Received 1 October 2023; accepted 20 October 2023. Date of publication 1 November 2023; date of current version 22 November 2023.

*Digital Object Identifier 10.1109/OJCOMS.2023.3329420*

# Convergence of AI and MEC for Autonomous IoT Service Provisioning and Assurance in B5G

**KHIZAR ABBAS[1], YEONGPIL CHO[1], ALI NAUMAN[2], PRINCE WAQAS KHAN[3], TALHA AHMED KHAN[4], AND KOTESWARARAO KONDEPU[5] (Senior Member, IEEE)**

[1]Department of Computer Science, Hanyang University, Seoul 04763, South Korea
[2]Department of Information and Communication Engineering, Yeungnam University, Gyeongsan 38541, South Korea
[3]Department of Industrial and Management Systems Engineering, West Virginia University, Morgantown, WV 26506, USA
[4]Institute for Communication Systems, University of Surrey, GU2 7XH Guildford, U.K.
[5]Department of Computer Science and Engineering, Indian Institute of Technology Dharwad, Dharwad 580011, India

CORRESPONDING AUTHOR: Y. CHO (e-mail: ypcho@hanyang.ac.kr)

This work was supported by the National Research Foundation of Korea (NRF) Grant funded by the Korea Government (MSIT) under Grant NRF-2022R1A4A1032361.

**ABSTRACT** With the exponential growth of Internet of Things (IoT) devices, IoT has become a transformative technology with applications spanning various domains. It encompasses a wide range of public and industrial vertical services that come with diverse and stringent Quality of Service (QoS) requirements. Traditional networks often struggle to meet the demands of these diverse IoT services. As a result, the introduction of 5G and Beyond 5G (B5G) networks holds promise in accommodating these diverse IoT services through network slicing technology. Network slicing involves partitioning a single physical network infrastructure into multiple logically isolated networks and ensures dedicated resources to each service as per QoS requirements. Additionally, Multi-Access Edge Computing (MEC) in B5G networks presents an innovative solution to facilitate low-latency communication for IoT services. However, the automatic provisioning and management of end-to-end (e2e) network slicing for IoT services across multi-domain infrastructures pose significant challenges, including manual error-prone resource configuration, network slice template preparation, and human intervention. This paper proposes an automated Artificial Intelligence (AI) and MEC-enabled solution for provisioning and managing network slice resources across multiple domains specifically tailored for IoT services. Our solution provides an abstraction layer that generates slice templates for each domain and automates the deployment of resources based on the specified QoS requirements. It automates the slice resource configuration process, reduces human intervention, and manages the complete lifecycle of IoT slices. We have conducted several tests with our system, creating multiple IoT slices, and have observed stable performance in slice design, resource provisioning, slice isolation, and management.

**INDEX TERMS** IoT, beyond 5G networks, MEC, SDN, network slicing, AI for 5G, service automation and management.

## I. INTRODUCTION

INTERNET of Things (IoT) has emerged as a versatile technology that holds immense potential to address a wide range of societal challenges [1]. McKinsey report predicts a substantial economic impact, ranging from 3.9 trillion to 11.1 trillion dollars annually by 2025, generated by the IoT industry [2]. With the continuous proliferation of IoT devices, such as sensors, home appliances, actuators, cameras, drones, and medical IoT devices, these interconnected devices can engage in networked interactions. This connectivity has found applications in various domains, including smart grid, smart healthcare, home automation, medical assistance, emergency services, and traffic management [3]. However, as the IoT industry expands, there is an increasing demand

for low-latency connectivity and high bandwidth to support real-time IoT applications effectively. Traditional centralized cloud-based IoT solutions face challenges in meeting diverse and stringent Quality of Service (QoS) requirements. So, 5G and Beyond 5G (B5G) networks are designed on the service-oriented pattern and aim to accommodate these diverse novel services, such as IoT real-time services with better quality of experience QoE [4]. These B5G networks enhance network throughput and support a wide range of highly innovative services with different QoS requirements, including virtual reality, augmented reality, metaverse, critical IoT, smart healthcare, smart city, and holographic applications. These services pave the way for autonomous vehicles, the Industrial Internet of Things (IIoT), and Industry 5.0 applications, revolutionizing our daily lives [5]. To fulfill the demands of these services, the B5G network must leverage network-slicing technology to provide instant and seamless connectivity, extensive broadband coverage, efficiency, reliability, and impeccable network availability.

Network slicing in B5G networks emerges as a viable solution to accommodate the diverse and differentiated QoS requirements of IoT services, enabling enhanced connectivity and improved performance. Network Slicing supports these diverse QoS IoT services by providing dedicated network resources to each service according to the Service Level Agreement (SLA). Network slicing is partitioning the physical network into multiple logical and isolated networks. Network slicing is made possible through the use of Software-Defined Networking (SDN) and Network Function Virtualization (NFV), which drive a transformative and innovative shift in future networking approaches. SDN and NFV allow the creation of scalable and flexible multiple network slices over a physical network [6], [7], [8]. Looking ahead, the future B5G services (i.e., robotics communication services, multi-sensory services, neuroscience intelligence (brain interaction with computer) services, autonomous industrial operations, etc.) will impose even more demanding QoS prerequisites, such as high throughput of 1 TB/s, latency less than a millisecond, and seven nines reliability [9], [10]. As a result, it is essential for B5G mobile networks to provide more stringent QoS assurances to fulfill the demands of disruptive services and applications on the horizon. It also needs to integrate edge and cloud resources to cater to high-performance requirements across different industrial verticals.

Edge computing is crucial in ensuring low latency for IoT time-critical services, particularly URLLC services in the B5G context. So, the intelligent integration of edge computing technologies is very beneficial to ensure better QoE for low latency, better throughput, and reliability [11]. However, the European Telecommunications Standards Institute (ETSI) has introduced an innovative technology named Multi-Access Edge Computing (MEC) for low-latency IoT services [12]. MEC is expected to support various 5G use cases by providing value-added services to end users. Apart from serving as an execution environment for edge-based applications, MEC offers multiple services related to end-users and RAN contexts, such as context-aware applications, location information, user information, and geographical edge applications [13], [14].

The automatic orchestration and management of diverse IoT services in a multi-domain environment such as MEC, RAN, core, and transport network domains is crucial and challenging for B5G service providers. There are many solutions developed based on NFV management and orchestration (NFV-MANO) to support the automatic orchestration and management of 5G and beyond services, including IoT services. However, all mechanisms require slice templates and configurations in a specific format and use manual configuration procedures for the deployment of resources. It is also complex because each domain requires different policy configurations to activate the resources such as core orchestrator, MEC orchestrator, RAN, and transport controller. However, automatic provisioning and managing the lifecycle of complete e2e IoT slicing with the MEC environment is still very challenging.

Nowadays, many industrial solutions have incorporated Artificial Intelligence (AI) technologies to proactively manage and ensure B5G services. In the 5G, the Third Generation Partnership Project (3GPP) has introduced an AI-enabled Network Data Analytics Function (NWDAF) to provide network intelligence [15]. NWDAF leverages AI methods to offer proactive resource management and assurance. Automating network slicing and leveraging AI technologies play a vital role in efficiently provisioning and managing services in a multi-domain network infrastructure. By employing proactive resource management and utilizing intelligent analytics, networks can effectively meet the requirements of IoT-diverse and evolving service demands.

### A. RESEARCH MOTIVATION

As aforementioned, the core objective of the 5G and beyond network is to support diverse IoT services with specific QoS needs. However, manually configuring these services is prone to errors, time-consuming, and requires significant expertise. Additionally, manually allocating resources across multiple domains like core, RAN, transport, and edge to establish e2e network slicing is not an optimal solution. So, the research community and standardization bodies are striving to automate the orchestration and management of network resources. While several solutions have emerged for automating resource deployment, they often need to catch up when managing resources across multiple domains. Moreover, efficient and automated service design is also crucial, as each domain may require different policies and configurations to activate network resources. For instance, IoT encompasses a wide range of services with varying and strict QoS demands, making allocating dedicated and shared resources to each IoT service complex. Additionally, many IoT services fall under the URLLC category, which necessitates ultra-low latency communication that can be challenging for B5G network operators to provide. Integration with Multi-Access

Edge Computing (MEC) becomes essential to address these latency issues. Due to that, designing and orchestrating IoT services over a multidomain network infrastructure supported by MEC is inherently complex. On the other side, AI-driven approaches show promise in managing, automating, and controlling various network operations. To navigate these complexities, there is a pressing need for an automated orchestration solution that leverages MEC and AI to provision and manage services across multidomain network infrastructures efficiently.

### B. RESEARCH CONTRIBUTIONS

To overcome the problem related to accommodating IoT diverse services and providing dedicated resources in an automated fashion, we have proposed a MEC-assisted efficient service provision mechanism. Following are the summarized objectives of our work:

- An advanced framework is presented that leverages SDN, NFV, MEC, NFV MANO, and AI technologies for intelligent provisioning and management of diverse IoT and B5G services automatically. It enables the provisioning of multi-domain resources, including RAN, MEC, core, and transport domains, and facilitates the activation of e2e network slices based on SLAs.
- The Proposed mechanism follows a closed-loop and one-touch mechanism to manage the lifecycle of e2e IoT services.
- Allows the orchestration of critical low-latency IoT services by integrating the MEC capabilities.
- provides openness to implementing AI and data processing and analytics mechanisms for efficient automation and proactive management of network resources.
- Several tests have been carried out to validate system performance by designing, instantiating, and activating multiple e2e IoT slices, and it performs well in the context of resource allocation, stability, flexibility, data rate, and response delay.
- We have also discussed background technologies for better understanding, such as 5G networks, MEC for low latency communication, benefits of e2e network slicing for IoT services, 3GPP orchestration and management capabilities, and AI applications for providing intelligence and proactive control over B5G networks.

### C. PAPER ORGANIZATION

The remainder of the article is structured as follows. Section II presents the background technologies and related literature, such as the 5G network, MEC and its components, e2e network slicing for IoT use cases, management and orchestration technologies for B5G networks, and AI-assisted automation and proactive management. Section III explains the proposed MEC-enabled IoT service provisioning and management system with its components for B5G services. The implemented testbed and experimental results and analysis are discussed in Section IV. Section V presents the conclusive remarks with future directions of the proposed work.

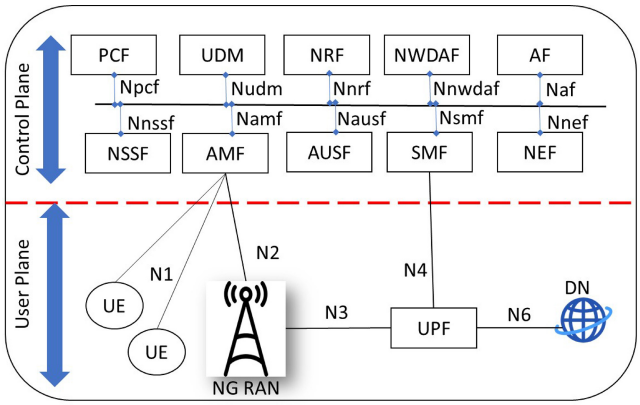



FIGURE 1. 5G SBA architecture.

## II. BACKGROUND TECHNOLOGIES

This section explains the background technologies such as the 5G network, MEC, and its components, benefits of network slicing for IoT applications, and automation and management of network slicing for diverse IoT services.

### A. 5G ARCHITECTURE

3GPP has designed a 5G system as a Service Based Architecture (SBA) for accommodating novel and diverse requirements services. Figure 1 illustrates the SBA architecture outlining the control and user/data plane communication. The lower data plane elements handle user data transportation, while the control plane comprises various core network Virtual Network Functions (VNFs) for handling various control plane tasks. These core VNFs include an Access and Mobility Management Function (AMF), Session Management Function (SMF), User Plane Function (UPF), and many more. Detailed functionalities and responsibilities of these VNFs are presented in 3GPP TS 23.501 [16], [17]. Decoupling the user plane from the control plane in the 5G system allows scalable, cost-effective, and flexible deployment of VNFs. By adopting the SBA, the 5G network optimizes network performance, enables dynamic resource allocation, and supports emerging applications and services with diverse requirements. It provides a foundation for leveraging SDN and NFV technologies, ensuring a flexible and scalable infrastructure for the evolving needs of B5G networks.

### B. MULTI-ACCESS EDGE COMPUTING

ETSI proposed the idea of utilizing distributed computing closer to User Equipment (UE) to attain faster response, low latency, and alleviate network congestion, named MEC. MEC in 5G and beyond networks plays a crucial role in bringing cloud computing capabilities to the edge [12]. By processing data near the end devices or users, MEC enables the network to achieve URLLC communication for mission-critical IoT

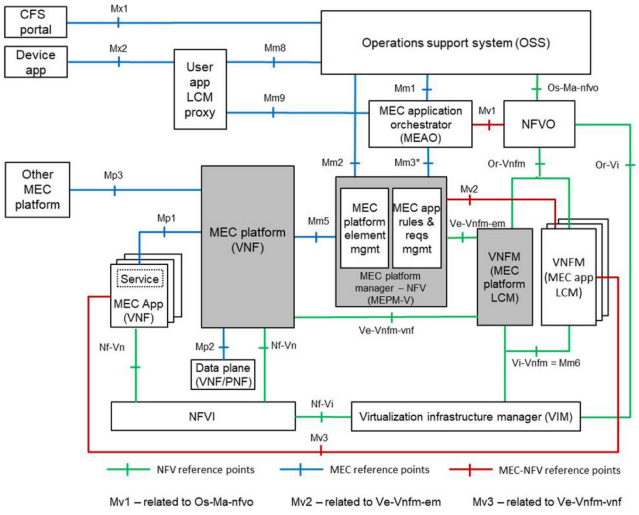



**FIGURE 2.** NFV-enabled MEC architecture for B5G services.

services. Additionally, MEC offers the advantage of reducing data transfer costs by processing data locally. Moreover, MEC brings several other benefits, such as the optimization of mobile resources by offloading computational-intensive services to the edge network, improved QoS assurance and QoE for end-users, and context-aware services by transforming RAN (gNodeBs) nodes into intelligent service hubs.

The ETSI MEC reference architecture comprises a MEC management entity and MEC hosts. The MEC host is a fundamental component of MEC architecture that consists of the MEC Platform (MEP), Virtualization Infrastructure (VI), and MEC applications. VI includes storage and computing resources for MEC applications. The MEC host enables the execution of MEC apps and provides necessary resources and services. Conversely, the MEP manages traffic routing among applications, services, and RAN networks within the MEC host. It configures the DNS proxy, creates an environment for MEC services, and performs essential tasks for MEC app operation. Virtualization techniques ensure isolation and efficient utilization of resources with specific network services. Furthermore, the MEC management entity comprises MEC host-level and MEC system-level management. In MEC system-level management, the MEC Orchestrator (MEO) is a central decision-making entity that has a broad view of services, MEC hosts, topology, and resources. The MEO selects the most appropriate MEC host based on QoS requirements and performs system-level tasks through the OSS and user application LCM proxy. The OSS handles user requests, enables or terminates applications, and relocates MEC applications. The LCM proxy entity assists in managing the complete lifecycle of user applications, working in conjunction with the MEO and OSS [12], [13].

The MEC host-level module contains Virtualization Infrastructure Manager (VIM) and MEC Platform Manager (MEPM). The MEPM provides critical information about the MEP to the MEO, performs MEC application lifecycle management, handles rules, MEC services authorization, and configures DNS. In addition, the VIM provides virtualization support for hosting VMs and facilitating rapid application deployment. It controls the allocation and management of virtualized resources within the virtual infrastructure.

Figure 3 presents an updated NFV-enabled MEC architecture from the last high-level architecture, which provides a common NFV environment to all MEC entities. It included NFV-enabled components involving running the MEP and MEPM as VNFs. The MEO is changed to the Mobile Edge Application Orchestrator (MEAO) while maintaining its functionality but utilizing the NFVO to commission MEC apps, MEP, and MEPM. This ensures that all resource management and instantiation processes follow well-designed NFV interfaces. As a result, edge resources can be managed as standard computational and storage resources and managed by the single VIM [13].

### C. NETWORK SLICING FOR IOT SERVICES

Network slicing is a technique of partitioning a single physical network into multiple virtual and isolated networks. Network slicing can be commonly divided into massive Machine-Type Communications (mMTC) service for IIoT automation and smart city use cases, enhanced Mobile Broadband (eMBB) service for tasks like ultra-HD video streaming, and Ultra-Reliable Low-Latency Communications (URLLC) for applications such as critical IoT, autonomous vehicle, remote surgery, and medical emergency [18]. In the network slicing context, IoT applications can be classified into four main areas based on their specific requirements: massive IoT services, short-range IoT services, critical IoT services, and low-range IoT services, as presented in Figure 4. These IoT services have many applications in the real world, such as industrial automation, critical IoT or smart healthcare, smart homes, transportation, energy systems, military applications, massive IoT, UAV applications, gaming, and many more [5], [6]. Moreover, network slicing offers dedicated resources, ensuring minimal interference from other applications and enhancing resource utilization, especially during high-density traffic periods. It also provides isolation for healthcare, factory, grid, and military applications, ensuring security and tailored communication needs. In the context of 5G, it enables UAV operations and data collection. Network slicing promises secure, high-throughput, and low-latency communication for diverse IoT services.

### D. NETWORK SLICING MANAGEMENT AND ORCHESTRATION (MANO)

To provide differentiated services to 5G consumers, the MANO of the e2e network is an essential activity for MNOs. Several mobile network standardization bodies have defined the specifications for e2e network slice MANO, such as ETSI, Internet Engineering Task Force (IETF), 3GPP, Fifth Generation Partnership Project (5GPP), ITU, Next

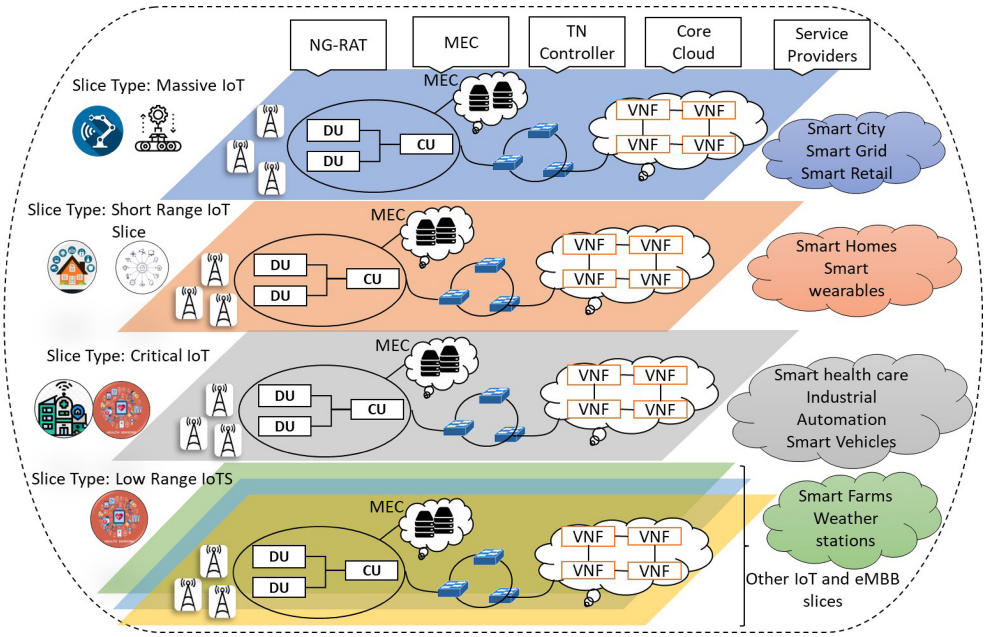



**FIGURE 3.** Overview of e2e network slicing including RAN, MEC, Core, transport capabilities for IoT use cases.

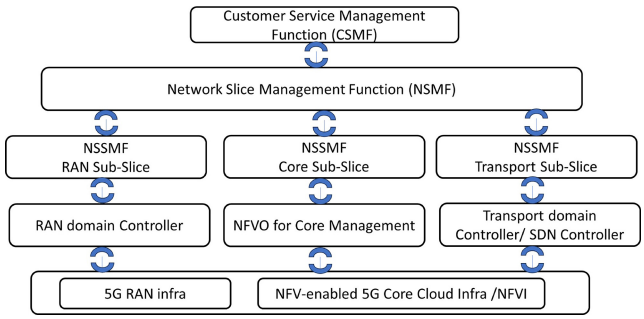



**FIGURE 4.** 3GPP high-level e2e network management and orchestration architecture for 5G services.

Generation Mobile Network (NGMN), etc. [16], [19]. ETSI has introduced the NFV-MANO platform, which automates the deployment of VNFs in an efficient way. MANO comprises an NFV Orchestration (NFVO) layer, VNF Managers (VNFMs), and Virtual Infrastructure Managers (VIMs). The NFVO is the orchestration entity that can manage the lifecycle of the network services with the cooperation of VNFM and VIM. It is responsible for the deployment of appropriate resources and establishes the connection. The MNOs input network service configurations through the Operation Support System /Business Support System (OSS/BSS), and NFVO deploys and activates the resource over the physical infrastructure with the help of VNFM and VIM. NFVO has multiple VNMFS and VIMs for the automation and management of resources.

The 3GPP has introduced an architecture for managing and orchestrating the service-oriented 5G networks. This system consists of Communication Service Management Function (CSMF), Network Slice Management Function (NSMF), and Network Slice Subnet Management Function (NSSMF), as presented in Figure 5. The CSMF acts as a central management entity for managing and deploying network slices. It is responsible for creating the network slice requests and sending them to NSMF for further operations. The MNOs used CSMF functions to plan, design, and activate the network slices. Conversely, NSMF translates the slice requirements and generates domain-specific configurations. Further, those configurations are forwarded to NSSMF for the deployment of the network slice instances. Each domain has separate NSMF, e.g., RAN, core, edge. Also, NSSMF ultimately manages each slice instance. Moreover, the CSMF receives network slice requests with QoS requirements from the customers and forwards those requests to NSMF. NSMF converts slice QoS into policies and activates the resources with the cooperation of NSSMF. The slice instances are appropriately monitored, and CSMF performs autoscaling of the network resources whenever needed [19], [20].

## III. RELATED WORK

In this section, we delve into the existing research concerning IoT services across diverse domains and examine how network slicing can effectively tackle the challenges posed by these IoT services. We also explore the relevant body of work on the automation of IoT network slicing and the utilization of AI approaches for network management.

### A. NETWORK SLICING FOR IOT USE CASES

#### 1) SLICING FOR SMART HEALTHCARE OR CRITICAL APPLICATIONS

The traditional telecommunication network is inadequate to meet the healthcare industry's diverse communication requirements of IoT use cases. The 5G offers higher data

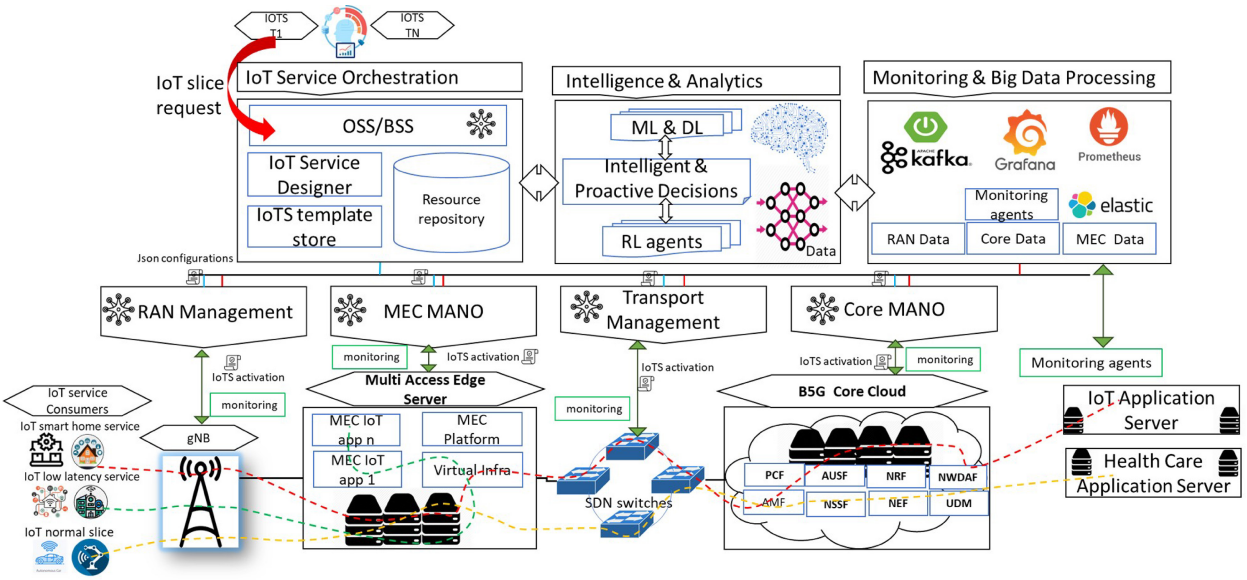



**FIGURE 5.** Detailed architecture of MEC-enabled service provisioning and management for diverse IoT applications.

rates, low latency, and faster response times, enabling remote monitoring and examination of patients. The reliability of communication services is crucial in healthcare applications, especially in critical procedures like remote surgeries. Network slicing provides an effective solution by ensuring dedicated resources and minimizing disturbances caused by resource variations from other applications. Privacy and security are top priorities in smart healthcare services, and network slicing enables the activation of specialized NFs to enhance privacy and security within healthcare slices. Network slice isolation ensures the confidentiality of healthcare traffic and prevents its visibility to other application slices [21].

### 2) SLICING FOR SMART TRANSPORTATION

In the realm of smart transportation, IoT plays a crucial role in various aspects, such as vehicle-to-infrastructure communication (V2I), vehicle-to-vehicle communication (V2V), vehicle-to-everything communication (V2X), vehicle infotainment, and autonomous driving. These services have unique connectivity and QoS requirements that a common network cannot adequately meet. For example, autonomous driving requires extremely low-latency communication. To cater to the diverse needs of V2X applications in a cost-effective manner, network slicing proves to be the ideal solution. Besides, it is very challenging for Mobile Network Operators (MNOs) to provide services and entertain multi-tenant service providers such as vehicle manufacturers or companies, road authorities, municipalities, etc., in the V2X scenario [22]. So, it becomes more complex to support multi-tenants through the infrastructure provided by the multiple MNOs. Adopting well-designed slice templates for each tenant can efficiently fulfill the QoS requirements of these multi-tenants. So, network slicing ensures the provisioning of dedicated resources for each service, such as V2X use cases, and it also ensures complete isolation from other use-case slices. So, It also overcomes and entertains the increasing high-density traffic demand issues by dynamically deploying VNFs on peak time (peak hours) or location type, enhancing network resource utilization efficiency. Moreover, dedicated security network slicing can be implemented to ensure robust security measures for different V2X applications.

### 3) SLICING FOR SMART INDUSTRIAL AUTOMATION

Traditional networks cannot meet the wide range of communication requirements of IIoT applications, such as low latency, high data rates, and reliability. Network slicing offers a cost-effective solution by utilizing a single network. By adjusting the network functions and slice configurations, network slicing can cater to various requirements, including security, mobility, latency, and bandwidth. Dedicated slices can be allocated for large-scale factory environments to address specific communication needs, such as automated machine operations, factory management, and automated maintenance and diagnostics [23]. These dedicated slices enable effective connectivity among sensors, actuators, workers, and machines while ensuring secure and optimized operations.

### 4) SLICING FOR SMART CITY APPLICATIONS

IoT devices in homes exhibit a wide range of heterogeneity, from basic ones with limited energy to more powerful ones requiring continuous power supply. Additionally, security levels vary from insecure devices to highly secure ones. Security devices, such as door locks and surveillance cameras, directly impact home security, making it crucial to prevent unauthorized access and potential harm. Network slicing offers a solution by implementing dedicated security

VNFs and complete slice isolation to smart home slices, mitigating negative effects caused by compromised devices [33]. In smart city applications, various IoT devices like traffic management, smart parking, road lights, and waste management systems are interconnected for centralized management. Many of the devices from these systems are battery-powered and resource-constrained. So, secure and energy-efficient type communication services are essential to ensure the well-being and lifestyle of smart city residents. Network slicing proves to be a valuable technology in meeting the communication requirements of these devices [34]. So, by providing lightweight VNFs in a dedicated slice with complete isolation, effective communication is achieved while maintaining security and preserving the integrity of the overall network.

#### 5) SLICING FOR GAMING

Traditional networks are inadequate to meet the strict QoS requirements, such as less than 20ms (ideally 5 to 9ms) for AR/VR gaming applications. Security and privacy are very important in applications like remote surgeries and virtual meetings due to their direct handling of sensitive information. AR/VR services demand high data rates for 360-degree videos requiring around 25Mbits/s of bandwidth, which increases significantly with higher video quality [35]. These diverse QoS requirements can be accommodated through network slicing. The novel 5G eMBB service scenario is specifically designed to support such applications.

#### 6) SLICING FOR SMART ENERGY APPLICATIONS

A smart grid encompasses numerous connected components in a large area. These components have different QoS requirements, serving various applications such as power line monitoring, smart homes, and smart vehicles. Existing network infrastructure faces challenges in providing reliable connectivity, managing the massive data generated, and fulfilling these strict requirements [36]. So, these challenges can be effectively addressed in a cost-efficient manner by allocating dedicated slices to each application in a smart grid environment. The security of smart grid applications is also a major concern, which can have severe consequences for an entire country, disrupting people's daily lives and damaging electrical assets. Network slicing offers a viable approach to incorporate security functions for grid systems, enabling the segregation of grid-based traffic and enhancing overall security.

#### 7) SLICING FOR UAVS APPLICATIONS

Nowadays, UAVs have emerged as a crucial technology for our daily lives. Communication with UAVs involves two types of data: control data and UAVs-generated data. The control data for managing the UAV operations and generated data is gathered from UAV's connected devices (sensors, cameras) for future usage. These data types have distinct network requirements that must be met in a cost-effective manner. Control data necessitates high reliability and low latency to ensure effective UAV control. On the other hand, generated data often involves transmitting a significant volume of data, which needs a high data rate [37]. So, 5G slicing in 5G addresses these challenges by enabling slice isolation and providing high data rate slices to UAVs to perform UAV operations and collect huge amounts of data from the field for future usage.

#### 8) SLICING FOR MASSIVE IOT APPLICATIONS

Network slicing offers a promising solution to meet the network requirements of massive IoT services, also known as mMTC in 5G, which involves billions of low-power devices transmitting small amounts of data. For example, in weather monitoring and smart farming applications, where IoT devices are deployed in rural areas with limited network capacity and coverage, these constrained devices need seamless connectivity to send collected data and preserve battery life [38], [39]. So, the power usage of constrained IoT devices can be optimized by using a dedicated slice with optimized and lightweight VNFs.

#### 9) SLICING FOR MILITARY APPLICATIONS

Ensuring secure communication in battlefield scenarios, such as soldier-to-control center communication and transmission of data from IoT devices, is of utmost importance. Utilizing a public network with wide coverage for military communication offers a cost-effective solution but a big security threat [40]. So, a dedicated, isolated, and secure network slice with proper security VNFs can provide secure communication to military applications.

IoT has made significant advancements in the smart retail industry, covering various areas such as smart vending machines and supply-chain management. It is considered a prominent use case for MEC. Network slicing is essential for enabling technologies like AR/VR that can profoundly impact customer decision-making in smart retail [41]. The application of smart wearables extends across domains like smart cities, healthcare, and gaming. Network slicing combined with MEC offers a solution to address critical challenges in communications of smart wearables, including limited battery and computing capabilities [42]. Smart supply chain management is recognized as a leading IoT application involving tasks such as tracking goods during transportation and exchanging inventory information. 5G mMTC slice can effectively support use cases in the smart supply chain. So, from the above discussion, network slicing seems a promising solution for diverse IoT services that provide completely isolated resources to enhance security and ensure high throughput and low latency communication.

Table 1 presents a comparative analysis of our system with the relevant techniques in the literature, which shows the superiority of our system for efficient IoT service provisioning. As explained, our system performs complete

**TABLE 1.** Comparative analysis of proposed mechanism with different approaches in the literature.

| Paper | Objective | E2E Slicing | E2E O&M | MEC support | AI Intelligence |
|---|---|---|---|---|---|
| [24] | Developed a robust network slicing and fog computing-supported authentication system designed to enhance the efficiency and security of 5G-empowered IoT services. | ✓ | ✗ | ✗ | ✗ |
| [25] | Presented a dynamic network slicing scheme for multitenant heterogeneous cloud RANs, considering tenant priorities, QoS assurance, interference management, and resource allocation. | ✓ | ✗ | ✗ | ✗ |
| [26] | This paper focuses on RAN resource sharing. It explores the benefits of enabling non-orthogonal resource sharing in uplink communications among mMTC, eMBB, and URLLC devices communicating with a central BS. | ✓ | ✗ | ✗ | ✗ |
| [27] | Introduces the Distributed Autonomous Slice Management and Orchestration (DASMO) concepts for overcoming scalability issues. In addition, this paper outlines the DASMO architecture and delves into the scalability aspects of DASMO monitoring. | ✓ | ✓ | ✗ | ✗ |
| [28] | Presented an SDN-NFV-enabled network slicing mechanism for smart grid applications. | ✓ | ✓ | ✗ | ✗ |
| [29] | Introduces a network-slicing architecture to enhance reliability in smart healthcare applications. The architecture employs application fingerprinting to efficiently tailor network resources and meet the specific reliability needs of each smart healthcare application. | ✓ | ✗ | ✗ | ✓ |
| [30] | Presented a customized network slicing mechanism for V2X services. This slicing solution encompass the allocation and slicing of RAN and core network resources, along with the configuration of end-device functionality to support a wide range of V2X use cases. | ✓ | ✓ | ✗ | ✗ |
| [31] | Demonstrated a network slicing mechanism for efficient and flexible service provisioning in IIoT networks. | ✓ | ✓ | ✗ | ✗ |
| [32] | This work focuses on SDN-based orchestration of network slicing for industrial use cases with strict and varying QoS requirements, specifically wind power plant networks. | ✓ | ✓ | ✗ | ✗ |
| Our Work | introduced an automation framework that harnesses the capabilities of SDN, NFV, AI, and MEC to streamline the provisioning and management of IoT critical and industrial services, optimizing efficiency and performance. | ✓ | ✓ | ✓ | ✓ |

automation and management of e2e IoT service. However, most systems in the literature focused on specific aspects such as core slicing, RAN slicing, or slice isolation. Our mechanism provides complete lifecycle management support of IoT service using AI, SDN-NFV, MANO, and MEC support.

### B. ORCHESTRATION AND MANAGEMENT OF IOT SERVICES

Several open-source orchestration platforms were implemented based on the 3GPP and ETSI standards. These platforms support network slicing and automation. Some well-known orchestration platforms are Tacker, open network automation platform (ONAP), Open Network Foundation (ONF), COMEC, JOX, M-CORD, SONATA, OPNFV, 5G NORMA, Cloudify, OpenBaton, OpenStack HEAT, and Open-O [43], [44], [45], [46], [47], [48], [49]. The primary aim of these platforms is to enable programmability to automate the network resources deployment over the infrastructure. The network administrators define the policy configurations for the deployment of the resources. The OpenStack platform is used to deploy the VNFs, and SDN-based controllers are used for chaining the VNFs. So, these existing orchestration platforms require specific and complex network configurations to activate resources.

The Open-Source MANO (OSM) orchestration platform has been developed based on ETSI NFV-MANO specifications that can efficiently orchestrate and manage the lifecycle of the network services. It has an integrated cloud platform, OpenStack, as a VIM and supports SDN controllers [50]. Moreover, ETSI has also introduced a Zero-Touch Service Management system (ZSM) for the complete automation of the network. It is an entirely closed-loop system that does not need any human involvement when in execution mode. It considers different AI, Machine Learning (ML), and big data approaches for proactively managing the network. It uses ML models to learn the user traffic patterns from the network and performs future predictions. It performs data analytics and extracts the network's trends, patterns, and behavior. Based on the predictions of the ML models, ZSM can proactively prepare the resources and perform autoscaling of the VNFs resources. It can manage Physical Network Functions (PNFs), VNFs, and physical infrastructure [51], [52].

### C. AI FOR NETWORK SLICING AUTOMATION AND PROACTIVE MANAGEMENT

Recently, AI has emerged as a promising technology for enhancing various network automation tasks within the context of B5G networks. Standardization bodies and researchers have increasingly adopted AI methods to enable proactive network management and assurance [53], [54]. For example, 3GPP introduced NWDAF in 5G for performing AI-enabled network intelligence. On the other side, ZSM also provides openness for implementing AI algorithms for network automation. IETF Intent-based networking (IBN) technology also introduced AI approaches for network automation. These AI and ML techniques majorly encompass a range of approaches, including Supervised Learning (SL), Unsupervised Learning (UL), Reinforcement Learning (RL), Federated Learning (FL), and Transfer Learning (TL), geared toward network automation. Each paradigm involves an exploration/training phase to optimize the learning algorithm and an exploitation/prediction phase to make inferences on new inputs. UL does not rely on labeled data but trains on an unlabeled dataset. K-means, K-Nearest Neighbors (KNN), and Principal Component Analysis (PCA) are popular clustering and dimensionality reduction mechanisms. In SL, a labeled dataset is used to learn a function that maps inputs to expected outputs. Standard techniques like Artificial Neural Networks (ANN), Bayesian networks, Random Forests (RF), Decision Trees (DT), Support Vector Machines (SVMs), Recurrent Neural Networks (RNNs), and Convolutional Neural Networks (CNNs) are used to solve supervised learning problems [55]. Conversely, RL does not involve explicit training; the agent or decision-maker learns and adapts in real-time to maximize long-term rewards. RL is valuable in control problems, such as resource autoscaling and power control, where the agent must respond to changing environmental conditions. Q- Learning (QL), Asynchronous Advantage Actor-Critic (A3C), Deep Q-Network (DQN), and Deep Deterministic Policy Gradient (DDPG) are some popular RL algorithms used in controlling future network operations [56], [57].

Besides, FL is a novel technique of ML where various nodes or devices keep their data locally and are involved in the collaborative training of a shared model. In FL, the model is distributed to individual devices rather than transmitting raw data to a central server. Each device makes local updates to the model using its data, and these updates are later combined to enhance the global model. Furthermore, TL is a technique that utilizes knowledge obtained from one domain/task and applies it to another related domain/task. Rather than training a model from scratch, a pre-trained model is used as a foundation and fine-tuned on the target task with a smaller dataset [55], [58].

However, with AI/ML algorithms, B5G MNOs can achieve automatic service provisioning, network intelligence, and proactive management of critical network resources across the core, RAN, MEC, and transport domains.

## IV. PROPOSED ARCHITECTURE OF AI AND MEC ASSISTED IOT SERVICE PROVISIONING MECHANISM

Autonomously provisioning resources over edge cloud and multi-domain infrastructure is very challenging. As a result, the traditional manual configuration approaches for configuring e2e slices for IoT services over a heterogeneous environment are complex, error-prone, tedious, and inefficient. To this end, network automation is introduced as a software-based repeatable solution that simplifies control of multi-domain and various platforms. Due to that, this work presents an efficient and intelligent framework for the autonomous provisioning of network slices over multi-domain edge cloud network infrastructure. This AI and MEC-assisted closed-loop mechanism consists of IoT service orchestration, core cloud management, RAN management, MEC management, transport management, monitoring and big data processing, and AI-driven intelligence and analytics module. The IoT service orchestration module designs an

e2e network slice for IoT service according to the specified QoS requirements. The IoT service orchestration module has OSS/BSS system, an IoT service designer, an IoT slice template store, and a resource repository.

The Web portal or OSS/BSS system is an interactive entity where users input their IoT slice QoS requirements, such as data rate, delay, service type, and latency for slice activation. The IoT template store has various generic domain-specific templates for configuring domain resources. IoT service designer with IoT template store and resource repository generates underlying domain and platform-specific e2e IoT slice templates to create e2e slice over infra. After that, these created IoT slice templates are deployed with the help of network orchestrators and controllers for instantiating e2e slice over the core, MEC, and RAN infrastructure. Once the slices are activated at the physical infrastructure, the monitoring and big data processing system continuously collects resource utilization information from RAN, MEC, transport, and core cloud domains. Besides, the AI-driven intelligence and analytics module provides efficient and intelligent service assurance and management.

Our system efficiently deploys network slices over RAN, Core, and MEC domains, which follow a closed loop and one-touch mechanism. This implemented system performs a complete lifecycle of IoT slices by automatically designing, activating, monitoring, and deleting IoT services. Figure 6 depicts the proposed AI and MEC-assisted IoT service provision and assurance architecture. The working functionality of each component is presented below. The proposed system consists of the following components:

### A. OPERATION SUPPORT SYSTEM AND BUSINESS SUPPORT SYSTEM (OSS/BSS)

It is a Web-based portal where IoT service providers and users define their IoT slice QoS requirements. It exposes many Application Programming Interfaces (APIs) to communicate with other modules of the system and push the configurations to underlying modules and platforms for implementing e2e network slices over RAN, MEC, transport, and core domain.

### B. IOT SLICE TEMPLATE STORE

It is a slice template generation entity that contains predefined multiple types of IoT slice templates in the form of JSON script or YAML script for configuring multi-domain network resources. As each of the underlying platforms may require a different type of configuration; for example, the ONF-COMAC platform requires HELM, and OSM requires NST (Network Slice Templates), VNFD (virtual Network function descriptors), and NSD (Network Service Descriptors). Similarly, every RAN controller and WAN SDN controller has specific scripting languages. Therefore, considering the variety of configuration requirements, this network template store provides configuration according to the needs of the underlying platform.

### C. RESOURCE REPOSITORY

There are a variety of standard architectures and development options to orchestrate network slicing for IoT services. In addition to that, multiple vendor-specific VNFs may be orchestrated per the operator's requirements. Hence, the resource repository is a network resource database that contains information on different VNFs, physical infrastructure, deployed platforms, and network architectures. Using the information, the IoT slice designer can allocate specific VNFs and how those VNFs communicate with each other. Every slice creation request may require a chain of VNFs, including network configuration specifications, to achieve the required bandwidth.

### D. IOT SLICE DESIGNER

The central IoT service designer module receives high-level operator requirements and, based on that, associates E2E slice context. The context is designed using the slice template store and resource repository. In addition, slice demand satisfaction is achieved by different functions; firstly, we propose VNF placement on a specific compute node based on a specified slice's latency and resource requirements. Secondly, routing path and flow rules are managed to ensure network slice QoS provisioning for transport network QoS satisfaction. Finally, it handles each slice RAN specific resources and VNF compute (CPU, RAM, and Storage) resources; the resource allocation is performed based on GBR (Guaranteed Bandwidth Rate) and non-GBR schemes in accordance to slice requirements. In addition to that, it includes the management of monitoring and updating configurations. Each platform has its monitoring procedures. To this end, the IoT service designer attaches monitoring configurations with the slice configuration to monitor the performed control actions on runtime. Finally, after creating the e2e slice design, the e2e policy translator generates and distributes the specific platforms' policy configurations through REST API interfaces.

### E. B5G CORE NETWORK MANAGEMENT

It consists of service providers and cloud-based 5G core components. Based on the slice requirements, different services can be orchestrated on B5G core networks. In our implementation, we have used a very popular NFVO named Open-Source MANO (OSM) to manage and automate the 5G core network VNFs, such as AMF, UPF, SMF, etc. OSM is an advanced network orchestrator developed by the ETSI [50]. Its purpose is to empower MNOs in automating the deployment of e2e B5G network services. Based on the NFV specifications, OSM consists of several modules, including the NFVO, VIM, and VNFM. The platform offers a user-friendly GUI portal for efficient management, monitoring, and automated deployment of network resources. With Open-source MANO, users can easily deploy and manage e2e slices while effectively overseeing their lifecycle. To deploy the slice template over the core, the IoT service orchestration module transmits it to OSM via the REST API,

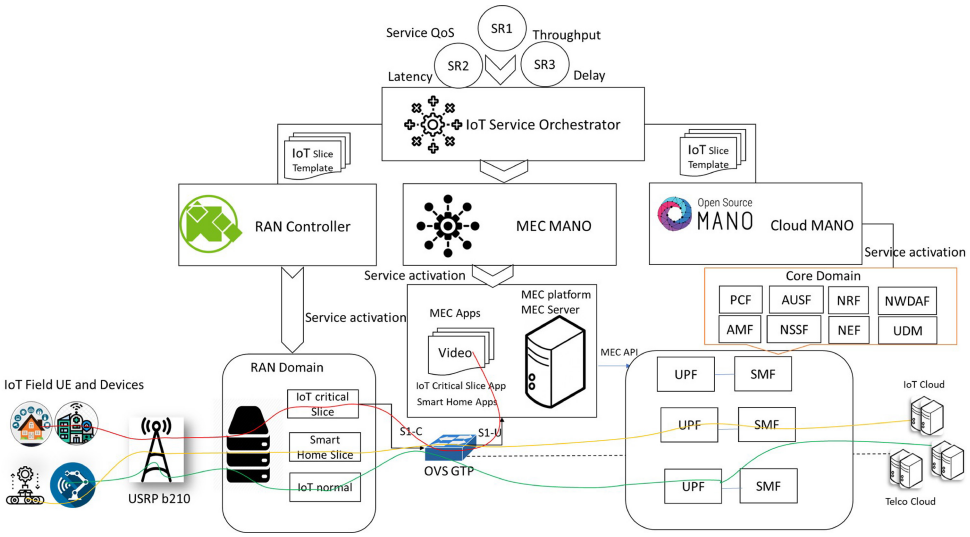



**FIGURE 6.** Implemented tested with all the components for IoT service provision and management.

enabling OSM to deploy the core network VNFs using the OpenStack-based VIM automatically.

### F. MULTI-ACCESS EDGE COMPUTING MANAGEMENT

MEC serves as a low-latency service provider and contains MEC applications, IoT low-latency applications, MEC platform, and other MEC components that provide services at the edge. In slice orchestration based on operator setting, many VNFs and services can be orchestrated to achieve slice latency demands. They receive the configuration from North-Bound APIs and orchestrate them using VIM (Virtual Infrastructure Managers)on the physical computing infrastructure. We have used low-latency MEC developed by the Mosaic group [59] to accommodate the low-latency IoT services such as critical IoT slices in our IoT network slicing testbed. The low latency MEC platform is an SDN-based platform comprised of the MEC platform, MEC application, and Abstraction. It provides complete programmability and flexibility to deploy and develop MEC applications and handles MEC operations for ensuing low latency and content-aware service provisioning for various network applications. Hence, low latency MEC is the best candidate to serve critical IoT services with low latency and fast response time.

### G. TRANSPORT NETWORK MANAGEMENT

The SD-WAN controller handles communication between different MEC sites and core networks. We have used the OpenDaylight SDN controller for managing routing tasks for establishing e2e IoT slicing in the transport network. It manages different routings according to the allocated and available resources to provide the specific bandwidth to the slices. The infrastructure and network distribution type performs segment routing, OSPF-based routing, and optimal path routing. It also receives the configuration from the northbound interface and activates them to physical networking devices such as SDN switches in our implementation.

### H. RAN MANAGEMENT

In other words, it is known as the RAN resource orchestrator and RAN domain controller. It enables control of network RAN resources. Each slice requires specific bandwidth, and the RAN controller serves as a control plane for handling RAN resource allocation. Depending on resource requirements, it allocates radio resources for each slice. In addition, it can control multiple RRUs (Radio Remote Units), and it receives configurations through the north-bound APIs and translates them to physical infrastructure. We have used the FlexRAN controller for managing RAN domain operation as well as slicing the RAN [60]. It performs static and dynamic RAN slicing. It accepts the RAN slice template in JSON string format.

### I. MONITORING AND BIG DATA PROCESSING

This module receives continuous monitoring information and analyzes critical events. Several monitoring agents are deployed on RAN, MEC, and core network resources for collecting data. The collected data of each module can be stored on different servers for future usage. Once the data is stored and we perform many data processing operations to make the data ready for the AI module for training multiple AI models to achieve various automation and management tasks. Based on dynamic thresholding and slice performance measures, it either invokes ML models to perform updates in underlying platforms or invokes direct updates in the underlying platforms through the IoT service orchestration module.

### J. AI-DRIVEN INTELLIGENCE AND ANALYTICS

Several domain-specific ML approaches can be added to this intelligence and analytics module for complete assurance

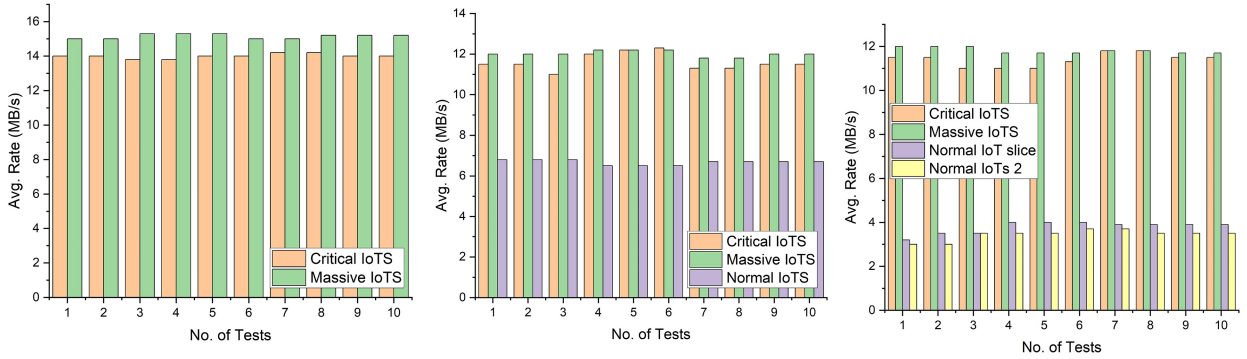



**FIGURE 7.** Experimental results of implemented IoT e2e slices with different QoS requirements a) illustrates the results of recorded downlink throughput recorded from two IoT slices b) shows the downlink throughput recorded from the deployed three IoT slices c) presents the throughput results achieved from four IoT slices.

and autonomic network slice orchestration. The SL, UL, FL, TL, and RL methods can be implemented to achieve multiple proactive assurance and management use cases such as DL-based network resource forecasting, attack detection from IoT slices, RL-based VNF placement, RL-based service function chaining, RL-based congestion control, and many more. AI seems to be the best option to provide intelligence to each RAN, MEC, core, and transport network domain and ensure better QoE to customers while preparing the resources proactively. AI can also assist in many network automation and management operations.

This part introduces the diverse use cases and applications of AI for core, RAN, MEC, and transport slicing for intelligent automation and management.

#### 1) AI FOR CORE NETWORK SLICING AUTOMATION

NFV technology enables MNOs to implement core network slicing cost-effectively and flexibly. VNFs constitute a significant part of core network slicing and can be deployed using VMs, containers, Kubernetes, and microservices. Multiple core VNFs can be selected, chained together to provide a specific service, and deployed at various locations per the QoS requirements. However, the lifecycle management of core VNFs is very crucial and challenging. AI approaches are very promising for performing lifecycle management of core VNFs in network slicing context. AI algorithms can be applied to perform VNFs resource usage prediction, anomaly detection, power management of core cloud, attack detection, service function chaining, autoscaling of VNFs, VNF placement, VNF scheduling, VNF configurations, VNF live migration, and efficient resource allocation.

#### 2) AI FOR MEC MANAGEMENT

MEC is a crucial technology to support B5G diverse requirement services, and it provides analytics, management, and computing support within or near RAN in close user proximity. The complexity arises in MEC from uncertain, multi-dimensional, and dynamic characteristics, making knowledge discovery, pattern learning, and decision-making optimization challenging. Traditional algorithms may have limitations in coping with such complexities of B5G networks. AI methods come to the rescue by extracting valuable insights from collected data and supporting various functions, such as prediction, optimization, and decision-making in the MEC environment. Lightweight AI algorithms can be utilized to develop intelligent edge applications because of the limited capabilities of MEC servers. For instance, RL-based model-free algorithms can be used for MEC resource management that does not rely on historical data. It learns from the environment in real-time to make suitable decisions. AI algorithms can handle various MEC operations such as efficient resource management, edge security, edge analytics, resource usage predictions, context-aware services, task offloading, caching optimization, and edge traffic management.

#### 3) AI FOR RAN MANAGEMENT

RAN slicing is a vital and complicated part of network slicing that fulfills QoS requirements. AI promises to perform proactive, optimized, and intelligent RAN management in an automated fashion. Multiple SL, UL, DL, RL, FL, and TL algorithms can be used in various tasks for RAN slicing management, such as RL-based spectrum management, mobility and handover management, beam forming and antenna optimization, carrier isolation, resource block management, spectrum sharing, priority and QoS-aware resource allocation, interference management, predictive maintenance, QoS assurance, energy efficiency, security, DL-based intelligent monitoring and prediction, and QoE assurance.

#### 4) AI FOR TRANSPORT MANAGEMENT

In B5G networks, transport slicing involves virtualizing network resources, such as nodes, links, and physical ports, enabling on-demand network transmission performance. AI brings automation, adaptability, and intelligence to transport slicing, enabling efficient resource utilization, better QoE, and increased agility in meeting the diverse QoS of B5G networks. AI can be applied to support intelligent and efficient transport slicing, such as RL-based dynamic bandwidth scaling, DL for intelligent traffic engineering and routing, multi-path resource allocation and optimization,

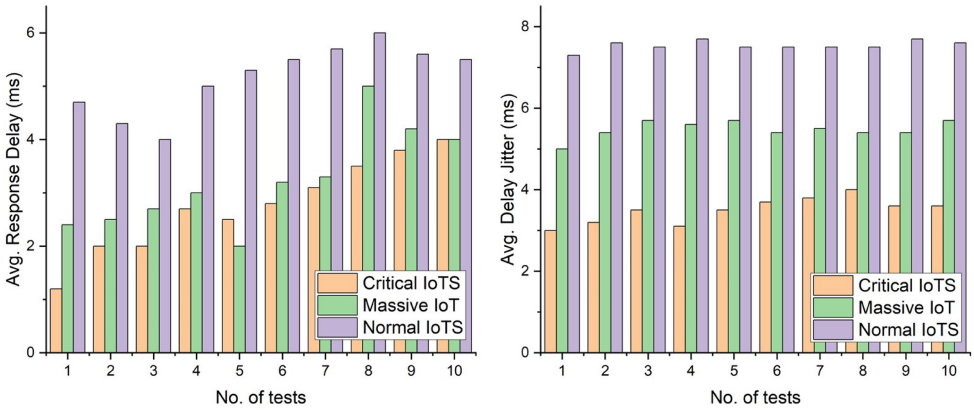



**FIGURE 8.** Presents the results of performance metrics such as average response delay and jitter delay achieved while deploying various e2e IoT slices a) recorded average response delay from the three activated slices b) achieved jitter delay from three deployed slices.

transport slice configurations, SLA assurance, and predictive maintenance and failure recovery.

#### 5) AI FOR HIGHER-LEVEL MANAGEMENT

The complete lifecycle management of network slicing involves slice designing, activation, monitoring, update and assurance, and deletion. AI can assist in performing multiple network slicing operations such as NLP-based intent translation, DRL for service designing, slice template preparation, DL for cost-aware slice capacity forecasting, SLA assurance, and multidomain SLA decomposition. So, integrating AI approaches is very important to perform complete lifecycle management of network slicing automatically.

Due to the aforementioned discussion, AI empowers B5G MNOs to perform network intelligence, efficient service provisioning, proactive management, and assurance within the context of e2e network slicing. For example, the SL approach, such as Long Short-Term Memory (LSTM), RF, and Support Vector Regression (SVR), can be used to predict the core slice VNFs utilization or data center capacity forecasts. The predicted resource usage results will be used for autoscaling decisions. In our scenario, the prediction results of the SL model on different time windows can be used by the IoT service orchestration module for making autoscaling decisions, such as scaling up or scaling down the core VNFs. So, the IoT service orchestrator decides the resource scaling and requests the core NFVO. Moreover, NFVO commands VNFM and VIMs to perform IoT slice core resource scaling.

## V. EXPERIMENTAL IMPLEMENTATION AND RESULTS

### A. EXPERIMENTAL TESTBED AND SETTINGS

Figure 8 illustrates our implemented e2e IoT slice provisioning and management testbed, featuring various components such as the IoT service orchestrator for designing IoT slice template, OSM for core and MEC VNFs and application deployment, FlexRAN controller for controlling and managing RAN slicing, low-latency MEC for IoT low latency service such as IoT critical slice in our experiments, SDN controller OpenDaylight for handling transport operations, OpenAirInterface (OAI) [61] 5G core VNFs, OAI gNB, and IoT UEs and devices and OAI UEs. The OAI is an open-source community that implements 5G core, RAN, and UE. Moreover, we have implemented low-latency MEC for providing low-latency services for IoT-critical applications. So, the IoT critical service can access the content from MEC applications, and normal IoT slices can serve from the IoT cloud. Our IoT service orchestrator application has been deployed to a PC with 16GB of RAM, a core I7 processor, a Windows 10 operating system (OS), and 512GB storage. This IoT service orchestration application is developed in HTML, Bootstrap, JavaScript, MYSQL database, and PHP and Python as programming languages. OSM with integrated OpenStack is deployed in a separate server with enough capabilities: 252 GB of memory, Ubuntu 18, 2 TB of storage, and 32 cores processor. OSM with OpenStack deploys the 5G core VNFs (UPF, AMF, SMF, etc.) for providing 5G core network capabilities to IoT slices.

Moreover, OAI gNB is deployed in a separate PC with Universal Software Radio Peripheral (USRP) B210 as a Software Defined Radio (SDR) for implementing RAN. OAI-simulated UEs and Raspberry Pi devices are used as IoT UEs for testing IoT slice e2e connectivity. On the other side, a low-latency MEC system with its components MEC platform and MEC application (as a video server and Web server) is also deployed in a server with 56GB memory, 512GB storage, and Ubuntu 18 LTS. SDN Open Virtual Switch (OVS) is deployed to provide paths (gateway) for accessing low-latency applications residing inside the MEC server. The MEC platform set up the forwarding rules inside OVS switches to communicate with MEC entities. Besides, the FlexRAN controller and SDN OpenDayLight controllers are deployed in two separate PCs with 16GB memory, core I7, Ubuntu 18, and 512GB storage. In our experimental setup, the 5G VNFs, RAN, and MEC components are interconnected using LTE and 5G network topology for establishing e2e network slices.

### B. RESULTS AND ANALYSIS

We have successfully designed and implemented a network automation platform specifically for e2e network slicing of IoT services. This platform offers automated functionalities for creating, updating, monitoring, and deleting network slice instances. To facilitate user interaction, we have developed a dashboard portal that allows users and MNOs to define the QoS requirements for IoT slices. The requested slice requirements are translated into domain-specific slice templates by our IoT service orchestration module. These templates are then deployed and activated over the infrastructure by the network orchestrator (OSM), FlexRAN controller, and SDN controller. The IoT service orchestrator communicates with the underlying domain orchestrators and controllers through a REST interface. Our IoT orchestration system is capable of managing and handling complex configuration generation tasks, and preparing slice templates that align with the underlying platforms. It serves as the primary orchestrator, managing IoT slice design, admission, and control mechanisms. The OSM orchestrator is responsible for deploying core network VNFs, the FlexRAN controller handles RAN slicing, the MEC NFVO handles MEC application deployment, and the OpenDayLight SDN controller manages transport network operations in our system.

Furthermore, our IoT service orchestration system prepares the slice templates for the OSM, FlexRAN controller, MEC, and transport controller. Our current testbed has focused on testing core and RAN slicing for IoT services. For this purpose, we have utilized OAI 5G core, OAI RAN, and OAI UE components to create an end-to-end network slicing testbed for IoT services. In addition, we have customized and developed radio resource management applications on top of the FlexRAN controller. This allows us to slice the RAN resources based on the RAN slice template received from the upper IoT service orchestration module. To establish an end-to-end connection, we have stitched together the RAN slice and core EPC VNFs, providing a dedicated AMF (or vMME in LTE networks) configuration in the template. Each slice is admitted using a unique public land mobile network (PLMN) ID and slice type, ensuring access to the specific slice. Through these efforts, we have created a comprehensive network-slicing framework tailored to IoT services, enabling efficient resource management and seamless connectivity across the network.

To assess the stability of our system, we conducted a series of tests involving three types of slices: critical IoT slice, massive IoT slice, and normal IoT slice. Figure 8a illustrates the first case where we deployed two static slices, a critical IoT slice and a massive IoT slice, utilizing 50% of the available RAN resources. We defined the QoS requirements through the OSS/BSS Web portal, and the IoT orchestrator automatically activated the end-to-end IoT slices across the infrastructure in collaboration with other components. As shown in Figure 8a, the downlink speed was measured for the deployed critical IoT slice and massive IoT slice. Both slices exhibited similar throughput, reaching a maximum of 15MB/s and 15.5MB/s, respectively, owing to equal resource allocation based on the SLA.

In Figure 8b, the average downlink throughput of the three slices was recorded during multiple tests with varying QoS requirements. The critical IoT slice, massive IoT slice, and normal IoT slice were deployed with 40%, 40%, and 20% of the gNB resource capacity, respectively. The critical IoT slice achieved a maximum average downlink throughput of 12.4MB/s, while the massive IoT slice reached a maximum of 12.5MB/s. Furthermore, the normal IoT slice exhibited a maximum throughput speed of 7MB/s. Figure 8c presents the average downlink throughput achieved when activating four IoT slices: critical IoT slice, massive IoT slice, normal IoT slice, and normal IoT slice 2, utilizing 40%, 40%, 10%, and 10% of the RAN resource capacity, respectively. Both the critical IoT slice and massive IoT slice, deployed with identical QoS requirements, achieved an average downlink throughput of approximately 12MB/s during multiple tests. On the other side, normal IoT slice 1 and slice 2 were deployed with the same resources and achieved a maximum of 4MB/s and 3.7MB/s average downlink data rates, respectively.

Figure 8 shows the performance metrics obtained from three different types of IoT services with varying QoS requirements, including average response delay and average jitter delay. In order to fulfill the low latency slice requirements, we employed a priority mechanism to prioritize slices with reduced delays and faster response times. To cater to low-latency communication needs, we designated the critical IoT slice as the most sensitive slice, demanding minimal latency and rapid response times. This critical IoT slice is served by low-latency MEC applications, where a Web server is deployed for access via IoT UEs. The massive IoT slice is assigned as the second-priority slice, also supported by the MEC system, while the IoT normal slice is considered a best-effort slice with no strict latency demands.

Several tests were conducted to evaluate the activated IoT slices' performance and communication capabilities. Figure 8a presents the recorded average response delay for each deployed slice. It is evident that the IoT critical slice outperforms the massive and IoT normal slices, achieving a faster response time. It attains a minimum response delay of 1.2ms and a maximum response delay of less than 4ms, even in worst-case scenarios. Conversely, the massive IoT and normal slices exhibit maximum response delays of 5ms and 6ms, respectively. Overall, all IoT slices yielded satisfactory results within our testbed, which was deployed using open-source solutions.

Figure 8b depicts the average jitter's results recorded during multiple tests on the three deployed slices. As previously mentioned, the critical IoT slice, designated as the priority slice, demonstrates a lower jitter delay compared to the other slices. The critical IoT slice exhibits a maximum jitter delay of 4ms, the massive IoT slice reaches a maximum of 5.7ms, and the normal IoT slice records a maximum of 7.7ms jitter delay. In the best-case scenarios, these

slices achieve minimum jitter delays of 3ms, 5ms, and 7.2ms, respectively. These critical, massive, and normal IoT slices showcase stable performance by providing high bandwidth, faster response times, and low latency, fulfilling the requirements of most IoT applications. Overall, our IoT service provisioning mechanism performs satisfactorily by creating end-to-end network slicing for IoT services with specific QoS requirements. It ensures dedicated core, MEC, and RAN resources for each IoT service, ensuring complete isolation and meeting their individual needs.

## VI. LIMITATIONS AND FUTURE CHALLENGES

Multidomain network slicing for IoT, while promising, presents several notable limitations. First and foremost, our system places high importance on the dynamic allocation of resources in a multi-tenant environment. To tackle this, we are developing an efficient mechanism utilizing RL. This cutting-edge approach will autonomously oversee slice admission and control, simplifying the entire process. Secondly, optimizing the sharing of Radio Access Network (RAN) resources is a significant challenge in future networks. Our strategy involves the creation of an AI-powered solution that automates the allocation and distribution of RAN resources. This automation will be finely tuned to meet IoT services' specific Quality of Service (QoS) demands. Moving on, achieving zero-touch and seamless IoT service provisioning and assurance necessitates automatically translating network policies. We are actively harnessing the power of natural language processing (NLP) techniques to meet this requirement. This innovative approach will effortlessly translate the requirements of IoT users into network policies. Additionally, Ensuring the security of IoT data and devices across multidomain is a complex endeavor, with vulnerabilities possibly emerging at interconnection points. So, a zero-trust network security mechanism is needed to ensure a secure and reliable environment for B5G users. Additional challenges include scalability concerns, management overhead, and consistent Quality of Service (QoS) assurance. Lastly, we are integrating and developing AI-based solutions to provide complete network intelligence and lifecycle management of e2e IoT services. These limitations underscore the need for innovative solutions and efficient management strategies to harness the full potential of multidomain network slicing for IoT services.

## VII. CONCLUSION

To automate the provisioning and management of IoT services, we have implemented a well-designed AI and MEC-enabled architecture that assists network service providers in deploying multidomain network resources efficiently and flexibly. Users and service providers can input their QoS requirements and design their IoT network slice through the OSS/BSS portal. To facilitate resource deployment, the IoT service orchestration module generates the IoT slice template for the RAN, core, MEC, and transport domains. Management and orchestration entities, such as NFVO for MEC and core, RAN controller, and transport controller, activate and deploy the resources based on the received IoT slice template across each domain infrastructure. Our system adopts a one-touch approach for activating e2e network slices for IoT services. Furthermore, we have integrated the MEC system to provide low-latency communication for critical IoT services. Our system automates IoT service design, activation, monitoring, and resource depletion. We have conducted various tests by activating multiple IoT slices with different QoS requirements over the multidomain resources, demonstrating stable and efficient performance in terms of flexibility, throughput, slice isolation, delay, and reliability. Therefore, our system efficiently enables the automatic provisioning and management of e2e slices for IoT services. In the future, we plan to enhance our framework by introducing more AI-driven automation use cases and implementing a blockchain-based zero-trust security mechanism to provide enhanced security for IoT slicing.

## REFERENCES


[1] B. Ahlgren, M. Hidell, and E. C.-H. Ngai, "Internet of Things for smart cities: Interoperability and open data," *IEEE Internet Comput.*, vol. 20, no. 6, pp. 52–56, Nov./Dec. 2016.

[2] R. Dobbs, J. Manyika, and J. Woetzel, *The Internet of Things: Mapping the Value Beyond the Hype*, McKinsey & Company, New York, NY, USA, 2015.

[3] C. Sobin, "A survey on architecture, protocols and challenges in IoT," *Wireless Pers. Commun.*, vol. 112, no. 3, pp. 1383–1429, 2020.

[4] G. A. Akpakwu, B. J. Silva, G. P. Hancke, and A. M. Abu-Mahfouz, "A survey on 5G networks for the Internet of Things: Communication technologies and challenges," *IEEE Access*, vol. 6, pp. 3619–3647, 2017.

[5] S. Wijethilaka and M. Liyanage, "Survey on network slicing for Internet of Things realization in 5G networks," *IEEE Commun. Surveys Tuts.*, vol. 23, no. 2, pp. 957–994, 2nd Quart., 2021.

[6] I. Afolabi, T. Taleb, K. Samdanis, A. Ksentini, and H. Flinck, "Network slicing and softwarization: A survey on principles, enabling technologies, and solutions," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 3, pp. 2429–2453, 3rd Quart., 2018.

[7] Z. Abou El Houda, B. Brik, and L. Khoukhi, "Ensemble learning for intrusion detection in SDN-based zero touch smart grid systems," in *Proc. IEEE 47th Conf. Local Comput. Netw. (LCN)*, 2022, pp. 149–156.

[8] K. Abbas et al., "An efficient SDN-based LTE-WiFi spectrum aggregation system for heterogeneous 5G networks," *Trans. Emerg. Telecommun. Technol.*, vol. 33, no. 4, 2022, Art. no. e3943.

[9] F. Guo, F. R. Yu, H. Zhang, X. Li, H. Ji, and V. C. Leung, "Enabling massive IoT toward 6G: A comprehensive survey," *IEEE Internet Things J.*, vol. 8, no. 15, pp. 11891–11915, Aug. 2021.

[10] B. Han, W. Jiang, M. A. Habibi, and H. D. Schotten, "An abstracted survey on 6G: Drivers, requirements, efforts, and enablers," 2021, *arXiv:2101.01062*.

[11] P. Cruz, N. Achir, and A. C. Viana, "On the edge of the deployment: A survey on multi-access edge computing," *ACM Comput. Surveys*, vol. 55, no. 5, pp. 1–34, 2022.

[12] "Multi-access edge computing (MEC); framework and reference architecture," ETSI, Sophia Antipolis, France, ETSI GS MEC 003 V3.1.1, 2019.

[13] F. Giust, X. Costa-Perez, and A. Reznik, "Multi-access edge computing: An overview of ETSI MEC ISG," *IEEE 5G Tech Focus*, vol. 1, no. 4, p. 4, 2017.

[14] Z. Abou El Houda, B. Brik, A. Ksentini, and L. Khoukhi, "A MEC-based architecture to secure IoT applications using federated deep learning," *IEEE Internet Things Mag.*, vol. 6, no. 1, pp. 60–63, Mar. 2023.

[15] J. Cao et al., "A survey on security aspects for 3GPP 5G networks," *IEEE Commun. Surveys Tuts.*, vol. 22, no. 1, pp. 170–195, 1st Quart., 2020.

[16] "System architecture for the 5G system," 3GPP, Sophia Antipolis, France, 3GPP Rep. TS 23.501 V15.3.0, 2018.

[17] G. Brown, "Service-based architecture for 5G core networks," Shenzhen, China, Huawei, White Paper, 2017.

[18] S. Zhang, "An overview of network slicing for 5G," *IEEE Wireless Commun.*, vol. 26, no. 3, pp. 111–117, Jun. 2019.

[19] I. Afolabi, T. Taleb, P. A. Frangoudis, M. Bagaa, and A. Ksentini, "Network slicing-based customization of 5G mobile services," *IEEE Netw.*, vol. 33, no. 5, pp. 134–141, Sep./Oct. 2019.

[20] K. Abbas, T. A. Khan, M. Afaq, and W.-C. Song, "Network slice lifecycle management for 5G mobile networks: An intent-based networking approach," *IEEE Access*, vol. 9, pp. 80128–80146, 2019.

[21] A. H. Celdrán, M. G. Pérez, F. J. G. Clemente, F. Ippoliti, and G. M. Pérez, "Dynamic network slicing management of multimedia scenarios for future remote healthcare," *Multimedia Tools Appl.*, vol. 78, pp. 24707–24737, Sep. 2019.

[22] C. Campolo, A. Molinaro, A. Iera, R. R. Fontes, and C. E. Rothenberg, "Towards 5G network slicing for the V2X ecosystem," in *Proc. 4th IEEE Conf. Netw. Softw. Workshops (NetSoft)*, 2018, pp. 400–405.

[23] A. E. Kalør, R. Guillaume, J. J. Nielsen, A. Mueller, and P. Popovski, "Network slicing in industry 4.0 applications: Abstraction methods and end-to-end analysis," *IEEE Trans. Ind. Informat.*, vol. 14, no. 12, pp. 5419–5427, Dec. 2018.

[24] J. Ni, X. Lin, and X. S. Shen, "Efficient and secure service-oriented authentication supporting network slicing for 5G-enabled IoT," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 3, pp. 644–657, Mar. 2018.

[25] Y. L. Lee, J. Loo, T. C. Chuah, and L.-C. Wang, "Dynamic network slicing for multitenant heterogeneous cloud radio access networks," *IEEE Trans. Wireless Commun.*, vol. 17, no. 4, pp. 2146–2161, Apr. 2018.

[26] P. Popovski, K. F. Trillingsgaard, O. Simeone, and G. Durisi, "5G wireless network slicing for eMBB, URLLC, and mMTC: A communication-theoretic view," *IEEE Access*, vol. 6, pp. 55765–55779, 2018.

[27] S. Kukliński and L. Tomaszewski, "DASMO: A scalable approach to network slices management and orchestration," in *Proc. IEEE/IFIP Netw. Operations Manage. Symp.*, 2018, pp. 1–6.

[28] F. Kurtz, C. Bektas, N. Dorsch, and C. Wietfeld, "Network slicing for critical communications in shared 5G infrastructures-an empirical evaluation," in *Proc. 4th IEEE Conf. Netw. Softw. Workshops (NetSoft)*, 2018, pp. 393–399.

[29] A. Vergutz, G. Noubir, and M. Nogueira, "Reliability for smart healthcare: A network slicing perspective," *IEEE Netw.*, vol. 34, no. 4, pp. 91–97, Jul./Aug. 2020.

[30] C. Campolo, A. Molinaro, A. Iera, and F. Menichella, "5G network slicing for vehicle-to-everything services," *IEEE Wireless Commun.*, vol. 24, no. 6, pp. 38–45, Dec. 2017.

[31] H. Wu, I. A. Tsokalo, D. Kuss, H. Salah, L. Pingel, and F. H. Fitzek, "Demonstration of network slicing for flexible conditional monitoring in Industrial IoT networks," in *Proc. 16th IEEE Annu. Consum. Commun. Netw. Conf. (CCNC)*, 2019, pp. 1–2.

[32] V. Theodorou, K. V. Katsaros, A. Roos, E. Sakic, and V. Kulkarni, "Cross-domain network slicing for industrial applications," in *Proc. Eur. Conf. Netw. Commun. (EuCNC)*, 2018, pp. 209–213.

[33] E. Kapassa et al., "An innovative ehealth system powered by 5G network slicing," in *Proc. 6th Int. Conf. Internet Things Syst. Manag. Security (IOTSMS)*, 2019, pp. 7–12.

[34] B. Dzogovic, B. Santos, J. Noll, B. Feng, and T. Van Do, "Enabling smart home with 5G network slicing," in *Proc. IEEE 4th Int. Conf. Comput. Commun. Syst. (ICCCS)*, 2019, pp. 543–548.

[35] B. Pokric et al., "Augmented reality enabled IoT services for environmental monitoring utilising serious gaming concept," *J. Wireless Mob. Netw. Ubiquitous Comput. Depend. Appl.*, vol. 6, no. 1, pp. 37–55, 2015.

[36] A. I. Sarwat, A. Sundararajan, and I. Parvez, "Trends and future directions of research for smart grid IoT sensor networks," in *Proc. Int. Symp. Sensor Netw. Syst. Security Adv. Comput. Netw. Appl.*, 2018, pp. 45–61.

[37] M. Mozaffari, W. Saad, M. Bennis, Y.-H. Nam, and M. Debbah, "A tutorial on UAVs for wireless networks: Applications, challenges, and open problems," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 3, pp. 2334–2360, 3rd Quart., 2019.

[38] P. Yang, X. Xi, T. Q. Quek, J. Chen, X. Cao, and D. Wu, "RAN slicing for massive IoT and bursty URLLC service multiplexing: Analysis and optimization," *IEEE Internet Things J.*, vol. 8, no. 18, pp. 14258–14275, Sep. 2021.

[39] S. S. Lekshmi, M. Anjana, B. B. Nair, D. Raj, and S. Ponnekanti, "Framework for generic design of massive IoT slice in 5G," in *Proc. Int. Conf. Wireless Commun. Signal Process. Netw. (WiSPNET)*, 2019, pp. 523–529.

[40] K. Wrona, "Securing the Internet of Things a military perspective," in *Proc. IEEE 2nd World Forum Internet Things (WF-IoT)*, 2015, pp. 502–507.

[41] S. G. Dacko, "Enabling smart retail settings via mobile augmented reality shopping apps," *Technol. Forecast. Soc. Change*, vol. 124, pp. 243–256, Nov. 2017.

[42] H. Sun, Z. Zhang, R. Q. Hu, and Y. Qian, "Wearable communications in 5G: Challenges and enabling technologies," *IEEE Veh. Technol. Mag.*, vol. 13, no. 3, pp. 100–109, Sep. 2018.

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

[44] C. Rotsos et al., "Network service orchestration standardization: A technology survey," *Comput. Stand. Interfaces*, vol. 54, pp. 203–215, Nov. 2017.

[45] "ONAP: Open networking automation platform." Accessed: May 20, 2020. [Online]. Available: https://www.onap.org/

[46] OpenBaton. "OpenBaton: NFV MANO-based framework." Accessed: May 25, 2020. [Online]. Available: https://openbaton.github.io/

[47] K. Katsalis, N. Nikaein, and A. Huang, "JOX: An event-driven orchestrator for 5G network slicing," in *Proc. IEEE/IFIP Netw. Operat. Manag. Symp.*, 2018, pp. 1–9.

[48] (Cloudify, Herzliya, Israel). *Cloudify: A Open Source Network Orchestrator*. Accessed: Jun. 3, 2020. [Online]. Available: https://cloudify.co/

[49] K. Abbas, T. A. Khan, M. Afaq, and W.-C. Song, "Ensemble learning-based network data analytics for network slice orchestration and management: An intent-based networking mechanism," in *Proc. IEEE/IFIP Netw. Operat. Manag. Symp.*, 2022, pp. 1–5.

[50] "Open source mano." Accessed: Jun. 1, 2020. [Online]. Available: https://osm-download.etsi.org/ftp/Documentation/201902-osm-scope-white-paper/#!02-osm-scope-and-functionality.md

[51] M. Liyanage et al., "A survey on zero touch network and service management (ZSM) for 5G and beyond networks," *J. Netw. Comput. Appl.*, vol. 203, Jul. 2022, Art. no. 103362.

[52] C. Benzaid and T. Taleb, "AI-driven zero touch network and service management in 5G and beyond: Challenges and research directions," *IEEE Netw.*, vol. 34, no. 2, pp. 186–194, Mar./Apr. 2020.

[53] S. Zhang and D. Zhu, "Towards artificial intelligence enabled 6G: State of the art, challenges, and opportunities," *Comput. Netw.*, vol. 183, Dec. 2020, Art. no. 107556.

[54] M. K. Shehzad, L. Rose, M. M. Butt, I. Z. Kovács, M. Assaad, and M. Guizani, "Artificial intelligence for 6G networks: Technology advancement and standardization," *IEEE Veh. Technol. Mag.*, vol. 17, no. 3, pp. 16–25, Sep. 2022.

[55] J. Wang, J. Liu, J. Li, and N. Kato, "Artificial intelligence-assisted network slicing: Network assurance and service provisioning in 6G," *IEEE Veh. Technol. Mag.*, vol. 18, no. 1, pp. 49–58, Mar. 2023.

[56] H. Yang, A. Alphones, Z. Xiong, D. Niyato, J. Zhao, and K. Wu, "Artificial-intelligence-enabled intelligent 6G networks," *IEEE Netw.*, vol. 34, no. 6, pp. 272–280, Nov./Dec. 2020.

[57] K. Abbas, J. Hong, N. Van Tu, J.-H. Yoo, and J. W.-K. Hong, "Autonomous DRL-based energy efficient VM consolidation for cloud data centers," *Phys. Commun.*, vol. 55, Dec. 2022, Art. no. 101925.

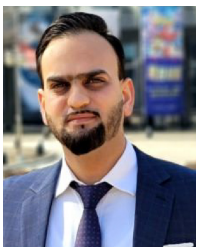
[58] W. Hammedi, B. Brik, and S. M. Senouci, "Federated deep learning-based framework to avoid collisions between inland ships," in *Proc. Int. Wireless Commun. Mobile Comput. (IWCMC)*, 2022, pp. 967–972.

[59] A. Huang and N. Nikaein, "LL-MEC a SDN-based MEC platform," in *Proc. 23rd Annu. Int. Conf. Mobile Comput. Netw.*, 2017, pp. 483–485.

[60] X. Foukas, N. Nikaein, M. M. Kassem, M. K. Marina, and K. Kontovasilis, "FlexRAN: A flexible and programmable platform for software-defined radio access networks," in *Proc. 12th Int. Conf. Emerg. Netw. EXp. Technol.*, 2016, pp. 427–441.
[61] (OpenAirInterface, Alpes-Cote d'Azur, France). "OAI: An open-source community." Accessed: May 10, 2020. [Online]. Available: https://www.openairinterface.org/

**KHIZAR ABBAS** received the B.S. degree in software engineering from the Government College University Faisalabad (GCUF), Pakistan, in 2014, the M.S. degree in computer science from the University of Agriculture Faisalabad, Pakistan, in 2017, and the Ph.D. degree in computer engineering from Jeju National University, South Korea, in 2022. He is currently a Postdoctoral Researcher with the System Security Lab, Department of Computer Science, Hanyang University, Seoul, South Korea. Before Joining Hanyang University, he worked as a Postdoctoral Researcher with the Distributed Processing and Network Management Laboratory, Department of Computer Science and Engineering, POSTECH, South Korea. Prior to this, he worked as a Visiting Lecturer with the Department of Computer Science, GCUF. His research interests include software-defined networks, beyond 5G, network slicing, network function virtualization, mobile-edge computing, network orchestration and management, network security, blockchain for networks, artificial intelligence for B5G networks, machine learning, and reinforcement learning.

**YEONGPIL CHO** received the B.S. degree in electrical engineering from POSTECH, South Korea, in 2010, and the Ph.D. degree in electrical and computer engineering from Seoul National University, South Korea, in 2018. He is currently a Professor with the Department of Computer Science, Hanyang University, Seoul, South Korea. His research interest includes system security against various types of threats.



**ALI NAUMAN** received the M.Sc. degree in wireless communications from the Institute of Space Technology, Islamabad, Pakistan, in 2016, and the Ph.D. degree in information and communication engineering from Yeungnam University, Republic of Korea, in 2022, where he is currently working as an Assistant Professor with the Department of Information and Communication. He has contributed to five patents and authored/coauthored three book chapters and more than 20 technical articles in leading journals and peer-reviewed conferences. The main domain of his research is in the field of artificial intelligence-enabled wireless networks for tactile healthcare, multimedia, and industry 5.0. The research interest also includes resource allocation for 5G and beyond-5G networks, device-to-device communication, Internet of Everything, URLLC, tactile Internet, and artificial intelligence.

**PRINCE WAQAS KHAN** received the master's degree in computer science from the University of Agriculture, Faisalabad, Pakistan, in 2017, and the Ph.D. degree from the Machine Learning Laboratory, Department of Computer Engineering, Jeju National University, Jeju, South Korea, in 2023. He is a Postdoctoral Researcher with the Department of Industrial and Management Systems Engineering, West Virginia University, Morgantown, WV, USA. He worked as an Assistant Professor with the School of Computing, Gachon University, Seongnam, South Korea. Before this, he worked as a Lecturer with the Department of Computer Science, University of Agriculture. He has also gained research experience as a Researcher with the Chongqing Key Laboratory of Cyberspace and Information Security, Chongqing University of Posts and Telecommunications, Chongqing, China. His research interests include artificial intelligence, machine learning, image processing, blockchain, and Internet of Things.

**TALHA AHMED KHAN** received the B.S. degree in computer science from the FAST National University of Computer and Emerging Sciences (FAST NUCES), Pakistan, and the M.S. and Ph.D. in computer engineering from Jeju National University, South Korea, in 2019 and 2023, respectively. He is a Research Fellow with the Institute for Communication Systems, University of Surrey, Guildford, U.K. He worked as an Assistant Professor with the Computer Science Department, Air University, Islamabad, Pakistan. His research interests include SDN, NFV, 5G mobile networks, intent-based networking, network orchestration, mobile edge computing, and VNF development.

**KOTESWARARAO KONDEPU** received the Ph.D. degree in computer science and engineering from the Institute for Advanced Studies Lucca, Italy, in 2012. He is currently an Assistant Professor with the Computer Science and Engineering Department, IIT Dharwad, India. His research interests include 5G, softwarization and virtualization of mobile networks, optical networks design, wired-wireless access convergence, communication networks reliability, energy efficient schemes in communication networks, and sparse senor networks.