Server: QUA
− Script
Script: ..\..\..\Triangle\DemoY1\AppUserFlow\ExoPlayer3.json
Continue on Error ☐
− Waits
Wait for Elements ☑
Polling Period: 1000 ms
Timeout ☑ 20000 ms

Figure 3: TAP test plan template for executing TRIANGLE test cases.

(i) Accept test campaign execution requests from the Portal.

(ii) Compose the TAP test plans required to run a test campaign and its test cases

(iii) Execute the TAP test plans

(iv) Upload the results of the execution to the Portal

To carry out these functions, Orcomposutor needs to communicate with the REST API of the Portal and with the OML database.

The request to execute a test campaign only includes the identifier of that campaign. Orcomposutor uses that id to request more information about the test campaign to the backend using its REST API. This information is used to determine which test case or test cases must be executed with TAP. A test campaign might include more than one test case; in that case, the Orcomposutor will prepare the execution of more than one TAP test plan.

TAP test plans contain several external parameters and test plan references that must be filled in before execution

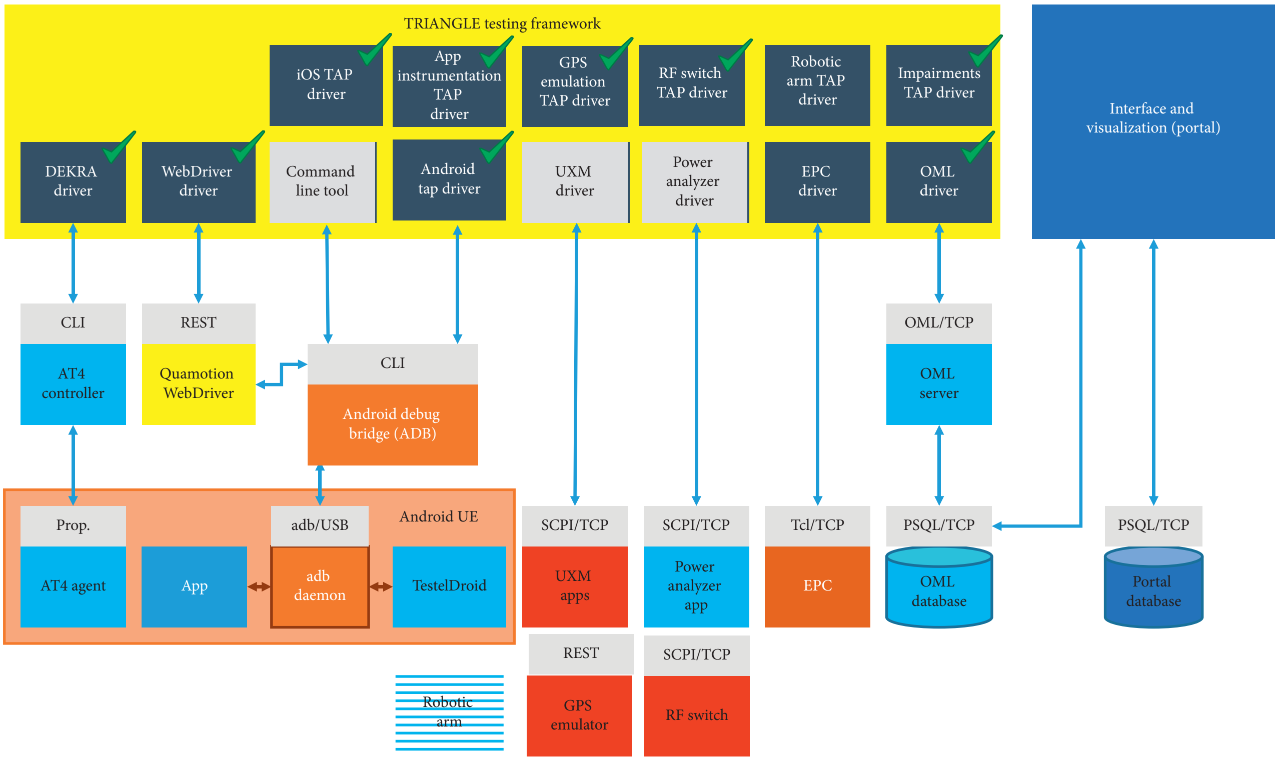


Figure 4: Testbed infrastructure automation based on TAP plugins.

can start. Orcomposutor must retrieve the appropriate information from the backend REST API in order to fill in these blanks (Figure 3), such as the id of the device used in the test case or the network scenario. We call the selection of this parameters and referenced TAP test plans the composition of the test plan.

Once the TAP test plan has been composed, it can be executed with the TAP CLI (command line interface). The Orcomposutor will store the TAP logs, as well as internal logs for diagnostic purposes.

4. Testbed Automation

In this section, we introduce the TAP plugins implemented to control each one of the components of the testbed. The plugins contains the test steps used in the TAP test plan shown in Figure 3. In Section 3 was introduced TAP and its usage as sequencer of actions to configure and control the components. Figure 4 shows the software interfaces (TAP plugins) developed to communicate the testing framework and the infrastructure components.

In TAP, an instrument is a logical entity that encapsulates the interaction with a physical instrument. At the very least, a TAP instrument must define all the necessary logic for connecting and disconnecting the TAP host machine with the real instrument, and it is a best practice to define methods for performing every possible (or required) actions that the instrument can execute.

For example, a TAP instrument for controlling a real power supply via SCPI must include two methods (open and close) that create and release the connection with the instrument and can have two extra methods for setting the voltage and current.

Instruments are extensively used in TAP steps, which expose the instrument functionality to the end user. Continuing with the previous example, we could define two steps for this instrument: one that turns off the power output, setting voltage and current to 0, and another that sets the voltage and the current using the values selected by the user in the test step configuration. TAP will automatically use the open and close methods from the instrument when required.

Many components offer a SCPI interface to receive commands. TAP provides support for writing drivers for SCPI-based components, which facilitates the work of adding more components. This is the case of the drivers for the UXM and Power Analyzer apps running on their corresponding hardware units. In both cases, the SCPI commands are delivered through a TCP connection.

For Android apps, the ADB command-line tool is a fundamental component. ADB can be used to send commands to apps running on the UEs or automate certain actions such as switching airplane mode ON and OFF to force the attachment of the UE to the base station. Some of the UE automation tools that are part of the testbed, such as Quamotion WebDriver, use ADB to perform their function.

The Quamotion WebDriver provides a REST API to manage the apps on the Android device and perform user actions, such as tapping or swiping. These commands are delivered to the Android UE using ADB. TestelDroid is also managed through ADB.

DEKRA TACS4 tool is managed through the proprietary control interface provided by the tool.

The EPC can be controlled through a fixed set of TCL (tool command language) scripts that use a TCL scripting API provided by the EPC components emulators. The TAP driver will execute these scripts, which will then send the appropriate commands to the EPC through the TCL API.

The measurements from all the probes are collected and sent to a central OML (ORBIT measurement framework and library) server, which uses a custom OML protocol. While some tools may send measurements directly to the OML server, the TAP orchestrator will use a driver that will allow sending measurements to the OML server from TAP, in two ways. First, for tools that generate results in CSV (comma-separated values) files, the driver will collect these files and send them as measurements to the OML server. Second, the driver will implement the standard TAP mechanism for handling results from drivers, so that drivers which are already well integrated with TAP can publish them to the OML server without additional work.

5. Conclusions

One of the main contributions of the TRIANGLE project is the design and development of a testing framework that automates the end-to-end configuration and control of a mobile communication testbed. The testing framework is based on TAP (test automation platform), a powerful editor of test sequences that also includes an SDK for the development of plugins for components offering a control interface.

The testing framework has proven to be flexible and sustainable by integrating a large number of components different in nature such as RAN (radio access network) equipment, a core network, instruments for measuring power consumption, mobile devices, and software tools acting as probes.

Moreover, the TRIANGLE testing framework is going to be adopted in the 5GENESIS project [17] in the experiment life cycle manager component of the coordination layer of the 5GENESIS facility. 5GENESIS is a project devoted to the realization of a 5G platform that will allow the validation of the major 5G key performance indicators (KPIs). 5GENESIS project is integrated by five platforms: Athens, Berlin, Limassol, Malaga, and Surrey.

Data Availability

This paper is about one of the main outputs of the Triangle project, the end-to-end automation of the TRIANGLE testbed. More detailed information about the the project can be found at https://www.triangle-project.eu/project-old/deliverables/. In particular, automation is described in WP3 deliverables. UMA is the WP3 leader and the partner in charge of the automation task.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

The TRIANGLE project was funded by the European Union's Horizon 2020 Research and Innovation Program (grant agreement no. 688712). The 5GENESIS project was funded by the European Union's Horizon 2020 research and innovation programme under grant agreement No. 815178.

References

[1] W. Lin, H. Zeng, H. Gao, H. Miao, and X. Wang, "Test sequence reduction of wireless protocol conformance testing to internet of things," *Security and Communication Networks*, vol. 2018, Article ID 3723691, 13 pages, 2018.

[2] A. F. Cattoni, G. C. Madueño, M. Dieudonne et al., "An end-to-end testing ecosystem for 5G the TRIANGLE testing house testbed," *Journal of Green Engineering*, vol. 6, no. 3, pp. 285–316, 2016.

[3] A. D. Zayas, L. Panizo, J. Baños, C. Cárdenas, and M. Dieudonne, "QoE evaluation: the TRIANGLE testbed approach," *Wireless Communications and Mobile Computing*, vol. 2018, Article ID 6202854, 12 pages, 2018.

[4] B. Blywis, M. Guenes, F. Juraschek, and J. H. Schiller, "Trends, advances, and challenges in testbed-based wireless mesh network research," *Mobile Networks and Applications*, vol. 15, no. 3, pp. 315–329, 2010.

[5] A. Zubow and R. Sombrutzki, "A low-cost MIMO mesh testbed based on 802.11n," in *Proceedings of the 2012 IEEE Wireless Communications and Networking Conference (WCNC)*, pp. 3171–3176, Paris, France, April 2012.

[6] R. Marco Alaez, J. M. Alcaraz Calero, Q. Wang et al., "Open-source based testbed for multioperator 4G/5G infrastructure sharing in virtual environments," *Wireless Communications and Mobile Computing*, vol. 2017, Article ID 1984314, 11 pages, 2017.

[7] C.-K. Jao, C.-Y. Wang, T.-Y. Yeh et al., "WiSE: a system-level simulator for 5G mobile networks," *IEEE Wireless Communications*, vol. 25, no. 2, pp. 4–7, 2018.

[8] EU H2020, "TRIANGLE project, deliverable D. 2.2 final report on the formalization of the certification process, requirements and use cases," https://www.triangle-project.eu/project-old/deliverables/, 2017.

[9] E7515A UXM, "Wireless test set getting started guide," 2017, https://www.keysight.com/main/gated.jspx?lb=1&gatedId=2459161&cc=US&lc=eng&parentContId=2371785&parentContType=ct&parentNid=-32909.0&fileType=VIEWABLE.

[10] TACS4 Performance Testing Platform, https://performance.tacs4.com/.

[11] A. Álvarez, A. Díaz, P. Merino, and F. J. Rivas, "Field measurements of mobile services with Android smartphones," in *Proceedings of the 2012 IEEE Consumer Communications and Networking Conference (CCNC)*, pp. 105–109, Las Vegas, NV, USA, January 2012.

[12] Quamotion WebDriver, http://docs.quamotion.mobi/webdriver/.

[13] A. Bruns, A. Kornstadt, and D. Wichmann, "Web application tests with Selenium," *IEEE Software*, vol. 26, no. 5, pp. 88–91, 2009.

[14] A. Díaz-Zayas, A. Salmerón Moreno, G. García Pascual, and P. Merino Gómez, "TRIANGLE portal: an user-friendly web interface for remote experimentation," in *Smart Industry & Smart Education*, M. Auer and R. Langmann, Eds., Springer, Cham, Switzerland, 2019.

[15] ETSI Standard ES 201 873-1 V3.4.1, The Testing and Test Control Notation Version 3; Part 1: TTCN-3 Core Language, 2008, Sophia Antipolis, France, European Telecommunications Standards Institute (ETSI).

[16] M. Singh, M. Ott, I. Seskar, and P. Kamat, "ORBIT measurements framework and library (OML), motivations, design, implementation, and features," in *Proceedings of the First International Conference on Testbeds and Research Infrastructures for the Development of Networks and Communities (Tridentcom 2005)*, Trento, Italy, February 2005.

[17] H. Koumaras, D. Tsolkas, G. Gardikis et al., "5GENESIS: the genesis of a flexible 5G facility," in *Proceedings of the 23rd International Workshop on Computer Aided Modeling and Design of Communication Links and Networks (CAMAD)*, pp. 17–19, Barcelona, Spain, September 2018.

Hindawi
Mobile Information Systems
Volume 2019, Article ID 4370847, 12 pages
https://doi.org/10.1155/2019/4370847

Research Article

Independence and Fairness Analysis of 5G mmWave Operators Utilizing Spectrum Sharing Approach

Mothana L. Attiah, [1,2] A. A. M. Isa, [1] Zahriladha Zakaria, [1] M. K. Abdulhameed, [1] Mowafak K. Mohsen, [1] and Ahmed M. Dinar [1]

[1] *Centre for Telecommunication Research and Innovation (CeTRI), Faculty of Electronic and Computer Engineering (FKeKK), Universiti Teknikal Malaysia Melaka (UTeM), Durian Tunggal, Malaysia*
[2] *Department of Computer Engineering, Electrical Engineering Technical College, Middle Technical University, Baghdad, Iraq*

Correspondence should be addressed to Mothana L. Attiah; mothana.utem@gmail.com and Zahriladha Zakaria; zahriladha@utem.edu.my

Received 16 October 2018; Accepted 10 January 2019; Published 14 February 2019

Guest Editor: Pinar Kirci

The spectrum sharing approach (SSA) has emerged as a cost-efficient solution for the enhancement of spectrum utilization to meet the stringent requirements of 5G systems. However, the realization of SSA in 5G mmWave cellular networks from technical and regulatory perspectives could be challenging. Therefore, in this paper, an analytical framework involving a flexible hybrid mmWave SSA is presented to assess the effectiveness of SSA and investigate its influence on network functionality in terms of independence and fairness among operators. Two mmWave frequencies (28 GHz and 73 GHz) are used with different spectrum bandwidths. Various access models have been presented for adoption by four independent mobile network operators that incorporate three types of spectrum allocation (exclusive, semipooled, and fully pooled access). Furthermore, an adaptive multi-state mmWave cell selection scheme is proposed to associate typical users with the tagged mmWave base stations that provide a great signal-to-interference plus noise ratio, thereby maintaining reliable connections and enriching user experience. Numerical results show that the proposed strategy achieves considerable improvement in terms of fairness and independence among operators, which paves the way for further research activities that would provide better insight and encourage mobile network operators to rely on SSA.

1. Introduction

Future mobile data usage and traffic growth are driven by diverse and innovative technologies and services, such as smart cities, health care, autonomous driving, augmented reality, virtual reality, and Internet of things [1]. However, today's extremely limited spectrum bandwidth at low frequencies (<6 GHz) can no longer accommodate novel and rapidly evolving applications [2]. This challenge underlines the need for freeing up additional spectrum to cope with the stringent requirements of bandwidth-hungry applications. In response to the bounded amount of spectrum, mmWave frequency bands have been recently introduced as an attractive enabling technology to address the spectrum shortage [3] because it has an ample amount of available spectrum with a multigigahertz range [4, 5]. Despite such wide spectrum range, it is still not unlimited if other services that utilize the same bands are considered [6]. In addition, the inefficiency of spectrum utilization is expected to occur remarkably in mmWave bands if a large chunk of spectrum is exclusively granted to a single independent mobile network operator (IMNO) [7]. Accordingly, spectrum sharing approach (SSA) is an option that can overcome such issue in a cost-effective manner [8]. However, pursuing such approach in 5G mmWave cellular networks is a major decision that requires extensive research to study its effectiveness and implications from technical, regulatory, and economic perspectives. Achieving considerable improvement in spectrum utilization via SSA without sacrificing the merits associated with traditional spectrum allocation (e.g., exclusive access) remains a major challenge that should be solved in a joint manner [9].

Several studies on the assessment of SSA implementation in mmWave communications have recently been conducted. In [10], a realistic indoor propagation channel and antenna models for mmWave networks were adopted along with link-specific context coordination to simulate an internetwork spectrum sharing strategy. Preliminary results demonstrated the viability of the proposed strategy in maximizing the throughput and eliminating the interference among operators.

Similarly in [6], many aspects regarding the technical enablers of SSA were addressed (e.g., beam directionality, base station (BS) density, and coordination) to study their influence on a set of network functionalities. Results proved the possibility of this approach in improving spectrum usage compared with the exclusive access spectrum allocation. In addition, the importance of interoperator coordination, especially for cell edge users, was indicated. Different from the above studies, the potential success of uncoordinated SSA was investigated in [5, 6, 11–13]. These studies showed the effectiveness of mmWave characteristics along with the directional beamforming technique in reducing cross-operator interference among multi-IMNOs, thereby eliminating the need for coordination among operators.

On the contrary, the authors in [14] confirmed that without coordination, sharing the spectrum among multi-IMNOs with different mmWave BS (mBS) deployment densities remains a great challenge, particularly when the mBS density of the interfering operator is higher than the operator with low mBS density. This observation may discourage the low-density operator to share their spectrum unless a low interference level is maintained among the multi-IMNOs. Moreover, in [15], two different network densities (i.e., fixed individual and fixed combined) with two mmWave cellular operators were suggested to model multi-IMNOs with colocated BSs; these multi-IMNOs can be reproduced and extended to any set of operators that allow straightforward analysis of key performance metrics (e.g., SINR). The analysis showed that infrastructure and spectrum sharing is more convenient for high-rate applications rather than low-rate ones.

In the present work, we extend the prior studies detailed above and our work in [8]. New assumptions are considered regarding the utilization of the hybrid mmWave spectrum sharing access (HMSSA) strategy, different path-loss models (commonly used in the literature), network planning, and agility improvements of operators with an acceptable level of mBS density. We also propose two access models for adoption by multi-IMNOs. To the best of the authors' knowledge, this study is the first to provide an analysis and deep discussion of two major challenges that face the successful realization of spectrum sharing among multiple mmWave entities. These two challenges are independent and fair, which may discourage operators to share their spectrum unless an acceptable trade-off is attained.

2. System Models

In this section, the proposed analytical framework is divided into four parts to simulate and apply the proposed HMSSA strategy accurately. Details are as follows.

2.1. Network Model. To serve a recognizable area, we consider two tiers of multi-IMNOs given by M, and each operator's network m^{th} has two spectrum bandwidths based on two carrier frequencies (28 and 73 GHz) given by c. Without loss of generality, let $W^{\text{m,c}}$ denote the total spectrum that is allocated to each operator m^{th}.

Let S_{m} be a set of mBSs of operator m^{th} and $S = \{S_1 \cup S_1 \ldots \cup S_{\text{m}}\}$ be a set of all mBSs in the network. However, all operators have their own mBSs S_{m} that can operate optionally at the two aforementioned mmWave carrier frequencies (28 GHz and 73 GHz). Notably, all mBSs are densely deployed and distributed as grid-based in an overlapping area that provides high coverage and QoS to a large number of user equipments (UEs), such that the simulation area is 1.2 km × 1.2 km. Additionally, all mBSs and UEs are assumed to be powered by multiple antenna (8 × 8). Each mBS has the right to grant a part of its allocated spectrum $W^{\text{m,l}}$ exclusively for users that belong to its operator in the lower mmWave band (28 GHz) and share a part of its allocated spectrum $W^{\text{m,h}}$ semiorthogonally or fully orthogonally to the users that belong to that operator or to other operators in the higher mmWave band (73 GHz). Let u denote a set of outdoor UEs and $u = \{u_1 \cup u_1 \ldots \cup u_{\text{m}}\}$, where u_{m} is a set of users of all operators. Each $u^{\text{th,m}}$ is served by a set of mBSs S_{m}, which either belong to the same or to different network operators based on spectrum regulation and link quality.

2.2. Mathematical Model. In this study, two types of mathematical expressions have been considered. The first is related to basic mobile communications, and the second is related to the mmWave communication system. They are derived and rewritten to model the proposed strategy and the baseline environments optimally. In the context of determining the special behavior of the overall hybrid mmWave spectrum sharing system, capturing one or more snapshots helps in gaining more insight on such approach and its implications on user experience and operator's revenue. We consider the commonly used close-in reference distance path-loss model [16–18] to calculate the received signal power at the receiving antenna:

$$PL(d_{\text{us}})^{\text{m,c}} = PL_{\text{fs}}(d_{\text{o}}) + 10 \times \gamma \times \log_{10}\left(\frac{d_{\text{us}}}{d_{\text{o}}}\right) + x_{\sigma}, \quad (1)$$

where $PL(d_{\text{us}})^{\text{m,c}}$ denotes the radio propagation path loss in dB; d_{us} denotes the separation distance in meters; d_{o} denotes the close-in free space reference distance (1 m); $PL_{\text{fs}}(d_{\text{o}})$ denotes the initial path loss in dB, which can be calculated using equation (2); γ denotes the radio propagation path-loss exponent; and x_{σ} denotes the zero mean Gaussian random variable with standard deviation in dB represented by (σ), given that a 10 dB shadowing margin is used in this work.

$$PL_{\mathrm{fs}}\left(d_{\mathrm{o}}\right) = 20 \times \log_{10}\left(\frac{4 \times \pi \times d_{\mathrm{o}}}{\lambda}\right), \tag{2}$$

where λ denotes the wavelength of the carrier wave. The radio propagation path-loss exponent and wavelength of both mmWave frequencies (28 GHz and 73 GHz) are listed in Table 1.

After applying equation (1), the average received signal power at the receiver can be calculated as follows [19]:

$$\mathrm{Pr} = P_{\mathrm{t}} + G_{\mathrm{t}} + G_{\mathrm{t}} - PL. \tag{3}$$

However, equation (3) is rewritten to handle the hybrid configuration brought about by the utilization of hybrid mBS deployment as follows:

$$\mathrm{Pr}_{\mathrm{us}}^{\mathrm{m,c}} = \mathrm{P}_{\mathrm{t}}^{\mathrm{m,c}} + \mathrm{G}_{\mathrm{t}}^{\mathrm{m,c}} + \mathrm{G}_{\mathrm{r}}^{\mathrm{m,c}} - PL_{\mathrm{us}}^{\mathrm{m,c}}, \tag{4}$$

where $\mathrm{Pr}_{\mathrm{us}}^{\mathrm{m,c}}$ and $\mathrm{P}_{\mathrm{t}}^{\mathrm{m,c}}$ are the received and transmitted power of mBS $S^{\mathrm{th,m}}$, respectively, that is, the mBS owned by operator m^{th} and operated at mmWave carrier frequency c and $\mathrm{G}_{\mathrm{t}}^{\mathrm{m,c}}$ and $\mathrm{G}_{\mathrm{r}}^{\mathrm{m,c}}$ are the linear gains of the transmitter and the receiver antennas in dBi, respectively.

To assess the feasibility of the proposed HMSSA strategy and characterize the performance of each operator of the multi-IMNOs, we consider the coverage probability as an indicator when SINR > threshold. For example, user $u^{\mathrm{th,m}}$, who associates with mBS $S^{\mathrm{th,m}}$ that is owned by the same or different operator m^{th} and shares or exclusively grants a particular portion of their spectrum available in either 28 GHz or 73 GHz carrier frequencies (c), is in outage if the SINR of that user is below than zero. The SINR of user $u^{\mathrm{th,m}}$ can be calculated as follows [20]:

$$\mathfrak{X}_{\mathrm{us}}^{\mathrm{m,c}} = \frac{\mathrm{Pr}_{\mathrm{us}}^{\mathrm{m,c}}}{\sum_{n=1}^{N} I_{\mathrm{us}}^{\mathrm{m,c}} + \eta^{\mathrm{m,c}}}, \tag{5}$$

where $\mathfrak{X}_{\mathrm{us}}^{\mathrm{m,c}}$ denotes the SINR and $\sum_{n=1}^{N} I_{\mathrm{us}}^{\mathrm{m,c}}$ denotes the aggregated interference received by receiver $u^{\mathrm{th,m}}$ from all neighboring mBSs that operate at the same frequency band except for the serving mBS $S^{\mathrm{th,m}}$, regardless if they belong to the same IMNO. Specifically, we assume that only a single beam comes from each mBS $S^{\mathrm{th,m}}$ that interferes receiver $u^{\mathrm{th,m}}$. $\eta^{\mathrm{m,c}}$ denotes the additive white noise power of operator m^{th} for a carrier frequency c and is obtained as follows [19]:

$$\eta^{\mathrm{m,c}} = 10 \times \log_{10}\left(\mathrm{KT}_{\mathrm{sys}}\right) + 10 \times \log_{10} W_{\mathrm{m,c}} + \mathrm{NF}^{\mathrm{m,c}}, \tag{6}$$

where $10 \times \log_{10}(KT_{\mathrm{sys}})$ for a given system temperature (17°C) equal to −174 dBm/Hz and NF_{u} denotes the noise figure of the $u^{\mathrm{th,m}}$ with a value of 6 dB. The calculated values of the SINR $\mathfrak{X}_{\mathrm{us}}^{\mathrm{m,c}}$ provide further user channel capacity calculation; thus, the average rate of user $u^{\mathrm{th,m}}$ can be calculated using the Shannon capacity theory as follows:

$$R_{\mathrm{us}}^{\mathrm{m,c}} = \Phi_{\mathrm{s}}^{\mathrm{m,c}} \times \left(\frac{W^{\mathrm{m,c}}}{u_{\mathrm{s}}\mathrm{th}}\right) \times \log_2\left(1 + X_{\mathrm{us}}^{\mathrm{m,c}}\right), \tag{7}$$

where $\Phi_{\mathrm{s}}^{\mathrm{m,c}}$ denotes the number of antenna elements in the connected mBS $S^{\mathrm{th,m}}$; $W^{\mathrm{m,c}}$ is the total amount of spectrum bandwidth of the specified m^{th}; $R_{\mathrm{us}}^{\mathrm{m,c}}$ denotes the channel capacity of the $u^{\mathrm{th,m}}$ channel; and u_{S}th denotes the number of users connected to the tagged $S^{\mathrm{th,m}}$.

Table 1: Path-loss exponent and wavelength parameters.

Frequency bands (GHz)	γ (dB)	λ (mm)
28	3.4	10.71
73	3.3	4.106

2.3. HMSSA Strategy Configurations. In this section, we present the most important configurations of the proposed HMSSA strategy and its models in detail. Four multi-IMNOs are considered and distributed throughout the simulation area of 1.2 km × 1.2 km. A square grid-based cell deployment topology is used to ensure high-quality network coverage and mimic the quickest possible cell deployment, such as installing cells on street lamp posts. Two access models are suggested for utilization by the four operators. Each operator shares a part of its own allocated spectrum $W^{\mathrm{m,c}}$ with other operators and exclusively grants the remaining part to its own subscribers $u^{\mathrm{th,m}}$, as detailed as follows:

(i) *Model 1.* In this model, we assume that the same spectrum bandwidth (1 GHz) allocated to the four operators at the low frequency 28 GHz ($W^{\mathrm{m},\lambda}$) and at the high frequency 73 GHz ($W^{(\mathrm{m,h})}$). However, the spectrum at the low-frequency band of 28 GHz ($W^{\mathrm{m},\lambda}$) is divided evenly into four parts, each with 250 MHz to be granted exclusively to one of the four multi-IMNOs to avoid cochannel interference phenomenon with other adjacent IMNOs. Meanwhile, the spectrum at the high-frequency band of 73 GHz ($W^{\mathrm{m,h}}$) is divided into two portions, each with 500 MHz. The first part is open (pooled/shared) for all multi-IMNOs, whereas the second part is divided into two portions, each portion is assigned as semipooled/shared by two multi-IMNOs. For example, the first part (250 GHz) is granted to OP1 and OP4, and the second part (250 GHz) is granted to OP2 and OP3 as shown in Figure 1.

(ii) *Model 2.* In this model, we assume two different sets of spectrum, that is, $W^{\mathrm{m},\lambda} = 1$ GHz at the low-frequency band of 28 GHz, and $W^{\mathrm{m,h}} = 1.5$ GHz at the high-frequency band of 73 GHz. The spectrum assignment is similar to that in Model 1 for the low-frequency band of 28 GHz. However, at the high-frequency band of 73 GHz, the spectrum is divided into two parts; one with 1 GHz and the second with 500 MHz. The first part is evenly divided into four parts, and each is granted to IMNOs with exclusive access only to its subscribers $u^{\mathrm{th,m}}$. In this assignment, the cochannel interference is nonexistent. The remaining amount (500 MHz) of the spectrum at 73 GHz is shared/pooled among the four multi-IMNOs. However, in the case of open-access mode, cochannel interference will exist among all adjacent operators, as shown in Figure 2.

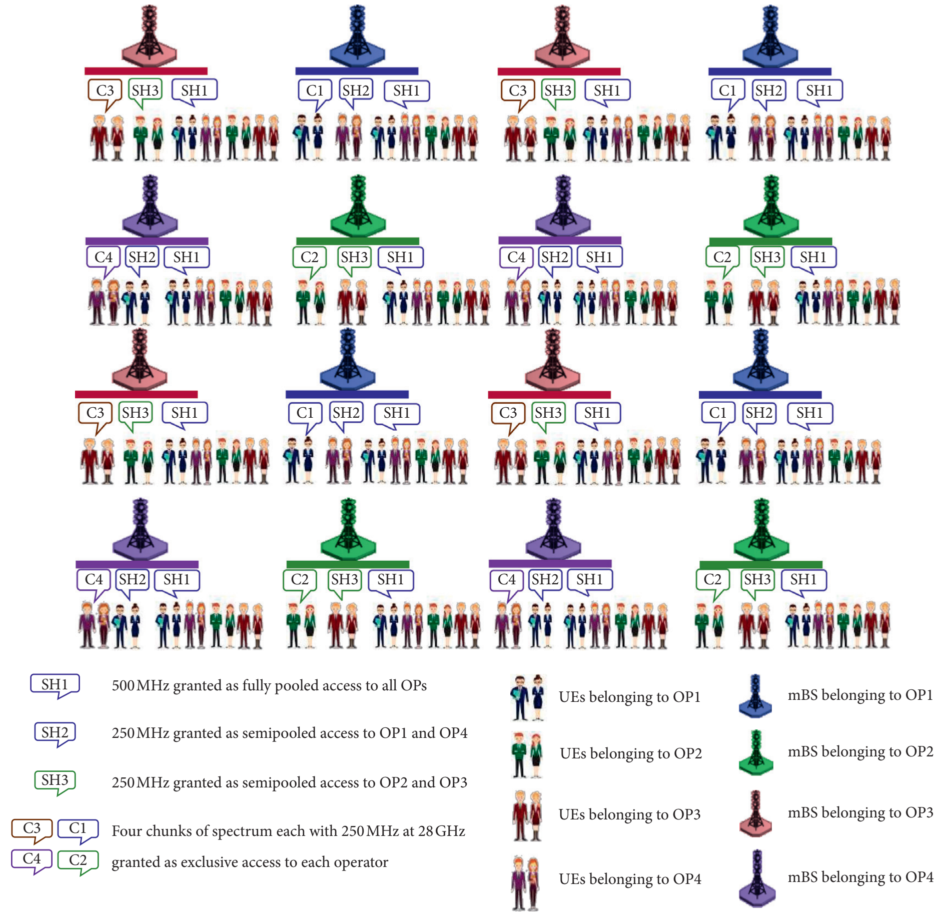

C3
SH3
SH1
C1
SH2
SH1
C4
SH2
SH1
C2
SH3
SH1
SH1
500 MHz granted as fully pooled access to all OPs
SH2
250 MHz granted as semipooled access to OP1 and OP4
SH3
250 MHz granted as semipooled access to OP2 and OP3
C3 C1
C4 C2
Four chunks of spectrum each with 250 MHz at 28 GHz granted as exclusive access to each operator
UEs belonging to OP1
UEs belonging to OP2
UEs belonging to OP3
UEs belonging to OP4
mBS belonging to OP1
mBS belonging to OP2
mBS belonging to OP3
mBS belonging to OP4

Figure 1: HMSSA Model 1.

2.4. UE-mmWave BS Association Scheme. In the proposed network configurations, rental or colocated-based mBS mode is suitable for adoption in of HMSSA strategy. In the first mode, each operator allows the rental of a part of its resources and infrastructures that are necessary for enabling efficient spectrum sharing among the multi-IMNOs. In the second mode, each operator has its own mBSs, which are hosted by other operators, provided that it is supplied with a part of the host's resources, location, cooling, and power supply.

In case of user and mBS association, the UEs that are subscribed to operator m^{th} have the right to associate with the mBS $S^{\text{th},m}$ that belongs to that operator or to other operators who share their resources based on the aforementioned modes (i.e., rent or colocated mode). In view of the proposed access strategy under Model 1, without loss of generality, three options are available for the users to associate with an mBS, which are described as follows:

(i) UEs can associate with an mBS that offers an exclusive access to 250 MHz at 28 GHz that belongs to the same operator.

(ii) UEs that are owned by one of a particular pair (OP1 and OP4 or OP2 and OP3, as assumed in this work) can associate with an mBS that belongs to the same or to the second operator of the same pair, which offers a semipooled access of 250 MHz at 73 GHz and vice versa.

(iii) UEs can associate with an mBS that belongs to OP1, OP2, OP3, or OP4, which offers a fully shared/pooled access of 500 MHz of the spectrum.

In Model 2, the UEs that are subscribed to the operator m^{th} have the right to associate with mBS $S^{\text{th},m}$ that belongs to that operator or to a different operator sharing the same frequency band based on the same constraints and options in Model 1; otherwise, the UEs that are owned by one of a particular pair can only associate with an mBS that belongs to the second operator of the same pair that offers an exclusive access of 250 MHz at 73 GHz and vice versa. In this case, the interference will be lower than that in Model 1, which utilizes semipooled spectrum access.

The user and cell association decisions are performed by using our proposed scheme, namely, AMMC-S, which relies on providing an optimal cell selection based on the offered signal quality as a function of SINR. For example, $u^{\text{th},1}$ is located closer to the four mBSs (i.e. $S^{\text{th},1}$, $S^{\text{th},2}$, $S^{\text{th},3}$, and $S^{\text{th},4}$)that belong to the four operators, as shown in Figures 3(a)–3(b). The $u^{\text{th},1}$ associates adaptively to $S^{\text{th},1}$ based on an exclusive

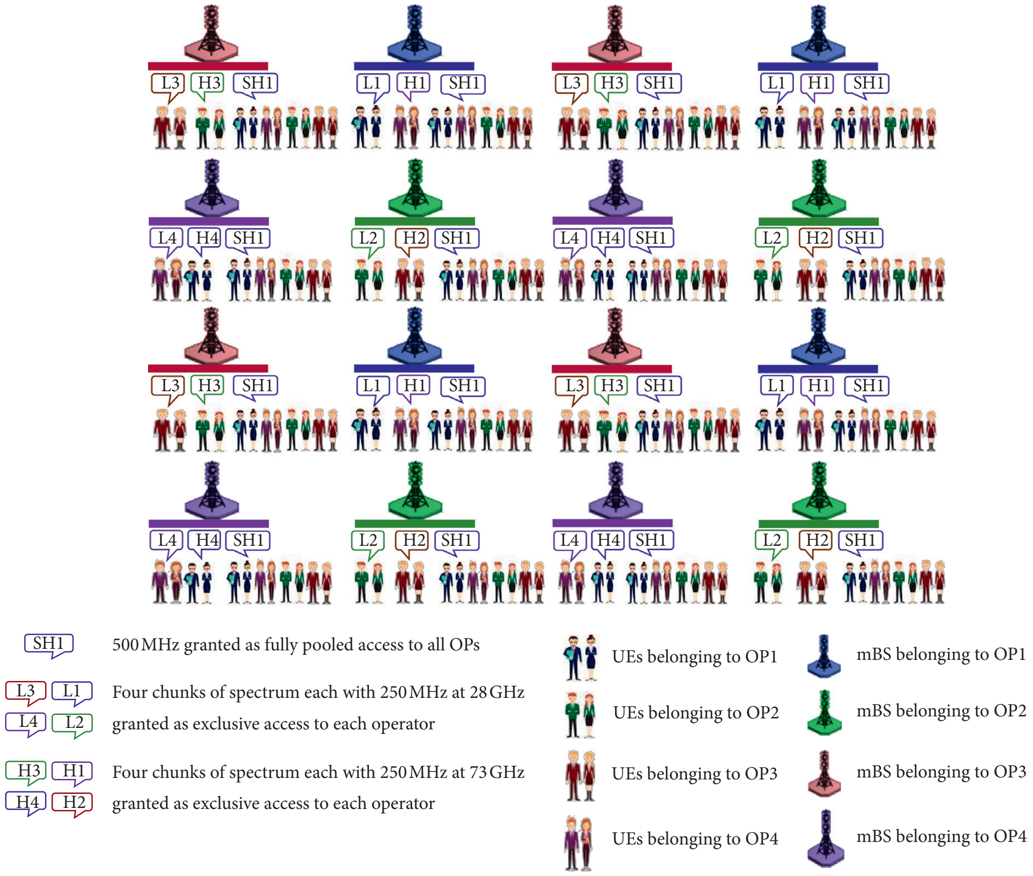


Figure 2: HMSSA Model 2.

access of 250 MHz at a 28 GHz carrier frequency that provides the highest SINR value to user $u^{\text{th},1}$ as illustrated in Algorithm 1.

3. Results and Discussion

In this section, the performance of the proposed HMSSA strategy is assessed numerically in a typical mmWave scenario that supports two hybrid access models based on mBS distribution and spectrum allocation. Two key performance metrics (i.e., outage probability and average rate distributions) are considered in the evaluation and assessment process. These performance metrics are tailored for the assessment of operator's independence and fairness, which is the main goal of this study. The related assumptions and simulation parameters are set, as shown in Table 2.

3.1. SINR Distributions. SINR represents a key system interference indicator to account for system interference and analyze its effect on network functionality. Typically, this is obtained by dividing the average received signal power by the sum between the noise power and the interfering power at the UE location as illustrated in equation (5). The lower SINR value the higher level of interference experienced by the UE from the adjacent mBSs. On the other hand, the SINR level is a measure to determine system coverage of the wireless network which represents one of the most distinct parameters in the future 5G use cases; thus, studying its influence on 5G systems is required to assess system performance. In this context, the SINR distributions of the proposed strategy with respect to the two models have been studied, as detailed in the following subsections.

3.1.1. HMSSA Model 1. Figure 4 shows the outage probability of the four operators (i.e., OP1, OP2, OP3, and OP4) based on different allocated bandwidth percentiles (5%, 50%, and 95%) utilizing HMSSA under Model 1. The SINR distributions are averaged over a sufficient number of iterations to achieve the desired accuracy. Notably, the outage probability of the users that have exclusive access to the spectrum (250 MHz) at 28 GHz carrier frequency is lower than that of the semipooled and fully pooled spectrum access at a 73 GHz carrier frequency. This phenomenon is caused by the fact that semipooled access and fully pooled spectrum access are semiopen or fully open; hence, the amount of interference is larger than that in the exclusive spectrum access. The semipooled access strategy operates seven adjacent mBSs, whereas the fully pooled strategy operates fifteen. By contrast, only three mBSs operate in the exclusive access, except the serving mBS (see Figure 1). However, the location of user $u^{\text{th,m}}$ in terms of mBS $S^{\text{th,m}}$ generally plays a dominant role in minimizing outage probability. The fully

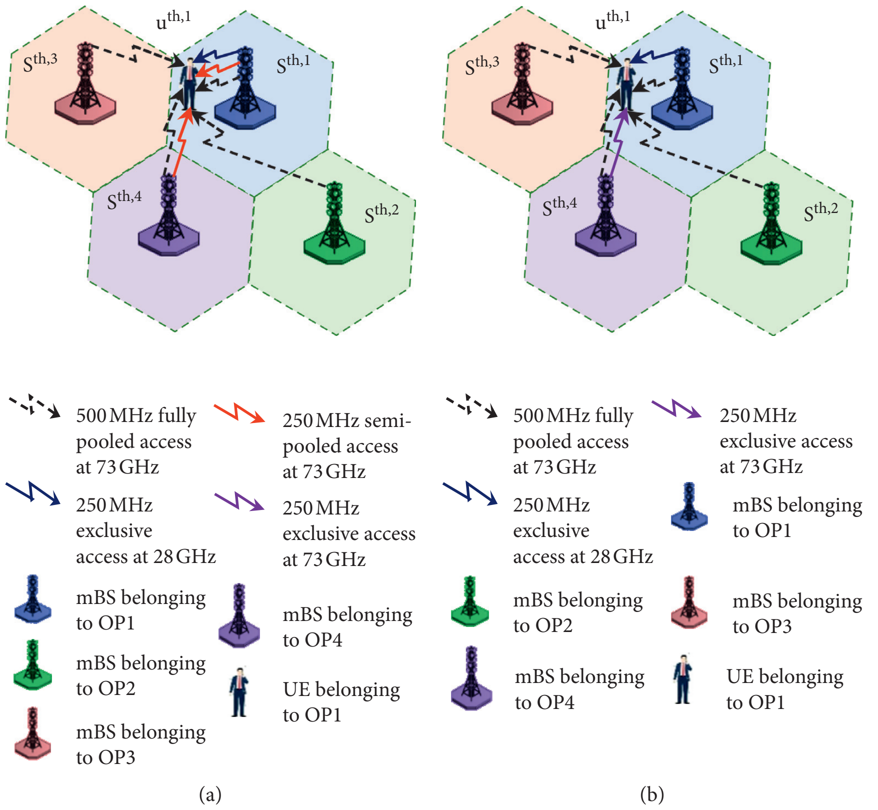


Figure 3: User-association options (a) HMSSA Model 1 (b) HMSSA Model 2.

```
Input: Set the initial parameters of $\forall m^{\mathrm{th}} \in M$, $\forall S^{\mathrm{th,m}} \in S$, $\forall u^{\mathrm{th,m}} \in u$, $\forall W^{\mathrm{m},\lambda}$ and $W^{\mathrm{m,h}} \in W^{\mathrm{m,c}}$, $\mathrm{P}_{\mathrm{t}}^{\mathrm{m,c}}$, $\mathrm{NF}_{\mathrm{u}}$
Deploy $S$, $u$ of all operators throughout the simulation area ($1.2\,\mathrm{km} \times 1.2\,\mathrm{km}$);
for $\forall m^{\mathrm{th}} \in M$ and $\forall u^{\mathrm{th,m}} \in u$ do
    Compute the distance $\forall u^{\mathrm{th,m}}$ in terms of $\forall S^{\mathrm{th,m}}$ that belong to the same or different mobile network operator (MNO);
    Compute $PL_{\mathrm{fs}}(d_{\mathrm{o}})$ $PL(d_{\mathrm{us}})^{\mathrm{m,c}}$, and $\mathrm{Pr}_{\mathrm{us}}^{\mathrm{m,c}}$ of $\forall u^{\mathrm{th,m}}$ according to equations (1), (2), and (4);
    Compute $\mathfrak{X}_{\mathrm{us}}^{\mathrm{m,c}}$ of $\forall u^{\mathrm{th,m}}$ in terms of $\forall S^{\mathrm{th,m}}$ that belong to the same or different MNO using equation (5);
    Associates $\forall u^{\mathrm{th,m}}$ to the tagged $S^{\mathrm{th,m}}$ that offers maximum $\mathfrak{X}_{\mathrm{us}}^{\mathrm{m,c}}$;
    Compute $R_{\mathrm{us}}^{\mathrm{m,c}}$ of $\forall u^{\mathrm{th,m}}$ considering the spectrum amount weather $W^{\mathrm{m},\lambda}$ or $W^{\mathrm{m,h}}$ and the spectrum access strategy (i.e., exclusive, semipooled, and fully pooled access) according to equation (7);
end for
Compute the outage probability of each operator as a function of SIN$R$;
Compute the average rate distributions of each operator $\mathrm{Avg}R^{\mathrm{m}}$, where $m = \{1, 2, 3\ldots M\}$;
Apply standard deviation formula using equation (10) for fairness assessment;
Output: Outage probability, Average rate distributions
```

Algorithm 1: Pseudocode of the HMSSA strategy and AMMC-S scheme implementation.

pooled spectrum access outperforms the semipooled spectrum access in some iterations, which occurs when the users are closer to an mBS $S^{\mathrm{th,m}}$ that belongs to different operator and only offers fully pooled access. For example, user $u^{\mathrm{th,1}}$ that subscribes to OP1, which is located extremely close to mBSs $S^{\mathrm{th,2}}$ and $S^{\mathrm{th,3}}$ owned by OP2 and OP3, will have a choice to associate with either $S^{\mathrm{th,2}}$ and $S^{\mathrm{th,3}}$, which have fully pooled spectrum access. Accordingly, the outage probability of the fully pooled spectrum access becomes lower than of the semipooled spectrum access.

In the proposed HMSSA strategy under Model 1, an additional flexible degree of freedom is utilized to bring advantages from all the available mBSs that operate at different carrier frequencies and spectrum assignments. Therefore, the outage probability is reduced considerably with SINR more than 3 dB of the cell edge user, which outperforms the most related works in [7, 11, 12, 15]. This result can be translated to an enhancement in the performance of the cell edge users. Hence, the coverage and data rate can be improved. Furthermore, the number of mBSs is also decreased, where only 16 mBSs are needed to cover a

Table 2: HMSSA and AMMC-S simulation parameters.

Parameters	Settings
mmWave BSs layout	Grid-based cell deployment
mmWave BSs density	16
# of operator	4
UE layout	Uniform random distribution
UE density	160 users
Simulation area	1.2 km × 1.2 km
Intersite distance (ISD)	300 m
mBS carrier frequency	28 GHz and 73 GHz
mBS transmit power	30 dB
Variant of white Gaussian noise	−174 dBm/Hz
mBS bandwidth	Model 1: 1 GHz for 28 GHz and 73 GHz Model 2: 1 GHz for 28 GHz and 1.5 GHz for 73 GHz

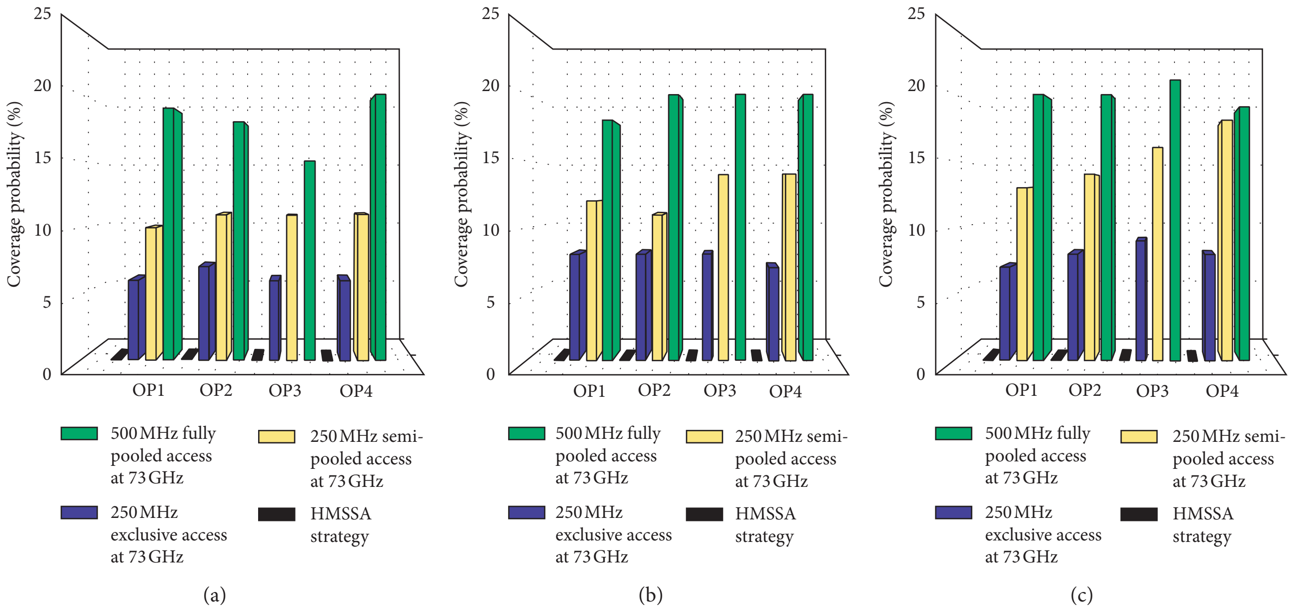
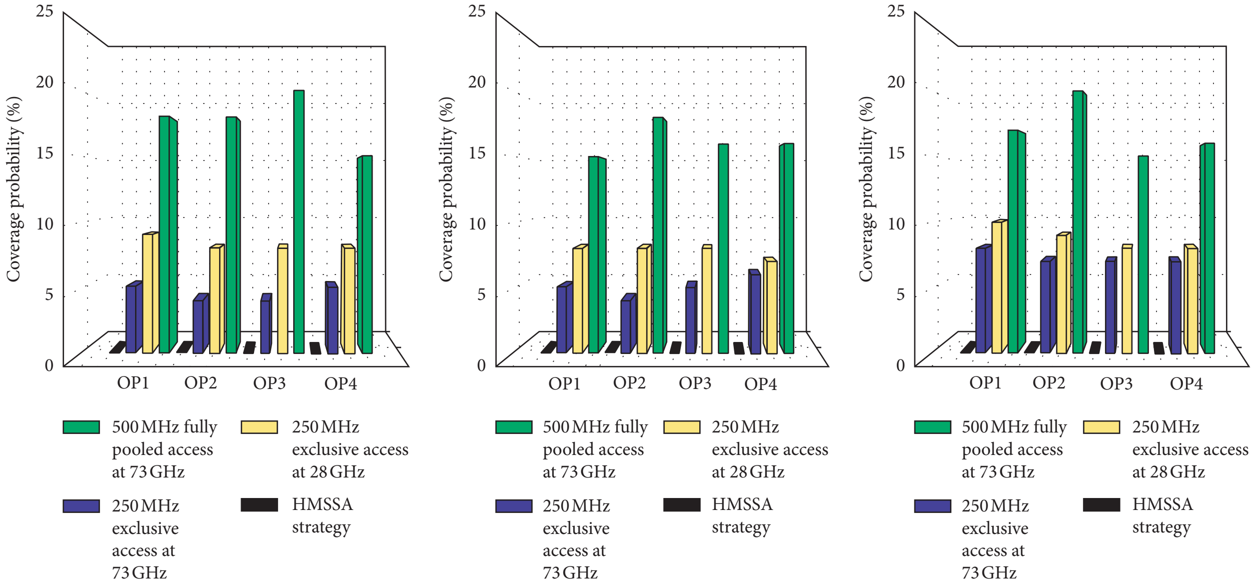


Figure 4: Outage probability percentage for all operators with different percentiles. (a) 5%, (b) 50%, and (c) 95% (Model 1).

1.2 km × 1.2 km area with good coverage which account less than the state of the arts [7, 11, 12, 15]. The outage probability percentages of OP1, OP2, OP3, and OP4 are zero (0%), as shown in Figures 4(a)–4(c). This significant maximization in the (SINR) performance is due to the hybrid spectrum portioning way that enables a flexible hybrid spectrum access strategies and allows for the availability of multiple links with different signal quality within a given transmission range. By adopting the proposed AMMC-S scheme, the UE association with a link that carries the highest SINR can be guaranteed and hence maintain ultrareliable level of connectivity which accounts as one of the most stringent future 5G constraints.

3.1.2. HMSSA Model 2. Model 2 is similar to Model 1. Except for the allocated spectrum amount. Moreover, in Model 2, each user can be associated with any mBS belongs to the same operator or to different operators based on one of the two choices, that is, exclusive access to 250 MHz at 28 GHz and fully shared/pooled access to 500 MHz of the spectrum at 73 GHz carrier frequency or exclusive access to 250 MHz at 73 GHz and fully pooled access to 500 MHz of the spectrum at 73 GHz carrier frequency. Such restrictions in Model 2 help to improve the outage probability of the semipooled spectrum access. The outage probability of all operators that utilize the proposed strategy are kept zero (0%), as shown in Figures 5(a)–5(c), with some improvement in the SINR distributions (>6 dB). The obtained improvement in this model widens the gap with other spectrum access strategies (i.e., exclusive, fully pooled), thereby adding 3 dB to the cell edge users (compared with Model 1). This phenomenon is caused by the fact that the additional amount of spectrum at 73 GHz reduces the interference between the mBSs that operate at such frequency as the number of adjacent mBSs that operate in the same bands is reduced. Fifteen adjacent mBSs are operated by the fully pooled access strategy, whereas only three adjacent mBSs are



(a)

(b)

(c)

Figure 5: Outage probability percentage for all operators with different percentiles. (a) 5%, (b) 50%, and (c) 95% (Model 2).

operated by exclusive access at carrier frequencies of 28 GHz and 73 GHz for each operator, except for the serving mBS (Figure 2).

Another finding related to the utilization of HMSSA strategy is its ability in reducing the number of mBSs to the half and providing a cost-effective solution for enhancing the spectrum utilization and reducing the CO_2 emissions; thus, introducing an environment-friendly wireless communication.

3.2. Average Rate Distributions. In this section, the average rate of all users that belong to the four operators is analyzed based on Monte Carlo simulations. A total of 160 users for each operator are deployed randomly throughout the simulation area. An average of ten users per mBS is assumed in this work. The channel capacity calculation of each UE is performed using Shannon's law illustrated in equation (7). The proposed HMSSA strategy for models 1 and 2 with their spectrum assignments are taken into consideration in this calculation. Figures 6(a) and 6(b) show that the average rate distributions of the proposed strategy for the four operators under models 1 and 2, respectively.

As previously mentioned, the main difference between models 1 and 2 is the allocated spectrum amount at 73 GHz carrier frequency. Such additional amount provides more flexibility to the operators to allocate a part of their spectrum exclusively to enrich the user experience. However, it is shown in Figure 6(b) that the average rate distributions for all operators slightly increases by an average of 7 MHz, 40 MHz, and 13 MHz for the three percentiles of the granted amount of spectrum (5th, 50th, and 95th), respectively. Such observation indicates that granting a large amount of bandwidth to the operator does not necessarily results in an increasing in the average rate. The reason is that the nature characteristic of mmWave signal could significantly impact the system performance if there is no action taken during the UA process. This confirmed the necessity of presenting an efficient UE-mBS association scheme coupling with the adoption of steerable directional antennas at both mBS and UE to strengthen the viability of mmWave wireless communications.

Figure 7 shows the UE rate enhancement of the proposed semipooled and HMSSA strategy (models 1 and 2) compared to the baseline standalone deployment scenario with respect to different percentile rates (5th, 50th, and 95th) and with some system configurations that are illustrated in Table 3.

Three scenarios are applied for the evaluation procedure, the baseline standalone deployment system with 16 mBSs for each operator. In this scenario, a particular UE that belongs to an operator (i.e., OP1) has the right to associate with only the mBS that belongs to its own operator. While in the semipooled scenario, 16 mBSs are divided into two groups; the first group with eight mBSs operate at 28 GHz carrier frequency and the second group with eight mBSs operate at 73 GHz carrier frequency, where the UEs have the right to associate with mBS that operates at 28 GHz carrier frequency or with mBS that operates at 73 GHz carrier frequency that belongs to its own operator or to its own pair operator based on the highest SINR. In case of HMSSA strategy, UEs can associate with mBS that belongs to its own or to different operator through an integrated option utilizing exclusive, semipooled, and fully pooled spectrum access in hybrid manner.

According to the implementation and evaluation of the above scenarios, it is notably that the proposed semipooled and HMSSA strategy (Model 1) enhances the average rate of the users by more than 143% and 193%, respectively;

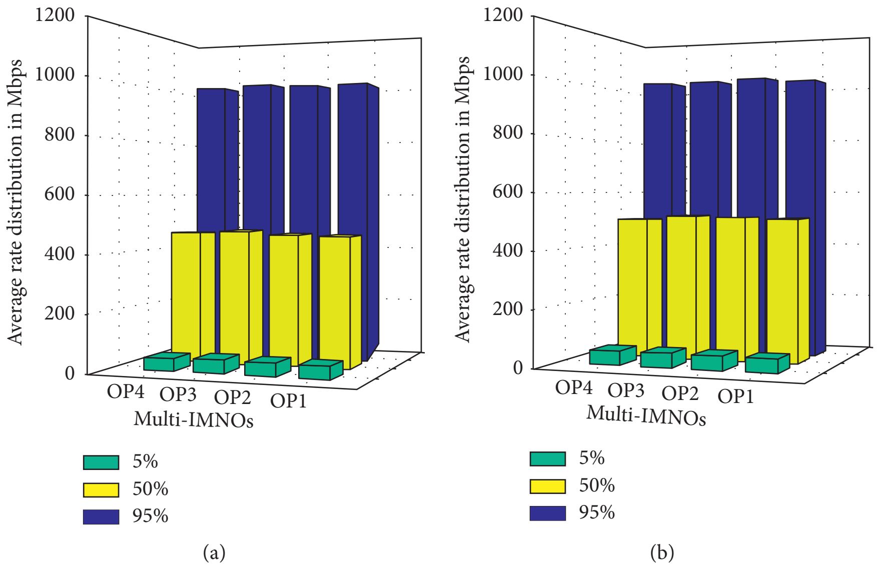


FIGURE 6: Average rate distributions of the four operators utilizing HMSSA strategy: (a) Model 1 and (b) Model 2 with different percentile rates.

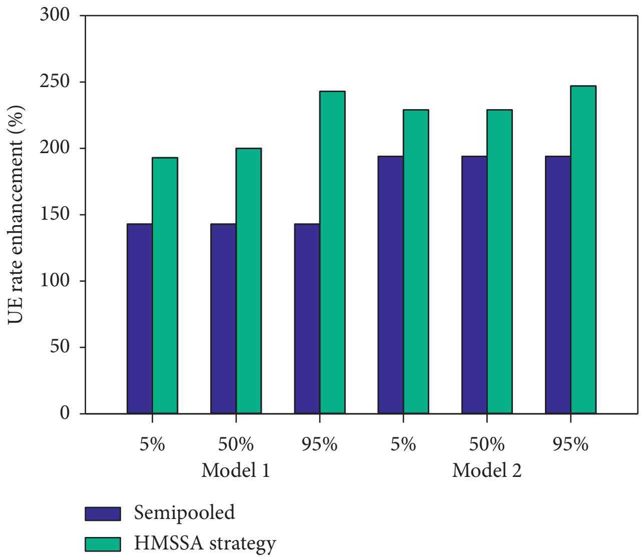


FIGURE 7: Semipooled and HMSSA strategy performance compared to the baseline standalone deployment system (with exclusive access at 28 GHz). X-label indicates the 5th, 50th, and 95th percentiles of the granted amount of spectrum bandwidth under Model 1 and Model 2 configurations.

whereas under (Model 2) configurations, the proposed semipooled and HMSSA strategy enhances the average rate of the users by more than 194% and 229%, respectively. The increase in the UE enhancement rate under Model 2 configurations can be attributed to the extra amount of the allocated bandwidth to the participated operators, specifically in the performance of semipooled (500 MHz at 28 GHz and 750 GHz at 73 GHz for each pair (i.e., OP1 and OP4)), as shown in Table 3.

These observations indicate that the utilization of such hybrid dynamic spectrum access strategy will pave the way for non-standalone cell deployment with non-standalone licensed spectrum access because of its ultraflexibility and capability that offers an optimal UE-mBS association that helps in maximizing the user experience.

Another important observation is that increasing the amount of allocated spectrum bandwidth at 73 GHz carrier frequency to operate as another exclusive right access for UEs under (Model 2) assumptions does not lead to much improvement in the UE rate. This can be attributed to the fact that UEs tend to associate with mBS that operates under exclusive right access at 28 GHz or 73 GHz which has the highest SINR than mBS that operates under semipooled and fully pooled spectrum access strategy. This UE behavior results in much increasing in the cell load and hence ruins the benefits of such extra amount of the allocated bandwidth at the higher carrier frequency (73 GHz).

To sum up, the reported enhancement in the performance of UE rate can be considered as an encouraging step to enable the success of SSA in 5G mmWave cellular networks with less mBSs density and small amount of spectrum bandwidth compared to the most related works in [7, 11, 12, 15].

3.3. Independence and Fairness Assessment. Assessing the operator's independence and fairness based on the signal quality (outage probability) and the average rate distributions of particular subscribers that belong to an operator m^{th} is very important to promote the operators to adopt SSA.

Particularly, in this work, characterizing OP1 as an independent operator implies that its performance is not influenced by other operators (e.g., OP2, OP3, and OP4).

Additionally, the term "fairness" is defined as the ability to handle all operators equally or in a manner that all operators are treated without bias.

Table 3: Baseline system, semipooled, and HMSSA strategy configurations for UE rate evaluation process.

Scenario	Spectrum access strategy	Carrier frequency	Granted amount of bandwidth	mBS deployment configuration
Baseline	Exclusive access	28 GHz	250 MHz	Standalone deployment
Model 1	Semipooled	28 GHz and 73 GHz	500 MHz at both 28 GHz and 73 GHz for each pair (i.e. OP1 and OP4)	Dual deployment
	HMSSA strategy	28 GHz and 73 GHz	1 GHz at 28 GHz and 73 GHz	Hybrid deployment
Model 2	Semipooled	28 GHz and 73 GHz	500 MHz at 28 GHz and 750 GHz at 73 GHz for each pair (i.e. OP1 and OP4)	Dual deployment
	HMSSA strategy	28 GHz and 73 GHz	1 GHz at 28 GHz and 1.5 GHz 73G Hz	Hybrid deployment

Remark 1. The coverage or average rate probability of user $u^{\mathrm{th,m}}$ who associates with operator m^{th} is independent if the coverage or average rate probability of another user does not affect the coverage or average rate probability of user $u^{\mathrm{th,m}}$, which can be expressed as follows:

$$\mathbb{P}\left(\mathbb{P}_{\mathrm{us}}^{m=1\ldots M,\mathrm{c}}\right) = \mathbb{P}_{\mathrm{us}}^{1,\mathrm{c}} \cdot \mathbb{P}_{\mathrm{us}}^{2,\mathrm{c}} \cdot \mathbb{P}_{\mathrm{us}}^{3,\mathrm{c}} \cdot \mathbb{P}_{\mathrm{us}}^{4,\mathrm{c}} \cdots \mathbb{P}_{\mathrm{us}}^{M,\mathrm{c}}, \qquad (8a)$$

where, $\mathbb{P}\,(\mathbb{P}_{\mathrm{us}}^{m=1\cdots\mathrm{M,c}})$ is the coverage or average rate probability of user $u^{\mathrm{th,m}}$ that associates with operator m^{th}.

Considering $M = 4$,

$$\mathbb{P}\left(\mathbb{P}_{\mathrm{us}}^{1,c} \cap \mathbb{P}_{\mathrm{us}}^{2,c} \cap \mathbb{P}_{\mathrm{us}}^{3,c} \cap \mathbb{P}_{\mathrm{us}}^{4,c}\right) = \mathbb{P}_{\mathrm{us}}^{1,\mathrm{c}} \cdot \mathbb{P}_{\mathrm{us}}^{2,\mathrm{c}} \cdot \mathbb{P}_{\mathrm{us}}^{3,\mathrm{c}} \cdot \mathbb{P}_{\mathrm{us}}^{4,\mathrm{c}}. \qquad (8b)$$

More specifically, either coverage or average rate probability of any operator (OP1 and OP2 as an example) is independent if and only if

$$\mathbb{P}\left(\frac{\mathbb{P}_{\mathrm{us}}^{2,\mathrm{c}}}{\mathbb{P}_{\mathrm{us}}^{1,\mathrm{c}}}\right) = \left(\frac{\mathbb{P}\left(\mathbb{P}_{\mathrm{us}}^{1,\mathrm{c}} \cap \mathbb{P}_{\mathrm{us}}^{2,\mathrm{c}}\right)}{\mathbb{P}_{\mathrm{us}}^{1,\mathrm{c}}}\right) = \mathbb{P}_{\mathrm{us}}^{2,\mathrm{c}}. \qquad (8c)$$

This condition can be applied for other operators to assess their independence.

By substituting the coverage probability of each user $u^{\mathrm{th,m}}$ in equation (8b),

$$\mathbb{P}\left(\mathfrak{X}_{\mathrm{us}}^{1,\mathrm{c}}\right) \cap \mathbb{P}\left(\mathfrak{X}_{\mathrm{us}}^{2,\mathrm{c}}\right) \cap \mathbb{P}\left(\mathfrak{X}_{\mathrm{us}}^{3,\mathrm{c}}\right) \cap \mathbb{P}\left(\mathfrak{X}_{\mathrm{us}}^{4,\mathrm{c}}\right) = \left(\mathfrak{X}_{\mathrm{us}}^{1,\mathrm{c}}\right) \cdot \mathbb{P}\left(\mathfrak{X}_{\mathrm{us}}^{2,\mathrm{c}}\right) \cdot \mathbb{P}\left(\mathfrak{X}_{\mathrm{us}}^{3,\mathrm{c}}\right) \cdot \mathbb{P}\left(\mathfrak{X}_{\mathrm{us}}^{4,\mathrm{c}}\right). \qquad (8d)$$

Recall equation (5). The coverage probability of each user $u^{\mathrm{th,m}}$ as a function of SINR entirely depends on the received signal power and the amount of interference from other adjacent mBSs that operate at the same band. User orientation in terms of the deployed mBS and the spectrum access strategy are key components in determining the received signal power and the amount of interference, respectively.

As $u^{\mathrm{th,m}}$ can only associate with the tagged mBS that offers a high SINR regardless of which operator it belongs to, in order to maximize its channel capacity. Moreover, four operators are considered in this study ($M = 4$); each operator has four mBSs. Each mBS has three different spectrum assignments (exclusive at 28 GHz, semipooled at 73 GHz, and fully pooled at 73 GHz) in Model 1 and (exclusive at 28 GHz, exclusive at 73 GHz, and fully pooled at 73 GHz) in Model 2. Therefore, equation (7) can be rewritten as

$$R_{\mathrm{us}}^{\mathrm{m},c} = \Phi_{\mathrm{us}}^{\mathrm{m},c} \times \left(\frac{W^{\mathrm{m},c}}{u_{\mathrm{s\,th}}}\right) \times \log_2\left(1 + \max\left(\max\left(\mathfrak{X}_{\mathrm{us}}^{1,c=1,2,3}\right) \cup \cdot \max\left(\mathfrak{X}_{\mathrm{us}}^{2,c=1,2,3}\right) \cup \max\left(\mathfrak{X}_{\mathrm{us}}^{3,c=1,2,3}\right) \cup \max\left(\mathfrak{X}_{\mathrm{us}}^{4,c=1,2,3}\right)\right)\right), \qquad (9)$$

where $c\ = 1, 2, 3$ denotes the spectrum assignments of each mBS. On the basis of equation (9), the average rate of each user depends on the SINR value regardless of which operator it belongs. Therefore, the utilization of the proposed strategy achieves a high degree of independence in terms of both performance metrics (i.e., coverage or average rate probability).

In terms of fairness, standard deviation formula is utilized to assess the differences among the operators that share the spectrum in terms of average rate distributions.

Remark 2. The average rate percentages of all operators are relatively close to one another. The small margin in the average rate probability among all operators indicates that the resources are evenly allocated to the users regardless of which operator the users belong to.

The standard deviation of the average rate of a set of operators is expressed as follows:

$$\mathrm{SD}^{\mathrm{ALL}} = \sqrt{\frac{\sum\left|\mathrm{Avg}R^{\mathrm{m}} - \mu\right|^2}{m}}, \qquad (10)$$

where $\mathrm{SD}^{\mathrm{ALL}}$ denotes the standard deviation of the average rate of M operators. Without loss of generality, $\mathrm{Avg}R^{\mathrm{m}}$ denotes the average rate of operator m^{th}. μ is the mean of the average rate values of M operators, which is represented by summing up all the average rates of a set of operators divided by the number of operators M.

As shown in Table 4, the proposed HMSSA strategy is successful in terms of equity in resource allocation, in which the maximum margin of the average rate does not exceed 6.4727 Mbps. The HMSSA strategy margin in terms of the exclusive or semipooled access and fully pooled access, along with their percentages can be ignored in comparison with the high data rate experienced by the users belonging to the operators. Thus, operators are encouraged to rely on such strategy, which has proven its proportional fairness in terms of resource allocation. The small margin results from the user positioning in terms of the deployed mBS and not from the rules of the proposed HMSSA strategy, in which UEs that

Table 4: Margin percentage and standard deviation of the proposed HMSSA strategy (Model 1 and Model 2).

HMSSA configurations	Percentiles (%)	HMSSA margin in terms of		SD^{ALL} (Mbps)	Average rate (Mbps)			
		Fully pooled access (%)	Exclusive or semipooled access (%)		OP1	OP2	OP3	OP4
Model 1	95	0.845	1.691	4.0169	1005.5	1004.4	1013.1	1005.5
	50	2.548	5.097	6.3713	457	462	472.3	463.5
	5	7.662	15.2	1.9155	44.3	45.8	47.7	43.3
Model 2	95	1.362	2.366	6.4727	1010.6	1026.3	1019.7	1017.5
	50	2.082	4.165	5.2066	506.2	509.9	513.9	501.7
	5	1.876	3.752	0.4690	50.1	51.2	50.6	50.9

belong to the different operators are deployed randomly and independently. Accordingly, the competition among multiple operators in terms of service delivery will be conducted in a proportionally fair manner with the existence of the hybrid SSA.

4. Conclusions

In this study, we investigate the implementation of a flexible HMSSA strategy by analyzing various practical aspects, such as spectrum access strategies, various rate percentiles, and two mmWave frequency bands with different characteristics and spectrum bandwidth. An optimization framework was developed to enable operators to harvest the gains from several considerations, such as hybrid spectrum integration, resource sharing strategy, as well as user-mBS association. Moreover, a detailed analytical and discussion is presented to assess independence and fairness among operators under the proposed HMSSA strategy assumptions. The numerical results show that the integration of a hybrid spectrum (i.e., exclusive, semipooled, and fully pooled) strategy can provide a considerable solution to overcome mutual interference issues, thereby reducing outage probability to zero with ($SINR>3\,\text{dB}$) and the number of mBSs to the half providing capital expenditure (CapEx) and operating expenditure (OpEx) savings. Furthermore, compared with exclusive access, the utilization of the proposed strategy is generally beneficial for guaranteeing an acceptable level of operator's independence and fair spectrum usage and maximizing the UE rate more than two folds. Moreover, utilizing such strategy aids in enabling a rapid creation of new wireless applications in a cost-effective manner. In future studies, we will expand these investigations to more complex scenarios, considering the adoption of spectrum access system and licensed shared access spectrum sharing models. UE-mBS association advancement will be part of the future work to improve mBS selection for enabling the SSA to meet the boldest 5G constraints.

Data Availability

The data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this article.

Acknowledgments

The authors gratefully acknowledge the UTeM Zamalah Scheme, Universiti Teknikal Malaysia Melaka (UTeM), and the supports from the Centre for Research and Innovation Management (CRIM), Centre of Excellence, Universiti Teknikal Malaysia Melaka (UTeM).

References

[1] P. Demestichas, A. Georgakopoulos, K. Tsagkaris, and S. Kotrotsos, "Intelligent 5G networks: managing 5G wireless/mobile broadband," *IEEE Vehicular Technology Magazine*, vol. 10, no. 3, pp. 41–50, 2015.

[2] M. Matalatala, M. Deruyck, E. Tanghe, L. Martens, and W. Joseph, "Performance evaluation of 5G millimeter-wave cellular access networks using a capacity-based network deployment tool," *Mobile Information Systems*, vol. 2017, Article ID 3406074, 21 pages, 2017.

[3] F. Wei and W.-x. Zou, "Suboptimal network coding subgraph algorithms for 5G minimum-cost multicast networks," *Frontiers of Information Technology & Electronic Engineering*, vol. 19, no. 5, pp. 662–673, 2018.

[4] T. S. Rappaport, S. Shu Sun, R. Mayzus et al., "Millimeter wave mobile communications for 5G cellular: it will work!," *IEEE Access*, vol. 1, pp. 335–349, 2013.

[5] S. Rangan, T. S. Rappaport, E. Erkip, F. Gomez-Cuba, T. S. Rappaport, and E. Erkip, "Millimeter-Wave cellular wireless networks: potentials and challenges," *Proceedings of the IEEE*, vol. 102, no. 3, pp. 366–385, 2015.

[6] F. Boccardi, H. Shokri-Ghadikolaei, G. Fodor et al., "Spectrum pooling in MmWave networks: opportunities, challenges, and enablers," *IEEE Communications Magazine*, vol. 54, no. 11, pp. 33–39, 2016.

[7] M. Rebato, F. Boccardi, M. Mezzavilla, S. Rangan, and M. Zorzi, "Hybrid spectrum sharing in mmWave cellular networks," *IEEE Transactions on Cognitive Communications and Networking*, vol. 3, no. 2, pp. 155–168, 2017.

[8] M. L. Attiah, A. A. M. Isa, Z. Zakaria, M. Ismail, R. Nordin, and N. F. Abdullah, "Coverage probability optimisation by utilizing flexible hybrid mmWave spectrum slicing – sharing access strategy for 5G cellular systems," *Journal of Telecommunication, Electronic and Computer Engineering (JTEC)*, vol. 10, no. 2, pp. 91–98, 2018.

[9] S. Pandit and G. Singh, *Spectrum Sharing in Cognitive Radio Networks*, Springer-International Publisher, Gewerbestrasse, Cham, Switzerland, 1st edition, 2017.

[10] G. Li, T. Irnich, and C. Shi, "Coordination context-based spectrum sharing for 5G millimeter-wave networks," in *Proceedings of 2014 9th International Conference on Cognitive*

Radio Oriented Wireless Networks and Communications (CROWNCOM), pp. 32–38, Oulu, Finland, June 2014.

[11] M. Rebato, M. Mezzavilla, S. Rangan, and M. Zorzi, "Resource sharing in 5G mmWave cellular networks," in *Proceedings of Millimeter-wave Networking Workshop (mmNet 2016)*, pp. 271–276, San Francisco, CA, USA, April 2016.

[12] A. K. Gupta, J. G. Andrews, and R. W. Heath, "On the feasibility of sharing spectrum licenses in mmWave cellular systems," *IEEE Transactions on Communications*, vol. 64, no. 9, pp. 3981–3995, 2016.

[13] H. Shokri-Ghadikolaei, F. Boccardi, C. Fischione, G. Fodor, and M. Zorzi, "Spectrum sharing in mmWave cellular networks via cell association, coordination, and beamforming," *IEEE Journal On Selected Areas in Communications*, vol. 34, no. 11, pp. 2902–2917, 2016.

[14] J. Park, J. G. Andrews, and R. W. Heath, "Inter-operator base station coordination in spectrum-shared millimeter wave cellular networks," *IEEE Transactions on Cognitive Communications and Networking*, vol. 4, no. 3, pp. 513–528, 2017.

[15] R. Jurdi, A. K. Gupta, J. G. Andrews, and R. W. Heath, "Modeling infrastructure sharing in mmWave networks with shared spectrum licenses," *IEEE Transactions on Cognitive Communications and Networking*, vol. 4, no. 2, pp. 328–343, 2018.

[16] H. Teng, K. Zhang, M. Dong, M. Zhao, and T. Wang, "Adaptive transmission range based topology control scheme for fast and reliable data collection," *Wireless Communications and Mobile Computing*, vol. 2018, article 4172049, 21 pages, 2018.

[17] T. S. Rappaport, F. Gutierrez, E. Ben-Dor, J. N. Murdock, Y. Qiao, and J. I. Tamir, "Broadband millimeter-wave propagation measurements and models using adaptive-beam antennas for outdoor Urban cellular communications," *IEEE Transactions on Antennas and Propagation*, vol. 61, no. 4, pp. 1850–1859, 2013.

[18] G. R. Maccartney and T. S. Rappaport, "73 GHz millimeter wave propagation measurements for outdoor urban mobile and backhaul communications in New York City," in *Proceedings of 2014 IEEE International Conference on Communications, ICC 2014*, pp. 4862–4867, Sydney, Australia, June 2014.

[19] T. S. Rappaport, J. N. Murdock, and F. Gutierrez, "State of the art in 60-GHz integrated circuits and systems for wireless communications," *Proceedings of the IEEE*, vol. 99, no. 8, pp. 1390–1436, 2011.

[20] N. Bhushan, J. Li, D. Malladi et al., "Network densification: the dominant theme for wireless evolution into 5G," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 82–89, 2014.