DOCTORAL THESIS

Admission Control Optimisation for QoS and QoE Enhancement in Future Networks

Author:

Abida Perveen

A thesis submitted in fulfilment of the requirements for the degree of Doctor of Philosophy

to the

School of Computing and Digital Technology

BIRMINGHAM CITY UNIVERSITY

February 2, 2022

Declaration of Authorship

I, Abida Perveen, declare that this thesis titled, “Admission Control Optimisation for QoS and QoE Enhancement in Future Networks” and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: Abida Perveen

Date: 02-02-2022

Abstract

Recent exponential growth in demand for traffic heterogeneity support and the number of associated devices has considerably increased demand for network resources and induced numerous challenges for the networks, such as bottleneck congestion, and inefficient admission control and resource allocation. Challenges such as these degrade network Quality of Service (QoS) and user-perceived Quality of Experience (QoE). This work studies admission control from various perspectives. For example, two novel single-objective optimisation-based admission control models, *Dynamica Slice Allocation and Admission Control* (DSAAC) and *Signalling and Admission Control* (SAC), are presented to enhance future limited-capacity network Grade of Service (GoS), and for control signalling optimisation, respectively. DSAAC is an integrated model whereby a cost-estimation function based on user demand and network capacity quantifies resource allocation among users. Moreover, to maximise resource utility, adjustable minimum and maximum slice resource bounds have also been derived. In the case of user blocking from the primary slice due to congestion or resource scarcity, a set of optimisation algorithms on inter-slice admission control and resource allocation and adaptability of slice elasticity have been proposed.

A novel SAC model uses an unsupervised learning technique (i.e. Ranking-based clustering) for optimal clustering based on users' homogeneous demand characteristics to minimise signalling redundancy in the access network. The redundant signalling reduction reduces the additional burden on the network in terms of unnecessary resource utilisation and computational time. Moreover, dynamically reconfigurable QoE-based slice performance bounds are also derived in the SAC model from multiple demand characteristics for clustered user admission to the optimal network. A set of optimisation algorithms are also proposed to attain efficient slice allocation and users' QoE enhancement via assessing the capability of slice QoE elasticity. An enhancement of the SAC model is proposed through a novel multi-objective optimisation model named *Edge Redundancy Minimisation and Admission Control* (E-RMAC). A novel E-RMAC model for the first time considers the issue of redundant signalling between the edge and core networks. This model minimises redundant signalling using two classical unsupervised learning algorithms, K-mean- and Ranking-based clustering, and maximises the efficiency of the link (bandwidth resources) between the edge and core networks.

For multi-operator environments such as Open-RAN, a novel *Forecasting and Admission Control* (FAC) model for tenant-aware network selection and configuration is proposed. The model features a dynamic demand-estimation scheme embedded with fuzzy-logic-based optimisation for optimal network selection and admission control. FAC for the first time considers the coexistence of the various heterogeneous cellular technologies (2G, 3G,4G, and 5G) and their integration to enhance overall network throughput by efficient resource allocation and utilisation within a multi-operator environment. A QoS/QoE-based service monitoring feature is also presented to update the demand estimates with the support of a forecasting modifier.

The provided service monitoring feature helps resource allocation to tenants, approximately closer to the actual demand of the tenants, to improve tenant-acquired QoE and overall network performance. Foremost, a novel and dynamic admission control model named *Slice Congestion and Admission Control* (SCAC) is also presented in this thesis. SCAC employs machine learning (i.e. unsupervised, reinforcement, and transfer learning) and multi-objective optimisation techniques (i.e. Non-dominated Sorting Genetic Algorithm II) to minimise bottleneck and intra-slice congestion. Knowledge transfer among requests in form of coefficients has been employed for the first time for optimal slice requests queuing. A unified cost-estimation function is also derived in this model for slice selection to ensure fairness among slice request admission. In view of instantaneous network circumstances and load, a reinforcement learning-based admission control policy is established for taking appropriate action on guaranteed soft and best-effort slice requests admissions. Intra-slice, as well as inter-slice resource allocation, along with the adaptability of slice elasticity, are also proposed for maximising slice acceptance ratio and resource utilisation.

Extensive simulation results are obtained and compared with similar models found in the literature. The proposed E-RMAC model is 35% superior at reducing redundant signalling between the edge and core networks compared to recent work. The E-RMAC model reduces the complexity from $\mathcal{O}(U)$ to $\mathcal{O}(R)$ for service signalling and $\mathcal{O}(N)$ for resource signalling. This represents a significant saving in the uplink control plane signalling and link capacity compared to the results found in the existing literature. Similarly, the SCAC model reduces bottleneck congestion by approximately 56% over the entire load compared to ground truth and increases the slice acceptance ratio. Inter-slice admission and resource allocation offer admission gain of 25% and 51% over cooperative slice- and intra-slice-based admission control and resource allocation, respectively. Detailed analysis of the results obtained suggests that the proposed models can efficiently manage future heterogeneous traffic flow in terms of enhanced throughput, maximum network resources utilisation, better admission gain, and congestion control.

List of Publications

J01: **Abida Perveen**, Raouf Abozariba, Mohammad Patwary, and Adel Aneiba, "Dynamic traffic forecasting and fuzzy-based optimised admission control in federated 5G-open RAN networks", 2021 Springer Journal of Neural Computing and Applications. [*Published*]

J02: **Abida Perveen**, Raouf Abozariba, Mohammad Patwary, and Adel Aneiba,"SCAC: Machine Learning based Slice Congestion and Admission Control in future Networks", 2022 IEEE Journal of Wireless Communications and Mobile Computing. [*In process to submit*]

C01: **Abida Perveen**, Mohammad Patwary, and Adel Aneiba, "Dynamically reconfigurable slice allocation and admission control within 5G wireless networks", 2019 IEEE 89th Vehicular Technology Conference. [*Published*]

C02: **Abida Perveen**, Mohammad Patwary, and Adel Aneiba, "End-use Aware Optimised Control Signalling for User Admission within 5G and Beyond Networks", 2020 IEEE International Conference on Communications. [*Published*]

C03: **Abida Perveen**, Raouf Abozariba, Mohammad Patwary, Adel Aneiba, and Anish Jindal, "Clustering-based Redundancy Minimisation for Edge Computing in Future Core Networks", 2021 IEEE 5G World Forum. [*Published*]

Acknowledgements

Firstly, I like to say special thanks to my Director of Studies "Dr Raouf Abozariba", second supervisor "Dr Adel Aneiba" and external supervisor "Professor Mohammad Patwary" for supporting me in this research with ideas and guidance. I have immensely benefited from their knowledge.

I am greatly thankful to my family and friends for their support and encouragement throughout my studies and career. Without their love and care, I would not have been able to complete this degree.

I like to say special thanks to my department “Computer Engineering and Build Environment”- Birmingham City University for their support and management in completing this research on time.

Finally, I would like to say thanks to the examiners for spending their time in reviewing this thesis.

Contents

List of Figures

List of Algorithms

List of Tables

List of Abbreviations

1G	**1**st **G**eneration
5G	**5**th **G**eneration
AI	**A**rtificial **I**ntelligence
CP	**C**ontrol **P**lane
CU	**C**entralised Control **U**nit
DN	**D**ata **N**etwork
DP	**D**ata **P**lane
DU	**D**istributed **U**nit
ML	**M**achine **L**earning
NP	**N**on-deterministic **P**olynomial-time
RL	**R**einforcement **L**earning
RU	**R**adio **U**nit
UE	**U**ser **E**quipment
AMF	**A**ccess & **M**obility **M**anagement **F**unction
DCN	**D**edicated **C**ore **N**etworks
E2E	**E**nd-to-**E**nd
eNB	**e**nhanced **N**ode **B**
EPC	**E**volved **P**acket **C**ore
GoS	**G**rade-**o**f-**S**ervice
HSS	**H**ome **S**ubscriber **S**erver
LTE	**Long**-**T**erm **E**volution
MAC	**M**edia **A**ccess **C**ontrol
MEC	**M**obile **E**dge **C**omputing
MME	**M**obility **M**anagement **E**ntity
MNO	**M**obile **N**etwork **O**perator

NFs	**N**etwork **F**unctions
NFV	**N**etwork **F**unction **V**irtualization
PCF	**P**olicy **C**ontrol **F**unction
PGW	**P**acket **G**ateway
QoE	**Q**uality **o**f **E**xperience
QoS	**Q**uality **o**f **S**ervice
RAT	**R**adio **A**ccess **T**echnology
RRU	**R**adio **R**esource **U**nit
SDN	**S**oftware **D**efined **N**etworking
SGW	**S**erving **G**ateway
SMF	**S**ession **M**anagement **F**unction
UDM	**U**nified **D**ata **M**anagement
UPF	**U**ser **P**lane **F**unction
3GPP	**T**hird-**G**eneration **P**artnership **P**roject
eMBB	**e**nhanced **M**obile **B**roadband
mMTC	**m**assive **M**achine **T**ype **C**ommunications
MANO	**M**anagement and **O**rchestration
MVNO	**M**obile **V**irtual **N**etwork **O**perator
NGMN	**N**ext-**G**eneration **M**obile **N**etworks
NSSF	**N**etwork **S**lice **S**election **F**unction
O-RAN	**O**pen **R**adio **A**ccess **N**etwork
URLLC	**U**ltra-**R**eliable **L**ow-**L**atency **C**ommunications

This thesis is dedicated to Prof. Mohammad Patwary, my parents, and my late sister Almas Shaheen.

Chapter 1

Introduction

1.1 Background Study

From the innovation of the first-generation (1G) cellular communication system in 1979 to the current, fifth-generation (5G) system, networks have been transformed from pure voice communication to Multimedia- and Artificial Intelligence- (AI) enabled smart communication (Gupta and Jha, 2015). Due to this continuous technological advancement, a wide range of innovative products and services from various sectors, such as the Internet of Thing (IoT), smart homes, industrial automation, healthcare, education and transportation, have noticeably improved our lifestyle, society, business and industrial operations. However, these emerging services are characterised by an extremely diverse set of requirements. For example, enhanced Mobile Broadband (eMBB) services require higher data rate and bandwidth, Ultra-Reliable Low-Latency Communications (URLLC) services require minimum latency and higher reliability, and massive Machine-Type Communications (mMTC) services have higher connectivity demand. On the other side, supporting these emerging services from the existing network is challenging due to inadequate traffic management and resource scarcity (Foukas et al., 2017). The next-generation wireless networks (5G and beyond) promise to meet diverse demand for services by incorporating the concepts of advanced traffic and resource management. Therefore, a call to propose new and novel research efforts in 5G and beyond network design and operational approaches for efficient admission control has activated the research community around the globe (ZHANG, TAO, and ZHANG, 2016; Panwar, Sharma, and Singh, 2016).

The rapid growth of technologies and emerging industry demands have considerably increased network complexity by stretching network resources and their virtualization (Gupta and Jha, 2015; Mudassir et al., 2019). Due to the increase in network complexity, the third-generation partnership project (3GPP) has proposed a novel and flexible architecture called "Network Slicing" (3GPP, 2018c). Network slicing is an end-to-end (E2E) architecture that acts as the key enabler for the new business model, supporting heterogeneous applications, and is a vital feature of the next generation. In this architecture, a single physical network is logically split into multiple virtual networks (network slices) by a dedicated or shared E2E set of logical network functions (network instances). Similarly, network slicing accommodates diverse use cases through unified physical infrastructure and shared network resources. Based on communication needs, 3GPP defined three main use cases: eMBB (enhanced Mobile Broadband), URLLC (Ultra-Reliable Low-Latency Communications), and mMTC (massive Machine-Type Communications). The demand for eMBB may require high data rates, such as required for watching HD video, or low data rates, such as required for sending a text message on WhatsApp. For URLLC, the demand is for a highly reliable connection with very low latency, such as required for self-driving cars. In the case of a massive IoT network, a huge number of devices need network connectivity for communication.

For network virtualization, network slicing uses the concept of Software Defined Networking (SDN), whereby the control and data planes are separated from each other. Isolation of the network planes reduces complexity and makes the network more manageable due to the ease of customisation of the virtual network instances. The virtual network instances are customised to balance the device's emerging heterogeneous demand with available network resources and to achieve maximum utility (Alliance, 2016; Kaloxylos, 2018; Barakabitze et al., 2020). The most common advantages of SDN are the implementation of network automation, agility, programmability and policy-driven network supervision. Therefore, SDN, along with Network Function Virtualization (NFV), is assumed to be a promising solution in the provisioning of network slicing features in 5G and beyond networks (Ordonez-Lucena et al., 2017; Khan et al., 2020a).

However, with the various enabling technologies available to future networks, efficient admission control is still an open issue that ensures the provisioning of better user Quality of Experience (QoE) and network Quality of Service (QoS) demand in an integrated environment, particularly when the load on the network changes frequently due to huge demand for connectivity, high-speed mobile users, and exponentially increasing diverse user-device specific requirements (Su et al., 2019). The design and implementation of the admission control algorithm relies on specific strategies chosen by the network operator (or provider). Such strategies explain how network instances would be customised to meet the incoming demand to achieve the defined goals of the network; for example, revenue maximisation, congestion control, and enhanced network performance, as well as user satisfaction with the network. The emphasis of the thesis is to investigate existing admission control strategies and to propose a novel model that utilises advanced Machine Learning (ML) techniques, such as unsupervised learning, reinforcement learning, and transfer learning, to efficiently support future use cases and acquire defined goals from the network.

1.2 Research Problem

Due to the rapid advancement in wireless technologies and the support of various simultaneous heterogeneous services, traffic demand is growing exponentially. Such demand results in inefficient user admission to networks and inefficient resource utilisation. This, in turn, makes network and resource management a more challenging task for network and service providers, as they are bound to the provisioning of guaranteed bandwidth and data rate by an agreed service level agreement (SLA). Considering the growing complexity of networks, previous network architectures (4G/LTE, Wi-Fi, 3G, 2G and 1G) are more suitable for human and data communication than for device communication and future heterogeneous traffic support (Moradi et al., 2018; Vashi et al., 2017; Ojijo and Falowo, 2020). Therefore, previous cellular networks face three major challenges: (1) lack of customisation and programmability in the core and data planes, (2) users dropping from the network due to congestion, and (3) large forwarding delays and signalling overheads in the control plane (Gupta and Jha, 2015; Caballero et al., 2018; Khan et al., 2020b).

For example, one of the problems in 4G EPS (Evolved Packet System) is the "Always ON" bearer for signalling purposes. This is a default bearer established at the time of user registration. Once a user is registered with the requested slice, this bearer will be established for signalling. Due to signalling overheads, this default bearer is the cause of an additional burden on the core network and resources (Trivisonno et al., 2018). For specific service provisioning, 3GPP proposed Dedicated Core networks (DECOR) and enhanced Dedicated Core architecture (eDECOR) for 4G/LTE in Releases 13 and 14 (3GPP, 2014; 3GPP, 2016), a service-oriented architecture that assigns dedicated core networks with specific characteristics to specific users/devices.

One problem with DECOR is re-routing, whereby User Equipment (UE) is not permitted to select the Mobility Management Entity (MME) of 4G/LTE. UE first attaches to the default MME. Then, based on the attaching request, the default MME redirects the user to the requested MME. This creates unnecessary signalling from eNodeB (eNB) to the default MME and Home Subscriber Server (HSS), and then back to the default MME and eNB to redirect signals to the target MME. In the case of a massive number of demands for connectivity, huge signalling overheads in the network occur that result in additional delays in communication and more users dropping due to congestion on the network, which in turn reduces the Grade-of-Service (GoS) of the serving network. Given this problem, UE assistance information was introduced for requested Dedicated Core Networks (DCN) selection in eDECOR. Moreover, the 4G core network also suffers from inefficient allocation of resources, core burden due to the massive amount of signalling, and complex control plane protocol stack, such as Long Term Evolution (LTE) protocol. This is due to a lack of proper customisation and programmability of the network functions on the core and data planes, as well as inefficient network policies (Gohil, Modi, and Patel, 2013; Kaloxylos, 2018; Agiwal et al., 2021). Therefore, there is a need for fine-grained customisation of network functions for access and the core network to enhance network performance and user experience. In release 15, 3GPP proposed the novel concept of network slicing with more scope for customisation and programmability of network functions (3GPP, 2018c). Although network slicing is reliant on virtualization for the provisioning of parallel services to the industrial market, virtualization itself relies on the availability of physical resources, whereby few physical resources are borrowed to implement partial or full virtualization. Initially, the researcher's focus was on only the virtualization of core network resources. However, for the provisioning of E2E heterogeneous use cases, slicing and virtualization of RAN resources are also essential. RAN and core network resources both need to be efficiently sliced into distinct instances to serve a variety of use cases in future networks. Therefore, a significant amount of research into the virtualization of access and core network instances is still required to fulfil the promises of 5G and beyond networks. Such promises include the coexistence of various heterogeneous technologies and their integration, ultra-low latency, high throughput and greater reliability. The major challenges in keeping these promises arise through rapidly changing user demand and service-specific characteristics that can lead to inappropriate admission control and resource allocation due to poor management policies in the network. Such as no policies are available on signalling redundancy reduction in the access and control plane of the 4G EPS network (Trivisonno et al., 2018). The frequently changing user demand and service-specific characteristics need to be investigated in-depth, especially for the support of the multi-tenant environment, where there is a need for dynamic customisation of network functions and resources (Le et al., 2016; Kaloxylos, 2018; Ojijo and Falowo, 2020). Therefore, the major contributions presented in this thesis are to propose optimised admission control using ML techniques, such as unsupervised learning, reinforcement learning and transfer learning, that accommodate future demand along with enhanced network QoS and user QoE.

1.3 Research Questions

Given the problem statement, the research questions are as follows:

- How to enhance Grade-of-Service (GoS) in future networks with limited or without scaling up network capacity to support a massive demand for heterogeneous traffic?

- Can minimising signalling redundancy in access and core networks help admission and resource utilisation without degrading network QoS and user-desired QoE demand within a dense environment?
- Can the coexistence of the various heterogeneous cellular technologies (2G, 3G,4G, and 5G) and their integration help to enhance overall network throughput through efficient resource allocation and utilisation within a multi-operator environment?
- How effective are ML techniques (unsupervised learning, reinforcement learning and transfer learning) and optimisation in bottleneck (as well as inter-slice) congestion and admission control in 5G and beyond networks to support mMTC and eMBB traffic demand?

1.4 Research Scope

The scope of this research is to provide appropriate answers to the above-mentioned research questions. While doing so, the crucial factors that need to be considered, as shown in Figure 1.1, are as follows:
- To cope with a diverse set of user and network requirements, throughput maximisation is a primary consideration in wireless communication systems. Throughput determines the effectiveness of the deployed network. 5G promises to provide end-to-end latency of less than 1 ms (Gupta and Jha, 2015; Ojijo and Falowo, 2020). Therefore, the major emphasis in a network is on guaranteed throughput that should not diminish while providing this latency. In a network, throughput is measured by achieved link efficiency and latency, whereby greater link efficiency at lower latency results in better throughput and less congestion. This, in turn, also improves the admission efficiency (i.e. GoS) in a network (Parvez et al., 2018). Therefore, one of the scopes of this thesis is to investigate the impact of massive heterogeneous demand for connectivity during admission control within a multi-operator- and multi-technology-enabled environment to improve network throughput by designing efficient admission control and resource allocation algorithms.

- An increase in demand for device connectivity helps the network operator to earn more revenue from maximum network resource utilisation in the deployed network. However, to efficiently accommodate the massive number of heterogeneous demands is not a simple task for the network operator due to continuously increasing network complexity. Various approaches to efficient resource allocation and utilisation were addressed in (Han et al., 2019; Su et al., 2019). However, greed on the part of the operator can see the network saturated with excessive resource allocation to users. In this case, network QoS and user QoE would be impacted due to inefficient resource utilisation (Sciancalepore et al., 2017; Jiang et al., 2016b). In this thesis, the scope is to propose an optimal network and resource utilisation model that also takes network QoS and user-perceived QoE into consideration during admission control and resource allocation.

- In a dense environment, enhancing admission gain and maintaining user satisfaction is challenging for the network providers due to limited network capacity. A solution for increasing network capacity through temporary resource scaling and sharing has been addressed in the literature (Gavrilovska, Rakovic, and Denkovski, 2018; Gutierrez-Estevez et al., 2018; Ojijo and Falowo, 2020). To increase admission gain and user satisfaction, resource scaling and sharing between providers often results in significant improvement in GoS. However, this approach is costly for the network providers, as they often have a limited budget to operate, and network revenue would be affected by making decisions on resource scaling and sharing. To minimise the cost and improve the revenue of the network, the network providers need to make an on-demand and precise decision about dynamically borrowing additional resources from the neighbouring network providers (Gutierrez-Estevez et al., 2019). Therefore, one of the plans in this research is to propose a model that optimises resource sharing and utilises the existing resources of various heterogeneous technologies in a way that improves admission gain (i.e. GoS) and user satisfaction in future networks.

- The increasing demand for services from future use-cases is starting to drive new control signalling traffic, primarily due to a rapid increase in the number of devices, both individuals and machines. This trend will have a significant impact on network performance in terms of increased network complexity and congestion, due to control signalling storms in the access and core network (Al-Fuqaha et al., 2015; Trivisonno et al., 2018). The scope is to propose a signalling redundancy minimisation model that reduces congestion within the access and core network to improve admission control and resource management within dense wireless networks.

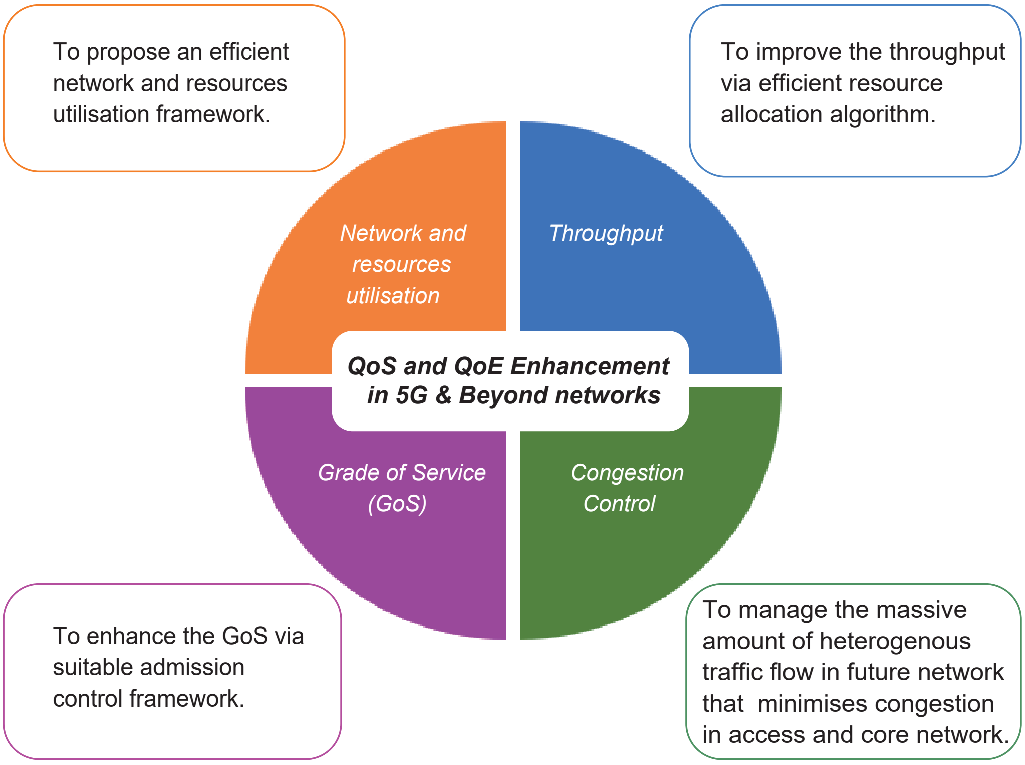

To propose an efficient network and resources utilisation framework.
To improve the throughput via efficient resource allocation algorithm.
Network and resources utilisation
Throughput
QoS and QoE Enhancement in 5G & Beyond networks
Grade of Service (GoS)
Congestion Control
To enhance the GoS via suitable admission control framework.
To manage the massive amount of heterogenous traffic flow in future network that minimises congestion in access and core network.

FIGURE 1.1: Research questions, scope and aim

1.5 Research Aim and Objectives

The aim of this research is to investigate the use of modern optimisation and machine learning (ML) techniques for dynamically admitting the heterogeneous traffic flow in future networks (i.e., 5G and beyond) to enhance overall user experience (QoE) and network performance (QoS) in terms of increased throughput, maximum network and resource utilisation, better admission gain and user satisfaction level, and congestion control, as shown in Figure 1.1. Accordingly, the research objectives to be achieved given the above-mentioned scope are itemised below:

1. To review and investigate the state of the art in admission control of heterogeneous traffic flow in future networks (i.e., 5G and beyond).
2. To identify and analyse the most relevant admission control strategies in the 5G and beyond networks to propose enhancements in future network management.
3. To design and develop novel admission control models for single objectives using optimisation and unsupervised machine learning techniques.
4. To design and develop novel admission control models for multiple objectives using optimisation, reinforcement learning, and transfer learning techniques.
5. To evaluate the robustness of the proposed models by comparing them with the existing models found in the literature.

1.6 Research Methodology

The purpose of the proposed research is to allow a massive amount of traffic flow in future wireless networks with the assurance of heterogeneous service provisioning and better QoE. Given the aim and above-mentioned objectives, the research onion approach is considered in this thesis to demonstrate the overall research methodology, as shown in Fig 1.2. Accordingly, the presented methodology has the following phases:

- **Phase I**: Research Philosophy:-Recent research study and analysis
- **Phase II**: Research Approach:-Modelling and network design specification
- **Phase III**: Research Techniques and Procedures:- Optimisation and machine learning techniques
- **Phase IV**: Research Strategy and Choice:-Simulation setup, results, and analysis
- **Phase V**: Research Time Horizon:-Thesis write-up, publications, and presentation in pursuit of a PhD

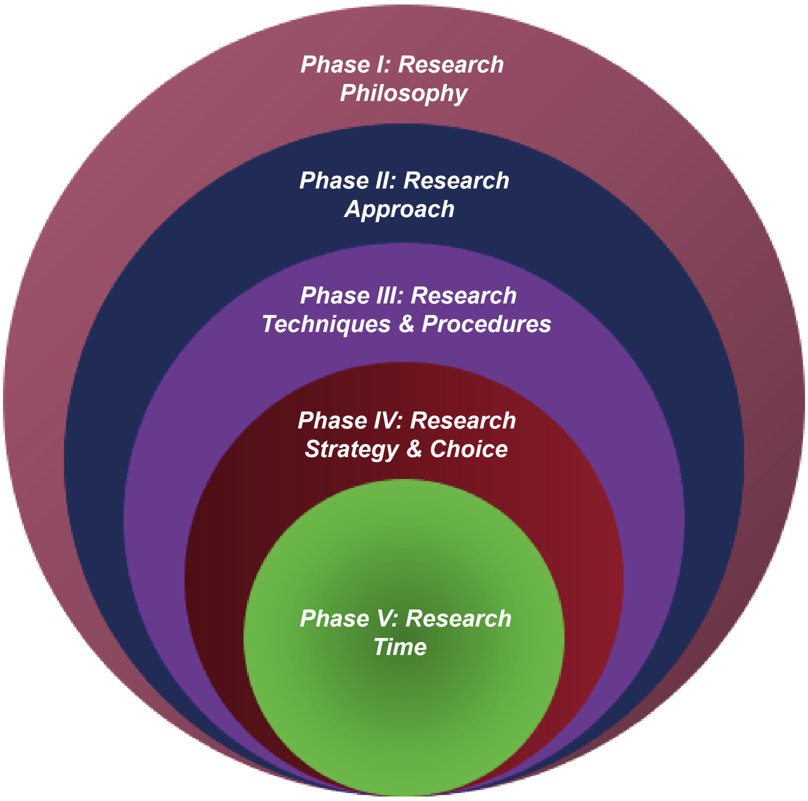

Phase I: Research Philosophy
Phase II: Research Approach
Phase III: Research Techniques & Procedures
Phase IV: Research Strategy & Choice
Phase V: Research Time

FIGURE 1.2: Research methodology

Phase I of the methodology begins with a background study to formulate the research theme, problem statement and research questions. The research aim and objectives are defined on the given problem statement and research questions. A detailed survey of related literature is then conducted. Moreover, network and resource requirements are analysed for modelling. After surveying various relevant research works, and their models, analysing results and their evaluation, improved theoretical system models are proposed, along with the presentation of the network design on various scenarios in phase II. Under mathematical problem modelling and numerical analysis, a set of equations and algorithms are established by using optimisation and machine learning techniques (unsupervised learning, reinforcement learning and transfer learning) in phase III of this research. In phase IV, using a given set of performance evaluation parameters, a simulation environment is established in MATLAB to obtain the required results for analysis. The required data for the evaluation parameters are acquired from the 5G implementation guideline and 3GPP R15 standard (3GPP, 2018c; GSMA, 2019). The robustness of the proposed model is analysed through comparison with the results of various relevant models in the existing literature. The research contributions derived from the proposed models are disseminated via well-known conferences and journal publications. Finally, on the attainment of suitable solutions to all research questions and objectives, the thesis is written up and presented in pursuit of a PhD in phase V.

1.7 Contribution to Knowledge

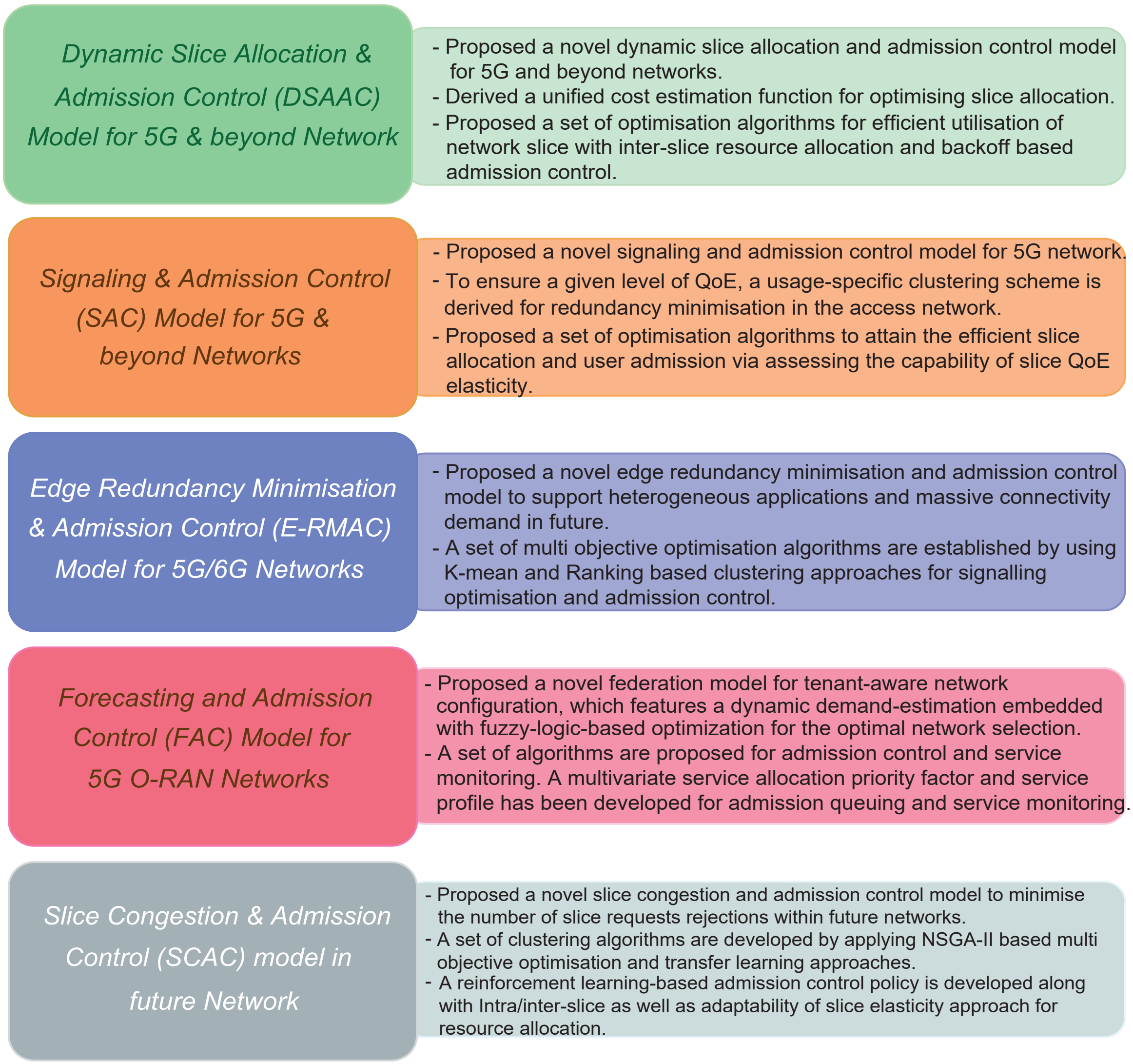
The major contributions of this thesis, as illustrated in Figure 1.3, are summarised as follows:

- To serve heterogeneous traffic demand with efficient resource utilisation and network GoS enhancement, a novel dynamic slice allocation and admission control (DSAAC) model for 5G and beyond networks has been presented in this thesis and disseminated internationally through the publication ***C01*** (Perveen, Patwary, and Aneiba, 2019). In this model, an integrated user-application-specific demand characteristics and network characteristics have been considered for admission control optimisation. Such characteristics include required bandwidth, data rate and priority. Moreover, a cost estimation function has been derived for optimising slice allocation and to quantify resource allocation decision metrics that is valid for both the static and dynamic nature of the user and network characteristics. A set of algorithms for efficient utilisation of network slice with inter-slice resource allocation and back-off-based admission control has also been proposed in DSAAC.

- To accommodate more user-specific traffic data, a novel signalling and admission control (SAC) model using optimisation and clustering techniques has been proposed in this thesis and disseminated internationally through the publication ***C02*** (Perveen, Patwary, and Aneiba, 2020). In this model, it is proposed that pre-clustering end-use analysis, usage-specific clustering, and clustering based on end-use application and device-specific resource demand be exploited. To ensure a given level of QoE, a usage-specific clustering scheme has been derived for redundancy minimisation in the access network. In doing so, dynamically reconfigurable QoE-based slice performance bounds have also been considered for user admission to the network. Moreover, a set of algorithms to attain efficient slice allocation and resource utilisation via assessing the capability of slice QoE elasticity has also been established.

- A novel edge redundancy minimisation and admission control (E-RMAC) model has been proposed in this thesis to support future networks in terms of minimising the aforementioned control signalling redundancy and congestion. This research work is disseminated internationally through the publication ***C03*** (Perveen et al., 2021a). E-RMAC model is the enhancement of the SAC model. In this model, two popular unsupervised learning techniques (K-mean- and Ranking-based clustering) and multi-objective optimisation (Non-dominated sorting genetic algorithm II) are employed to reduce core network signalling redundancy. By using these techniques, cluster-based signal and admission control algorithms have been established to maximise link efficiency between the edge and core networks.
- A novel dynamic traffic forecasting and admission control (FAC) model for a federated O-RAN environment has been proposed in this thesis and published in ***J01*** (Perveen et al., 2021b). This model predicts future traffic demand for efficient admission control and resource allocation. In this model, a fully reconfigurable admission control model using fuzzy-logic optimisation by drawing on the information on user demand and network capacity has been proposed for the optimal network selection. After optimal network selection, a set of algorithms are also proposed for admission control and service monitoring. Whereby, a multivariate service allocation priority factor has been developed for admission queuing. Moreover, a service profile has been built on admissions for service monitoring to ensure efficient resource utilisation and better user-perceived QoE.
- A novel slice congestion and admission control (SCAC) model has been presented to minimise the number of slice requests rejection that occurs due to bottlenecks and intra-slice congestion in the network. This model consists of a slice demand analysis and classification (SDAC), a demand clustering and

queuing (DCQ), and an admission and resource management (ARM). Two popular unsupervised learning algorithms, Ranking and K-mean clustering algorithms, along with multi-objective optimisation and transfer learning, have been employed for slice request queuing. A unified cost-estimation function is also derived for slice selection to ensure fairness among slice requests. Given instantaneous network circumstances and load, a reinforcement learning-based admission control policy is also established for taking appropriate action on the guaranteed soft and best-effort slice request admissions. Intra slice, as well as inter-slice, resource allocation, along with the adaptability of slice elasticity, are also proposed in this model for maximising slice acceptance ratio and resource utilisation.



Dynamic Slice Allocation & Admission Control (DSAAC) Model for 5G & beyond Network

- Proposed a novel dynamic slice allocation and admission control model for 5G and beyond networks.
- Derived a unified cost estimation function for optimising slice allocation.
- Proposed a set of optimisation algorithms for efficient utilisation of network slice with inter-slice resource allocation and backoff based admission control.

Signaling & Admission Control (SAC) Model for 5G & beyond Networks

- Proposed a novel signaling and admission control model for 5G network.
- To ensure a given level of QoE, a usage-specific clustering scheme is derived for redundancy minimisation in the access network.
- Proposed a set of optimisation algorithms to attain the efficient slice allocation and user admission via assessing the capability of slice QoE elasticity.

Edge Redundancy Minimisation & Admission Control (E-RMAC) Model for 5G/6G Networks

- Proposed a novel edge redundancy minimisation and admission control model to support heterogeneous applications and massive connectivity demand in future.
- A set of multi objective optimisation algorithms are established by using K-mean and Ranking based clustering approaches for signalling optimisation and admission control.

Forecasting and Admission Control (FAC) Model for 5G O-RAN Networks

- Proposed a novel federation model for tenant-aware network configuration, which features a dynamic demand-estimation embedded with fuzzy-logic-based optimization for the optimal network selection.
- A set of algorithms are proposed for admission control and service monitoring. A multivariate service allocation priority factor and service profile has been developed for admission queuing and service monitoring.

Slice Congestion & Admission Control (SCAC) model in future Network

- Proposed a novel slice congestion and admission control model to minimise the number of slice requests rejections within future networks.
- A set of clustering algorithms are developed by applying NSGA-II based multi objective optimisation and transfer learning approaches.
- A reinforcement learning-based admission control policy is developed along with Intra/inter-slice as well as adaptability of slice elasticity approach for resource allocation.

FIGURE 1.3: Contribution to knowledge

1.8 Organisation of Thesis

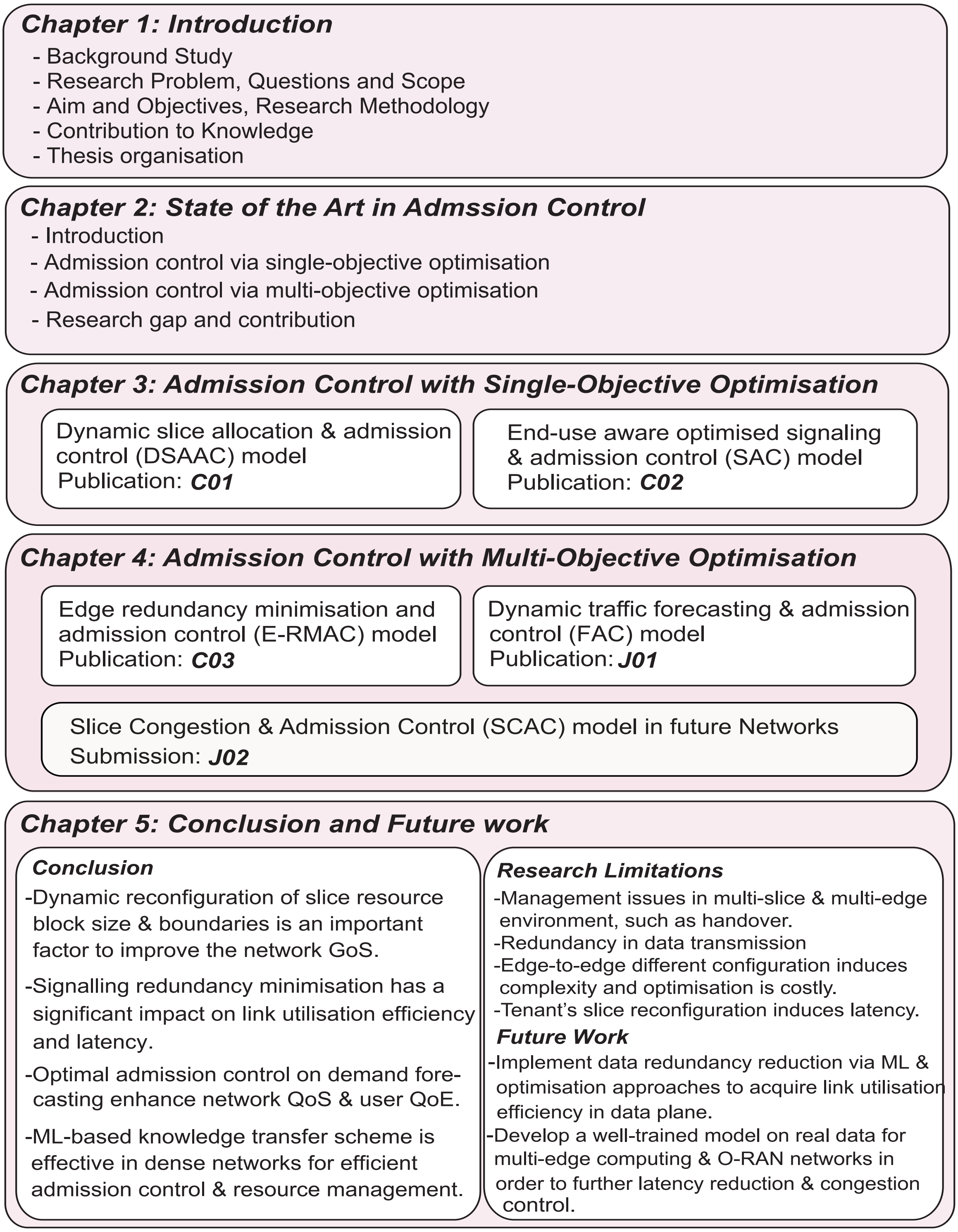
This thesis is organised as follows (see. Figure 1.4):

- Chapter 2 presents a comprehensive survey of the state of the art in admission control for traffic and resource management in wireless networks. The main contributions of the thesis are discussed in detail in Chapters 3 and 4.

- In Chapter 3, two novel single objective optimisation models for slice and user admission in 5G and beyond networks are proposed. Introduction on admission control with single objective optimisation models is described in Section 3.1. In Section 3.2, an integrated user application, as well as a network-specific, characteristics-based dynamic slice allocation and admission control model is proposed for 5G wireless networks. Section 3.3 describes a cluster-based optimised control signalling and admission control model to accommodate more user-specific data traffic. The concluding remarks on the chapter are presented in Section 3.4.

- Chapter 4 extends the work presented in Chapter 3 with the help of multi-objective optimisation, whereby three novel models have been proposed for user admission in edge and federated 5G-OpenRAN networks. Introduction on admission control with multi-objective optimisation is described in Section 4.1. In Section 4.2, a novel NSGA-II algorithm-based signal-clustering and admission-control model is presented for signalling redundancy reduction in future edge networks. In Section 4.3 on dynamic demand forecasting, a fully reconfigurable fuzzy-logic-based admission control model is presented to accommodate high-density traffic demand in an open radio access network. By using multi-objective optimisation (i.e. NSGA-II) and machine learning (i.e. unsupervised learning, reinforcement learning, and transfer learning) techniques, a novel slice congestion and admission control model is presented in Section 4.4. This model is proposed for bottleneck and intra-slice congestion control in 5G and beyond networks. The concluding remarks on the chapter are presented in Section 4.5.

- Finally, I concluded the presented research with a detailed analysis of the achieved research objectives in Chapter 5. Moreover, limitations in the proposed models and possible directions for future research are also presented in detail.



Chapter 1: Introduction

- Background Study
- Research Problem, Questions and Scope
- Aim and Objectives, Research Methodology
- Contribution to Knowledge
- Thesis organisation

Chapter 2: State of the Art in Admssion Control

- Introduction
- Admission control via single-objective optimisation
- Admission control via multi-objective optimisation
- Research gap and contribution

Chapter 3: Admission Control with Single-Objective Optimisation

Dynamic slice allocation & admission control (DSAAC) model
Publication: C01

End-use aware optimised signaling & admission control (SAC) model
Publication: C02

Chapter 4: Admission Control with Multi-Objective Optimisation

Edge redundancy minimisation and admission control (E-RMAC) model
Publication: C03

Dynamic traffic forecasting & admission control (FAC) model
Publication: J01

Slice Congestion & Admission Control (SCAC) model in future Networks
Submission: J02

Chapter 5: Conclusion and Future work

Conclusion

-Dynamic reconfiguration of slice resource block size & boundaries is an important factor to improve the network GoS.

-Signalling redundancy minimisation has a significant impact on link utilisation efficiency and latency.

-Optimal admission control on demand forecasting enhance network QoS & user QoE.

-ML-based knowledge transfer scheme is effective in dense networks for efficient admission control & resource management.

Research Limitations

-Management issues in multi-slice & multi-edge environment, such as handover.
-Redundancy in data transmission
-Edge-to-edge different configuration induces complexity and optimisation is costly.
-Tenant's slice reconfiguration induces latency.

Future Work

-Implement data redundancy reduction via ML & optimisation approaches to acquire link utilisation efficiency in data plane.
-Develop a well-trained model on real data for multi-edge computing & O-RAN networks in order to further latency reduction & congestion control.

FIGURE 1.4: Thesis structure

Chapter 2

State of the Art in Admission Control

2.1 Introduction

5G wireless networks promise to provide a diverse range of services by incorporating advanced network and resource management techniques such as network function virtualisation (NFV), software-defined networking (SDN) and network slicing. These techniques, especially network slicing, provide resource customisation capabilities to balance emerging demand in the network against the available capacity. By next-generation mobile networks (NGMN) alliance (Alliance, 2016), the network slicing concept consists of 3 layers: Resource layer, Network Slice Instance Layer, and Service Instance Layer, as depicted in Figure 2.1. The resource layer holds and manages the resources within a resource pool. This layer shares resources among slices upon request through instances. The network slice instance layer contains the slice instance to establish an end-to-end connection to serve the specific application via sub-network slice instances. The sub-network slice instance contains a set of customised resources from the resource pool, which can be shared among various slice instances. Moreover, a slice can obtain the resources from multiple sub-network slice instances. The sub-network slice instance can also be shared among various active slices for resource allocation on demand. The capability of multi-slice connectivity for heterogeneous service provisioning makes the scenario more complex. The service instance layer contains application instances to serve a particular application from the respective network. Services can be provided by the network operator or by 3rd parties.

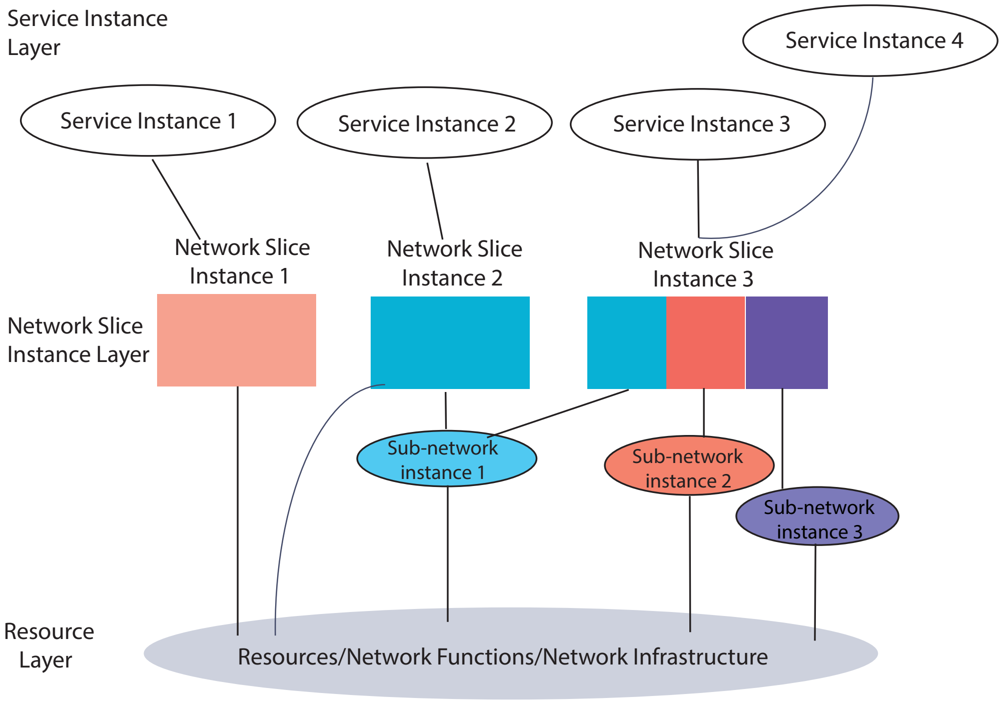


FIGURE 2.1: Network slicing conceptual architecture by NGMN (Alliance, 2016)

In the network, resource allocation and utilisation are influenced by decisions made by admission control algorithms. Admission control is a fundamental approach to achieving specific objectives that ensure efficient network traffic and resource management. These objectives include revenue optimisation, network Quality of Service (QoS) and user Quality of Experience (QoE) control, congestion control, and admission fairness (Han et al., 2018a; Kammoun et al., 2018; Caballero et al., 2018; Sun et al., 2019a; Ojijo and Falowo, 2020). Today, the primary concern of mobile network and service providers is to ensure adequate resource allocation and their maximum utilisation to earn more revenue by improving network performance or QoS, and to satisfy user QoE (Ge and Tan, 2014). The critical challenge encountered by the mobile network or service providers is to balance the emerging user demand against available resources in such a way that it does not create congestion. Accordingly, coping with emerging user demand in proportion to available resources and achieving their maximum utilisation are a guaranteed attraction for researchers today. Enormous advances in virtualisation technologies divert the attention of researchers and the industry to virtually maximise network capacity in proportion to demand to serve more users with their desired QoE and earn revenue (Jiang et al., 2016b; Pratap and Das, 2021). Accordingly, a comprehensive survey of the state of the art in admission control for traffic and resource management in wireless networks is presented in detail in the following sections.

2.2 Admission Control Strategies

Along with technological advancements, admission control becomes difficult in unforeseen scenarios, due to growing network complexity. On another side, having better network QoS is also essential to assure users of the provisioning of their required QoE. Therefore, efficient admission control in future networks that ensures better resource management without degrading user QoE and network QoS demand is still an open issue (Su et al., 2019). The design and implementation of an admission control algorithm relies on a specific strategy chosen by the network operator to achieve the above-mentioned objectives. The admission control approach can be simple with single objective optimisation (e.g. priority-based or first-come-first-served, random or greedy with single variate optimisation), or complex with multi-objective optimisation and ML-enabled approaches (Ojijo and Falowo, 2020), as discussed in detail in the following subsections.

2.2.1 Admission Control via Single-Objective Optimisation

The simplest strategy to consider the incoming request is the conventional approach with single objective optimisation for admission control. In these approaches, the strategy is to serve the requests as they arrive (first-come-first-served), or to admit the requests in order of priority, or to adopt a greedy approach to maximise revenue, or random approaches to ensure fairness among users during admission, as discussed below in detail.

The first-come-first-served or first-in-first-out (FIFO) strategy is the simplest approach to user admission to the network and ignores non-trivial constraints such as latency and bandwidth demand (Ojijo and Falowo, 2020). For example, the authors Han et al. proposed a utility-based admission control model for network slicing in 5G (Han et al., 2019). To serve the requests of multiple queues, first come first served (FCFS), last come first served (LCFS), random selection for service (RSS), and priority-based (PR) schemes have been investigated by the authors in their work. Similarly, Anand et al. proposed a signal-to-interference ratio (S/I) based call admission control model (Anand and Chockalingam, 2003). In this model, calls are been admitted by the network at the time of arrival based on their S/I ratio. Fenton method approximation has been used in this model for call blocking probability estimation. However, a newly admitted call can cause an outage to any ongoing calls due to the S/I ratio and resource scarcity in this model. Similarly, Walingo et al. proposed a connection admission control model for radio resource management in 5G (Walingo and Takawira, 2014). In this model, signal-to-interference ratio and delay parameters had been used to estimate call blocking probability. However, When there are no resources on the network, any request arriving, either critical or not, will be automatically rejected. Such admissions may lead to degradation of network performance due to poor traffic management. Moreover, to obey resource constraints, the requests must be within the bounds of the admission region (Bega et al., 2017). Nowadays, FIFO is not a frequently used admission control approach due to its lack of optimisation.

During admission control, network providers may give preference to requests that belong to a certain category; for example, to URLLC slice requests in 5G, where the requests have strict latency requirements. This type of request is considered to have higher priority for admission and is also expensive in terms of operational cost (Kammoun et al., 2018; Soliman and Leon-Garcia, 2016). Therefore, admission of high-priority requests onto the network results in higher revenue. By utilising a Reinforcement Learning (RL) method, a priority-based slice admission scheme is suggested by the authors in (Raza et al., 2018). The author's emphasis is on revenue maximisation while also considering the latency requirements of the access network. Compared to FIFO, this scheme increases the revenue earned from high-priority requests with superior QoE but also increases the rejection ratio of lower-priority requests. Thus, short-term contracts on high-priority request fulfilment may be less profitable compared to long-term contracts on medium- or low-priority requests (Ojijo and Falowo, 2020). An example could be a network provider considering a contract to accommodate requests for autonomous driving with higher priority while rejecting other requests. Contracts such as this may be scarce and limited to a certain region. On the other hand, providers endeavour to have longer-duration contracts at higher demand to generate more revenue. Similarly, Caballero et al. proposed a network slicing game-based admission control and resource allocation model for guaranteed rate services (Caballero et al., 2018). In this model, users have been assigned a weight based on their resource utilisation from a particular slice. The authors used the Nash equilibrium method for admission control in their work, that checks whether the slices can satisfy the rate requirements of all admitted users or not. However, some of the slices become saturated by this approach that results in unfairness in resource allocation among slices. Another approach proposed by Caballero et al. is static slicing for admission control and resource allocation. In this approach, each slice receives a fixed fraction of resources from the network resource pool, which is shared among its users. If there are no resources in the slice, the user will drop instantly from the network. Similarly, Jiang et al. proposed two novel approaches 5G Slice Allocation (5G-SA) and 5G Admission Control Slice Allocation (5G-AC-SA) in (Jiang, Condoluci, and Mahmoodi, 2016). 5G-SA consider intra/inter slice priority for resource allocation among slices. In this approach, resources are allocated to the users based on their demand and earned revenue in the slice. In this model, resources are allocated to users of high data rate demand in case of heavy load on the network, which degrades overall network QoS due to lower satisfaction of users with low data rate demand. 5G-AC-SA consider users' intra-slice priority and demanded QoE for slice allocation and admission control. This approach gives better fairness in resource allocation as compared to 5G-SA. However, in the case of massive connectivity demand, network QoS degrades due to more number of lower priority users and slices rejections from the network.

The authors of (Challa et al., 2019) proposed a greedy policy for admission control. In this strategy, a Partial Adaptive Greedy (PAGE) algorithm was deployed that maximised revenue and minimised SLA violation for customers. In this algorithm, a π policy is exploited. π is determined by a learning process through several iterations to meet the admission objective, and an example of such a policy is given in (Tang, Shim, and Quek, 2019). Online auction on available resources and greedy approaches are applied for resource allocation in (Liang et al., 2019). In this model, users are admitted to the network based on profit. The user, who can give more profit among other users to the network becomes the winner and has been served from the network. A bankruptcy game-based resource allocation algorithm for 5G Cloud-RAN slicing is proposed by the authors in (Jia et al., 2018). In this approach, user groups are created for admission to the network based on the Lloyd Shapley approach. In this work, a user would be a part of a group, if the user adds more benefits to the slice. However, a greedy-based admission control strategy may not always be optimal. Such a policy makes the greedy decision on the spot to achieve the objective, such as increasing the admitted requests to earn more revenue from the network. The admitted requests may affect network QoS and resource utilisation due to the congestion created by their massive number and long queuing delays (Yi, Wang, and Huang, 2018).

Random admission control helps to reduce unfairness occurring among users during the admission process. One such example is the Markov model for slice admission control proposed by the authors in (Han, Feng, and Schotten, 2018). In this work, the state transitions, such as pre-state to post-state and vice versa, are evaluated to reduce the computational complexity in the physical network. A fairly normal distribution can be achieved by applying this strategy over a long period to user admission; however, random admission control is not popular due to its lack of optimisation and management policy. Similarly, Lee et al. in (Lee et al., 2018), proposed a dynamic network slice management model for multi-tenant heterogeneous cloud-RANs. This model consists of two controls: an Upper-Level, which manages resources allocation, user association and admission; and a Lower-Level, which manages access network resources allocation among different users. Pablo et al. in (Caballero et al., 2018), considered a dynamic slice resource-sharing mechanism that shares resources among the elastic and inelastic nature of user traffic to achieve required network performance and revenue. Similarly, Zheng et al. in (Zheng et al., 2018), studied a simple and dynamic resource sharing model. This model allocates a "share" of pool resources to each tenant's slice based on demand and ensures efficient resource utilisation. In addition, from this share, each tenant assigns the resources to the admitted user via a slice share constrained proportionally fair (SCPF) scheme. The work in (Wu et al., 2018) developed a bio-inspired resources allocation model for 5G IoT applications. This model considers a diverse group of users, along with their homogeneous service and resource requirements, and their social relationships and behaviour update these characteristics periodically via a cellular automaton (CA) model. However, this approach results in creating unfairness on users' admission to the network due to their extremely diverse set of requirements. Due to increased network complexity, modern intelligent and multi-objective optimisation algorithms might be adopted in future wireless networks to deal with unfairness occurring among users in admission control (Adou, Markova, and Gudkova, 2018; Ojijo and Falowo, 2020).

2.2.2 Admission Control via Multi-Objective Optimisation

Multi-objective optimisation approaches have been successfully applied to provide optimal solutions to several difficult non-deterministic polynomial-time (NP) problems in wireless communication systems, such as spectrum allocation (Zhao et al., 2009; Gözüpek and Alagöz, 2011; Shami, El-Saleh, and Kareem, 2014), resource scheduling (Gu et al., 2015), channel assignment (Xu et al., 2012), indoor and outdoor tracking (Gharghan et al., 2015), and call admission control (Jain and Mittal, 2016). Compared to conventional approaches, multi-objective optimisation algorithms are known to find efficient solutions faster and more accurately, and they can produce a solution individually or in combination with other approaches.

In wireless/cellular communication, well-known examples of applied multi-objective optimisation algorithms are Swarm Intelligence (SI) and Genetic Algorithms (GA). Swarm Intelligence (SI) techniques for optimisation include Particle Swarm Optimisation (PSO) (Kennedy and Eberhart, 1995), Ant Colony Optimization (ACO) Dorigo and Di Caro, 1999, Dragonfly Algorithm (DA) (Mirjalili, 2016), Salp Swarm Algorithm (SSA) (Mirjalili et al., 2017) and Grey Wolf Optimizer (GWO) (Mirjalili, Mirjalili, and Lewis, 2014). PSO is among the top multi-objective optimisation approaches, due to its lower computational time requirement and simplicity compared to other approaches. Impressed by PSO, the authors proposed an enhanced swarm intelligence algorithm for a multi-objective joint power and admission control optimisation problem. Their optimisation algorithm is developed from two-phase PSO (TPPSO) and diversity global position binary PSO (DGP-BPSO) variants (El-Saleh et al., 2021). Similarly, Du, Jianbo, et al. proposed an enhanced optimal PSO-based radio resource allocation and admission control scheme in an LTE-A system. The proposed multi-objective optimisation algorithm is applied for resource block and power allocation with lower computation complexity. For modulation and coding, a Channel Quality Indicator (CQI)-based assignment scheme is adopted to obtain higher throughput from diverse channel conditions (Du et al., 2016). Similarly, Wang et al. adopted a Neighbourhood-Redispatch (NR) PSO-based approach for solving the mixed-integer programming problem of resource block assignment and power allocation in device-to-device communication (wang2017resource). Another, PSO-based resource allocation and admission control scheme for Software-Defined Heterogeneous Cellular Networks is proposed in (Gong et al., 2019). However, PSO suffers from premature convergence, because it attempts to obtain near-optimal solutions quickly. In addition, late phases of the search process with no further improvements are a burden on the network in terms of time consumption and unnecessary resource holding (Bhatia, Chauhan, and Yadav, 2021).

An enhanced swarm algorithm-based call admission control technique is proposed in (Suresh and Kumaratharan, 2021) to minimise call-blocking probability due to congestion in 5G cloud-based radio access networks (C-RAN). In this research, fuzzy parameters are optimised by applying the artificial fish swarm algorithm-based fuzzy inference system (FIS-AFSA) algorithm. A similar approach is proposed by Jain and Mittal in (Jain and Mittal, 2016) for handoff priority and handoff guarantee services in the cellular network. In this approach, optimal resources were allocated to requests over the dynamically adjusted threshold for admission control. Chowdhury et al. proposed an adaptive multi-level bandwidth-allocation scheme for non-real-time calls. This scheme diminishes the probability of call dropping by efficiently using available bandwidth (Chowdhury, Jang, and Haas, 2013). Khan et al. (Khan et al., 2021) presented a decoupled cell association method to solve the problem of resource allocation and admission control in 5G networks. They proposed an outer approximation algorithm (OAA) to acquire an optimal solution with lower computational complexity and better throughput.

GA is among the most important classes of evolutionary algorithms and has great potential to solve difficult optimisation problems that start with a search of a randomly generated population. The search proceeds, generation after generation, through biological operations. Such operations include selection, crossover, mutation and reproduction to evolve to the next generation (Mirjalili, 2019). The authors of (Sun, Lin, and Xu, 2018) proposed a two-level resource scheduling model for fog computing. They implemented clustering and the improved non-dominated sorting genetic algorithm II (NSGA-II) for optimum resource scheduling and admission control. In another effort, a multi-objective genetic algorithm-based optimisation framework was developed. This framework aims to obtain an optimal solution for spectrum allocation for Internet of Things applications (Han et al., 2018b). A hybrid-fuzzy logic-based genetic algorithm (H-FLGA) is proposed to solve a multi-objective resource optimisation problem in 5G vehicular networks. Adopting a service-oriented view, H-FLGA is implemented in the SDN controller for optimal admission control and resource allocation (Khan et al., 2019). Because of the higher computational complexity, these algorithms can be applied only to big data available on the cloud node. A hybrid GA and Binary PSO-based resource scheduling scheme is proposed for D2D multi-cast communication. Resource allocation is formulated as a min-max optimisation problem (Hamdi and Zaied, 2019). Such problems are generally known as NP-hard combinatorial problems, and typically require enormous amounts of searching to find an optimal solution.

In a multi-services scenario, a multi-population genetic algorithm is proposed to solve the resource allocation problem for D2D communications (Li et al., 2017). Bouali et al. proposed a fuzzy, multiple-attribute decision-making (MADM) approach for the best RAT selection (Bouali, Moessner, and Fitch, 2016). Similarly, Inaba et al. proposed a fuzzy call admission control model for multimedia networks (Inaba et al., 2015). In (Faris et al., 2019), the binary GA is implemented in combination with Random Weight Network (RWN) for identification of the most relevant features and spam detection. The proposed detection technique filters traffic during admission control, which prohibits spam messages from wasting resources such as storage, bandwidth, and productivity.

Multi-objective optimisation approaches generally provide a solution faster but sacrifice optimisation and accuracy by trapping in local minima or maxima. Thus, PSO- and GA-like multi-objective optimisation approaches may be adopted in cases where approximate solutions are sufficient for the learning process instead of more accurate solutions, which are computationally expensive. Such solutions are not mathematically intensive. Due to the higher computational complexity of optimal solutions, these approaches are partially suitable for future networks, and for existing mission-critical and latency-sensitive applications, where the promise is to provide latency under 1 millisecond. In the last few years, machine learning techniques, when used in combination with the optimisation algorithm, have proven their efficiency in solving various NP-hard problems of wireless communication, such as network admission control, resource allocation, radio resource management and channel estimation (Wang et al., 2020; Fourati, Maaloul, and Chaari, 2021).

The evolution of new technologies and standards (i.e. new radio (NR), and network slicing) is continuously increasing the network complexity of new use cases and service classes such as URLLC, mMTC and eMBB. These new technologies are designed to be flexible enough to meet future network service requirements and use cases. However, such flexibility increases network complexity through the growing number of core network control parameters, and increased complexity is forcing the implementation of fundamental changes in network operations. Recently, ML techniques have proven their efficiency in various domains of wireless communication. These techniques have the potential to increase the value of 5G and beyond networks, subject to proper integration into the system. As key components of future networks, they will play a significant role in customising the network at the technical level through ML-based network planning and service deployment, policy control, configuration, resource management and monitoring (Le et al., 2018; Mahmood et al., 2019; Challita, Ryden, and Tullberg, 2020; Nguyen et al., 2020).

A substantial amount of literature is available on the applications of ML techniques in wireless communication. For example, how network admission control, resource allocation, radio resource management and channel estimation can be leveraged from ML techniques is discussed in (Jiang et al., 2016a; Gündüz et al., 2019; Chen et al., 2019). However, the authors did not consider deployment and network design issues regarding how to implement ML techniques in applications of wireless networks. Among ML techniques, reinforcement learning (RL) has been frequently used in wireless networks for various purposes. The goal of RL techniques is to obtain an optimal policy that maximises the reward. Through a system of reward and penalty, RL reduces the complexity of the problems with strict constraints in the network. The reward is evaluated through the agent to proceed in network operations (Li et al., 2018; Maksymyuk et al., 2018; Elayoubi et al., 2019; Sun et al., 2019b).

Among RL approaches, the Markov Decision Process (MDP) is widely used to address various decision-making problems in dynamic wireless environments such as cognitive radio, spectrum management, power control and wireless security, and is determined by the state, action, transition probabilities and reward (Ye, Li, and Juang, 2019; Sun, Peng, and Mao, 2018). In MDP, policy iteration, value iteration, or a Q-learning approach can be applied to achieve optimal solutions. In policy iteration, the aim is to acquire the best policy that minimises the cost and maximises the reward. The value iteration approach evaluates the state–action pair to obtain an optimum value of the defined objectives. The Q-learning approach builds a lookup table of state–action pairs and rewards (Altman, 2000). Q-learning-based network selection, slice admission, and congestion control schemes have been proposed for 5G network control in (Han et al., 2018a; Wang, Su, and Liu, 2019; Antevski et al., 2020). Han et al. proposed a Q-learning-based admission and congestion control model for 5G network slicing (Han et al., 2018a). Similarly, Wang et al. proposed a novel network selection model by using Q-learning for 5G heterogeneous networks (Wang, Su, and Liu, 2019). A Q-learning strategy for the federation of 5G services is proposed in (Antevski et al., 2020). Although Q-learning is proven in convergence to an optimal solution, this approach is inefficient in a higher dimension state and action space. Building a Q-table for a large volume of data is memory intensive. Moreover, Q-learning suffers from slow convergence.

In an artificial neural network, highly effective solutions used to address high-dimensional problems are known as deep RL (DRL) or deep Q-learning (Shrestha and Mahmood, 2019; Santos et al., 2020). Moreover, deep Q-learning applies gradient descent to optimise the objective function, while using a deep neural network to approximate the Q function. DRL can be value-based or policy-based. Value-based DRL relies on deep Q-learning to find an optimal policy based on Q-function (Pouyanfar et al., 2018; Xiong et al., 2019). Li et al. in (Li et al., 2018) proposed a deep reinforcement learning-based model for resource management in network slicing. Similarly, Bega et al. proposed a deep learning-based model known as DeepCog for cognitive network management in sliced 5G networks (Bega et al., 2019). The authors in (Tang, Zhou, and Kato, 2020) proposed a deep reinforcement learning-based model for dynamic uplink/downlink resource allocation in high mobility 5G HetNet. By utilising the deep reinforcement learning approach, another intelligent resource slicing model for URLLC and eMBB traffic in the 5G and beyond network is proposed in (Alsenwi et al., 2021). An end-to-end network slicing model based on deep Q-learning for a 5G network is proposed by the authors in (Li, Zhu, and Liu, 2020). However, if the available data is highly correlated, and the Q-function is estimated from a nonlinear function approximator, then DRL can diverge to unsuitability.

Recent studies have revealed that conventional ML approaches have shortcomings in solving future network problems, especially emergency and mission-critical problems. This is because of the diverse characteristics and requirements of future wireless networks such as dynamic environment, high mobility, interference and diverse connections. Conventional ML approaches are usually trained in specific scenarios with a significantly huge amount of data. ML algorithms are highly data-intensive, where the size and duality of data matter, and the quality of the data determines the required processing time and computational complexity. For example, higher-dimensional data will require more time and is computationally costly. Moreover, a huge quantity of high-quality, raw data sent to the central node for training and processing creates congestion in the network, which might not be acceptable in latency-sensitive scenarios. Wireless environments may have significant variations from one scenario to another (e.g. user mobility and changes in data demand (Kato et al., 2020; Chen et al., 2019). The impact on ML performance due to varying network circumstances can hinder its applicability in future wireless networks. Therefore, because of the underlying network conditions, ample tuning is required in existing ML techniques to obtain a better result.

In network design and configuration, the design of an optimisation algorithm is complex, due to various network parameters and their associated constraints. To address these kinds of issues, Transfer Learning (TL) among advanced ML techniques has recently emerged as an effective solution, where knowledge is transferred from one optimised task or problem to solve another related or similar problem (Cook, Feuz, and Krishnan, 2013). TL has various advantages over conventional ML approaches. For example, the learning process in TL is faster due to the use of pre-trained models or policies, and knowledge sharing between tasks. Compared to traditional approaches, knowledge transfer in TL reduces the computing demand and congestion created in the network due to the huge amount of data. Just enhanced quality and quantity of training data is used in TL, which also provides data privacy protection (Zhuang et al., 2020). A significant amount of research into the applications of TL in wireless networks is available in the literature. For example, a novel TL-based paradigm for dynamic spectrum allocation and topology management of radio networks is proposed by authors in (Zhao et al., 2013). The knowledge learned through spectrum allocation is converted through their proposed priority algorithm and applied to topology management. During their research, Zhao et al. investigated the use of the K-means clustering approach for optimal spectrum and load management of mobile broadband networks (Zhao et al., 2015), whereby coefficients acquired from Q-parameters after demand clustering were transferred from spectrum allocation to broadband load management. Parera et al. proposed a transfer-based model for resource utilisation in wireless networks (Parera et al., 2020). The authors exploited deep learning and TL algorithms for dynamic resource allocation and efficient network control. Wagle et al. proposed three transfer learning algorithms for radio frequency allocation in wireless cellular networks (Wagle and Frew, 2012). The objective of their proposed TL algorithms is to identify the similarities in demand from the original data set to extract pertinent information, which was used in the target data set to achieve efficient radio frequency allocation. Zeng et al. proposed a deep TL-based traffic prediction framework for wireless cellular networks (Zeng et al., 2020). The authors proposed a spatial-temporal cross-domain neural network model (STC-N) in this work. STC-N model uses cross-domain data along with a regional fusion TL strategy to improve the accuracy of future traffic prediction. TL-and DRL-based mode selection and resource management models for fog RAN, V2V communication and 5G networks are proposed in

(Sun, Peng, and Mao, 2018; Zhang et al., 2019b; Dong et al., 2020). Similarly, Parera et al. proposed a TL model for channel quality prediction of a given frequency carrier in wireless networks (Parera et al., 2019). In this work, convolutional neural networks and long short-term memory networks have been considered as TL tasks. The existing research did not provide an ML-based solution for bottlenecks or intra-slice congestion problems to ensure efficient admission control in future networks.

Impressed by the effectiveness of TL in solving the problems of wireless networks, the ML-enabled optimal admission control and resource management model is presented in Chapter 4 of this thesis for bottleneck and intra-slice congestion control in 5G and beyond networks. The goal of this approach is to manage the demand proportionally with available capacity using two unsupervised learning-based clustering and optimisation approaches for congestion control. In view of the eMBB network's complexity, knowledge learned by implementing optimisation of mMTC traffic demand for clustering is implemented to eMBB traffic demand to reduce bottleneck congestion. RL-based admission control and resource management have also been proposed using intra-slice and inter-slice resource allocation, along with adaptability of slice elasticity, to maximise admission gain and resource utilisation by reducing the slice request rejection ratio.

2.3 Research Gap and Contributions

Nowadays, conventional single objective admission control strategies are not as popular, due to a lack of optimisation and the presence of unfairness during admission control. However, due to their simplicity and cost effectiveness, they are still in use for indoor communication in small networks (e.g. airports, bus stops, train stations, and shopping malls) (Kaloxylos, 2018; Le et al., 2016; Ojijo and Falowo, 2020). In a defined area, how 5G or future networks manage resources and unexpected demand with enhanced network QoS and user QoE is still an open issue (Gohil, Modi, and Patel, 2013; Khan et al., 2020b). Rapidly changing traffic flow and its associated heterogeneous demand can cause saturation in a network due to resource scarcity. For example, greed for earning more revenue and inefficient admission control cause unfairness in resource allocation among users, which overwhelms the network and creates congestion due to redundant signalling and data. To address these problems, enhanced conventional approaches are a better option for small networks to find a suitable solution with low computation complexity, cost, and time requirement. Therefore, a novel DSAAC model for 5G and beyond networks has been presented in Chapter 3 to service diverse traffic demand with efficient resource utilisation and network GoS enhancement. This approach takes into account network and user-application-specific demand characteristics for admission control. These characteristics include the required bandwidth, data rate, and priority. In addition, a cost estimation function that can quantify resource allocation decision metrics for both static and dynamic user and network characteristics has been developed. DSAAC also presents a set of optimisation algorithms for effective network slice utilisation, inter-slice resource allocation, and back-off-based admission control. Similarly, a novel SAC model utilising optimisation and clustering approaches has also been presented in Chapter 3 to support more user-specific traffic data. In this model, it is proposed that pre-clustering end-use analysis, usage-specific clustering, and clustering based on end-use application and device-specific resource demand be exploited. A usage-specific clustering strategy has been developed for the access network's signalling redundancy minimisation that guarantees a given level of QoE. For user admission to the network, dynamically reconfigurable QoE-based slice performance bounds have also been taken into account. Additionally, a set of optimisation algorithms for achieving effective resource utilisation and slice allocation through evaluating the elasticity of slice QoE bounds has also been devised.

Multi-variate optimisation admission control has been frequently applied in wireless networks. This is because of their ability to adapt to dynamically reconfigure the network according to pre-defined optimisation objective(s). However, these approaches require a great deal of time for training and processing (Mirjalili, 2019; Han et al., 2018b). This is because a large amount of data is required for training purposes from the demand. Data contains simple signalling requests, as well as high-quality data, such as data for augmented reality, 4K images etc. In the case of a huge volume of network traffic flow, redundancy in demand is a burden on the network in terms of congestion created. Moreover, the time spent on training on the data from the demand would generate additional latency, which creates large forwarding delays in communication and user dropping due to congestion, which may not be acceptable in latency-critical scenarios in future networks (Gupta and Jha, 2015; Caballero et al., 2018; Mahmood et al., 2019). In view of these constraints, redundancy reduction in the demand at the edge of the network can help to minimise congestion and latency in access and the core network. To assist future networks in minimising the aforementioned control signalling redundancy and congestion, a novel E-RMAC model has been presented in Chapter 4. E-RMAC model is the enhancement of the SAC model. In this model, two popular unsupervised learning techniques (K-mean- and Ranking-based clustering) and multi-objective optimisation (Non-dominated sorting genetic algorithm II) are employed to reduce core network signalling redundancy. By using these techniques, cluster-based signal and admission control algorithms have been established to maximise link efficiency between the edge and core networks.

Moreover, running optimisations for forecasting demand is a better option, due to time constraints in an integrated network environment such as O-RAN. A novel dynamic traffic forecasting and admission control (FAC) model for a federated O-RAN environment has also been presented in Chapter 4, that predicts future traffic demand for efficient admission control and resource allocation. In this model, a fully reconfigurable admission control model using fuzzy-logic optimisation by drawing on the information on user demand and network capacity has been proposed for the optimal network selection. After optimal network selection, a set of algorithms are also proposed for admission control and service monitoring. Whereby, a multivariate service allocation priority factor has been developed for admission queuing. Moreover, a service profile has been built on admissions for service monitoring. These approaches are developed in the proposed models for maximising network QoS and user-perceived QoE, as well as congestion control caused by the massive number of connectivity demands in future networks.

Increasing bottleneck congestion and rejection ratio due to the massive number of connectivity demands in future networks, especially in network slicing, is a critical problem that needs attention (Dandachi et al., 2019). The use of conventional and optimisation approaches to address these issues has several constraints, such as data dependency, computational complexity and cost. The existing models for network slicing provide only brief guidelines on design and architecture and do not provide a specific solution to issues such as these. To address this problem, along with efficient admission control and resource management, the implementation of the advanced ML approach is a better solution. Among advanced ML approaches, RL and TL have proved highly effective at providing optimal solutions to various wireless network problems (Nguyen et al., 2021; Tan et al., 2018), as discussed earlier. Therefore, a novel SCAC model has been presented in Chapter 4 to minimise the number of slice requests rejection that occurs due to bottlenecks and intra-slice congestion in the network. This model consists of a slice demand analysis and classification (SDAC), a demand clustering and queuing (DCQ), and an admission and resource management (ARM). Two popular unsupervised learning algorithms, Ranking and K-mean clustering algorithms, along with multi-objective optimisation and transfer learning, have been employed for slice request queuing. A unified cost-estimation function is also derived for slice selection to ensure fairness among slice requests. Given instantaneous network circumstances and load, a reinforcement learning-based admission control policy is also established for taking appropriate action on the guaranteed soft and best-effort slice request admissions. Intra slice, as well as inter-slice, resource allocation, along with the adaptability of slice elasticity, are also proposed in this model for maximising slice acceptance ratio and resource utilisation.

Chapter 3

Admission Control with Single–Objective Optimisation

3.1 Introduction

In the last few years, continuous advancement in wireless technologies has enabled small industries to quickly enter the market at a low cost. These industries are helping various sectors with their innovative products and services such as in healthcare, industrial automation, transportation, education, etc. (Da Xu, He, and Li, 2014; Gupta and Jha, 2015). However, to meet the continuously rising demand of the small industries, a cost-effective and simple solution is required in network admission control and resources management (Ojijo and Falowo, 2020). Therefore, due to simplicity and lower computational complexity, two novel single-objective optimisation-based admission control and resources management models are presented in this chapter. In Section 3.2, a dynamic slice allocation and admission control model, otherwise known as the DSAAC, is proposed to ensure better network QoS (or GoS) and user-perceived QoE by using a QoE-based resource allocation metric. Next, an optimised signalling and admission control model, also known as the SAC, is proposed in Section 3.3 for signalling redundancy minimisation in the access network. In this model, a ranking-based clustering approach is applied to the demand for optimal admission control and resource allocation. Finally, a summary of the chapter is given in Section 3.4.

3.2 Dynamic Slice Allocation and Admission Control (DSAAC) Model

Nowadays, a significant number of industries are adopting novel business models by delivering innovative products and services via the internet of things (IoT) and other wireless-enabled technologies. These industries bring social and economic benefits to various sectors such as education, industrial automation, transportation, and healthcare (Da Xu, He, and Li, 2014; Kaloxylos, 2018). Consequently, in recent years, this trend reports a significantly huge volume of connected devices, which is termed as an ultra-dense wireless network. *Ericsson Mobility* report on 5G development predicts that there will be approximately 550 million 5G subscriptions by 2025 (Gozalvez, 2017). This trend will induce significant challenges in the network, leading to increased network complexity, congestion and user dropping due to massive traffic, control signalling storms in the access and core network, and resource inefficiency in terms of usage (Brown, 2012; Hicham, Abghour, and Ouzzif, 2014; Al-Fuqaha et al., 2015; Gupta and Jha, 2015; Bhandari, Sharma, and Wang, 2018).

Efficient admission control helps network operators and service providers to maximise resource utilisation and earn more revenue. In this context, a novel concept proposed by the 3rd generation partnership project (3GPP) is network slicing (3GPP, 2018c). This is a vital feature of future networks for virtual network deployment enhancing the existing infrastructure to accommodate heterogeneous QoE requirements of various customers. Software defined network (SDN), along with network function virtualisation (NFV), is a promising solution for the provisioning of dedicated services in cloud-based networks (Ofcom, 2017; Ordonez-Lucena et al., 2017). However, virtualisation for resource allocation in network slicing to ensure better network quality of service (QoS) and user quality of experience (QoE) is more challenging compared to cloud-based networks. These challenges include (1) slice admission to the network and user admission to the slice with required QoE, (2) dense traffic management and congestion control on the slice, (3) optimal and dynamic resources allocation and their efficient utilisation, and (4) varied network conditions such as critical and latency-sensitive networks (Kaloxylos, 2018; Su et al., 2019).

A significant amount of research is being carried out all around the world to address these challenges. For example, the Markov Model has been proposed in (Han, Feng, and Schotten, 2018) for the support of MNO synchronous slice admission in 5G Networks. In this work, the state transitions, such as pre-state to post-state and vice versa, are evaluated to reduce the computational complexity in the physical network. Lee et al., in (Lee et al., 2018), proposed a dynamic network slice management model for multi-tenant heterogeneous cloud-RANs. This model consists of two controls: an Upper-Level, which manages resources allocation, user association and admission; and a Lower-Level, which manages access network resources allocation among different users. Pablo et al., in (Caballero et al., 2018), considered a dynamic slice resource-sharing mechanism that shares resources among the elastic and inelastic nature of user traffic to achieve required network performance and revenue. In this model, users have been assigned a weight based on their resource utilisation from a particular slice. The authors used the Nash equilibrium method for admission control in their work, that checks whether the slices can satisfy the rate requirements of all admitted users or not. However, some of the slices become saturated by this approach that results in unfairness in resource allocation among slices. Another approach proposed by Caballero et al. is static slicing for admission control and resource allocation. In this approach, each slice receives a fixed fraction of resources from the network resource pool, which is shared among its users. If there are no resources in the slice, the user will drop instantly from the network. Similarly, Zheng et al., in (Zheng et al., 2018), studied a simple and dynamic resource sharing model. This model allocates a "share" of pool resources to each tenant's slice based on demand and ensures efficient resource utilisation. In addition, from this share, each tenant assigns the resources to the admitted user via a slice share constrained proportionally fair (SCPF) scheme. The work in (Wu et al., 2018) developed a bio-inspired resources allocation model for 5G IoT applications. This model considers a diverse group of users, along with their homogeneous service and resource requirements, and their social relationships and behaviour update these characteristics periodically via a cellular automation (CA) model. Random admission control helps to reduce unfairness occurring among users during the admission process. One such example is the Markov model for slice admission control (Han, Feng, and Schotten, 2018). A fairly normal distribution can be achieved by applying this strategy over a long period during user admission. However, random admission control is not popular due to a lack of optimisation and management policy. Other authors (Challa et al., 2019) have proposed a greedy policy for admission control. In this strategy, a partial adaptive greedy (PAGE) algorithm was deployed that maximised revenue and minimised SLA violation for customers. Similarly, Anand et al. proposed a signal-to-interference ratio (S/I) based call admission control model (Anand and Chockalingam, 2003). In this model, calls are been admitted by the network at the time of arrival based on their S/I ratio. Fenton method approximation has been used in this model for call blocking probability estimation. However, a newly admitted call can cause an outage to any ongoing calls due to the S/I ratio and resource scarcity in this model. Similarly, Walingo et al. proposed a connection admission control model for radio resource management in 5G (Walingo and Takawira, 2014). In this model, signal-to-interference ratio and delay parameters had been used to estimate call blocking probability. However, When there are no resources on the network, any request arriving, either critical or not, will be automatically rejected. Such admissions may lead to degradation of network performance due to poor traffic management.

The existing work adopts greediness by considering only a few access or core network parameters for admission control and resource allocation. But practically, there are influences of many known and unknown parameters in the dynamic admission and allocation process, especially in the case of a massive amount of connectivity demand, where the network can get saturated with inefficient admission control, resource allocation and management policies that result in inducing congestion and latencies into the provisioning of heterogeneous applications (Andrews et al., 2014; Alliance, 2015). Therefore, this research work aims to facilitate future traffic demand with the assurance of better QoE in the provisioning of heterogeneous services from 5G networks. Given the above-mentioned challenges, an intelligent slice-management solution is proposed in this section. The major contributions of this work are as follows:

- Proposed a novel architectural model for dynamic slice allocation and admission control in 5G and beyond networks.

- To serve future wireless networks with better QoS and user QoE, a unified cost estimation function has been derived from the normal equation for optimising slice allocation and admission control in the proposed model.

- For efficient network slice utilisation, a set of optimisation algorithms with intra-slice allocation elasticity, inter-slice handover, and a user back-off-scheme-based admission control and resources allocation has been proposed.

- Robustness of the proposed admission control and resources allocation algorithms are analysed through GoS, network utility, mean delay and throughput. Moreover, the results obtained are compared with those of existing models found in the literature.

3.2.1 DSAAC System Model

In this work, the NGMN (Alliance, 2016) network slicing conceptual architecture, as seen in Figure 2.1, is enhanced by the proposed Dynamic Slice Allocation and Admission Control (DSAAC) model. The proposed model of network slicing consists of four major layers: a *Resource layer*, a *Network layer*, an *Application or Service layer*, and a *User layer*, as shown in Figure 3.1. The resource layer holds and manages the resources within a resource pool. This layer shares resources among slices upon request through instances. The network layer contains the slice instance to establish an end-to-end connection to serve the specific application via subnetwork slice instances. The subnetwork slice instance contains a set of customised resources from the resource pool, which can be shared among various slice instances. For example, a slice such as s_2 can support heterogeneous applications or services requested by a single user group or from multiple groups. This means sharing of resources is allowed by the network management policies in that particular instance. In addition, a user group can be connected to multiple slices simultaneously; for example, user group s_{21} is also connected to application instance s_{11} in the network layer. Moreover, a slice can obtain the resources from multiple subnetwork slice instances. The

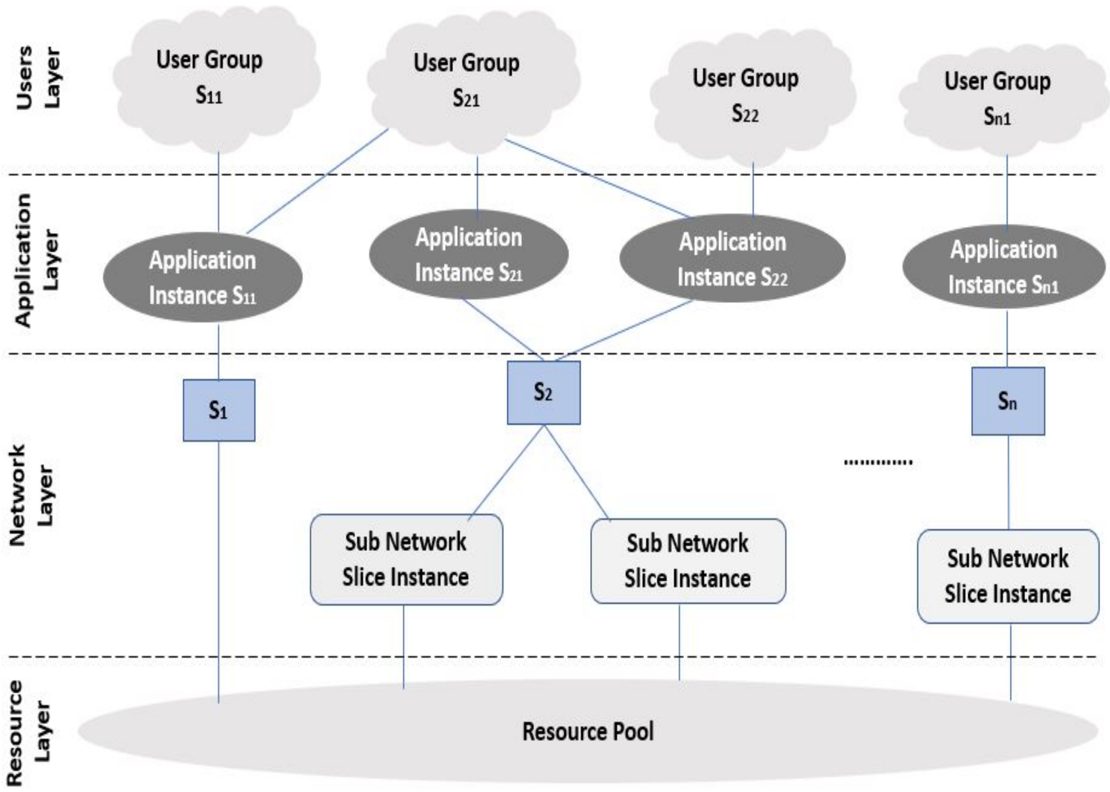


FIGURE 3.1: Slice interaction schematic diagram of the proposed DSAAC model

subnetwork slice instance can also be shared among various active slices for resource allocation on demand. The capability of multi-slice connectivity for heterogeneous service provisioning makes the scenario more complex. The application or service layer contains application instances to serve a particular application from the respective network.

To acquire uniform slice allocation and resource utilisation, the user layer, (a sublayer of the application layer in the NGMN concept), is separated from the application layer in the proposed model. Disaggregation of the user layer from the application layer is acquired with respect to different groups of users and their homogeneous characteristics and association with the verticals, applications or network operators. In view of this, E2E slice management and orchestration via the proposed DSAAC model is shown in Figure 3.2. The admission controller is responsible for the optimal slice selection and user admission to the network by the exchange of information between access (i.e. random access network (RAN)) and core control network functions (i.e. access and mobility management function (AMF) and network slice selection function (NSSF)). After slice selection, a slice instance is created from the sub-slice instances based on the management and orchestration policies and agreed service level agreement (SLA) with the tenant. On successful creation of slice instances, the mobile network operator (MNO) populates network types for its users for service provisioning. After that, the session management function (SMF) establishes an E2E session between the user and data network for communication via the user plane function (UPF). Considering the proposed model, the following subsection elaborates on the proposed assessment scheme of slice demand and leads toward the design of the optimisation algorithms for admission control and intra/inter-slice-based resources allocation.

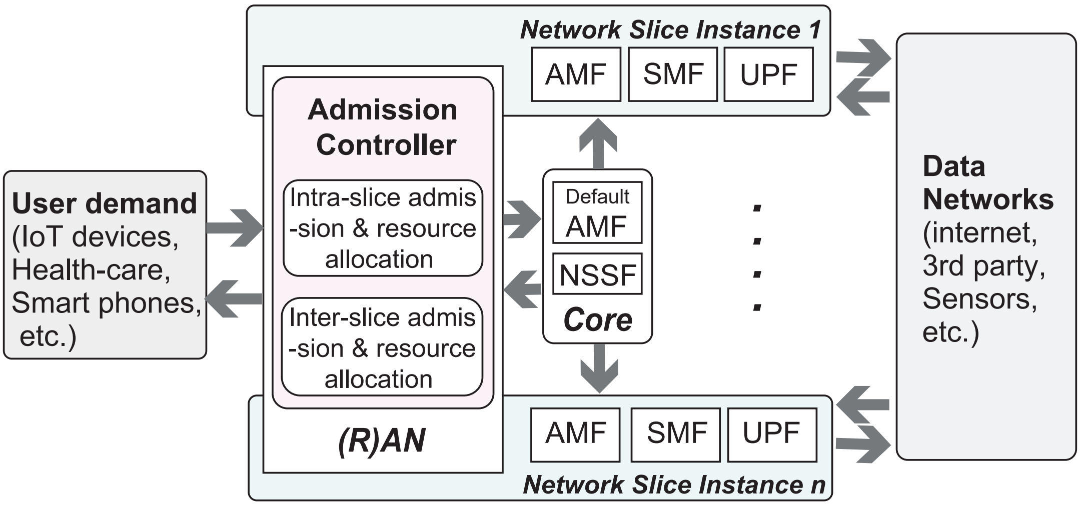

Network Slice Instance 1
AMF
SMF
UPF
Admission Controller
Intra-slice admis -sion & resource allocation
Inter-slice admis -sion & resource allocation
(R)AN
User demand (IoT devices, Health-care, Smart phones, etc.)
Default AMF
NSSF
Core
AMF
SMF
UPF
Network Slice Instance n
Data Networks (internet, 3rd party, Sensors, etc.)

FIGURE 3.2: End-to-End slice management and orchestration via the proposed DSAAC model

Network Design and Statistics: In this work, a cellular network with M number of users denoted as $\mathbf{u} = \{u_1, u_2, \ldots, u_M\}$ is considered, as illustrated in Table 3.1. It is assumed that each user's behaviour is independent and different from other users in the network. Moreover, each user can support up to k number of heterogeneous applications simultaneously, denoting user-specific application set $\Lambda = \{\Lambda_1, \Lambda_2, \ldots, \Lambda_k\}$. A set of user-application demands, as well as real-time, network-specific collaborative data (e.g. user-demanded QoE and available network capacity), is supposed to be available in AMF. Given the user-network-specific data, each application is supposed to have N required characteristics that are predetermined within an AMF repository. As users can have access to multiple applications simultaneously, these measurable demand characteristics might be different

TABLE 3.1: DSAAC model key symbols and definitions

Symbols	Definitions
$\mathbf{u}$	Set of M number of users
S	Set of slices in the network
$\mathbf{A}$	Demand matrix of M users
$\mathbf{w}$	network weights for slice selection
$\mathbf{v}$	Cost estimation function for slice selection
λ_{s_u}	Slice upper configuration bound
λ_{s_l}	Slice lower configuration bound
R_A^s	sth slice available resources
R_m	mth user required resources
$P_{b(m)}$	Blocking probability of the mth user from the slice
$U_m(R_m)$	mth user utility
U^s	sth slice utility
U^{Net}	Network utility on set $\mathcal{S}$

from one application to another application and quantified within a range of values between a_j^{min} and a_j^{max}, where a_j^{min} and a_j^{max} are the minimum and maximum jth demand from the slice, respectively. Once a resource request is received from the user, AMF assesses that resource request and populates the respective coefficient within the user-specific rows in the matrix $\mathbf{A}$, as shown in Eq.(1). At a given time, the matrix $\mathbf{A}$ row-wise statistical characteristics are represented by their mean (i.e. $\mu_1, \mu_2, \ldots, \mu_N$) and variance (i.e. $\sigma_1, \sigma_2, \ldots, \sigma_N$) respectively.

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1N} \\ a_{21} & a_{22} & a_{23} & \cdots & a_{2N} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{M1} & a_{M2} & a_{M3} & \cdots & a_{MN} \end{bmatrix}. \tag{3.1}$$

Let us assume that k=1 for simplicity. When a user u_m requests access to the network for service provisioning of its kth application with few desired characteristics, it issues a request denoted as $\mathbf{a}_{mn} = [a_{m1}, a_{m2}, \ldots, a_{mn}]$, where $u_m \in \mathbf{u}$, $k \in \Lambda$, $\mathbf{a}_{mn} \in \mathbf{A}$. The vector $\mathbf{a}_{mn}$ contains statistics related to user demand such as bandwidth, priority, latency, etc. Based on the active network current load and slice priorities, etc., the network weighting factors denoted as $w_1, w_2, \ldots, w_N$ for each user's demanded characteristics are also defined. This is to acquire a reflection of the network status in advance of the user demand for service provisioning. Accordingly, the **w** vector is expressed as

$$\mathbf{w} = [w_1, w_2, \ldots, w_N]^T . \tag{3.2}$$

The network weighting factors are dynamic in nature, due to changes in network circumstances. For example, $w_1, w_2, \ldots, w_N$ values might change periodically or non-periodically with respect to the critical and environmental conditions of the particular slice s; where $s \in S$, and $S = \{s_1, s_2, \ldots, s_n\}$.

3.2.2 Proposed DSAAC Model Schema

Dynamic network reconfiguration is essential to accommodate the massive amount of future heterogeneous traffic. The flexibility provided in the 5G network slicing feature is to obtain the user/application-requested QoE with the capability of dynamic resources allocation in a slice. Thus, it is essential to have flexibility in resource allocation from available resources in the slice pool to adapt to user application requirements (Khan et al., 2020b). On-demand resource aggregation technique for admission control has been adopted in this work, which provides resource elasticity along with slice reconfiguration dynamically when needed. For this aggregation model, a framework is defined with slice upper and lower bounds. In addition, a set of generalised expressions are presented below that define the elasticity of N number of demand characteristics for M number of active users in matrix **A**. By taking into consideration the truncated Gaussian distribution of the user demand on time t, as in (Brichet and Simonian, 1998), lets denote the minimum aggregate slice resource requirement for sth slice with jth characteristics column as $R^s_{j(Min)}$. Since γ_j is the minimum jth resource required to serve a single user, μ_j is the mean jth resource requirement from a single user application, Δ_j is the difference of the minimum jth resource demand from the mean, and σ_j is the variance of jth resource demand. Then, the minimum jth resource requirement for M number of users in the sth slice is calculated by the following equation.

$$R^s_{j(Min)} = \mu_j M^s + \Delta_j \sqrt{M^s} , \tag{3.3}$$

where,

$$\Delta_j = \gamma_j - \mu_j\,. \tag{3.4}$$

Now, the minimum sth slice resource characteristics bound over N resources is represented by

$$R^s_{Min} = [R^s_{1(Min)}, R^s_{2(Min)}, \ldots, R^s_{N(Min)}]\,. \tag{3.5}$$

Similarly, slice maximum jth resource requirement for M users, as in (Brichet and Simonian, 1998), is calculated as follows:

$$R^s_{j(Max)} = \sum_{m=1}^{M} \mu_j + \sqrt{\gamma_j \sum_{m=1}^{M} \sigma_j^m}\,. \tag{3.6}$$

By now, the maximum sth slice resource characteristics bound over N resources is expressed as:

$$R^s_{Max} = [R^s_{1(Max)}, R^s_{2(Max)}, \ldots, R^s_{N(Max)}]\,. \tag{3.7}$$

3.2.2.1 User Admission Control: In the dense network environment, inefficient admission control is a significant issue that increases the blocking probabilities of the request from the network due to resource scarcity (Khan et al., 2020b). This problem can be modelled as a single objective optimisation problem, where the objective is to minimise the blocking probabilities, denoted as P_b, from the network via efficient admission control on set U_s. Mathematically it can be represented as

$$\begin{aligned} \min \quad & P_b, \\ \text{s.t.} \quad & \sum_{u_m \in U_s} a_m R_m \le s_c\,, \end{aligned} \tag{3.8}$$

where, $a = 1$ on a user's admission to the slice, otherwise zero. R_m is the guaranteed resource allocation to the mth user. Aggregate resource allocation on set U_s should not exceed slice capacity, s_c. In this work, a cumulative soft decision-making solution has been proposed to redirect the u_mth user to an appropriate slice for admission. This solution relies on the novel cost function estimation to make an intelligent decision on admission control dynamically, due to the frequently changing nature of the characteristics. The cost function is derived from the normal equation over the user-application-specific demand, network dynamic characteristics, and slice bounds with former information. Thus, the cost function, symbolised as **v**, can be estimated as:

$$\mathbf{v} = \mathbf{A}\mathbf{w} = [v_1, v_2, \ldots, v_M]^T \,. \tag{3.9}$$

This implies that the estimated cost function for mth user is:

$$v_m = \sum_{n=1}^{N} \mathbf{a}_{mn} \mathbf{w}_n \,, \tag{3.10}$$

where $\mathbf{a}_{mn} \in \mathbf{A}$ and $\mathbf{w}_n \in \mathbf{w}$. v_m is the cost estimation value of mth user, this represents the slice type for mth user admission and resource allocation. Each slice is designated with a reconfigurable bound set, denoted as $[\lambda_{s_l}, \lambda_{s_u}]$, to serve a user demand from the slice, where λ_{s_l} represents the slice lower bound and λ_{s_u} represents the slice upper bound. Accordingly, they are evaluated as:

$$\lambda_{s_l} = \frac{1}{S_n}(i) \,, \tag{3.11}$$

and

$$\lambda_{s_u} = \frac{1}{S_n}(i+1) \,, \tag{3.12}$$

where i is index and i $=\{0, 1, 2 ... S_n - 1\}$. and S_n represents the total number of active slices in set S.

3.2.2.2 Intra-Slice Admission and Resources Allocation: Algorithm 1 represents the proposed intra-slice resources allocation. On arrival of the user request, the computed user cost value is assessed against the slice cost bounds, as shown in Algorithm 1 and verified by the equation below:

$$f_s(v_m) = \begin{cases} \textit{Admit}, & \lambda_{s_l} < v_m \leq \lambda_{s_u} \\ \textit{Reassessed via (3.1)} & \text{otherwise} \end{cases} \,. \tag{3.13}$$

Subject to the availability of the resource on that slice, the user is admitted and accommodated with the desired resources. The slice resource pool would also be updated after user admission and resource allocation. In addition, the resource utilisation of a user's allocated resources is computed to obtain the overall slice and network utilisation for revenue estimation by the operator. Otherwise, in the case

Algorithm 1: Intra-slice admission and resources allocation

```
Input: $s \in S$, $R^s_A \neq 0$, $u_m \in \mathbf{u}$, and calculate $v_m$.
Output: $P_b$, $U_m(R_m)$, $\sum_{u_m \in U_s} U_m(R_m)$ and $\sum_{s \in S} U^s$.
begin
    if ($\lambda_{s_l} < v_m \leq \lambda_{s_u}$) then
        if ($R^s_A \geq R_m$) then
            Admit $u_m$
            Assign the resources
            Update available resources in $R^s_A$
            Update running resource pool with $u_m$ and $v_m$
            Sort running resource pool with respect to $v$ in descending order
            Compute $P_b$
            Calculate $U_m(R_m)$, $\sum_{u_m \in U_s} U_m(R_m)$ and $\sum_{s \in S} U^s$
        end
    else
        Back-off and Reassessed via (3.1)
    end
end
```

of an out-of-bounds condition or no resources being available on the slice, this user might be backed off in the matrix **A** and reassessed by the slice with a higher cost value. This reassessment to accommodate user application is subject to a user's priority conditions and network GoS. The consideration may include the possibility of back-off with the time-shift nature of the application and resource availability. In the case of an inability to time-shift or there being no resources on the slice, the proposed allocation model assesses the possibility of inter-slice handover.

3.2.2.3 Inter-Slice Admission and Resources Allocation: In this model, the provision of inter-slice admission control and resource allocation using the roaming principle is considered. During inter-slice admission control and resources allocation, the sth slice acts as the primary candidate slice for the mth user. However, in the case of unavailability of resources in the primary slice, user admission is assessed by the neighbouring slice (i.e. $(s+1)$ or $(s-1)$). The lower and upper asset bounds (λ_{s_l} and λ_{s_u}) of the neighbouring slice will be temporally updated for that user only via:

$$\lambda_{s_u} = \lambda_{s_u} + \delta(\lambda_{(s+1)_u} - \lambda_{(s+1)_l}), \tag{3.14}$$

and

$$\lambda_{s_l} = \lambda_{s_l} - \delta(\lambda_{(s-1)_u} - \lambda_{(s-1)_l}) \,, \tag{3.15}$$

where δ varies from zero to 0.5 by the central limit theorem (Rosenblatt, 1956; Johnson, 2004; Fischer, 2010). Algorithm 2 and 3 represents the inter-slice admission control and resources allocation strategies for slice $(s-1)$ and $(s+1)$, respectively. Assuming capital expenditure (CAPEX) is proportional to the slice index; i.e. $CAPEX_{(s-1)} < CAPEX_{(s+1)}$.

Algorithm 2: Inter-slice admission and resources allocation from slice $(s-1)$

```
Input: $(s-1) \in S$,  $(s+1) \in S$,  $R_A^s = 0$ or $R_A^s < R_m$,  $u_m \in \mathbf{u}$,  update
  cost bounds (i.e. $\lambda_{s_l}, \lambda_{s_u}$), set s = (s-1), and calculate $v_m$.
begin
    if $(\lambda_{s_l} < v_m \leq \lambda_{s_u})$ then
        if $(R_A^s \geq R_m)$ then
            Compute Algo. 1
        else
            Handover $u_m$ to $(s+1)$
        end
    end
end
```

Algorithm 3: Inter-slice admission and resources allocation from slice $(s+1)$

```
Input: $(s+1) \in S$,  $R_A^s = 0$ or $R_A^s < R_m$,  $u_m \in \mathbf{u}$,  update cost bounds
  (i.e. $\lambda_{s_l}, \lambda_{s_u}$), set s = (s+1), and calculate $v_m$.
begin
    if $(\lambda_{s_l} < v_m \leq \lambda_{s_u})$ then
        if $(R_A^s \geq R_m)$ then
            Compute Algo. 1
        else
            Back-off and Reassessed via (3.1)
        end
    end
end
```

3.2.3 Performance Evaluation Measures

This section presents the measures for performance evaluation of the proposed work. The context of evaluation is set out to be an assessment of GoS, throughput, mean delay, and network utility. The proposed model provides additional flexibility with its inter-slice allocation elasticity and inter-slice handover for admission control. The performance evaluation measures developed in this section have been aligned with the work in the existing literature.

3.2.3.1 Grade-of-Service: Resource insufficiency on slice *s* might cause the *m*th user to be blocked or a handover to the neighbouring slice, subject to the availability of sufficient resources to serve the user. Hence the user blocking probability $P^{j}_{b(m)}$ from slice *s* with respect to the *j*th resource characteristic is derived by the probability density function or PDF, which is denoted as $f(x_j)$. In this work, overall demand for the *j*th resource is assumed to have exponential distribution on the network, which states that the *m*th user will be served if the required *j*th resource statistic is in the range $a_j^{s_{min}}$ to $a_j^{s_{max}}$, otherwise, the user is not served.

$$P^{j}_{b(m)}\{a_j^{s_{min}} < X_{j(m)} \leq a_j^{s_{max}}\} = 1 - \int_{a_j^{s_{min}}}^{a_j^{s_{max}}} f(x_j)dx_j\,, \tag{3.16}$$

where $f(x_j)$ is obtained by

$$f(x_j) = \frac{1}{\mu_j^s} e^{\frac{-x_j}{\mu_j^s}}\,. \tag{3.17}$$

User's admission or access probability with respect to the *j*th resource characteristics, denoted as $P^{j}_{adm(m)}$, to slice *s* is obtained by

$$P^{j}_{adm(m)} = 1 - P^{j}_{b(m)}\,. \tag{3.18}$$

Overall blocking probability of the *m*th user from the slice is $P_{b(m)}$. This can be obtained as

$$P_{b(m)} = \Pi_{j=1}^{N}(P^{j}_{b(m)}) = \Pi_{j=1}^{N}(1 - P^{j}_{adm(m)})\,. \tag{3.19}$$

Subsequently, the probability of the user being admitted to the network slice is $P_{adm(m)}$, which can be rewritten as

$$P_{adm(m)}\{\lambda_l^i < v_m \leq \lambda_u^i\} = \Pi_{j=1}^{N}(1 - P_{b(m)}^{j})\,. \tag{3.20}$$

Overall blocking probability over M number of users on slice s can be found as

$$P_b = \Pi_{m=1}^{M} P_{b(m)}\,. \tag{3.21}$$

3.2.3.2 Throughput: Likewise, average slice throughput, which is symbolised as η, is defined as a fraction of successful rate transmission with respect to blocking (P_b) and retransmission probabilities (p_{rtx}), respectively, over available data rate and n resources (Anand and Chockalingam, 2003). Accordingly, η in the proposed model can be obtained as

$$\eta = \frac{\ell_v\Gamma^{(v)}(1 - P_b) + \ell_d\Gamma^{(d)}(1 - P_b)(1 - p_{rtx})}{n\Gamma^{(d)}}\,, \tag{3.22}$$

where, ℓ_v, ℓ_d, $\Gamma^{(v)}$, $\Gamma^{(d)}$ represents the network load and data rate in terms of voice and data traffic, respectively.

3.2.3.3 Mean Delay: Mean delay, which is denoted as $\bar{D}_d$ in the data transmission, is defined as the product of system service time T_{ser}, blocking probability and the number of retransmissions per packet N_{rtx} (Anand and Chockalingam, 2003). Accordingly, $\bar{D}_d$ in the proposed model can be obtained as

$$\bar{D}_d = P_b T_{ser} N_{rtx}\,, \tag{3.23}$$

where, service time, T_{ser}, is the sum of the mean waiting time of a user request in a buffer; known as buffering time T_{buff} and its processing time T_p (Anand and Chockalingam, 2003). Hence, T_{Ser} can be achieved as:

$$T_{ser} = T_{buff} + T_p\,. \tag{3.24}$$

As long as the resources are available within the slice, the user request will be scheduled concurrently, and the waiting time will be zero.

3.2.3.4 Aggregate User and Slice Utility: For provisioning of a particular service, each user has a minimum resource demand or guaranteed rate requirement (as in Lee et al., 2018), which is denoted as R_m in the proposed model. The admitted users' aggregate and guaranteed resource requirement from the sth slice can be obtained as

$$\sum_{u_m \in U_s} a_m R_m \leq s_c \,, \tag{3.25}$$

where, $a = 1$ on a user's admission to the slice, otherwise zero. The aggregate value should not exceed slice capacity s_c with respect to the bounds, as discussed earlier. To this end, the user transmission resources rate R_m can be calculated as

$$R_m = \frac{\omega_m}{\sum_{u_m \in U} \omega_m} c_m \,, \tag{3.26}$$

where c_m is the achievable rate or peak rate of user m, which is the product of the user resource reservation factor and the aggregate resources of the slice. ω is the non-negative user share over slices. Hereafter, the mth user utility, which is denoted as U_m, with respect to R_m can be acquired as

$$U_m(R_m) = \varphi f_m(R_m) \,, \tag{3.27}$$

where, R_m is greater than the user minimum guaranteed rate requirement, γ_m, φ is the user priority, and $f_m(.)$ is the concave utility function, where $f_m(R_m) = \frac{R_m^{(1-\alpha)}}{(1-\alpha)}$ and $\alpha = 0$ subject to linear sum. Subsequently, based on the user utility, $U_m(R_m)$, the slice s utility, U^s, is the sum of individual utilities (as in Caballero et al., 2018). Thus

$$U^s = \sum_{u_m \in U_s} U_m(R_m) \,. \tag{3.28}$$

Based on the individual utility U^s, the overall network utility over slices from set S is

$$U^{Net} = \sum_{s \in S} U^s \,. \tag{3.29}$$

3.2.4 Performance Analysis and Results

To evaluate the robustness of the proposed model, a simulation environment is constructed in MATLAB software. In this simulation environment, a small network is considered with $S_n = 5$ number of slices, where the number of resource parameters considered in the matrix (**A**) is N=5. The parameters considered are required data rates =$[8, 100]$ kb/s, latency-sensitivity=$[10, 200]$ ms, priority=$[1, 5]$ ms, bandwidth = $[10, 100]$ MHz, and acceptable packet loss ratio =$[10^{-2}, 10^{-7}]$ (as in GSMA, 2019). $\ell_v = [1, 20]$ Erlangs, and $\ell_d = [1, 14]$ Erlangs are the considered range of network loads with regards to voice and data communication. The minimum considered data rates for the communication of voice and data in the network are $\Gamma^{(v)} = 8kb/s$ and $\Gamma^{(d)} = 16kb/s$, respectively (as considered by Walingo and Takawira, 2014; Caballero et al., 2018).

Figure 3.3 to Figure 3.6 illustrate the analytical performance of the proposed slice allocation and admission control model. The performance is assessed through GoS, network utilisation, as well as throughput, and mean delay. The results obtained are also compared with the results of existing models found in the literature. The existing models of a similar context are named as Connection Admission Control (CAC) (Walingo and Takawira, 2014), Signal to Interference ratio (SIR) (Anand and Chockalingam, 2003), Static Slicing (SS) and Central decision-based Network Slice (NES) allocation and admission control (Caballero et al., 2018).

Figure 3.3 shows the probability of a user being blocked upon admission to the network. The result obtained is compared with existing admission control models (i.e. SIR (Anand and Chockalingam, 2003), and CAC (Walingo and Takawira, 2014). It can be seen that the blocking probability of the proposed model, obtained from (3.21), is significantly low compared to SIR and CA. For example, at 7 Erlangs DSAAC attained a gain of 30% on CAC and 48% on SIR in terms of blocking probability of the network. Blocking probability increases with an increase in the network load. The achieved gain of DSAAC in terms of blocking probability of the network at 12 Erlangs on CAC is 27% and SIR is 39%. The reason behind this gain in blocking probability from the proposed model is the consideration of the cumulative probability of N resource demand characteristics within the matrix **A**, and reconfigurable

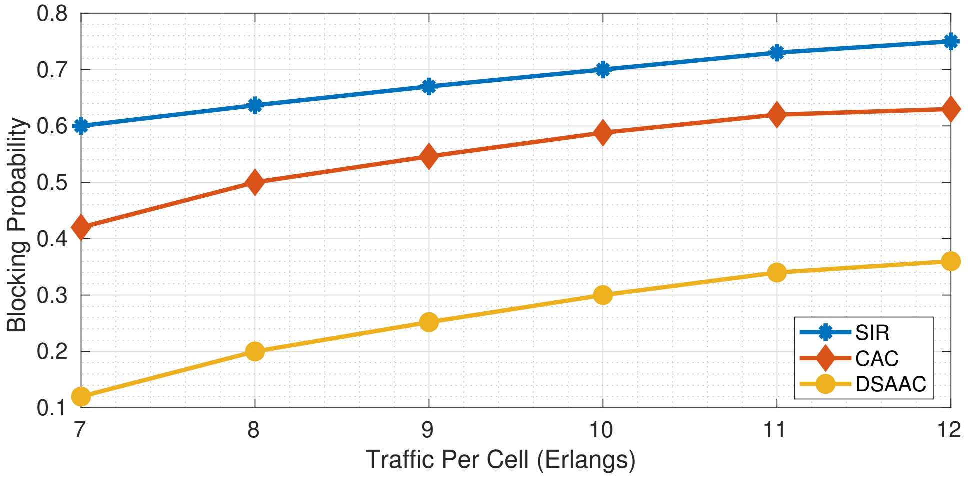


FIGURE 3.3: Network GoS comparison of DSAAC with SIR (Anand and Chockalingam, 2003), and CAC (Walingo and Takawira, 2014) based admission control models.

slice resource bounds for inter-slice admission and resources allocation. However, the existing models consider only the interference-to-signal (I/S) ratio and delay as admission parameters. Congestion occurs due to load more than networks capacity employing existing models, which increases the blocking probability of these networks due to resource scarcity. Hence, the consideration of multiple resource characteristics in the matrix **A**, ensures a wider range of flexibility in admission control and resources allocation.

Figure 3.4 shows the network utilisation, as obtained from (3.29), at various loads. The result obtained for utilisation for the proposed model is compared with the result of SS and NES models from the literature (Caballero et al., 2018). The network utilisation obtained from DSAAC is significantly greater compared to its counterparts. For example, at the lowest network load, the achieved gain in utilisation from the proposed model on NES is 1.08% and on SS is 4.3%. Based on efficient admission control, with an increase in network load, network utilisation also increases. Thus at 100% load, utilisation gain from the proposed model is 1.32% on NES and 5% on SS. The existing models implemented the greedy approach for user admission to the network with consideration of user rate requirements, weight and slice share. However, DSAAC considers the aggregate user resource demand from a slice (3.25). This consideration is for efficient admission control and slice utilisation with the flexibility of intra-slice, as well as inter-slice resource allocation. Hence, efficient slice utilisation from the proposed model leads to greater network utilisation.

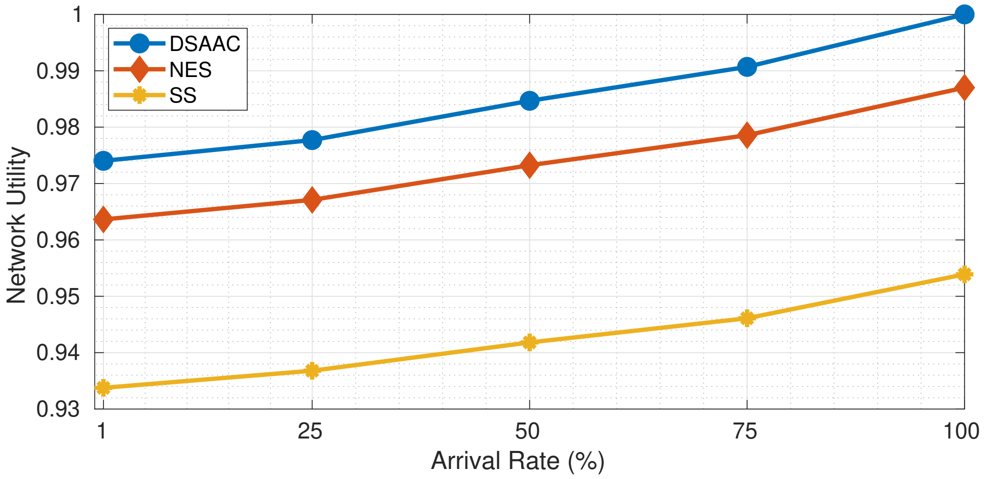


FIGURE 3.4: Network utilisation vs. traffic arrival with respect to aggregate resource requirement.

Figure 3.5, and Figure 3.6 illustrates the acquired mean delay and network throughput for various load. It can be seen that the proposed model performs efficiently over the increased load due to the lower number of user requests blocked from the network. Lower probabilities of blocking are achieved by the provided flexibility in the admission process, which is slice elasticity and inter-slice admission and resource provisioning. Thus, by reducing the blocking probabilities and number of retransmissions, the proposed model reduces mean delay and increases network throughput. Moreover, the mean delay in communication that is obtained by (3.23) in DSAAC is remarkably lower compared to its counterparts. For example, at 10 Erlangs, the mean delay by DSAAC is 0.043s, whereas by CAC and SIR it is 0.061s and 0.122s, respectively. Thus the achieved gain in mean delay from DSAAC is 41.86% on CAC and 64.75% on SIR, which results in better network QoS and user-demanded QoE.

The achieved throughput in Fig 3.6 increases with network load. However, it starts to decline in the case of a load that is greater than the network capacity, and congestion results. However, the throughput achieved by the proposed model is superior to CAC and SIR-based allocation models. For example, at 13 Erlangs, the throughput gain achieved by DSAAC on CAC is 8% and on SIR it is 13.2%. This throughput gain is due to lower blocking probabilities and the number of retransmissions. In summary, the proposed dynamic user admission and slice allocation

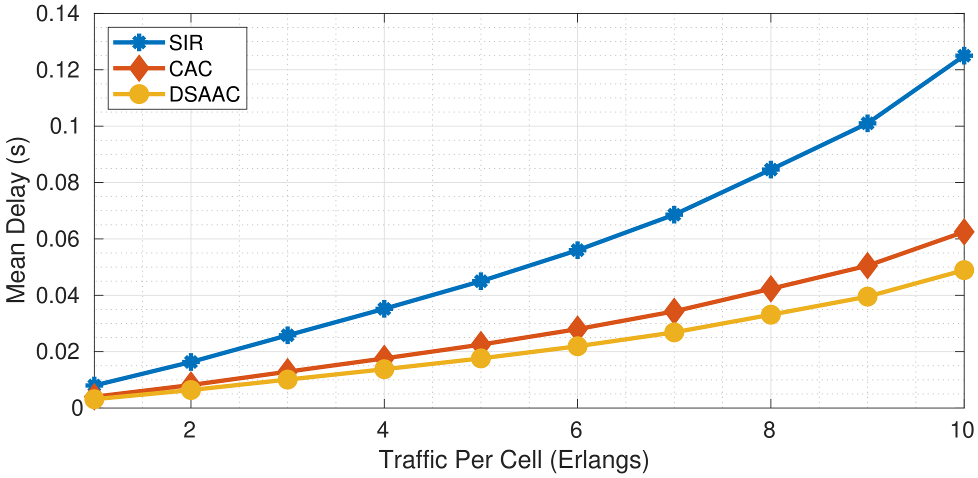


FIGURE 3.5: Mean delay measurements over the increased load.

model is better than the existing models in terms of lower blocking probability, efficient network utilisation, better throughput, and reduced mean delay for future wireless networks.

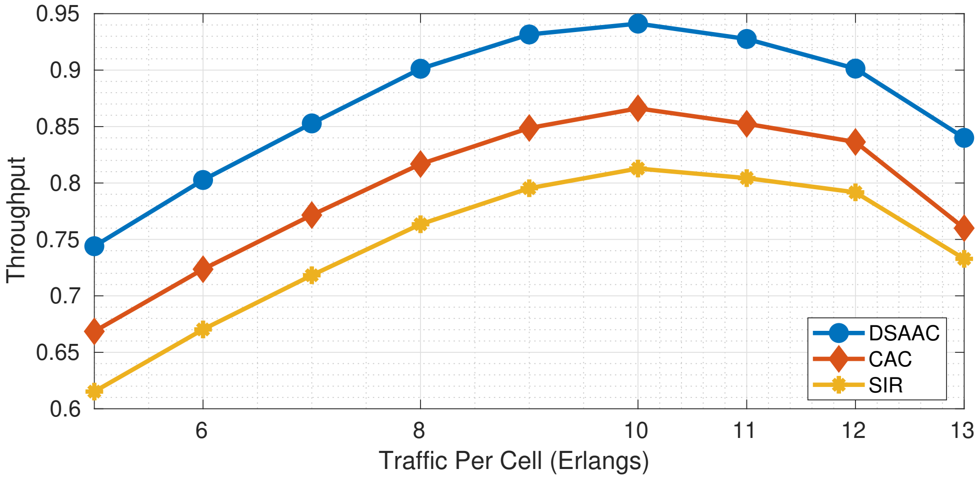


FIGURE 3.6: Average system throughput over the increased load.

3.3 Signalling and Admission Control (SAC) Model

Compared to previous cellular technologies, 5G technology has revolutionised wireless and cellular communication through the support of a massive number of connectivity demands, provisioning of much higher data rates, and lower latency. Lower latencies are essential for ubiquitous computing, including IP multimedia subsystem (IMS), VoLTE, VoWiFi, autonomous smart, and critical applications (Andrews et al., 2014; Alliance, 2015). Due to advances in technology, virtual industries have significantly impacted the networks because of increased usage through both individuals and machines. Therefore, application heterogeneity in the network is creating divergent signalling traffic that will generate a huge wave of communication (Aguwamba, 2020).

Moreover, the exponentially rising demand for network access cannot be ignored. Forecasts and market reports predict the proliferation of smart and heterogeneous services will lead to millions of devices being deployed, each requiring continuous network connectivity. *Ericsson Mobility* report on 5G development predicts that there will be approximately 550 million 5G subscriptions by the end of the decade (Gozalvez, 2017). This continuously increasing connectivity demand, along with the number of devices, may have a significant impact on user QoE and network performance because of increased network complexity, congestion, overloading, and control signalling storms in the access and core network (Brown, 2012; Al-Fuqaha et al., 2015; Gupta and Jha, 2015).

Today, the research community's emphasis is to resolve the issues through international standardisation, especially the massive amount of device connectivity and diameter signalling in latency-critical scenarios. For example, a novel solution known as *Diameter Protocol*, had been proposed for control plane signalling optimisation in LTE networks (Ewert, Norell, and Yamen, 2012). The *Diameter Signalling Controller (DSC)* is the principal component of this protocol. The controller operates to balance the load between the mobility management entity (MME) and the home subscriber server (HSS) of the LTE network on a priority basis. Another effort by Trivisonno et al. (Trivisonno et al., 2018) has been the development of an E2E connectivity model for signalling optimisation in a 5G IoT network. In this model, the IoT devices are gathered based on their homogeneous characteristics and association with a single base station to build a virtual device class. For each device class, a single default barrier is established in their work for control plane signalling reduction and to overcome the load limitations problem in the data plane. However, this research is limited by prioritisation among device classes, few resource characteristics for classification, and the holding of the barrier by the particular class for a long time. Thus, earlier research has neglected the massive number of device demands, their heterogeneous demand characteristics, and the need for consistent connectivity. Jiang et al. proposed two novel approaches 5G Slice Allocation (5G-SA) and 5G Admission Control Slice Allocation (5G-AC-SA) in (Jiang, Condoluci, and Mahmoodi, 2016). 5G-SA consider intra/inter slice priority for resource allocation among slices. In this approach, resources are allocated to the users based on their demand and earned revenue in the slice. In this model, resources are allocated to users of high data rate demand in case of heavy load on the network, which degrades overall network QoS due to lower satisfaction of users with low data rate demand. 5G-AC-SA consider users' intra-slice priority and demanded QoE for slice allocation and admission control. This approach gives better fairness in resource allocation as compared to 5G-SA. However, in the case of massive connectivity demand, network QoS degrades due to more number of lower priority users and slices rejection from the network.

Many solutions have been proposed in recent research to manage the massive amount of future connectivity from the limited network capacity; for example, the use of Millimetre Wave (Rappaport et al., 2013; Roh et al., 2014; Qiao et al., 2015), Intelligent Cognitive Radio (Wang et al., 2019; Yu, Lin, and Chen, 2019) in cellular networks, especially in unmanned aerial vehicles (UAVs) communication. These approaches have enormous potential for implementation in cellular wireless radio communication. This is because of the provisioning of lower latency, higher bandwidth and data rate, on-demand deployment, and reconfiguration flexibility. Unlike radio communication, they have numerous shortcomings, such as interference, propagation losses, limited coverage, greater power consumption, channel characterisation, and hardware limitations (Farooq and Rather, 2019; Zhang et al., 2019a; Arjoune and Kaabouch, 2019; Kakalou et al., 2017). Due to the limitations of these approaches, a substantial amount of research solutions are demanded in cellular wireless communication, especially for signalling optimisation and admission control in the case of the massive amount of future traffic demand. Hence, it is expected that research efforts in signalling can efficiently balance available capacity against the demand to ensure efficient admission control. This research work aims to provide an adequate admission control model that can control the signalling storm generated by the massive amount of future network heterogeneous traffic. The proposed approach first analyses the incoming network traffic signalling, as shown in Figure 3.7. From the acquired analysis, signalling requests are processed, where the requests are clustered based on user-application demand co-relation and homogeneous characteristics to ensure efficient admission control within slice QoE bounds. Eventually, the aim is to ensure efficient load balancing and resource utilisation, and reduce network complexity, cost and congestion, which will enhance the system GoS for the massive amount of future heterogeneous network traffic. Accordingly, the following major contributions of this research work are:

- Proposed a novel optimised signalling and admission control model for 5G and beyond networks.
- Derived a usage-specific demand analysis and clustering scheme based on user-application demand co-relation and homogeneous characteristics, to ensure a given level of network QoS and user QoE.
- Proposed slice QoE elasticity bounds from QoE-based resource allocation metric and a set of QoE aware optimised admission control algorithms. These algorithms achieve efficient slice allocation and user admission by assessing the capability of the slice QoE elasticity.
- Presented the theoretical basis of performance enhancement due to the utilisation of the proposed model. In addition, the robustness of the proposed model is analysed with admission gain, signalling reduction, and acquired QoE from the network, and compared with the existing models in the literature.

3.3.1 SAC System Model

The system model presented in section 3.2.1 is enhanced with the presented signalling and admission control, also known as SAC, model, as shown in Figure 3.7.

This model proposes a dynamic pre-clustering analysis of the incoming heterogeneous traffic characteristics. This analysis is to manage the massive amount of signalling load on slices to enhance slice resource utilisation and slice GoS. In the proposed model, gNodeB or (R)AN is a node to aggregate the signalling in the wireless network. In this research, the *Signalling and Admission Controller* is assumed to be an essential aspect of future networks for both multi-node and multivariate environments. The controller normalises and centralises the massive amount of signalling concerning the specific network requirements to ensure efficient large-scale network management. The controller is comprised of three major components: the Signalling (Pre_Cluster) analyser, the Signalling processing (Clustering) system, and the QoE-based signalling and admission controller, as shown in Figure 3.7. The function of the analyser is to analyse the signalling requests of the incoming traffic concerning their demand characteristics for processing and admission control. Afterwards, the signal processing system processes those signals and gathers them into a cluster based on their homogeneous demand and device heterogeneity. The resulting optimised service signals are examined by the core network entities (such as AMF and NSSF) to execute authentication and slice selection. After slice selection, a network slice instance is created from the sub-slice instances based on the management and orchestration policies and agreed service level agreement (SLA) with the tenant. Thus a particular network slice instance is configured based on the user-required

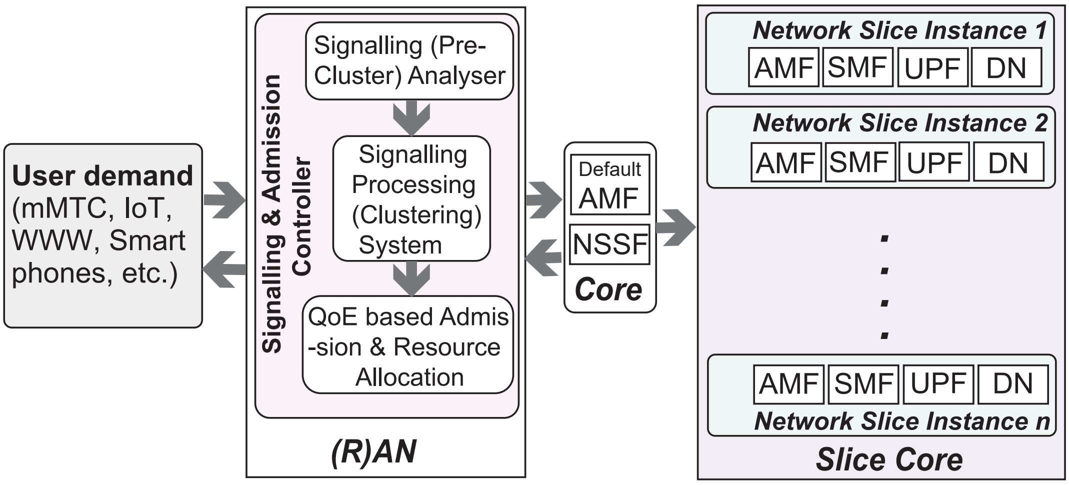

User demand (mMTC, IoT, WWW, Smart phones, etc.)
Signalling & Admission Controller
Signalling (Pre-Cluster) Analyser
Signalling Processing (Clustering) System
QoE based Admis -sion & Resource Allocation
(R)AN
Default AMF
NSSF
Core
Network Slice Instance 1
AMF
SMF
UPF
DN
Network Slice Instance 2
AMF
SMF
UPF
DN
AMF
SMF
UPF
DN
Network Slice Instance n
Slice Core

FIGURE 3.7: End-to-End service management and orchestration via the proposed signalling and admission control (SAC) model

QoE. On the successful configuration of slice instances, the mobile network operator (MNO) populates network types for its users for service provisioning. After that, the session management function (SMF) establishes an E2E session between the user and data network for communication via the user plane function (UPF).

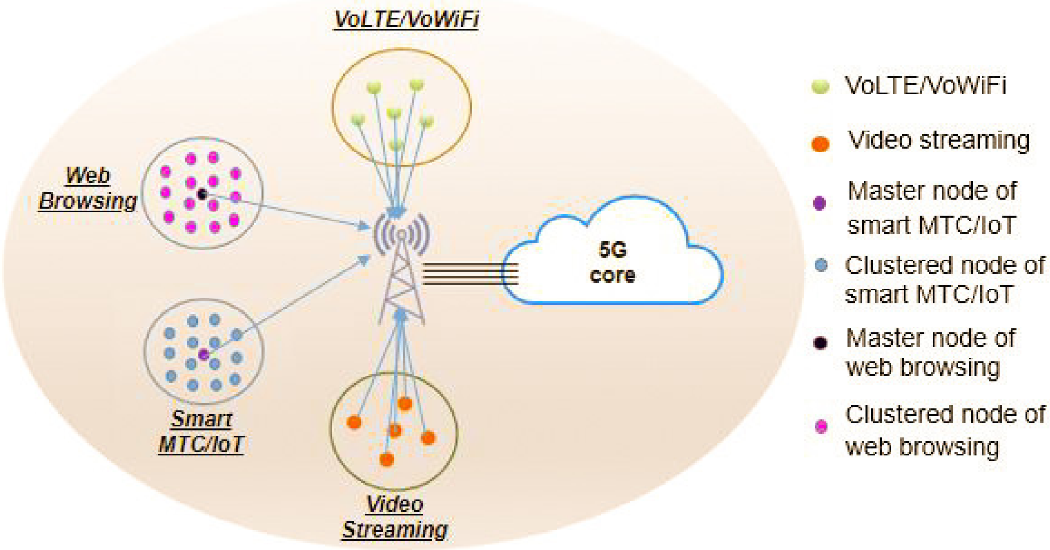

VoLTE/VoWiFi
Web Browsing
5G core
Smart MTC/IoT
Video Streaming
VoLTE/VoWiFi
Video streaming
Master node of smart MTC/IoT
Clustered node of smart MTC/IoT
Master node of web browsing
Clustered node of web browsing

FIGURE 3.8: Scenario of signalling optimisation via clustering

Network Design and Statistics: In this model, a small network is considered with M number of users, symbolised as a set $\mathbf{U} = \{U_1, U_2, \ldots, U_M\}$, as illustrated in Table 3.2. Let us assume that this small network can support K heterogeneous applications simultaneously, which is denoted as the user-specific application set $\Lambda = \{1, 2, \ldots, K\}$. Such applications include live streaming and video calling, using smart help points, and Web browsing to obtain the required information, as shown in Figure 3.8. Each application is supposed to have a set of specific resource characteristics, symbolised as set $\mathbf{J} = \{1, 2, 3, \ldots, N\}$. Such characteristics include data rate, bit error rate (BER), latency, demand density and service priority. Moreover, each resource characteristic has a set of acceptable states defined on the basis of users' preferences, denoted as $\mathbf{S} = \{St_1, St_2, St_3, \ldots, St_N\}$. Accordingly, the possible distinct service requests for a Λ_k heterogeneous application will be denoted as the set $\mathbf{Z}_k = \{Z_1, Z_2, Z_3, ...Z_N\}$, whereby the cardinality of $\mathbf{Z}_k$ (i.e. $|S|^{|J|}$) is the total possible service requests for a Λ_k application, where $\Lambda_k \in \Lambda$.

TABLE 3.2: SAC model key symbols and definitions

Symbols	Definitions
$\mathbf{U}$	Set of M number of users
$\mathbf{X}$	Demand matrix of M users
$\mathbf{X}_R$	Ranking-based clustered demand matrix
c_k	kth clustered request
r_m	mth user required resources
C^s_{up}	sth slice uplink capacity
$C^s_{up(sig)}$	sth slice uplink signalling capacity
$QoE^s_{k(\gamma)}$	sth slice application-specific minimum guaranteed QoE bounds
$QoE^s_{k(r)}$	sth slice application-specific maximum guaranteed QoE bounds
$Q^s_{m(r)}$	
	mth user-desired QoE
$Q^s_{m(\gamma)}$	mth user minimum agreed QoE

3.3.2 Proposed SAC Model Schema

To establish a connection for communication, each user sends a signalling request to the access network node. Users who belong to a similar application and have similar service demands send homogeneous signals to the network for service provisioning. This massive amount of homogeneous signalling generates a burden on the network functions, which induces greater latencies in communication due to congestion. While service provisioning of a massive amount of heterogeneous traffic, it is essential to ensure efficient network resource utilisation and that the network GoS does not drop below a certain level. Therefore, this work proposes a dynamic signalling optimisation and admission control architecture, known as SAC. The proposed architecture contains three main components: the Signalling (Pre_Cluster) analyser, the Signalling processing (clustering) system, and the QoE-based signalling and admission controller, as described in Figure 3.7. The functionality of each component in correspondence with the proposed architecture is explained in detail in the following section.

3.3.2.1 Signalling (Pre_Cluster) Analyser: Whenever a user u_m accesses the network with a connectivity request, a desired resource request is issued, symbolised as $\mathbf{x}_{mj} = [x_{m1}, x_{m2}, \dots, x_{mN}]$, where $u_m \in \mathbf{U}$, $\mathbf{x}_{mj} \in \mathbf{X}_k$ and $j \in \mathbf{J}$ from set $\mathbf{Z}_k$ for its Λ_k application. Moreover, for simplicity, two acceptable states are considered in set $\mathbf{S}$

for each application characteristic. With help of the signalling analyser, the respective application assessor (or the cluster head that can be in the RAN or the master node in the smart networks) assesses each resource characteristic of the user's request to populate a respective coefficient within the matrix $\mathbf{X}_k$ as a row index in (R)AN, as shown in (3.30).

$$\mathbf{X}_k = \begin{bmatrix} x_{11} & x_{12} & x_{13} & \cdots & x_{1N} \\ x_{21} & x_{22} & x_{23} & \cdots & x_{2N} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{M_k1} & x_{M_k2} & x_{M_k3} & \cdots & x_{M_kN} \end{bmatrix} . \tag{3.30}$$

For admission control, the signalling analyser also accesses the current network circumstances. On the uplink communication in a 5G network, the slice capacity C^s_{up} is the sum of total capacity reserved for signalling, represented as $C^s_{up(sig)}$, and data transmission, represented as $C^s_{up(data)}$, respectively.

$$C^s_{up} = C^s_{up(sig)} + C^s_{up(data)} \, . \tag{3.31}$$

In a 5G network, the observed slice capacity $C^s_{up(obs)}$ is also the sum of the total capacity utilised by the signalling and data transmission:

$$C^s_{up(obs)} = C^s_{up(obs\_sig)} + C^s_{up(obs\_data)} \, . \tag{3.32}$$

Now, $C^s_{up(obs\_sig)}$ over the application set Λ can be determined as follows:

$$C^s_{up(obs\_sig)} = \sum_{\Lambda_k \in \Lambda} |\mathbf{U}_k| r_{Sig(k)} = \sum_{\Lambda_k \in \Lambda} Rank(\mathbf{X}_k) r_{Sig(k)} \, , \tag{3.33}$$

where, $\mathbf{U}_k$ represents the set of users associated with the kth application, and $\mathbf{U}_k \subseteq \mathbf{U}$. $Rank(\mathbf{X}_k)$ determines the matrix $\mathbf{X}_k$ rank with regards to the kth application before clustering. $r_{Sig(k)}$ is the desired user resource demand for signalling that represents the NAS PDU in 5G and beyond networks (3GPP, 2018c). $C^s_{up(obs)}$ should not exceed the reserved slice capacity C^s_{up} but instead be equal to or less than C^s_{up}, as shown below:

$$C^s_{up} \geq C^s_{up(obs)} \, . \tag{3.34}$$

However, the observed slice capacity will increase exponentially with an increase in slice load, especially in the case of a massive traffic load on the slice.

3.3.2.2 Signalling Processing (clustering) System: Users who belong to a similar application and have similar service demands send homogeneous signals to the network for service provisioning. In case of a massive amount of connectivity demand, the observed slice capacity can exceed the reserved slice capacity. Moreover, the massive amount of homogeneous signalling generates a burden on the network through poor GoS, inefficient resource utilisation, and congestion, which induces greater latencies in communication. Therefore, while service provisioning of a massive amount of heterogeneous traffic, it is essential to ensure efficient network resource utilisation and that the network GoS does not drop below a certain level. This can be modelled as an optimisation problem. whereby the objective is to minimise $C^s_{up(obs)}$ in a way that it should not exceed C^s_{up}. Mathematically, it can be written as

$$\begin{aligned} \min \quad & C^s_{up(obs)} \leq C^s_{up}, \\ \text{s.t.} \quad \sum_{u_m \in \mathbf{U}} C^s_{up(obs\_sig)}(u_m) \quad & \leq \quad \sum_{u_m \in \mathbf{U}} C^s_{up(sig)}(u_m)\,, \\ \sum_{u_m \in \mathbf{U}} \alpha_{(u_m)} \quad & \leq \quad 1\,. \end{aligned} \tag{3.35}$$

The observed signalling capacity due to the users of set **U** should not exceed the overall reserved uplink signalling capacity. Moreover, α represents all the signalling demands from the users of set **U** should be admitted by the network.

In view of the massive volume of requests, the implementation of an efficient clustering approach is essential to reduce signalling overheads in the access network. Therefore, this work proposes a dynamic signalling processing system for incoming traffic, which will facilitate a massive amount of heterogeneous traffic with the potential of redundant signalling reduction through a ranking-based clustering approach. In the processing system, each subsequent user request is assessed based on users' application-specific service signalling and resource characteristics. This assessment implements a comparative analysis approach to determine whether or not the requests are homogeneous in demand with respect to device type and resource demand characteristics. If the mth user request, symbolised as $\mathbf{x}_{mj}$, is similar to the $(m-1)$th user request, symbolised as $\mathbf{x}_{(m-1)j}$, the application assessor will group them into a cluster, due to their homogeneous demand characteristics, as illustrated in Algorithm 4. Given that clustering, a respective coefficient will be populated in the rank matrix $\mathbf{X}_R$ as a single row entry for the respective cluster demand. Thus, the updated rank matrix $\mathbf{X}_{R_k}$ for the Λ_k application will be as follows:

$$\mathbf{X}_{R_k} = \begin{bmatrix} x_{1_k1} & x_{1_k2} & x_{1_k3} & \cdots & x_{1_kN} \\ x_{2_k1} & x_{2_k2} & x_{2_k3} & \cdots & x_{2_kN} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ x_{R_k1} & x_{R_k2} & x_{R_k3} & \cdots & x_{R_kN} \end{bmatrix}, \tag{3.36}$$

whereby, for M users of the kth application, rank R of the respective matrix $\mathbf{X}$ will be equal to or less than M (denoted as $R_k \leq M_k$) regarding the guaranteed soft, best-effort, and hard QoE user traffic demand.

Similarly, $\mathbf{A}$ matrix for K number of heterogeneous applications will be as follows:

$$\mathbf{A} = \begin{bmatrix} a_{1_11} & a_{1_12} & a_{1_13} & \cdots & a_{1_1N} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ a_{R_11} & a_{R_12} & a_{R_13} & \cdots & a_{R_1N} \\ a_{1_21} & a_{1_22} & a_{1_23} & \cdots & a_{1_2N} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ a_{R_21} & a_{R_22} & a_{R_23} & \cdots & a_{R_2N} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ a_{1_K1} & a_{1_K2} & a_{1_K3} & \cdots & a_{1_KN} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ a_{R_K1} & a_{R_K2} & a_{R_K3} & \cdots & a_{R_KN} \end{bmatrix}. \tag{3.37}$$

Based on their homogeneity, requests are grouped as a common scaled entity in the application cluster. Thus, after clustering, the clustered signalling capacity $C^s_{up(cls\_sig)}$ will be less than or equal to the $C^s_{up(obs\_sig)}$, which, in turn, reduces $C^s_{up(obs)}$ to make it approximately equal to or less than C^s_{up}:

$$C^s_{up(cls\_sig)} = \sum_{\Lambda_k \in \Lambda} Rank(\mathbf{X}_{R_k}) r_{Sig(k)}, \tag{3.38}$$

where the rank of the matrix $\mathbf{X}_{R_k}$, denoted as $Rank(\mathbf{X}_{R_k})$, determines the total number of possible signals of the kth application after clustering. Due to the homogeneous resource demand, it is also assumed that r_{Sig} is similar for the signalling requests of all applications. Accordingly, the matrix $\mathbf{A}$ rank is less than the number of users accessing the network at that instant for service provisioning of their heterogeneous applications from set Λ. Thus, (3.38) can be as rewritten as

$$C^{s}_{up(cls\_sig)} = r_{Sig} \sum_{\Lambda_k \in \Lambda} Rank(\mathbf{X}_{R_k}) = Rank(\mathbf{A}) r_{Sig} \,. \tag{3.39}$$

Likewise, redundancy reduction from the observed data transmission ($C^{s}_{up(obs\_data)}$) can also reduce observed uplink capacity.

Algorithm 4: Clustering for signalling optimisation

```
Input: $User\_List_k = \{x_1, x_2, x_3....x_{M_k}\}$ is in order, $Z_k$.
Output: $Cluster\_List_k = \{x_1, x_2, x_3....x_{R_k}\}$ is in order with respect to the user
  arrival time.
begin
    $count = 1$
    $User\_List_k \neq \emptyset$
    $Cluster\_List_k \longleftarrow \emptyset$
    Every element of $User\_List \in Z_k$
    for $(i = 0, i < User\_List_k.length, i++)$ do
        if $(i = 0)$ then
            Update $Cluster\_List_k[i] \longleftarrow User\_List_k[i]$
        end
        if $(i \neq 0)$ && $(User\_List[i] == User\_List[i-1])$ then
            Increment the count.
            Scale the value of $User\_List[i]$ by count.
            Cluster $User\_List[i]$ with $User\_List[i-1]$ with scaling factor
            Update $Cluster\_List_k[i-1] \longleftarrow User\_List[i]$
        else
            Update $Cluster\_List_k[i] \longleftarrow User\_List[i]$
        end
    end
end
```

From the Algorithm 4, in time 0 to t, the first user from the list $User\_List$ to access the access node for service provisioning of the applications (e.g. for smart MTC and Web browsing), will be considered a cluster head. With a guarantee to abide by the security policy, it is supposed that the cluster head has the right to assess user requests from similar device types and application characteristics in the neighbourhood. Moreover, based on homogeneity among requests of application k, requests will be included in a unique cluster for their admission to the respective slice and resources allocation. Thus, each cluster represents a unique application and device type. All the unique clusters from the application k are listed in the cluster list, denoted as $Cluster\_List$, for admission control. Now, a single signalling request on each cluster would be sent to the core network for multiple users.

3.3.2.3 QoE-based Admission Control and Resource Allocation: After signalling optimisation, now the desired QoE of the cluster request, denoted as $QoE^s_{(c_k)}$ (where $c_k \in Cluster\_List_k$) is assessed by the slice QoE bound set, which is symbolised as $[QoE^s_{k(\gamma)}, QoE^s_{k(r)}]$ for service provisioning of the kth application, as shown in Algorithm 5. $QoE^s_{k(\gamma)}$ and $QoE^s_{k(r)}$ are the respective slice application-specific minimum and maximum guaranteed QoE bounds for cluster request admission to the network. For QoE based admission control, $QoE^s_{(c_k)}$ is computed from the individual user-desired QoE within the cluster c. Whereby, user-desired QoE is acquired as:

$$Q^s_{m(r)} = (\frac{r^s_m}{R^s_{Max}}) + \beta_m + \iota_m + \rho_m \leq 1\,, \tag{3.40}$$

where, R^s_{Max} represents the maximum sth slice available resources, r^s_m represents the mth user-desired resource demand, ρ, ι and β represent the user priority, latency sensitivity, and acceptable user-application-specific BER, respectively. Their statistical values are normalised between zero and 1 for simplicity. When the network is experiencing a higher volume of traffic (e.g. during peak hours and special occasions), the network will become saturated such that it will not be able to allocate users' desired resources for their service provisioning. In instances such as this, the user will acquire the guaranteed QoE on the provisioning of its agreed minimum resource (denoted as γ^s_m) demand from the slice. This can be illustrated as:

$$Q^s_{m(\gamma)} = (\frac{\gamma^s_m}{R^s_{Min}}) + \beta_m + \iota_m + \rho_m < 1\,. \tag{3.41}$$

Significantly, now the slice QoE over set $\mathbb{U}_s$ with respect to user-desired and guaranteed resource demand can be expressed as follows:

$$Q^s_r = \sum_{m \in \mathbb{U}_s} Q^s_{m(r)} \leq 1 \,, \tag{3.42}$$

$$Q^s_\gamma = \sum_{m \in \mathbb{U}_s} Q^s_{m(\gamma)} \leq 1 \,. \tag{3.43}$$

Likewise, the overall network QoE (denoted as Q) over set $\mathbb{S}$ with respect to the guaranteed or desired QoE along with slice priority (i.e., ρ_s) can be represented by the equations:

$$Q_\gamma = \sum_{s \in \mathbb{S}} \left(Q^s_\gamma\right)^{\rho_s} \,, \tag{3.44}$$

$$Q_r = \sum_{s \in \mathbb{S}} \left(Q^s_r\right)^{\rho_s} \,, \tag{3.45}$$

$$Q = Q_r + Q_\gamma \leq 1 \,. \tag{3.46}$$

The network assigns resources based on a scheduling frame slot t from the total number of T scheduling frames. Therefore, the time average network QoE can be estimated as:

$$\mathbb{E}[Q] = \frac{1}{T} \sum_{t=1}^{T} Q^t \,. \tag{3.47}$$

If $QoE^s_{(c_k)}$ is within the slice guaranteed QoE bounds for serving the particular application, the resource requests of users in that cluster will be assessed to implement efficient resources allocation and admission control, as shown in Algorithm 5. Therefore, to accommodate the user request, the user-required resources, either demanded r or guaranteed γ resources, will be assessed based on slice uplink capacity (i.e. C^s_{up}). Based on available slice resources and the required QoE of the *cth* request, users of the *cth* cluster will be admitted to the admission queue, which is symbolised as $Admit\_List_k$ of the *kth* application. Otherwise, users of the *cth* cluster will be admitted to the backoff queue, symbolised as $Block\_List_k$, of the *kth* application for group backoff. Then, in the next slot, the *cth* cluster request will be reassessed, along with the updated QoE demand characteristics obtained from the estimates. From here on, the list of admitted users will be passed to the core control network functions (AMF, SMF, and UPF) for slice configuration, session establishment, and data transmission.

```
Algorithm 5: QoE-based admission control and resource allocation
Input: $Cluster\_List_k = \{x_1, x_2, x_3....x_{R_k}\}$ is in order with respect to the user
  arrival time. $Cluster\_List$ are in order with respect to the application
  priority. Chose $c_k$ as a $kth$ application cluster request (i.e.
  $c_k \in Cluster\_List_k$). $QoE^s_{(c_k)} \neq 0$, $QoE^s_{k(\gamma)} \neq 0$, $QoE^s_{k(r)} \neq 0$, $C^s_{up} \neq 0$.
  $Admit\_List_k = 0$, $Block\_List_k = 0$. slot $0 \longrightarrow t$.
Output: $Admit\_List_k \neq 0$, $Block\_List_k \geq 0$.
begin
    for $(i = 0, i < Cluster\_List_k.length, i++)$ do
        $c_k \longleftarrow Cluster\_List_k[i]$
        if $(QoE^s_{k(\gamma)} < QoE^s_{(c_k)} \leq QoE^s_{k(r)})$ then
            for $(j = 0, j < c_k.length, j++)$ do
                $m_k \longleftarrow c_k[j]$
                if $(m_{k(r)} \leq C^s_{up})$ or $(m_{k(\gamma)} \leq C^s_{up})$ then
                    Add mth request in admission queue at $Admit\_List_k[i]$
                    update $C^s_{up}$
                else
                    Add mth request in $back\_off$ queue at $Block\_List_k[i]$
                end
            end
        else
            Reassess and update $QoE^s_{(c_k)}$
        end
    end
    Send signalling via $Admit\_List_k$ towards AMF.
    Populate respective coefficient in matrix $A$ with respect to application
      priority and user arrival time.
    Send group $back\_off$ signal via $Block\_List_k$.
    Queue $back\_off$ users in $t \longrightarrow t + 1$ slot at front of $Cluster\_List_k$.
end
```

3.3.3 Performance Analysis and Results

The robustness of the proposed model is assessed through a simulated network environment that supports four different use cases: VoLTE/VoWiFi, video streaming, smart MTC, and Web browsing. Therefore, a dense virtual network with various system parameters has been developed in MatLab software for evaluation. This validation illustrates how the proposed model helps to reduce signalling redundancy that leads to reduced congestion and how admission control and QoE can be improved in a denser network. The parameters considered are acceptable user-application-specific BER ($\beta = [10^{-9}, 0]$), latency sensitivity ($\iota = [100, 400]$ ms), priority ($\rho = [1, 4]$), and desired data rate ($r = [1, 25]$ Mbps) (as given in GSMA, 2019). The results obtained, as shown in Figure 3.9 to Figure 3.12, illustrate the improved efficiency and robustness of the proposed model and are compared with the results of relevant models found in the literature (Jiang, Condoluci, and Mahmoodi, 2016; Trivisonno et al., 2018).

Figure 3.9 illustrates uplink signalling capacity $C_{up(sig)}$ vs. traffic load from 5G heterogeneous applications. It can be seen that the control signalling capacity obtained from (3.39) in the proposed model is significantly lower than that of the existing 5G-mIoT model, as given in (Trivisonno et al., 2018). The performance in terms of redundant signalling reduction achieved from the proposed ranking-based clustering approach is 96% over the entire range of $\mathbf{U}$. This performance is a result of each cluster representing a unique application and device type. So, the number of clusters will remain the same based on device type and the number of heterogeneous applications. However, based on the incoming load on the network, the number of requests within a cluster can vary. Hence, the signalling redundancy reduction from the cluster reduces the unnecessary and massive amount of control messages flowing into the network. Otherwise, the aforementioned huge volume of message flow into the network creates congestion and reduces network QoS.

Similarly, Figure 3.10 represents the admission gain (or the number of admissions) on service provisioning of various heterogeneous applications vs. their considered priority in a dense network. In this work, priority is determined with respect to the reliability demanded by the application. Hence, the considered order is VoLTE/VoWiFi > Live Streaming > MTC > Web Browsing. When the network is fully loaded, cluster requests from VoLTE/VoWiFi and Live Streaming applications take priority according to demanded QoE. However, a few requests belonging to MTC and Web Browsing were rejected due to resource scarcity. However, when the load is greater than the available capacity, the proposed model admits

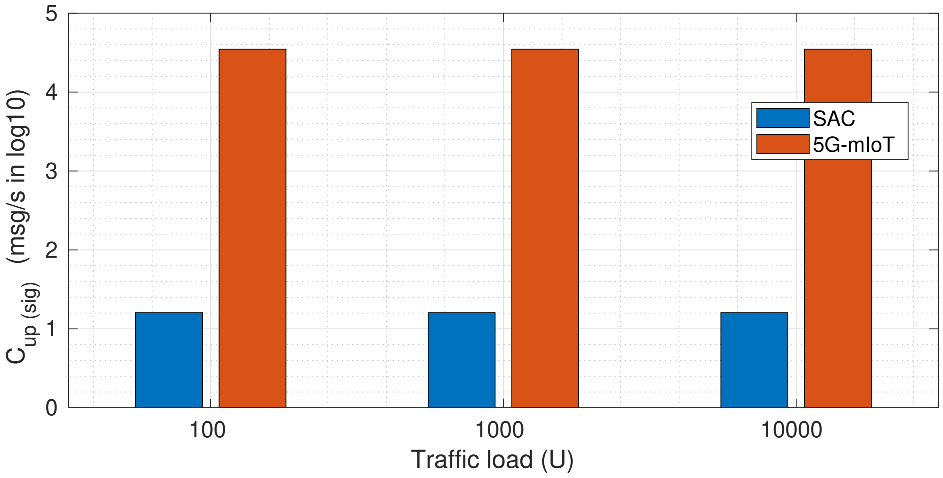


FIGURE 3.9: Computation of core signalling over varying user density

the request according to their guaranteed QoE bounds (between desired and minimum guaranteed QoE demand) to reduce the rejection ratio within the network. Thus, at $\mathbf{U} = 20000$, the admission gain or number of admissions obtained from VoLTE/VoWiFi users is greater than that of its counterparts (i.e. > 90%). However, the admission gains or the number of admissions obtained from live streaming, MTC, and Web browsing users are 90%, 83%, and 74%, respectively. This trend in gain is because of the different preferences of the application. Therefore, clustered users of an application with higher priority always take priority for admission to the network. Moreover, consistency in the achieved admission gain at heavy load is due to the consideration of clustering and adaptability of slice elasticity for resource allocation. Subject to availability of resources, slice elasticity guarantees a lower cluster user backoff from the slice in the case of massive network traffic load.

Figure 3.11 shows the average QoE obtained by (3.46) on various heterogeneous applications from 2000 users. The proposed SAC model performance was compared with that of its counterparts i.e. 5G Slice Allocation (i.e. 5G-SA) and 5G Admission Control Slice Allocation (i.e. 5G-AC-SA) models, as in (Jiang, Condoluci, and Mahmoodi, 2016). Compared to its counterparts, a significant difference is seen in the QoE achieved by the proposed model. Moreover, the achieved QoE of live streaming is better than that achieved through MTC and Web browsing requests, due to priority considerations. In the case of heavy load on the network, user requests

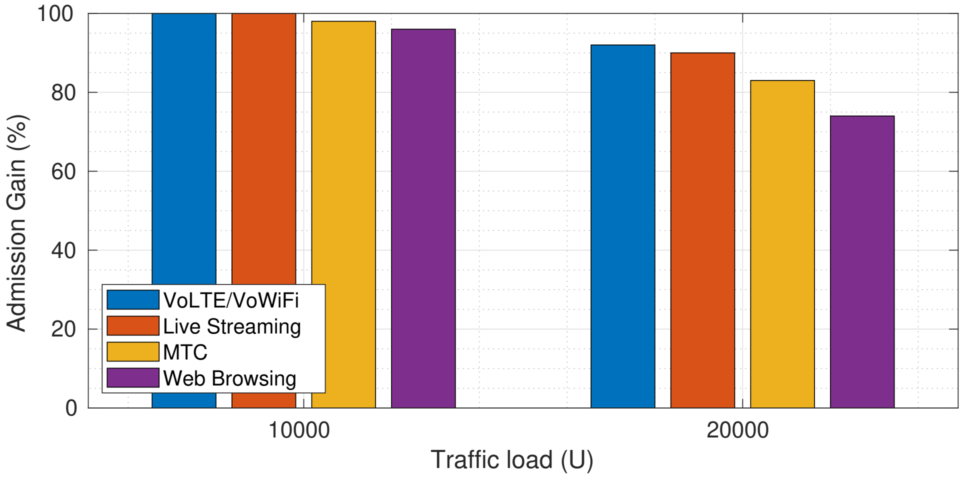


FIGURE 3.10: Computation of admission gain over the massive load

for live streaming need to be accommodated at their minimum guaranteed QoE demand Q_γ based on priority considerations, where $\gamma = 0.5$. In terms of operational cost, the particular service slice QoE bounds will be dynamically adjusted to admit user requests into the slice to lower the rejection ratio. This extension in bounds in turn increases the average QoE by reducing blocking probability. A similar trend can be seen in serving MTC and Web-browsing cluster users by the proposed model. However, the achieved QoE of MTC and Web-browsing users in existing models are reducing, due to resource scarcity in their slices and higher blocking probability. In the proposed model, each user in the cluster will be treated fairly by the slice through dynamic bounds adjustment in an unforeseen situation. Besides clustering, this adaptability eventually increases overall QoE and slice utilisation. Consequently, the proposed SAC model guarantees higher user admission and enhanced slice QoE compared to existing models.

Figure 3.12 demonstrates the QoE of the entire network at various traffic loads. The results obtained from the proposed model are also compared with those of its counterparts. It can be seen that at minimum guaranteed demand $\gamma_s = 0.5$, the gain in achieved QoE is high, that is 31.3% and 19.4% on 5G-SA and 5G-AC-SA based models upon the arrival of $\mathbf{U} = 100$. A similar trend can be seen at $\mathbf{U} = 30$, and $\mathbf{U} = 60$. This is because the proposed model admits users into the slice through clustering and slice elasticity to enhance achieved QoE. Therefore, all users, along with their desired QoE demand, will be admitted in the form of clusters and served from the slice subject to the availability of resources. However, user admission and

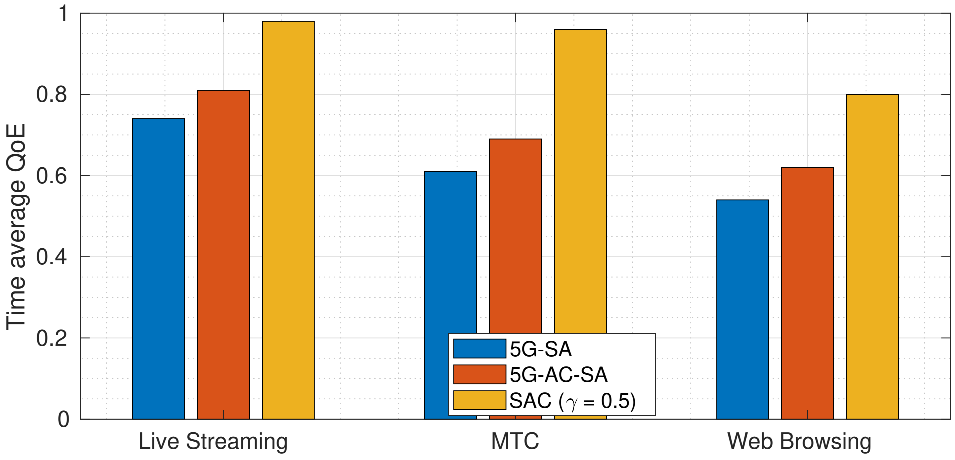


FIGURE 3.11: Average QoE over heterogeneous 5G applications

their associated QoE continuously decrease as traffic load increases because of the massive number of users requesting connectivity. Thus, an overloaded network creates competition among the users for admission and resource allocation, which leads to users being backed off due to congestion or resource deficiency. However, compared to existing models, the proposed model performs efficiently at increased load and shows meaningful QoE due to flexible QoE bounds.

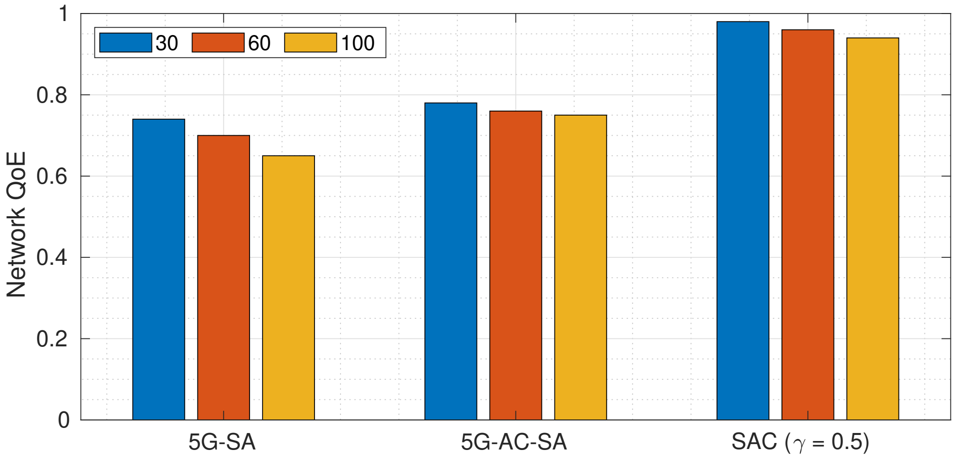


FIGURE 3.12: Computation of network QoE over 5G-SA/5G-AC-SA, and SAC.

To summarise, the proposed dynamically signalling analysis and QoE-based admission control model performs efficiently with regards to reduced redundant signalling in the access network, better resource and network utilisation, and better user QoE in a dense network.

3.4 Summary

In this chapter, integrated slice allocation, admission control and signalling redundancy reduction strategies have been presented. First, DSAAC model has been presented in Section 3.2. In this model, a decision matrix, along with the unified cost estimation function is proposed for dynamic slice allocation and admission control. This model considers varied user demands, as well as multiple real-time network resource characteristics for optimised admission control to enhance the network GoS. These characteristics include user and slice bandwidth, data rate, priority, latency sensitivity, and cost revenue. Moreover, to maximise resource utility, adjustable minimum and maximum slice resource bounds have also been derived. In the case of user blocking from the primary slice due to congestion or resource scarcity, inter-slice admission control and resource allocation and adaptability of slice elasticity have been proposed.

Moreover, an access network control signalling redundancy minimisation SAC model is presented in Section 3.3 for 5G and beyond networks. In this model, a three-stage approach involving pre-clustering analysis, usage-specific clustering, and a signalling optimisation and admission mechanism, has been introduced. This model deals with the usage and user-device-specific heterogeneity in a single-layer approach rather than a two-layer approach. The proposed unsupervised learning-based clustering approach reduces the additional burden on the network in terms of unnecessary resource utilisation and computational time by reducing redundancy in the signalling. Moreover, a set of optimisation algorithms are also proposed to attain efficient slice allocation and users' QoE enhancement via assessing the capability of slice QoE elasticity.

Eventually, the proposed models are evaluated in terms of GoS, network utility, mean delay, throughput, uplink signalling load, and admission gain. The results obtained are also compared with those of relevant models in the literature and suggest that the proposed DSAAC and SAC models outperform their existing counterparts. From the comparative results, it can be seen that a flexible but efficient decision metric can be obtained through the accumulation of user demand and network resource characteristics. The proposed models provide explicit definitions of the requirements of network slice characteristics, which leads to better admission control and resource utilisation to ensure enhanced network QoS and user-acquired QoE. Accordingly, a summary of an analysis of DSAAC and SAC models is also presented in Table 3.3.

TABLE 3.3: Summary of analysis of DSAAC and SAC Models

Analysis Measures	**DSAAC Model**	**SAC Model**
Admission objective	Objective of the DSAAC model is unified slice allocation and network GoS enhancement.	Objective of the SAC model is signalling redundancy minimisation and users' QoE enhancement.
Slice elasticity	Slice reconfigurable resource bounds for inter-slice admission and resource allocation	Slice reconfigurable QoE bounds for intra-slice admission and resource allocation
Tenancy	Multi-tenant support	Multi-tenant support
Slicing domain	E2E slice management and orchestration support	E2E slice management and orchestration support
Admission strategy	Single objective optimisation	Single objective optimisation
Optimisation Algorithm	Normal equation	Unsupervised learning
Admission domain	Intra and inter-slice admission and resource allocation	Intra slice admission and resource allocation
Admission efficiency	The average admission efficiency of the DSAAC model over a fully loaded network is 88.4% due to lower blocking probabilities from the network.	The average admission efficiency of the SAC model over a fully loaded edge network is 91.12% due to signalling redundancy minimisation in the edge access network.

Chapter 4

Admission Control with Multi-Objective Optimisation

4.1 Introduction

Continuous technological advancements and the massive amount of heterogeneous service support have made today's network environment more complex in terms of management and operation. Compared to conventional approaches, it has been recognised for decades that multi-objective optimisation algorithms can quickly and accurately find efficient solutions to the problems of complex networks. Multi-objective optimisation approaches have been successfully applied to provide optimal solutions to several non-deterministic polynomial-time (NP) hard problems in wireless communications systems, such as spectrum allocation (Zhao et al., 2009; Gözüpek and Alagöz, 2011; Shami, El-Saleh, and Kareem, 2014), resource scheduling (Gu et al., 2015), channel assignment (Xu et al., 2012), indoor and outdoor tracking (Gharghan et al., 2015), and call admission control (Jain and Mittal, 2016). Such approaches can produce a solution individually or be used in combination with other approaches such as machine learning. Therefore, due to their ability to solve complex problems, optimised admission control models have been presented in this chapter. In Section 4.2 of this chapter, an edge redundancy minimisation and admission control model, also known as the *E-RMAC* model, is presented for signalling optimisation to ensure better network QoS and user-demanded QoE using efficient admission control and resource allocation. A forecasting and fuzzy-logic-based admission control model, also known as the *FAC* model, is presented in Section 4.3

for resource allocation and admission control on the forecasted demand in the 5G open-RAN network. In this model, a non-dominated sorting genetic algorithm is employed for optimal network selection on the forecasted demand. Similarly, for bottleneck congestion control in 5G and beyond network, slice congestion and admission control (SCAC) model is presented in Section 4.4, that uses the optimisation and machine learning approaches such as unsupervised learning, reinforcement learning and transfer learning. Finally, a summary of the chapter is given in Section 4.5.

4.2 Edge Redundancy Minimisation and Admission Control (E-RMAC) Model

Mobile edge computing (MEC) is expected to be a promising key enabler for provisioning latency-sensitive heterogeneous services within future wireless networks; for example, autonomous drones for live streaming, autonomous vehicle control, and health-monitoring systems for emergencies. MEC was first introduced by the *European Telecommunications Standard Institute* (ETSI) in 2015 (Hu et al., 2015) and extends the capabilities of cloud computing by moving resources closer to the network edge to provide lower latency and higher reliability for highly demanded and latency-sensitive applications (Abbas et al., 2017; Pham et al., 2020; Qu et al., 2020). Therefore, it is expected that MEC will be an essential part of future network architecture to improve overall cellular network performance; for instance, AI-enabled edge architecture to fulfil the visions of 6G (e.g. seamless connectivity, ultra-low latency, ultra-high data rates, and reliability) (Russell and Norvig, 2010; Qu et al., 2020).

However, due to the continuously increasing demand for various heterogeneous services in wireless-cellular networks, MEC is facing numerous challenges in traffic flow management (Roman, Lopez, and Mambo, 2018). One such example is the need for efficient MEC network management that minimises latency in network data and control planes, and maximises E2E link efficiency. Recent research in (Liu and Zhang, 2018) and (Sun et al., 2020) applied the concepts of data offloading at edge or cloud nodes. During data transmission, data offloading is mainly for the service provisioning of critical applications from the edge network while keeping latency within acceptable bounds. In contrast, the current MEC provides limited or no access to the core control and management functions of the control plane in the cellular network (Roman, Lopez, and Mambo, 2018). The limited access of MEC to the core network functions (NFs) reduces overall network performance. For example, if device density is more than the capacity of the edge network, a huge volume of traffic from these devices flows into the core network. The serving network edge may collaborate with other edges in the neighbourhood to offload a certain amount of data onto them. However, the neighbouring network edges have a similar kind of signalling from their associated traffic. So, redundancy occurs from similar signalling, which leads to inefficient resource utilisation in the collaborated environment (Chen et al., 2018; Ullah et al., 2019). For example, 5G joint network slicing and edge computing model is proposed by the authors in (Xiang et al., 2019) for latency minimisation. Branch and bound method is been used by the authors for E2E slice creation. In their model, edges are ranked based on the link and computational capabilities along with least latency. Data is been offloaded to the neighbouring edges based on the latency requirement of the application. However, offloading signalling or data from primary edge also generate signalling and data redundancy in the neighbouring edge. In the cellular network, such issues have gained significant attention from the research community around the globe. For example, a novel solution proposed for signalling optimisation is *Diameter Protocol*, which carries out CP signalling of the LTE network (Ewert, Norell, and Yamen, 2012). Another effort to handle CP signalling redundancy is the E2E connectivity model proposed by the authors in (Trivisonno et al., 2018) for massive IoT in 5G networks.

Similarly, asymmetry in traffic flow and its management at the network edge causes congestion (Cao et al., 2019). For example, when a user connectivity request is received, the edge node accesses the user profile managed by a centralised core network user unified data management (UDM) function to retrieve the user data for the offered UDM services, including subscriber data management, authentication, and event exposure. The UDM offers services through numerous service operations. For example, a subscriber data management service is offered by *Get*, *Subscription*, *Unsubscription*, *Modify*, and *Notification* service operations (3GPP, 2020). For the provisioning of the offered services, these operations perform several signalling operations between the edge and core NFs. The core NFs also exchange user information with each other in case of modification in privileges or policies, or notification of a user's subscription or unsubscription from a particular application (Behrad et al., 2020). The massive volume of user requests could have similar signalling and response in the core network, which would create communication overhead on the link capacity of the core network. Such overheads might include 100% in a baseline LTE/EPC system, due to bearer establishment, and 40% in a 5G system, due to slicing and device-based classification (Trivisonno et al., 2018). This leads to degradation of overall network performance by inducing substantial latency and congestion, which is potentially intolerable for latency-sensitive applications (Emara, Filippou, and Sabella, 2018).

Considering these issues, a hash-based grouping scheme is proposed by the authors in (Hung, Hsieh, and Wang, 2017) for traffic flow management in the MEC system. Wang et al. (Wang and Cai, 2019) proposed an intelligent edge management and optimisation model for latency-critical applications of 5G networks. To reduce the communication overheads in 5G mIoT networks, Cao et al. proposed a fast-authentication and data transfer scheme (Cao et al., 2019). A bankruptcy game-based resource allocation algorithm for 5G Cloud-RAN slicing is proposed by the authors in (Jia et al., 2018). In this approach, user groups are created for admission to the network based on the Lloyd Shapley approach. In this work, a user would be a part of a group, if the user adds more benefits to the slice. However, a greedy-based admission control strategy may not always be optimal. Such a policy makes the greedy decision on the spot to achieve the objective, such as increasing the admitted requests to earn more revenue from the network but also creates a lot of congestion due to signaling redundancy. Moreover, equal ratio strategy (EO) and traffic proportion (TP) approaches are employed by the authors as a benchmark in this work. The equal ratio strategy means resources are allocated equally to different slices. This result in over and under resource utilisation in the network on frequently varying demand. The traffic proportion strategy means that resources are allocated to different slices in proportion according to their random requirements. These approaches degrade the network QoS through inefficient resource allocation. Existing research into MEC highlights the problem of latency minimisation through data offloading and traffic flow management between access and edge nodes. However, how to minimise latency and congestion between the edge and core control NFs of cellular networks is still an open issue.

To ensure efficient resource utilisation and traffic management, this work endeavours to prevent control signalling storms from entering the core network. This could be achieved by moving the essential NFs of the core network onto the edge for efficient admission control. The edge core functions would acquire limited privileges from the core of the cellular network for security assurance. Accordingly, the following major contributions of this work are:

- A novel edge architectural model known as edge redundancy minimisation and admission control (E-RMAC) has been proposed in this work to support massive amount of connectivity demand and to reduce signalling redundancy in future core networks.
- A k-mean- and ranking-based clustering approach has been implemented in this work, along with genetic optimisation for control signalling redundancy reduction and efficient admission control.
- The proposed model is assessed through performance evaluation measures such as latency, link efficiency, admission control, and fairness of resource allocation. The outcomes of the proposed model are compared with the outcomes of existing models found in the literature.

4.2.1 E-RMAC System Model

It is assumed that future networks would be more efficient at managing varied traffic flow according to network capacity. One example is MEC, which can manage network traffic dynamically with efficient admission control and resource utilisation (Pham et al., 2020). For CP signalling optimisation and efficient admission control, a novel edge architectural model has been presented in this work for future core networks, as illustrated in Figure 4.1. In this architectural model, the RAN is the aggregation point that receives signalling requests in addition to user-application-specific demand for resources. In this work, the service signalling and resource demand for latency-sensitive applications are sent to the core network through the traditional RAN or edge RAN. The edge controller of the RAN is a crucial entity that analyses and centralises application signalling and resource demand to ensure optimal network management at all times. The controller consists of three major components: pre-clustering demand analysis and categorisation, a demand processing (clustering) system, and admission control and resource allocation. As the name suggests, the demand analyser at the network edge analyses and categorises user-application-specific service signalling and resource demand for clustering. The proposed processing system of the edge controller processes service signalling and resource demands with respect to their homogeneous characteristics and clusters them for signalling optimisation and admission control. The clustered signals are then sent to the core network UDM to fetch user-service-specific profiles onto the edge for admission control and resource allocation. However, in the core UDM, user profiles are also clustered based on homogeneous control information and signalling response. The clustered responses in the UDM help to reduce control signalling redundancy among the core NFs. Otherwise, this may lead to resource inefficiency and congestion in the core network.

The essential core NFs are configured at the edge to support the massive volume of latency-sensitive applications with faster authentication, admission control, and resource allocation, as shown in Figure 4.1. The proposed core edge NFs are termed *Edge Access and Mobility Management Function* (eAMF), *Edge Session Management Function* (eSMF), *Edge Network Slice Selection Function* (eNSSF), *Edge Unified Data Management* (eUDM), *Edge User plane Function* (eUPF), and *Edge Serving/Packet Gateway* (eSGW/ePGW). Within the core network, these functions would be managed by the proposed *Edge Management Function* (EMF). With the assistance of the proposed EMF, the core edge NFs would acquire limited privileges from the core PCF to ensure network security (Roman, Lopez, and Mambo, 2018).

Key Notations and Description: In this work, the deployment of an edge network is considered to serve an urban area. It is assumed that the deployed edge network can serve U total number of users, symbolised as $\mathcal{U} = \{1, 2, \cdots, U\}$ of URLLC, mMTC,

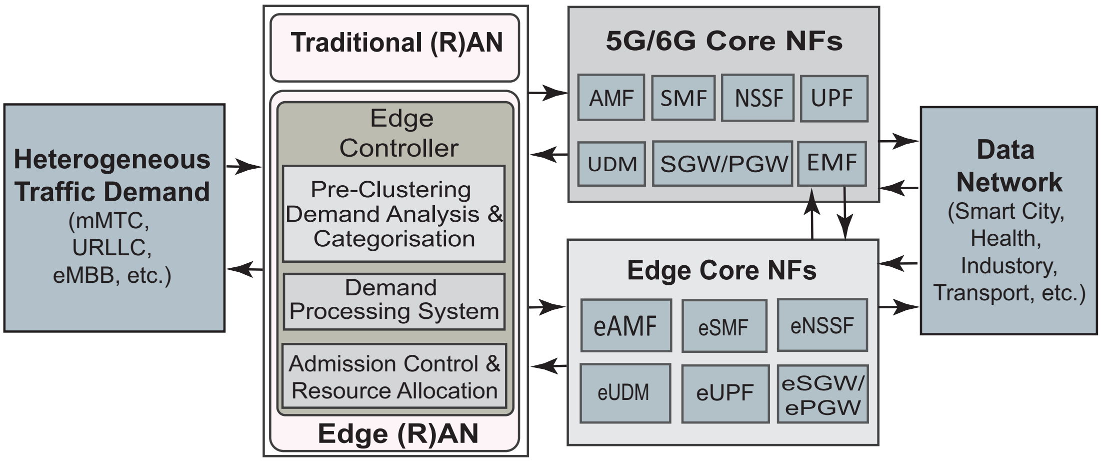

Heterogeneous Traffic Demand
(mMTC, URLLC, eMBB, etc.)
Traditional (R)AN
Edge Controller
Pre-Clustering Demand Analysis & Categorisation
Demand Processing System
Admission Control & Resource Allocation
Edge (R)AN
5G/6G Core NFs
AMF
SMF
NSSF
UPF
UDM
SGW/PGW
EMF
Edge Core NFs
eAMF
eSMF
eNSSF
eUDM
eUPF
eSGW/ ePGW
Data Network
(Smart City, Health, Industory, Transport, etc.)

FIGURE 4.1: Proposed edge architecture for future core networks

and eMBB applications. Each user associated with this network has M number of service and resource demand characteristics, represented by $\mathcal{M} = \{1, 2, \cdots, M\}$. The service demand, symbolised as a set $\mathcal{S} = \{1, 2, \cdots, S\}$, and $\mathcal{S} \subset \mathcal{M}$, determines the signalling request to the UDM-offered services. Moreover, these UDM services are offered by a set of service operations, represented by $\mathcal{P} = \{1, 2, \cdots, P\}$. Likewise, the resource demand, denoted as a set $\mathcal{L} = \{1, 2, \cdots, L\}$, determines the L number of resource request of a particular application, where $\mathcal{L} \subset \mathcal{M}$. Each of the resource characteristics belonging to $\mathcal{L}$ is independent and different from other resource characteristics based on the particular application. Such characteristics could be data rate, available bandwidth, acceptable jitter, packet loss, etc. Moreover, it is also assumed that each user can be connected simultaneously to a maximum of K heterogeneous applications, expressed as a user-specific application set $\Lambda = \{1, 2, \cdots, K\}$. Key symbols included in this work are illustrated, along with their definitions, in Table 4.1.

4.2.2 Proposed E-RMAC Model Schema

In this section, I am going to present a systematic scheme of the proposed E-RMAC model, for future core networks, which is shown in Figure 4.2 and also discussed in detail in the following subsections.

TABLE 4.1: E-RMAC model key symbols and definitions

Symbols	Definitions
$\mathcal{U}$	Set of users in the network
$\mathcal{M}$	Set of services and resource demands
$\mathcal{S}$	Set of UDM-offered service signalling
$\mathcal{P}$	Set of UDM-offered service operations
$\mathcal{L}$	Set of resource demand represents
Λ	Set of applications
$\mathbf{V}_e$	Edge user-demand matrix
$\mathbf{A}_r$	Resource-demand matrix
$\mathbf{A}_s$	Service matrix for signalling
C_{up}	Total uplink capacity
$C_{up(sig)}$	Reserved uplink signalling capacity
$C_{up(obs\_sig)}$	Observed uplink signalling capacity

4.2.2.1 Pre-clustering Demand Analysis and Categorisation

The incoming user demand for applications belonging to set Λ is continuously assessed by the edge node. This assessment is for optimal network management in latency-sensitive and massive-device-connectivity situations. Therefore, when U number of users access the edge network for service provisioning of the κth application ($\kappa \in \Lambda$), a demand matrix $\mathbf{V}_{e(U \times M)}$ is constructed for these users by the edge controller. Each element, v_{um}, belonging to $\mathbf{V}_e$ determines the user-application-specific service, as well as resource demand characteristics from set $\mathcal{M}$.

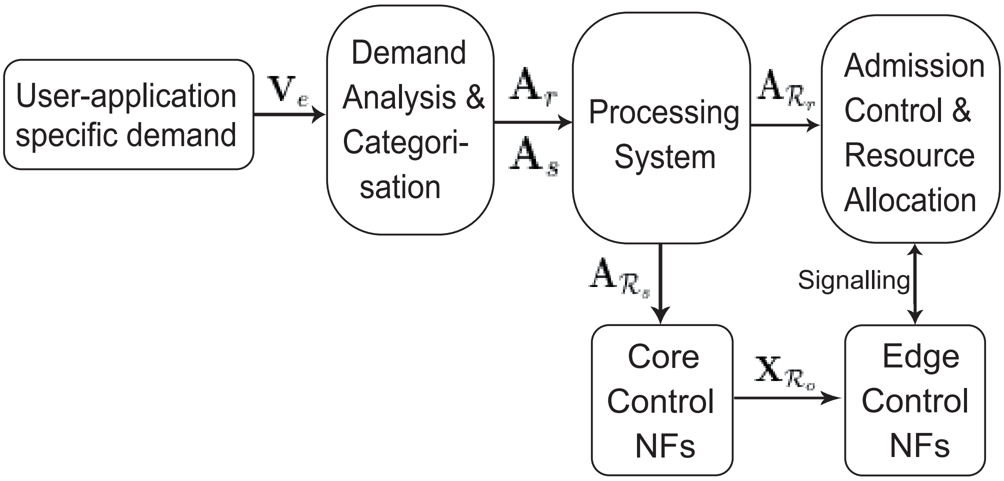

User-application specific demand
$\mathbf{V}_e$
Demand Analysis & Categori-sation
$\mathbf{A}_r$
$\mathbf{A}_s$
Processing System
$\mathbf{A}_{\mathcal{R}_r}$
Admission Control & Resource Allocation
$\mathbf{A}_{\mathcal{R}_s}$
Signalling
Core Control NFs
$\mathbf{X}_{\mathcal{R}_o}$
Edge Control NFs

FIGURE 4.2: Systematic diagram of the proposed E-RMAC model

The pre-clustering system detaches the signalling requests of the UDM service from the resource demand of the kth application available in $\mathbf{V}_e$ for clustering and signalling optimisation, as shown in Figure 4.2. For the UDM sth service signalling, the service operations are populated as a row entry in the service matrix $\mathbf{A}_{s(U\times P)}$ for that particular edge user, where $s_{up} \in \mathbf{A}_s$ represents user-application-specific service demand. Likewise, resource demand $\mathcal{L}$ for application κ contains a set of user-application-specific resource demand characteristics. The edge-user-required resource demand characteristics are populated as a row entry within the resource demand matrix $\mathbf{A}_{r(U\times L)}$, where $r_{ul} \in \mathbf{A}_r$ represents the user-application-specific resource demand. Isolated services and resource demand characteristics are passed to the processing system for signalling optimisation and efficient admission control, as explained in detail in the next subsection.

4.2.2.2 Demand Processing (Clustering) System for Signalling

Optimal resource allocation plays an important role in network QoS and user QoE. However, supporting the massive number of service demands in an edge network for device connectivity becomes challenging for the network itself, because of the limited link capacity (bandwidth) of the edge, latency sensitivity of the application, and offloading constraints (Emara, Filippou, and Sabella, 2018). To address this challenge, a demand processing system is presented in this section. With the help of optimisation and clustering approaches, the proposed demand processing system can serve a massive number of connectivity requests with efficient admission control and resource utilisation.

A) Clustering for Capacity Optimisation: A huge amount of control signalling traffic in the core network (more than core capacity) creates signalling overheads, which reduces network performance on ineffective resource utilisation (Hung, Hsieh, and Wang, 2017). For communication, the uplink network capacity, represented as C_{up}, is the sum of the total capacity reserved for transmission of signalling and data in the network, as follows:

$$C_{up} = C_{up(sig)} + C_{up(data)} \,. \tag{4.1}$$

In the network, when the demand on the κth application is heavy, the reserved capacity should be equal to or greater than the observed traffic capacity, represented as $C_{up(obs)}$. This observed capacity is the total capacity consumed during transmission by signalling and data for the κth application in the network, as illustrated below:

$$C_{up(obs)} = C_{up(obs\_sig)} + C_{up(obs\_data)} \leq C_{up}\,, \tag{4.2}$$

whereby, $C_{up(obs\_sig)}$ for sth service signalling and rth resources demand signalling for the set $\mathcal{U}$ can be determined as follows:

$$C_{up(obs\_sig)} = \sum_{u \in \mathcal{U}} \alpha_{(u,s)} s_{Sig(u)} + \sum_{u \in \mathcal{U}} \alpha_{(u,r)} r_{Sig(u)}\,, \tag{4.3}$$

whereby, α would be set to 1 only if the sth service or rth resource demand is granted, otherwise 0. The uth user-desired-service signalling ($s_{Sig(u)}$) and resource demand signalling $r_{Sig(u)}$ are measured in bits; for instance, NAS PDU in 5G and beyond networks (ETSI, 2020). On the network edge, the observed capacity increases exponentially with increasing application demand, which may cause inefficient resource utilisation. Thus, ineffective capacity utilisation could be modelled as an optimisation problem. The objective of the mentioned optimisation problem is to meet the signalling demand coming from $\mathcal{U}$ such that the overall uplink capacity, $C_{up(sig)}$, is utilised efficiently. This can be written mathematically as follows:

$$\begin{array}{lrcl} \min & \sum_{u=1}^{U} C_{up(obs\_sig)}, & & \\ \text{s.t.} & \displaystyle\sum_{u=1}^{U} C_{up(obs\_sig)} & \leq & C_{up(sig)}\,, \\ & \displaystyle\sum_{u=1}^{U} \alpha_{(u)} & \leq & 1\,, \end{array} \tag{4.4}$$

where, $u \in \mathcal{U}$, $|\mathcal{S}| = 1$ and $|\mathcal{L}| = 1$ for simplicity. The observed signalling capacity due to the users of set $\mathcal{U}$ should not exceed the overall reserved uplink signalling capacity. Moreover, all the signalling demands from the users of set $\mathcal{U}$ should be admitted by the edge controller. In view of the massive volume of requests and their latency constraints, the implementation of an efficient clustering approach is essential to fetch user profiles onto the edge, speed up the admission process, and reduce signalling overheads in the core network.

To efficiently process service and resource demand signalling, a ranking-based clustering technique has been adopted in this work. Ranking-based clustering is a simple and powerful approach used to compute the similarity index within the cluster (Saxena et al., 2017). The isolated services and resource demand signalling received from the pre-clustering system are grouped, based on their homogeneous demand characteristics, into clusters. This is to fetch the user profile to the edge. Thus, if the uth user service (or resource) demand, $\mathbf{s}_{up}$ (or $\mathbf{r}_{ul}$), is similar to $(u-1)$ user service (or resource) demand, $\mathbf{s}_{(u-1)p}$ (or $\mathbf{r}_{(u-1)l}$), the processing system will group them into a cluster. The respective coefficients for homogeneous signalling or clustered signalling are placed as a single entry for the particular cluster demand in a row of the ranking matrix. After clustering for the sth service and rth resource demand, the updated matrices are called ranking matrices and represented as $\mathbf{A}_{\mathcal{R}_s(R\times P)}$ and $\mathbf{A}_{\mathcal{R}_r(R\times L)}$, respectively. $\mathcal{R} = \{1, 2, 3, \cdots, R\}$ represents the ranking set for service or resource demand signalling with R possible individual clusters. The proposed clustering mechanism reduces the complexity from $\mathcal{O}(U)$ to $\mathcal{O}(R)$, because the respective rank of the matrix, either $\mathbf{A}_{\mathcal{R}_s}$ or $\mathbf{A}_{\mathcal{R}_r}$, will be less than U number of users requests belonging to the demand matrix $\mathbf{V}$ for the κth application. Hence, the massive service signalling or resource demand clustering on the edge helps to reduce overload and congestion in the network by efficient network resource utilisation. Therefore, after ranking-based clustering, the clustered signalling would reduce $C_{up(obs\_sig)}$ to make it approximately equal to or less than $C_{up(sig)}$, as follows:

$$C_{up(obs\_sig)} = Rank(\mathbf{A}_{\mathcal{R}_s})s_{Sig(u)} + Rank(\mathbf{A}_{\mathcal{R}_r})r_{Sig(u)}\,, \tag{4.5}$$

where, after clustering $Rank(\mathbf{A}_{\mathcal{R}_s})$ and $Rank(\mathbf{A}_{\mathcal{R}_r})$ determines the individual cluster signals for the sth service and rth resource demand of the κth application. Now, the ranking-based service signalling is sent to the core network UDM to provide the particular service. Likewise, the ranking-based resource demand signalling for the κth application is sent to the core NFs of the edge for admission control and resource allocation.

Once the core UDM receives the service signalling, it re-clusters user service responses, based on their homogeneity, into a response matrix, represented as $\mathbf{X}_{\mathcal{R}_o(R\times P)}$. Thus, the matrix $\mathbf{X}$ respective rank would be less than $\mathbf{V}$ but equal to or greater than $\mathbf{A}_{\mathcal{R}_s}$. In addition, on UDM-offered services, $\mathcal{S}$, a service profile would be built for clustered users in the network core for ease of signalling and information exchange among the core NFs. The UDM processing of user service signalling (e.g. authorisation, authentication, subscription operations, etc.) is illustrated in Algorithm 6. For a particular service operation, each user privilege would be acquired from the user profile available in the edge UDM. $\mathcal{S}\_List_\kappa$ represents the list of service s clusters maintained by the edge for κth application, where $s_\tau \in \mathcal{S}\_List_\kappa$ and $\tau \in \mathcal{R}$. If all cluster s_τ users have the same access, they would be added to the service list, *Service_List*$_\kappa$ with a unique group ID, *G_id*. Otherwise, users of s_τ would be subclustered into $\mathcal{SS}\_List_\kappa$ on possible responses and added to the *Service_List*$_\kappa$ or *Reject_List*$_\kappa$ for admission control.

B) Clustering for Latency Optimisation: Due to redundant signalling in a dense edge network, the traditional approaches for resource allocation cause ineffective resource utilisation, which results in congestion and latency in communication (Roman, Lopez, and Mambo, 2018). Hence, resource allocation becomes challenging for the edge operators in latency-sensitive scenarios with massive connectivity demand and limited network capacity. The presented latency-minimisation problem can be modelled as an optimisation problem. The objective of the mentioned optimisation problem is to minimise the mean latency, symbolised as $T(\mathcal{N})$, of N optimal number of clusters from a set $\mathcal{N} = \{1, 2, 3, \cdots, N\}$. This can be written mathematically as follows:

$$\begin{aligned} \min \quad & T(\mathcal{N}), \\ \text{s.t.} \quad & \sum_{u=1}^{U} \beta_{(u,r)} \gamma_{(u,r)} \quad \leq \quad \Upsilon_{(r)}\,, \\ & \sum_{u=1}^{U} \sum_{n=1}^{N} \mathcal{U}_{(u,n)} \quad = \quad U\,, \end{aligned} \tag{4.6}$$

where, $u \in \mathcal{U}$, $n \in \mathcal{N}$, $r \in \mathbf{A}_r$. $\beta_{(u,l)} = 1$, only if the demanded rth resource is allocated to the uth user, otherwise 0. γ is the quantity of resource r. The aggregate resources allocated to the set $\mathcal{U}$ should not exceed the total available resources, Υ, of the particular resource r. Each user from set $\mathcal{U}$ should belong to a particular cluster with regard to homogeneous resource demand.

In this work, two popular unsupervised learning algorithms, *k-mean* and *Ranking-based* clustering, have been applied for demand clustering. For the huge number of users and their heterogeneous demand, acquiring an optimal clustering solution of (4.4) and (4.6) with minimum latency and better link efficiency is essential for the

Algorithm 6: Service signalling over clustering

```
Input: Chose $s_\tau$ clustered service demand, $s_\tau \in \mathcal{S}\_List_\kappa$, where,
  $\mathcal{S}\_List_\kappa = \{s_1, s_2, s_3, \cdots, s_{R_s}\}$, $Service\_List_\kappa = Reject\_List_\kappa = \emptyset$.
Output: $Service\_List_\kappa \neq \emptyset$ & $|Reject\_List_\kappa| \geq 0$.
begin
    for ($i = 0, i < \mathcal{S}\_List_\kappa.length, i++$) do
        $s_\tau \longleftarrow \mathcal{S}\_List_\kappa[i]$
        Assign a $G\_id$ to $\mathcal{S}\_List_\kappa[i]$ users
        if (Check user privileges w.r.t. $s_\tau$ matches) then
            Add $G\_id$ of the cluster users of $s_\tau$ demand in $Service\_List_\kappa[i]$ of
              edge.
        else
            sub-cluster $s_\tau$ users into $\mathcal{SS}\_List_\kappa$ w.r.t. possible cluster service
              response.
            for ($j = 1, j < \mathcal{SS}\_List_\kappa.length, j++$) do
                $ss_\tau \longleftarrow \mathcal{SS}\_List_\kappa[j]$
                Assign a $G\_id$ to $\mathcal{SS}\_List_\kappa[j]$ users
                if (check $ss_\tau$ of each associated user) then
                    Add $G\_id$ of the cluster users of $ss_\tau$ demand in
                      $Service\_List_\kappa[i]$ of edge.
                else
                    Add $G\_id$ of the cluster users of $ss_\tau$ demand in
                      $Reject\_List_\kappa[i]$ of edge.
                end
            end
        end
    end
    Send $Service\_List_\kappa$ & $Reject\_List_\kappa$ towards edge core NFs for admission
      control.
end
```

edge network. However, simultaneously acquiring an optimal solution of (4.4) and (4.6) is an NP-hard problem. This is because to acquire link efficiency, if the number of clusters is reduced, latency would also increase due to the increasing number of requests within the cluster. However, (4.4) and (4.6) show convexity in behaviour to acquire global minima during optimisation. Thus, a basic but adequate optimisation method called the *Non-dominated Sorting Genetic Algorithm II* (NSGA-II) has been adopted in this work. By NSGA-II, each of the clusters belonging to set $\mathcal{N}$ would have an appropriate genetic representation on the mean rth resource demand. On each cluster k-mean resource demand, U total number of users are scattered into N number of clusters. The similarity index, denoted as δ, among users of a cluster is acquired by the ranking-based approach. The main objective is to minimise mean latency, $T(\mathcal{N})$, in N number of clusters with respect to the rth resource allocation. The latency experienced can be obtained by obtaining each cluster latency from its associated users acquired resource r, such as:

$$\Delta_n = \sum_{u=1}^{U_n} \frac{\gamma_{(u,r)}}{\Upsilon_{(r)}} - \delta_n \sum_{u=1}^{U_n} \frac{\gamma_{(u,r)}}{\Upsilon_{(r)}} \, , \tag{4.7}$$

where, $\delta_n = [0, 1]$ for cluster n. Now, the mean latency $T(\mathcal{N})$ on N optimal clusters is:

$$T(\mathcal{N}) = \frac{1}{N} \sum_{n=1}^{N} \Delta_n \, . \tag{4.8}$$

Resource demand of users of the nth cluster would be multiplexed into an aggregate resource demand, γ_n, in the MAC layer. Now, along with a brief header for associated user identification, the MAC layer would send the frame to the physical layer for transmission (ETSI, 2020). In the core network, the received MAC layer frame with the optimal number of clusters and their aggregate resource demand would be placed onto the resource list, denoted as R_List_κ for admission control and resource allocation by the eAMF and eNSSF.

4.2.2.3 Admission Control and Resource Allocation

In the core network, each γ_n resource demand belonging to the R_List_κ would be assessed with respect to the available edge capacity, denoted as Υ_e, for cluster user admission control and resource allocation by the eAMF and eNSSF, as illustrated in Algorithm 7. If γ_n demand is within the guaranteed edge QoE bounds, the associated clustered users would be added to the admission queue, represented as $Admit\_List_\kappa$, of κth application. Otherwise, the nth cluster users will go through sub-clustering with regards to the edge available capacity. Hence, the subcluster aggregate demand, $\gamma\gamma_n$, would be populated onto the $\mathcal{RR}\_List_\kappa$. Hereafter, $\gamma\gamma_n$ would be assessed, and the user would be admitted into $Admit\_List_\kappa$ for resource allocation. Otherwise, the subclustered users placed in $Offload\_List$ would be offloaded to the neighbouring edge for resource allocation. Now, the admitted user list would be delivered to the eSMF, eSGW/ePGW, and eUPF for the establishment of a connection, selection of gateway and data transmission.

From $Admit\_List_\kappa$, each cluster demand would be executed locally or offloaded to a neighbouring edge node. The unidirectional E2E latency, represented as T_{ee}, is computed as the sum of transmission time ($T_{(tx)} = \frac{\beta_\gamma}{\mathcal{B}}$), queuing time ($T_{(qu)} = \Delta$), and execution time ($T_{(ex)} = \frac{c_{\gamma n}}{C_{(exe)}}$) of each edge node (Emara, Filippou, and Sabella, 2018). $C_{(exe)}$ is the available computation capacity (CPU cycles per second) of the edge node. c_{γ_n} determines the required computational capacity of γ_n demand. Δ determines the average user queuing latency in a cluster. β_γ and $\mathcal{B}$ are the corresponding data size in bits and available data rate, respectively, in bits per second.

$$T_{ee} = \omega_e \big(T^e_{(qu)} + T^e_{(ex)} + T^e_{(tx)} \big) , \tag{4.9}$$

whereby, the admission index ω would be set to 1 if the task had been admitted to the edge node for further operations, otherwise zero. The degree of network resource utilisation represents fairness in resource allocation among users admitted into the network. Throughput, denoted as $\eta(u)$, acquired by the uth user, is a product of the tolerable latency probability ($p_{T_{ee}}$) and resource allocation probability (p_r). Thus, fairness of resource allocation among set $\mathcal{U}$ users, by Jain's fairness equation (Jain, Durresi, and Babic, 1999), is obtained as:

$$\mathcal{F}_\eta = \frac{\left(\sum_{u \in \mathcal{U}} \eta(u)\right)^2}{U \times \sum_{u \in \mathcal{U}} (\eta(u))^2} . \tag{4.10}$$

Algorithm 7: Admission control with clustering and optimisation

```
Initialisation: Build $\mathcal{R}\_List_\kappa = \{\gamma_1, \gamma_2, \ldots, \gamma_N\}$ via K-mean and
  Ranking-based clustering. Chose $\gamma_n$ as a $kth$ application $n$th cluster
  demand ($\gamma_n \in \mathcal{R}\_List_\kappa$), $\Upsilon_e \neq 0$, $Admit\_List_\kappa = 0$.
for ($i = 0, i < \mathcal{R}\_List_\kappa.length, i++$) do
    $\gamma_n \longleftarrow \mathcal{R}\_List_\kappa[i]$
    Assign a group ID to $\mathcal{R}\_List_\kappa[i]$ users
    if ($\gamma_n \leq \Upsilon_e$) then
        Add the cluster users of $\gamma_n$ demand in $Admit\_List_\kappa[i]$ of edge.
        Update $\Upsilon_e$.
        Compute $\Delta$ of cluster $n$
        Compute $C_{up(obs\_sig)}$ with cluster $n$
    else
        sub-cluster $\gamma_n$ users into $\mathcal{RR}\_List_\kappa$ w.r.t. $\Upsilon_e$
        for ($j = 1, j < \mathcal{RR}\_List_\kappa.length, j++$) do
            $\gamma\gamma_n \longleftarrow \mathcal{RR}\_List_\kappa[j]$
            Assign a sub-group ID to $\mathcal{RR}\_List_\kappa[i]$ users
            if ($\gamma\gamma_n \leq \Upsilon_e$) then
                Add the subcluster users of $\gamma\gamma_n$ demand in $Admit\_List_\kappa[i]$ of
                  edge.
                Update $\Upsilon_e$.
                Compute $\Delta$ of cluster $n$
                Compute $C_{up(obs\_sig)}$ with cluster $n$
            else
                Add the subcluster users of $\gamma\gamma_n$ demand in $Offload\_List[i]$ for
                  offloading to the neighbouring edge network
            end
        end
    end
end
Find opt. $T(\mathcal{N})$ and $C_{up(obs\_sig)}$ via NSGA-II.
Send $Admit\_List_\kappa$ towards edge core NFs.
Send offload demands to the neighbouring edges from $Offload\_List$.
Compute $T_{ee}$ and $\mathcal{F}_\eta$ over $\mathcal{U}$.
```

4.2.3 Performance Analysis and Results

For evaluation of the proposed model's performance, a set of analytical results has been presented in this section. The parameters considered for the numerical analysis in MatLab are $U = [500, 10^4]$, $C_e = 32$ GB, $C_{up} = 500$ MHz, $\gamma_{MTC} = [0.064, 1]$ Mb/s, $\gamma_{URL} = [1, 5]$ Mb/s and $\gamma_{MBB} = [25, 100]$ Mb/s, respectively (as obtained from GSMA, 2019). The results obtained from the proposed model are compared with the results of several alternative approaches from (Trivisonno et al., 2018; Hung, Hsieh, and Wang, 2017; Xiang et al., 2019 and Jia et al., 2018).

4.2.3.1 Impact of Clustering on Latency: Figure 4.3 determines the mean uplink latency measurements for various traffic densities. In this model, the results are acquired through mathematical analysis and compared with hash-based (Hung, Hsieh, and Wang, 2017), and joint network slicing and mobile edge computing (JSNC) (Xiang et al., 2019) models. At the beginning of the results, it can be seen that the achieved latency, as obtained from (4.9), is low due to a lower traffic load in each approach. However, the achieved latency increases with traffic load in all approaches, and their performance gain in terms of acquired latencies also varies. The achieved uplink latency is 1.3 ms at demand $\mathcal{U}_e = 2500$, which is significantly lower than that of JSNC (i.e. 2.2 ms) and hash-based model (i.e. 2.5 ms). This noticeable difference in acquired latency compared to existing models is due to the proposed k-mean- and ranking-based clustering approach for user admission and resource allocation. The proposed approach effectively reduces the signalling overheads on the edge capacity in a dense network and causes relatively low signalling latency. However, the existing models uses the concept of offloading on the neighbouring edges with least latency, as explained earlier. Thus, the k-mean- and ranking-based clustering approach in the proposed model results in better resource management and utilisation.

4.2.3.2 Impact of Clustering on Capacity: Figure 4.4 shows the achieved link efficiency in terms of bandwidth utilisation at the given traffic density. It can be observed that the achieved link efficiency obtained from (4.5) in the proposed model is approximately 95% at full load (i.e. $U = 10^4$), which is markedly higher than that

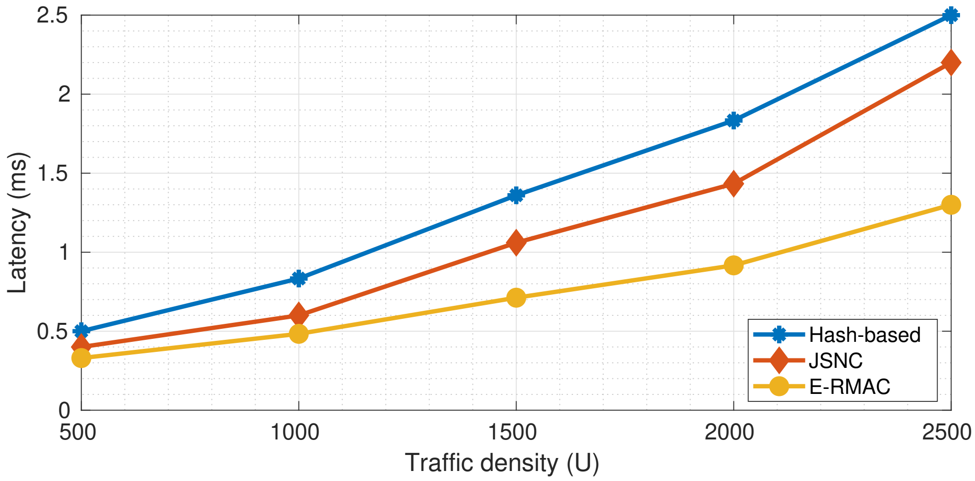


FIGURE 4.3: latency measurements on varying traffic density, $\gamma_{MTC} = [0.064, 1]$ Mb/s, $\gamma_{URL} = [1, 5]$ Mb/s and $\gamma_{MBB} = [25, 100]$ Mb/s.

of the existing model, 5G-mIoT, as given in (Trivisonno et al., 2018). 5G-mIoT employed a random approach for clustering the requests into various physical device classes based on their arrival time and associated base station. The achieved gain in terms of link utilisation efficiency in the proposed model is a result of the k-mean and ranking-based clustering approach, along with the applied optimisation. It is also seen that the proposed clustering approach would save more resources in the case of heavy traffic density compared to lower density on the network. The reason behind this saving is the maximum number of acceptable requests in a cluster that is acquired from optimisation. Thus, if more requests are within a cluster, more resources would be saved. The number of requests in a cluster is obtained from optimisation; therefore, the size of the cluster would not be affected by traffic density. The achieved gain of the proposed model on 5G-mIoT is 19% at $U = 10^3$ and 24% at $U = 10^4$. Hence, the proposed clustering approach would utilise the link efficiently by minimising the flow of redundant signalling into the network core in a dense environment.

4.2.3.3 Impact of Clustering on Admission Control: Figure 4.5 explains the user admission with and without clustering at various traffic loads. Clustered users are admitted into the network in order with respect to their tolerable latency. Application users with higher priority due to latency sensitivity have a preference over other

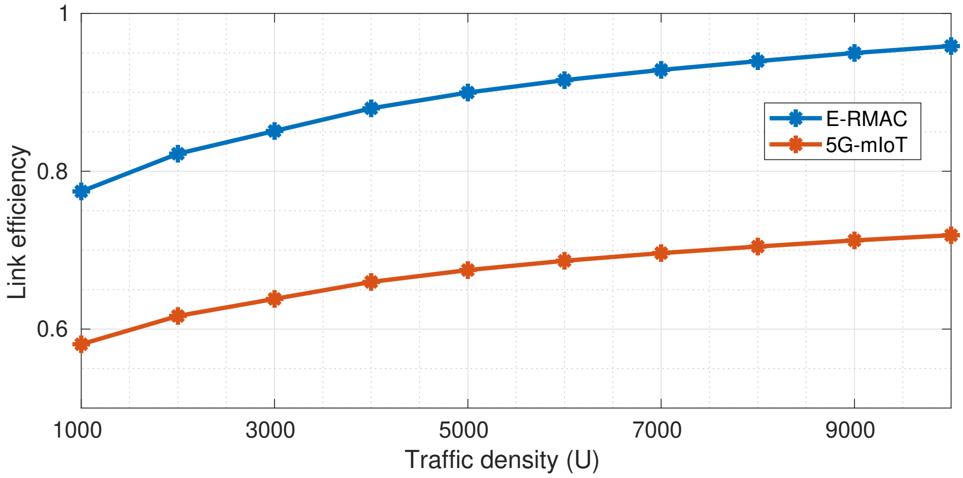


FIGURE 4.4: Computation of link efficiency due to varying traffic density, $C_{up(sig)} = 100$ MHz, $s_{sig} = 0.002$ MB/s, and $r_{sig} = 0.006$ MB/s.

application users for admission into the edge network. It can be seen that the admission efficiency obtained by clustered users is noticeable on the entire range of U (i.e. $U = [1000, 3000]$). The acquired admission efficiency is due to optimal resource utilisation, which reduces latency and congestion in the network to admit more users. However, user admission without clustering would be lower in each case, due to redundancy in signalling and congestion. In particular, a massive volume of demand resource scarcity in the edge network would create competition for admission and resource allocation. This results in users being offloaded onto the neighbouring edge network or backed off. Hence, in the case of massive traffic demand, the clustering approach guarantees fewer or even zero user clusters being offloaded from the edge to the neighbouring edges, subject to the availability of resources and latency consideration.

4.2.3.4 Impact of Clustering on Resource Utilisation: Figure 4.6 illustrates the fairness in resource allocation obtained from (4.10) versus time. A remarkable difference can be seen in the resource allocation of the proposed model compared to its counterparts (i.e. Bankruptcy Game (BG), Equal Ratio (EQ), and Traffic Proportion (TP) (Jia et al., 2018). The resource allocation index by the proposed model is approximately 1 compared to that of BG, EQ and TP, with their fairness indexes hovering around 0.99, 0.92, and 0.91. In Jia et al., 2018) approach, user groups are created for admission to the network based on the Lloyd Shapley approach. In this work, a user would be a part of a group, if the user adds more benefits to the slice. However, a

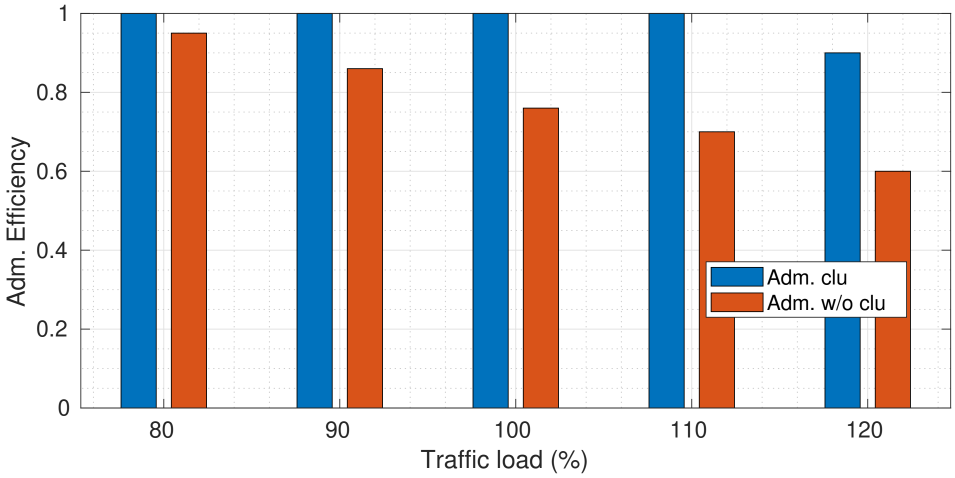


FIGURE 4.5: Admission efficiency on varying traffic density, $U = [1000, 3000]$.

greedy-based admission control strategy may not always be optimal. Such a policy makes the greedy decision on the spot to achieve the objective, such as increasing the admitted requests to earn more revenue from the network but also creates a lot of congestion due to signaling redundancy. The achieved fairness in the proposed E-RMAC model is by the admission of edge users in the form of clusters over their guaranteed or desired demand. Moreover, throughput is greater, due to small or no communication overheads on the link, which enhances the fairness index. Thus, upon arrival of the clustered request, the efficient resource allocation among clustered users leads to maximised resource utilisation, along with enhanced admission gain.

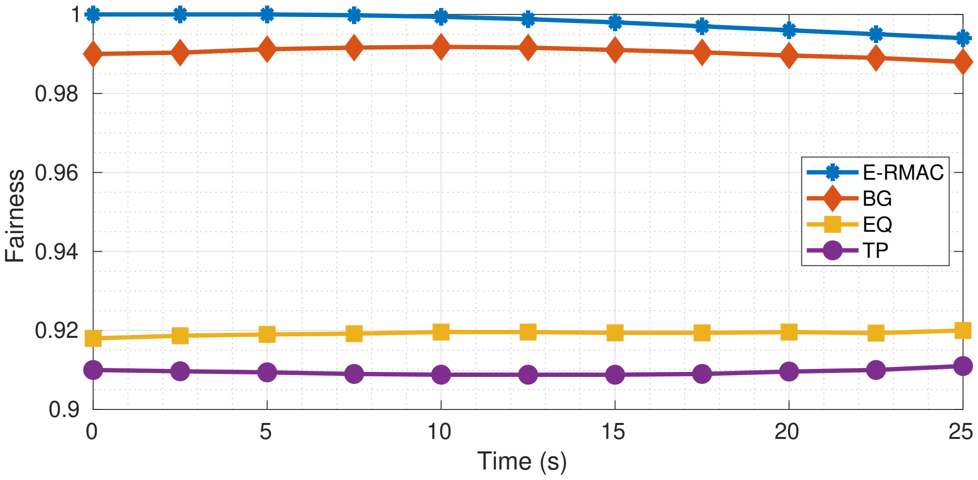


FIGURE 4.6: Resource allocation fairness on varying traffic density, $U = [500, 2500]$.

To summarise, the k-mean- and ranking-based clustering approach, along with optimisation, not only reduces signalling redundancy in the access and core networks but can also enhance admission gain and resource utilisation in future networks. Thus, it is expected that the proposed model will be a crucial part of future networks for traffic and network management. Similarly, in the case of a multi-operator environment, a novel forecasting and fuzzy-logic-based admission control and resource allocation model for O-RAN is presented in the following section.

4.3 Forecasting and Fuzzy-logic-based Admission Control (FAC) Model

The adaptive nature of 5G technology presents numerous opportunities to network operators to enhance system capacity and provide more efficient radio resource utilisation. This adaptivity is acquired partly by recent advancements in NFV, network slicing, and the coexistence of multiple RATs (Tseliou et al., 2016). One of the principal incentives behind redesigning cellular networks is to serve a plethora of devices with different requirements. For enhancing system capacity, numerous techniques are available, such as the use of ultra-dense small cell distribution (Habibi et al., 2019), millimetre waves (mmWave) (Busari et al., 2017; Wang et al., 2018; Uwaechia and Mahyuddin, 2020), new radio (NR) (Richart et al., 2016; Memisoglu et al., 2019; Camps Mur et al., 2020), and intelligent cognitive radio (Amjad, Rehmani, and Mao, 2018; Wang et al., 2019; Yu, Lin, and Chen, 2019; Ahmad et al., 2020). However, to guarantee a seamless multi-operator orchestration, limited emphasis is seen in the literature on the interpolation of current standards with existing ones. Such a seamless ecosystem is essential to provide efficient radio resource utilisation and enhanced QoE (Andrews et al., 2014; 3GPP, 2017).

The coexistence and cooperation of several heterogeneous RATs provides better performance through supporting higher data rates, efficiently accumulating system capacity, and reducing latency and packet loss. From the implementation perspective, network operators usually use existing network infrastructure to serve voice calling and Web browsing applications. This existing network infrastructure offers satisfactory services to conventional applications. Additionally, whenever users are outside 5G coverage, network operators expect that the availability of legacy RATs is essential to provide seamless end-user services (Banchs et al., 2015; Lee et al., 2018; Camps Mur et al., 2020). Thus, the coexistence of heterogeneous technologies is becoming the dominant feature of the current and future cellular wireless generations. In this respect, the novel concept of open RAN (also known as O-RAN) could be considered a complementary option to the new 5G RATs (Alliance, 2018; Gavrilovska, Rakovic, and Denkovski, 2020).

The tenant-based approach for network selection is a frequently used mechanism for providing heterogeneous RAT services (3GPP, 2012; Aryafar et al., 2013; 3GPP, 2013a; 3GPP, 2013b; Tseliou et al., 2016). The best RAT selection by the tenant from an available set creates latency. Moreover, this selection leads to network congestion, particularly when a large number of devices request access to the network. The inefficient admission control saturates the network, which is impractical for providing real-time services with stringent QoE demand (Bouali, Moessner, and Fitch, 2016; Yu, Lin, and Chen, 2019). Ineffective resource allocation and utilisation also have a significant impact on network QoS. In such situations, having an efficient model for the tenant to select the best access network from the available heterogeneous RATs remain an open issue. Therefore, in this work, a forecasting and admission control (aka. FAC) model is proposed to support the presented federated O-RAN architecture. This model has been built on dynamic traffic demand and expanded by particle filtering, followed by network selection using NSGA-II fuzzy-logic optimisation to address the challenges mentioned above.

The presented federated O-RAN is an extension of the novel O-RAN architecture. A federation controller is proposed for the O-RAN architecture to achieve the objectives of FAC model autonomously. The controller operates as a switch for selecting the optimal network from the available set of heterogeneous RATs based on the various heterogeneous networks, as well as user demand characteristics. These characteristics include available network bandwidth, latency sensitivity of the requested service, packet loss, required data rate, and signal strength. Accordingly, the following major contributions of the presented research are:

- A novel federation model known as forecasting and admission control (aka. FAC) model is proposed for tenant-aware network selection and configuration. This model features a dynamic demand-estimation scheme embedded with fuzzy-logic-based optimisation for optimal network selection.

- Two algorithms for network selection and admission control are proposed, in which a multivariate service allocation priority factor is established for admission queuing. On attainment of optimal admission, a service profile is also build for service monitoring.

- Performance of the proposed model is compared analytically with various state-of-the-art schemes for admission control and resource allocation to show how FAC is more efficient at providing better network QoS and user QoE.

4.3.1 Related work

5G promises support for service heterogeneity, on-demand service deployment, and coordination among various access network technologies. In recent years, several global research groups have undertaken work on integrating various heterogeneous technologies to provide more agile services to end-users. For example, 3GPP describes dual connectivity between 4G LTE and 5G NR in TR 38.804 Release 14. Similarly, 3GPP TR 37.900 Release 15 explains the deployment architecture for multiple RAT in a network. Radio frequency requirements were also identified in release 15 for implementing the Multi-Standard Radio (MSR) Base Station (3GPP, 2012; 3GPP, 2013a; 3GPP, 2013b; 3GPP, 2017; 3GPP, 2018a). The xRAN forum is another novel effort, which is an open-source alternative to the conventional RAN architecture. xRAN provides a solution by separating the data and control planes of network devices and opening intelligent interfaces among various RAN building blocks (Forum, 2016). Recently, a novel architecture called open radio access network (or O-RAN, aka V-RAN) has been proposed with the interoperability of various networks as its core principle. O-RAN is an emerging technology that incorporates virtualisation and intelligence in networks. One such example is the OpenRAN project by Telecom Infra, a software-driven architecture that evolved from Cloud RAN (C-RAN). It provides a solution based on the concepts of SDN and openness of general-purpose hardware (Wang, Roy, and Kelly, 2019).

The C-RAN Alliance and the xRAN forum merged into an O-RAN in 2018 to support the evolution of 5G and beyond networks (Alliance, 2018). O-RAN is a multi-vendor and interoperable technology that eliminates dependencies on a particular network deployment from scratch. This interoperable technology opens protocols and interfaces between various heterogeneous network components to incorporate intelligence into RAN that supports different deployment scenarios. More than 160 well-known contributors from small to large size companies, academic institutions, vendors, and network operators are participating in the standardisation of this technology (e.g. Nokia, Intel, Hewlett Packard Enterprise, Vodafone) (Nokia, 2020; Gavrilovska, Rakovic, and Denkovski, 2020; Niknam et al., 2020). In this thesis, the presented research work continues this trend by introducing a federation layer within an O-RAN architecture to enable dynamic traffic forecasting, efficient admission control, and service monitoring.

Along with efficient integration of multiple access technologies, the prediction of future network demand (using efficient forecasting techniques) is among operators' main challenges. To cope with volatile demand, network operators strive to make resource management and orchestration processes highly automated. To realise this, Sciancalepore et al. proposed a traffic-forecasting and slice-scheduling approach that employs the concepts of the Holt–Winters theory for admission control in 5G networks (Sciancalepore et al., 2017). To solve the geometric knapsack problem, two low-complexity algorithms were developed by the authors that ensure optimal slice admission and better QoE. Moreover, the authors proposed enhancements to their work for user mobility analysis on best-effort and guaranteed traffic, as given in (Sciancalepore, Costa-Perez, and Banchs, 2019). In this work, the signalling-based network slicing broker had been utilised for cellular network capacity forecasting. For an efficient transportation management system, Raikwar et al. also used the Holt–Winters method for predicting vehicular traffic demand in short- and long-term traffic windows (Raikwar et al., 2017). Tseliou et al. proposed the Monte Carlo–based traffic-forecasting model for on-demand resource allocation

TABLE 4.2: Table on existing research

Subject	Authors & Publications	Description
Support for heterogeneous connectivity	3GPP TR 23.234 (R-12), 38.804 (R-14), and 37.900 (R-15) (3GPP, 2012; 3GPP, 2017; 3GPP, 2018a) .	3GPP worked on interworking of cellular network and WLAN, dual connectivity of cellular users with 4G LTE and 5G NR, and Multi-RAT deployment architecture, as given in releases 12, 14 and 15.
	Telecom Infra (Wang, Roy, and Kelly, 2019), xRAN forum (Forum, 2016), O-RAN Alliance (Alliance, 2018), Open RAN technical report (Techplayon, 2019; Wireless, 2020; Nokia, 2020) Gavrilovska et al. (Gavrilovska, Rakovic, and Denkovski, 2020) and Niknam et al. (Niknam et al., 2020) .	O-RAN is a multi-vendor, interoperable product with neutral architecture to support various access technologies. O-RAN intelligently opens protocols and interfaces among various RAN components to integrate various operators' networks and supports different deployment scenarios with lower time to market and cost.
Demand Forecasting	Sciancalepore et al. (Sciancalepore et al., 2017; Sciancalepore, Costa-Perez, and Banchs, 2019), and Raikwar et al. (Raikwar et al., 2017) .	Techniques based on Holt–Winters theory are proposed by the authors for long- and short-term traffic demand forecasting to ensure efficient admission control and resource management in cellular networks.
	Tseliou et al. (Tseliou et al., 2016), Dudek et al. (Dudek, 2016; Dudek, 2019), and Hippert et al. (Hippert, Pedreira, and Souza, 2001) .	The authors implemented Monte Carlo–based prediction models for on-demand resource allocation in cellular and neural networks.
	Narmanlioglu et al. (Narmanlioglu et al., 2017), Miao et al. (Miao et al., 2016), and Zhang et al. (Zhang et al., 2017) .	A significant amount of work based on Bayesian techniques is presented by the authors to predict the number of active users and their distribution within the cellular network for localisation and resource allocation over handover.
	Madan et al. (Madan and Mangipudi, 2018), and Monteil et al. in (Monteil et al., 2020)	The authors implemented ARIMA RNN, DNN, and LSTM based approach for network traffic forecasting.
Fuzzy-logic-based network selection and resource allocation	Inaba et al. (Inaba et al., 2015), Bouali et al. (Bouali, Moessner, and Fitch, 2016), Goudarzi et al. (Goudarzi et al., 2019), and Kaloxylos et al. (Kaloxylos et al., 2014) .	The authors implemented the fuzzy-logic-based approach in their proposed hybrid model for efficient access network selection among heterogeneous networks.
	Khan et al. (Khan et al., 2019), Zeng et al. (Zeng et al., 2019), silva et al. (Silva et al., 2018), and Shrimali et al. (Shrimali, Bhadka, and Patel, 2018) .	The authors adopted fuzzy-logic and multi-criterion optimisation schemes, or algorithms such as a genetic algorithm, to propose their model for resource allocation in 5G cellular and vehicular networks.

to tenants in LTE networks. In this work, the authors integrated a multi-tenant slicing capacity broker into the 3GPP reference architecture for extracting short- and long-term variations in traffic patterns (Tseliou et al., 2016). A few other models using the Monte Carlo approach for short-term forecasting can be found in (Hippert, Pedreira, and Souza, 2001; Dudek, 2016; Dudek, 2019). To ensure efficient resource allocation, Narmanlioglu et al. proposed a Bayesian technique–based forecasting model to predict the active number of users in an LTE network (Narmanlioglu et al., 2017). Miao et al. proposed a multi-spatio-temporal model for forecasting cellular user traffic (Miao et al., 2016). Based on the Bayesian model and Markov chain Monte Carlo (MCMC) techniques, the authors in (Zhang et al., 2017) proposed a hybrid forecasting model to predict traffic distribution in a cellular network. Holt–Winters and Bayesian are simple yet work well over short time series in a linear system using prior information and assumptions about the user. These are basic exponential smoothing techniques. Similarly, Monte Carlo–based forecasting techniques rely on data from previous instances (Raza and Khosravi, 2015). However, these techniques are unsuitable for forecasting future demand in dense networks, especially in cases where there is limited or no prior information about network capacity and user demand.

The suggested method in (Madan and Mangipudi, 2018) decompose the network traffic into linear and non-linear components for forecasting. A discrete wavelet transformation is used by the authors for network traffic decomposition. After that, the Autoregressive Integrated Moving Average (ARIMA) and Recurrent Neural Network (RNN) models are applied for forecasting the linear and non-linear components. The final forecasts are acquired by averaging the forecasts of both components. Similarly, the authors in (Monteil et al., 2020) proposed a data-driven forecasting approach using long short-term memory (LSTM) and deep neural network (DNN) for resource reservation in a sliced 5G network. In this work, the authors employed ARIMA as a baseline model. In long time series data forecasting , ARIMA and LSTM also proved their efficiency by having less average error in actual and forecasted data. ARIMA uses linear regressions technique for forecasting the demand. This is because ARIMA assumes each demand trend to be constant over time. Moreover, a long historical horizon is necessary to run ARIMA on a particular problem. Running ARIMA, LSTM and DNN on a wide dataset are computationally expensive (Siami-Namini, Tavakoli, and Namin, 2018). Therefore, they are not suitable to solve the problems of wireless communication where latency matters. In this work, a hybrid Monte Carlo–based particle filtering technique have been implemented to predict future network traffic demand. This is because particle filtering technology has proved its superiority due to its non-dependency on previous data samples in non-Gaussian and nonlinear systems. Therefore, its multi-modal processing capability makes it suitable for a wide range of communication applications.

Fuzzy-logic optimisation is the most effective approach to dealing with information scarcity and uncertainties in available information about users. A significant amount of research can be found on fuzzy-logic optimisation. Bouali et al. implemented a fuzzy multiple-attribute decision-making (MADM) approach in their work to select the best RAT on the network's defined policies (Bouali, Moessner, and Fitch, 2016). Goudarzi et al. proposed a multi-point-algorithm-based hybrid model for the most suitable RAT selection from available heterogeneous networks. This model implements biogeography-based optimisation on probabilities obtained from a Markov decision process for RAT selection (Goudarzi et al., 2019). Kaloxylos et al. implemented a fuzzy-logic approach for an efficient RAT selection between (H)eNBs and Wi-Fi APs (Kaloxylos et al., 2014). This scheme addresses static and low-mobility users only. A fuzzy call admission control model is proposed by the authors in (Inaba et al., 2015) for wireless multimedia networks. Similarly, a hybrid fuzzy-logic-based genetic algorithm (H-FLGA) is proposed by the authors for resource allocation in 5G VANETs (Khan et al., 2019). The authors in (Zeng et al., 2019) proposed a fuzzy-logic-based multi-criterion model for resource allocation in the 5G NOMA system. Their proposed resource allocation algorithms are "serve channel-gain-based subchannel allocation" (denoted as SCG-SA) and "low-complexity, fuzzy-logic user-ranking-order-based joint resource allocation" (denoted as FLURO-JRA). Likewise, Silva et al. proposed a fuzzy-logic-based self-tuning model for resource allocation in dense cells (Silva et al., 2018). This model compared received signal strength with the threshold derived from the signal power, user velocity, and channel quality. In another work, a multi-objective optimisation model is proposed by Shrimali et al. in (Shrimali, Bhadka, and Patel, 2018) for resource allocation in cloud networks. This model utilises the concepts of fuzzy logic to generate coefficients on the defined multiple objectives. In this work, the implemented genetic algorithm uses these coefficients to generate Pareto optimal solutions. A bankruptcy game-based resource allocation algorithm for 5G Cloud-RAN slicing is proposed by the authors in (Jia et al., 2018). In this approach, user groups are created for admission to the network based on the Lloyd Shapley approach. In this work, a user would be a part of a group, if the user adds more benefits to the slice. However, a greedy-based admission control strategy may not always be optimal. Such a policy makes the greedy decision on the spot to achieve the objective, such as increasing the admitted requests to earn more revenue from the network but also creates a lot of congestion due to signaling redundancy. Moreover, equal ratio strategy (EO) and traffic proportion (TP) approaches are employed by the authors as a benchmark in this work. The equal ratio strategy means resources are allocated equally to different slices. This result in over and under resource utilisation in the network on frequently varying demand. The traffic proportion strategy means that resources are allocated to different slices in proportion according to their random requirements. These approaches degrade the network QoS through inefficient resource allocation. An online auction-based resource allocation model for 5G networks is proposed in (Liang et al., 2019). In this model, admission control is based on the user's profile and their bidding. The greedy and first-come-first-out approaches have been used by the authors to maximize the network's revenue. However, more users' bidding as compared to available resources in the network would drop the overall network QoS and user satisfaction from the network.

The existing research mainly considers only a few user-application specific requirements or network-specific statistical characteristics for best network selection and resource allocation. However, today's network is dynamic and more complex, due to numerous heterogeneous applications and their requirements, as well as the network's uncertain circumstances, especially in the case of coexistence of various heterogeneous RATs or O-RAN. In such scenarios, information about how uncertain the application requirements and network circumstances can be is crucial for fuzzy-logic operations. This is because higher uncertainty with regards to requirements and availability leads to inappropriate network selection, inefficient resource allocation, and agreed QoE degradation (Ghosh et al., 1998). The proposed model applies NSGA-II, coupled with the fuzzy-logic approach, to provide an optimal network selection based on user-forecasted demand, network capacity, and fitness policies to ensure efficient resource allocation and network management.

4.3.2 FAC System Model

To efficiently manage and serve a massive amount of heterogeneous traffic from the deployed multiple access networks is still an open issue (Gupta and Jha, 2015; Kaloxylos, 2018). A novel forecasting and admission control (FAC) model for federated O-RAN is presented in this section. In this model, a federation layer is proposed to enhance the features of O-RAN , as illustrated in Figure 4.7. The federation controller of the proposed layer operates as a switch for selecting an optimal network from the available heterogeneous networks with respect to user-application-specific demand, as discussed in detail in the following subsections.

4.3.2.1 Logical network architecture of federated O-RAN

O-RAN is designed with openness and intelligence. It is built by desegregating three key components (i.e. the Radio Unit (RU), the Distributed Unit (DU), and the Centralised Control Unit (CU)) of the traditional RAN by intelligently decoupling their virtualised software and hardware functionalities (Techplayon, 2019; Wireless, 2020). The establishment of open-standard protocols and interfaces between software and hardware components of the RAN eliminates vendor dependency on traditional networks. Moreover, O-RAN facilitates a wide range of heterogeneous services by transforming existing business models into a new paradigm, or launching new business models with a shorter time to market and lower cost (Alliance, 2018).

As illustrated in Figure 4.7, O-RAN consists of four functional building blocks: (1) *Orchestration and Automation*, (2) *RAN Intelligent Controller (RIC) near real time*, (3) *Multi-RAT Control Unit protocol stack*, and (4) *Distributed Unit* (DU) and *Remote Radio Unit* (RRU) (Techplayon, 2019). In contrast to the general RAN, O-RAN near-real-time and non-real-time controllers are decoupled due to strict latency requirements. In O-RAN, these controllers are placed as isolated layers and connected through an

A1 interface. An orchestration and automation function supervises non-real-time services such as configuration, network design, and policy management. This function also analyses the incoming traffic on the access network to model the training data for run-time executions by RIC near-real-time controller. The trained model, RAN database, and intelligent radio resource management unit in RIC near-real-time controller provide a reliable and robust execution platform for third-party applications.

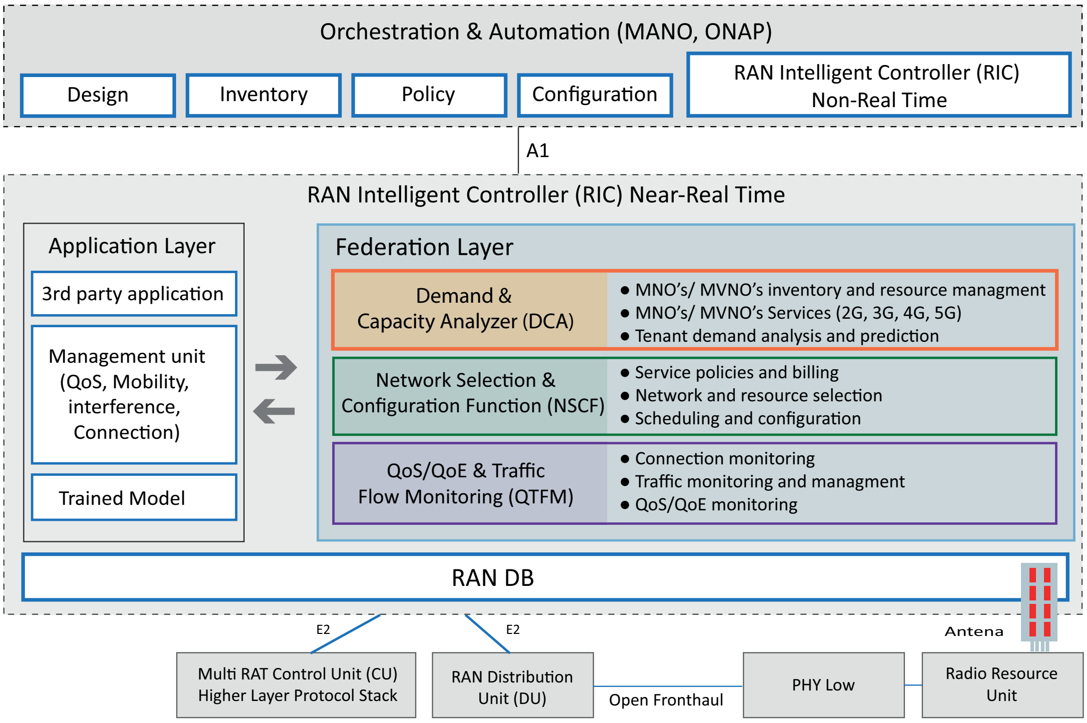

Orchestration & Automation (MANO, ONAP)
Design
Inventory
Policy
Configuration
RAN Intelligent Controller (RIC) Non-Real Time
A1
RAN Intelligent Controller (RIC) Near-Real Time
Application Layer
3rd party application
Management unit (QoS, Mobility, interference, Connection)
Trained Model
Federation Layer
Demand & Capacity Analyzer (DCA)
● MNO's/ MVNO's inventory and resource managment
● MNO's/ MVNO's Services (2G, 3G, 4G, 5G)
● Tenant demand analysis and prediction
Network Selection & Configuration Function (NSCF)
● Service policies and billing
● Network and resource selection
● Scheduling and configuration
QoS/QoE & Traffic Flow Monitoring (QTFM)
● Connection monitoring
● Traffic monitoring and managment
● QoS/QoE monitoring
RAN DB
E2
E2
Antena
Multi RAT Control Unit (CU) Higher Layer Protocol Stack
RAN Distribution Unit (DU)
Open Fronthaul
PHY Low
Radio Resource Unit

FIGURE 4.7: E2E network operations and management in federated O-RAN via FAC model

The main purpose of the proposed model is to efficiently utilise network resources through optimal network selection and user admission. The controller of the federation layer has three main functions: (1) *Demand and Capacity Analyser* (DCA), (2) *Network Selection and Configuration Function* (NSCF), and (3) *QoS/QoE and Traffic Flow Monitoring* (QTFM). The DCA analyses incoming traffic and available network capacity for demand forecasting to ensure optimal network selection, admission control, and resource allocation. The DCA also holds the MVNO resource inventory, which contains MVNO supporting services, content, and billing information. Based on forecasted demand and the set of available networks and resources, the NSCF selects an optimal network through fuzzy-logic optimisation for user admission and resource allocation. In fuzzy-logic optimisation, network selection relies on the suitability-factor derived from multiple decision parameters. On defined policies, this factor ensures that network QoS and user QoE continue to meet the agreed level. The selected MVNO receives service requests, chooses a gateway through SGW/PGW, and establishes an E2E connection through SMF for data transmission and management. The next step is to continuously monitor the admitted network traffic to ensure the granted QoE is within guaranteed bounds through efficient resource utilisation. If the allocated resources are over/underutilised and user-acquired QoE is less than the guaranteed bounds, the QTFM will trigger the analyser. The purpose of this triggering is to modify the demand predictions after observing the difference between forecasted and actual demand.

A multi-RAT CU protocol stack is installed on the virtualisation platform to process the heterogeneous wireless generation protocols. The DU and RRU functions of this model are responsible for baseband and Radio Frequency (RF) processing. These functional units are linked to the RIC near-real-time controller via the E2 interface in the O-RAN (Techplayon, 2019). This novel, vendor-neutral architecture would enable virtual industries to quickly upgrade their networks or deploy new networks to support various deployment scenarios and geographies.

4.3.2.2 E2E customised network configuration in federated O-RAN

Novel O-RAN architecture can more efficiently accommodate rising heterogeneous service demand in future networks than traditional RAN (Nokia, 2020). Therefore, industry professionals and researchers envisage that O-RAN will be an essential component of future wireless/cellular networks. O-RAN reduces vendor and operator dependency on conventional network deployment and operates by opening protocols and interfaces among various building blocks of the access network. Thus, the coexistence of heterogeneous technologies in O-RAN will facilitate an automated vendor network with reduced network operational cost and enhanced performance (Gavrilovska, Rakovic, and Denkovski, 2020).

When a device is turned on, it sends a control signal for E2E network connectivity and configuration to the O-RAN, as shown in Figure 4.8. The CU of the respective virtual access node (i.e. *virtual Base Transceiver Station* (vBTS), *virtual Node B*

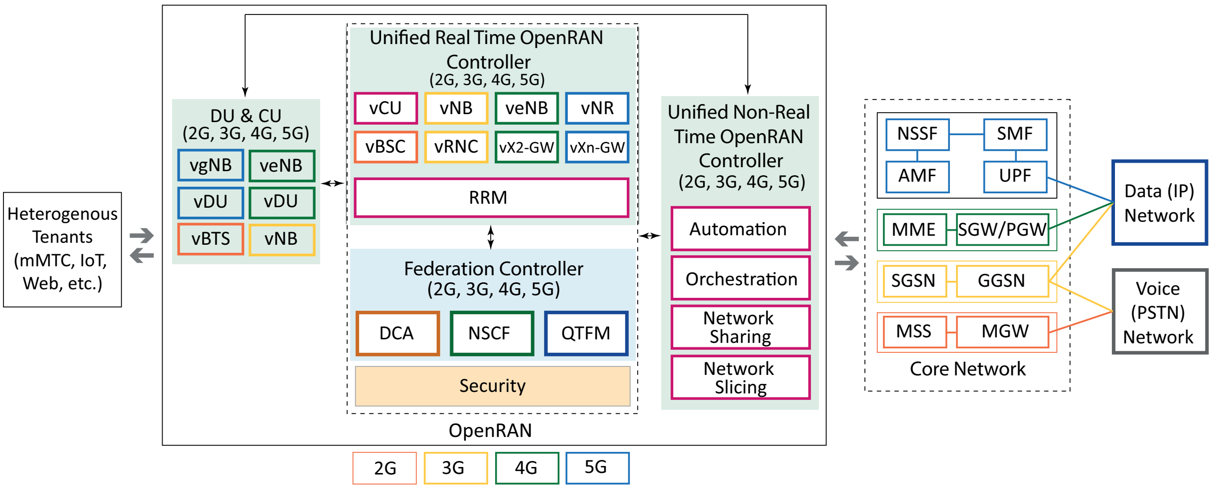


FIGURE 4.8: Federated O-RAN architecture illustrating the NFs for the customised network configuration

(vNB), *virtual evolved Node B* (veNB), *virtual Next Generation Node B* (vgNB)) receives the `Service Request` and `Registration Request` of the control signal. Hereafter, the CU sends the control signal to the unified real-time or non-real-time controller, based on the requested service sensitivity. The virtual access node requests tenant information from the corresponding repository in the core network, such as *Home Location Register* (HLR) and *Unified Data Management* (UDM). The tenant's subscription data is sent back and confirms whether or not the tenant is authorised for service from the network. After successful authorisation and authentication, tenant requests will be sent to the real-time O-RAN controller of the federation layer for admission control and resource allocation by the NSCF. However, tenant admission to the network would be through the conventional approach in the non-real-time O-RAN controller. Based on user preferences and demanded service network statistics, the NSCF selects an optimal network from the service operator list for the requested service. The optimal networks are in order in the list with regards to the forecasts generated by the DCA. After that, the federation controller sends the service request and the session ID to the corresponding NFs in the core to perform customised network configuration. These functions include *Mobile Switching Station* (MSS), *Serving GPRS Support Node* (SGSN), *Mobility Management Entity* (MME), and *Access and Mobility Management Function* (AMF). The ID contains the network function instance address, where the NAS message terminates (Choi and Park, 2017). Finally, the request forwards to the respective core entities (i.e. *Media Gateway* (MGW), *Gateway GPRS*

Support Node (GGSN), *Serving Gateway/Packet Gateway* (SGW/PGW), *Session Management Function* (SMF), *user plane function* (UPF)) for gateway selection, E2E session establishment for data or voice communication through DU, and service management and monitoring through QTFM.

TABLE 4.3: FAC model key symbols and definitions

Symbols & Definitions	
$\mathcal{U}$	Set of tenants in the network
$\mathcal{M}$	Set of MNOs
$\mathcal{V}$	Set of MVNOs
$\mathcal{S}$	Set of services
$\mathcal{S}\_Op$	List of service operators on $\mathcal{S}$
$\mathcal{N}$	Set of resources
τ	Tenant's forecasted demand
$d_{(n)}$	Aggregate nth resource demand
$R_{(n)}$	Aggregate nth resource allocation
$\mathcal{R}\_Op$	Service operator's available resources
τ_γ, τ_h	Acceptable tenant resource bounds
$\mathcal{Q}_{(\gamma)}, \mathcal{Q}_{(h)}$	Expected tenant QoE bounds
$B_{(\gamma)}, B_{(h)}$	Service network resource bounds
$\mathcal{S}_{\mathcal{Q}_\gamma}, \mathcal{S}_{\mathcal{Q}_h}$	Network's guaranteed QoS bounds

4.3.2.3 Network description

In this work, a network is considered with U number of tenants, denoted as a set $\mathcal{U} = \{1, 2, \ldots, U\}$, and M number of MNOs, denoted as a set $\mathcal{M} = \{1, 2, \ldots, M\}$, respectively. Each MNO supports up to V total number of MVNOs, represented by a set $\mathcal{V} = \{1, 2, \ldots, V\}$. Moreover, each MVNO is assumed to have N number of similar resources, indexed by a set $\mathcal{N} = \{1, 2, \ldots, N\}$. Let's assumed that each v is independent and different from other vs associated with the same MNO in terms of resource capacity, guaranteed network QoS, and billing information, where $v \in \mathcal{V}$. Information such as this is stored in the inventory matrix in the repository of DCA by the federation controller and symbolised as **P**,

$$\mathbf{P} = \begin{bmatrix} \mathbf{v}_{11} & \mathbf{v}_{12} & \mathbf{v}_{13} & \cdots & \mathbf{v}_{1V} \\ \mathbf{v}_{21} & \mathbf{v}_{22} & \mathbf{v}_{23} & \cdots & \mathbf{v}_{2V} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{v}_{M1} & \mathbf{v}_{M2} & \mathbf{v}_{M3} & \cdots & \mathbf{v}_{MV} \end{bmatrix}, \tag{4.11}$$

here $\mathbf{v}_{ij\,(1\times N)}$ is the resource vector, where, $i \in \mathcal{M}$ and $j \in \mathcal{V}$. Each operator provides services from 2G, 3G, 4G, and 5G, which is represented as a service set $\mathcal{S} = \{g_2, g_3, g_4, g_5\}$. In addition, these services represent a particular row entry in the mask matrix **G** and

$$\mathbf{G} = \begin{bmatrix} g_2\mathbf{v}_{11} & g_2\mathbf{v}_{12} & g_2\mathbf{v}_{13} & \cdots & g_2\mathbf{v}_{MV} \\ g_3\mathbf{v}_{11} & g_3\mathbf{v}_{12} & g_3\mathbf{v}_{13} & \cdots & g_3\mathbf{v}_{MV} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ g_5\mathbf{v}_{11} & g_5\mathbf{v}_{12} & g_5\mathbf{v}_{13} & \cdots & g_5\mathbf{v}_{MV} \end{bmatrix}. \tag{4.12}$$

If the operator supports service $g_\imath$, where $g_\imath \in \mathcal{S}$ and $\imath = \{2, 3, 4, 5\}$, then the entity is determined by 1 in **G**, otherwise zero. For service provisioning to the tenants with frequently varying demand in latency-sensitive or critical applications, 5G networks have an additional feature called network slicing. Each 5G network is assumed to have S total number of slices with both homogeneous and heterogeneous resource capacity from set $\mathcal{N}$. In that case, $\mathbf{v}_{ij}$ is $S \times N$ dimension matrix. This is because each slice has the potential of resource elasticity to support a varying number of connectivity requests. DCA holds this matrix for the provisioning of services to tenants over forecasting and optimal admission control.

4.3.3 Proposed FAC Model Schema

A *Forecasting and Admission Control* (FAC) model is presented in this section. This model applies the aforementioned sampling-based forecasting technique to obtain optimal network selection through three crucial components of the federation controller. These components are Demand and Capacity Analyser (DCA), Network selection and configuration function (NSCF), and QoS/QoE and Traffic Flow management (QTFM), as illustrated in Figure 4.9 and discussed in detail in the following subsections.

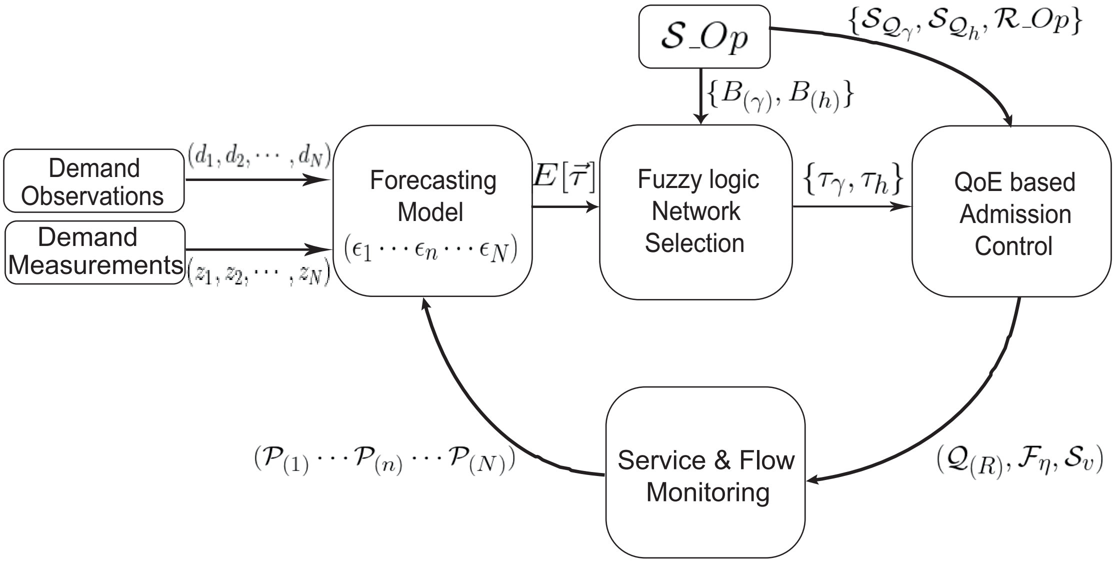

$\mathcal{S}\_Op$
$\{\mathcal{S}_{\mathcal{Q}_\gamma}, \mathcal{S}_{\mathcal{Q}_h}, \mathcal{R}\_Op\}$
$\{B_{(\gamma)}, B_{(h)}\}$
Demand Observations
$(d_1, d_2, \cdots, d_N)$
Demand Measurements
$(z_1, z_2, \cdots, z_N)$
Forecasting Model
$(\epsilon_1 \cdots \epsilon_n \cdots \epsilon_N)$
$E[\vec{\tau}]$
Fuzzy logic Network Selection
$\{\tau_\gamma, \tau_h\}$
QoE based Admission Control
$(\mathcal{P}_{(1)} \cdots \mathcal{P}_{(n)} \cdots \mathcal{P}_{(N)})$
Service & Flow Monitoring
$(\mathcal{Q}_{(R)}, \mathcal{F}_\eta, \mathcal{S}_v)$

FIGURE 4.9: Systematic diagram of the proposed Forecasting and Admission Control (FAC) model

4.3.3.1 Forecasting and demand characterisation

In traffic engineering, traffic analysis is the fastest approach to knowing the characteristics of future service demand in advance (Miao et al., 2016). More precise forecasting results in maximising user-acquired QoE, network QoS, and resource utilisation. In this work, a *Sequential Monte Carlo* (SMC)–based particle filtering technique has been implemented for forecasting future wireless network demand. DCA isolates tenant demand with respect to specific service requirement from the network and then observes the actual demand from the service network using the particle filter for future demand forecasting, as shown in Fig 4.9.

For observation, whenever the uth tenant accesses the network for service s provisioning; where $u \in \mathcal{U}$ and $s \in \mathcal{S}$, it issues a request denoted as $\mathbf{d}_{un} = [d_{u1}, d_{u2}, \ldots, d_{un}]$, where $\mathbf{d}_{un} \in \mathcal{D}$. The vector $\mathbf{d}_{un}$ contains the tenant-demand-specific characteristics. These characteristics include service holding time, physical resources, latency, priority, and revenue. The role of the controller is to assess the request to acquire tenant-specific application characteristics to populate the respective coefficient in the demand matrix $\mathcal{D}$ as a row entry, as shown below:

$$\mathcal{D} = \begin{bmatrix} d_{11} & d_{12} & d_{13} & \cdots & d_{1N} \\ d_{21} & d_{22} & d_{23} & \cdots & d_{2N} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ d_{U1} & d_{U2} & d_{U3} & \cdots & d_{UN} \end{bmatrix}. \tag{4.13}$$

After resource allocation to the tenant u for the service s from a particular network, the respective coefficients are populated in the allocation matrix $\mathcal{R}$ as a row entry. Hence, the resource allocation vector can be represented by $\mathbf{r}_{un} = [r_{u1}, r_{u2}, \ldots, r_{un}]$, where $\mathbf{r}_{un} \in \mathcal{R}$ and

$$\mathcal{R} = \begin{bmatrix} r_{11} & r_{12} & r_{13} & \cdots & r_{1N} \\ r_{21} & r_{22} & r_{23} & \cdots & r_{2N} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ r_{U1} & r_{U2} & r_{U3} & \cdots & r_{UN} \end{bmatrix}. \tag{4.14}$$

Initially, the nth resource demand of the uth tenant, symbolised as $d_{(n)}$, must be analysed through probability density function (PDF) on the observation time window, symbolised as t_{ow} (from $t - t_{obs}$ to t). Thus, the equation is:

$$d_{(n)} = \int_{t-t_{obs}}^{t} f(d_{n,t})\, d(d_{n,t}), \tag{4.15}$$

where, $d_{n,t} \in \mathcal{D}$, and $d_{(n)}$, as known by ground truth demand, is obtained through Gaussian distribution function, as shown in Figure 4.10. Next, the measured demand, symbolised as $z_{(n)}$, on the ground truth of the previous interval, along with the measurement noise covariance, symbolised as Y, in the system is obtained as:

$$z_{(n)} = \frac{d_{(n,t-1)}}{20} + \mathrm{Y}. \tag{4.16}$$

In view of required services and network capacity, observation of the allocated resource, denoted as R, to the tenants on t_{ow} can be obtained as

$$R_{(n)} = \int_{t-t_{obs}}^{t} f(r_{n,t})\, d(r_{n,t}), \tag{4.17}$$

where, $R_{(n)} \leq d_{(n)}$, $R_{(n)}$ is the observed nth resource allocation and $r_{n,t} \in \mathcal{R}$. Therefore, the initially acquired tenant QoE, symbolised as $\mathcal{Q}_{(n)}$, for the nth resource is obtained by

$$\mathcal{Q}_{(n)} = \frac{R_{(n)}}{d_{(n)}} \leq 1. \tag{4.18}$$

The forecasting model $f(\cdot)$ uses the particle filter estimates, denoted as $\hat{\tau}$, to forecast tenant demand, symbolised as τ, on $d_{(n)}$. Thus, the overall estimates over the forecasting window t_w (from $t+1$ to $t+f$), (see Figure 4.10), are obtained by

$$E[\tau_{(n)}] = \int_{t+1}^{t+f} f(\hat{\tau}_{n,t})\, d\hat{\tau}_{n,t}\,, \tag{4.19}$$

where

$$f(\hat{\tau}_{n,t}) = \frac{\hat{\tau}_{t-1}}{2} + \frac{25\hat{\tau}_t}{1+\hat{\tau}_t^2} + 8\cos\left(1.2(t-1)\right) + \epsilon_{(n)}\,, \tag{4.20}$$

and, $\frac{\hat{\tau}_{t-1}}{2} + \frac{25\hat{\tau}_t}{1+\hat{\tau}_t^2} + 8\cos\left(1.2(t-1)\right)$ provides the estimates from the posterior PDF (Walters and Ludwig, 1994), and $\epsilon_{(n)}$ is the noise covariance in the system. The uth tenant overall forecasted demand, symbolised as $E[\vec{\tau}]$, on the forecasting window t_w is obtained by $E[\tau_{(n)}]$ on set $\mathcal{N}$, where $|\mathcal{N}| \geq 1$. Accordingly, the estimated measured value (denoted as $\hat{z}_{(n)}$) can be obtained as follows:

$$\hat{z}_{(n)} = \frac{f(\hat{\tau}_{n,t})}{20}\,. \tag{4.21}$$

Based on $\hat{z}_{(n)}$ from the particle filter and $z_{(n)}$ from demand observation, the computed error, also known as noise covariance and denoted as $\epsilon_{(n)}$ (Allen, 1971), in the system is obtained by

$$\epsilon_{(n)} = \sqrt{\frac{1}{t_w^2} \sum_{t-t+1}^{t+f} (z_{(n,t)} - \hat{z}_{(n,t)})^2}\,. \tag{4.22}$$

This mean-squared error determines the covariance of the observation and estimates. Furthermore, $\epsilon_{(n)}$ updates in each iteration until the error in the prediction converge, where $z_{(n)} - \hat{z}_{(n)} \approx 0$.

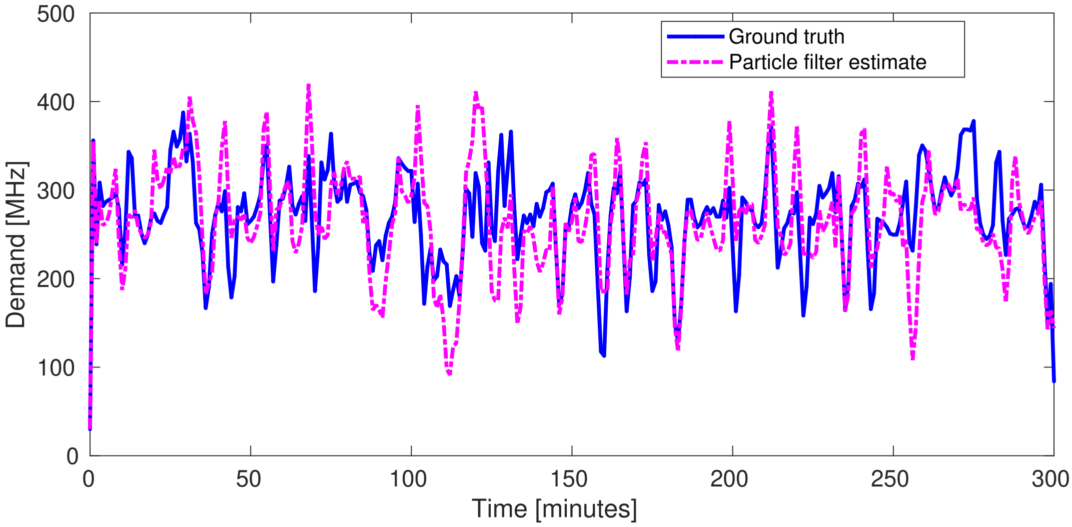


FIGURE 4.10: Long-term resource demand forecasting; $U = 1$ with $E[u_i] = 100$ users, $|\mathcal{N}| = 1$, $t_w = 300$ minutes, and $\tau_h = [10, 100]$ MHz

4.3.3.2 Fuzzy-logic and QoE-based admission control

The main goal of the proposed federation approach is to provide efficient admission control and resource utilisation in a heterogeneous RAN environment, which is based on precise tenant demand forecasting to enhance network throughput and tenant-acquired QoE. To maximise the network throughput and tenant acquired QoE on the forecasted demand is a NP hard problem. Mathematically, it can be written as

$$\begin{aligned} \max \quad & \mathcal{Q}, \eta \\ \text{s.t.} \quad & B_{(\gamma)} < \tau \leq B_{(h)}\,, \\ & \mathcal{S}\_Op \neq 0\,, \end{aligned} \tag{4.23}$$

Where, τ demand should be within the lowest and highest guaranteed resource bounds of the service network and the service network list $\mathcal{S}\_Op$ should not be empty. Based on the maximum $\mathcal{Q}$ and η, the optimised tenant-forecasted demand matrix, symbolised as **Ten**, would be constructed for admission control.

Therefore, NSGA-II has been considered in this work to select an optimum network for the tenant upon arrival. NSGA-II is a multi-objective optimisation method of quickly obtaining optimal and non-dominated solutions by using an explicit diversity preserving mechanism (Deb et al., 2002). The presented optimisation approach considers tenant QoE and throughput maximisation during network selection to obtain fair resource allocation and maximum utilisation, as explained in the following subsections.

Algorithm 8: NSGA-II based optimisation for network selection

```
Input: Forecasted demand ($\vec{\tau}$), number of generations ($gen$), population
  size ($\rho$), evaluation objectives ($\mathcal{Q}$, and $\eta$), network resource bounds ($B_{(\gamma)}$,
  $B_{(h)}$) and $B_{(\gamma)} < \tau \leq B_{(h)}$.
Output: Optimised $\mathbf{Ten}_u$ w.r.t. $\mathcal{Q}$ and $\eta$.
begin
    $P_0(\vec{\tau}_1, \vec{\tau}_2, \ldots, \vec{\tau}_\rho)$ = select $\vec{\tau}$ from resource bounds $(B_{(\gamma)}, B_{(h)})$ of networks
      from $\mathcal{S}\_Op$.
    $F_o(\vec{f}_1, \vec{f}_2, \ldots, \vec{f}_\rho)$ = evaluate objective $(P_0(\vec{\tau}_1, \vec{\tau}_2, \ldots, \vec{\tau}_\rho))$.
    Sort $P_0$ w.r.t. $F_0$.
    for $i = 1 \rightarrow gen$ do
        $P_{i,parent}(\vec{\tau}_1, \vec{\tau}_2, \ldots, \vec{\tau}_\rho)$ = select $(P_{i-1}(\vec{\tau}_1, \vec{\tau}_2, \ldots, \vec{\tau}_\rho))$.
        $P_{i,child}(\vec{\tau}_1, \vec{\tau}_2, \ldots, \vec{\tau}_{\frac{\rho}{2}})$ = crossover $(P_{i,parent}(\vec{\tau}_1, \vec{\tau}_2, \ldots, \vec{\tau}_\rho))$.
        $P_{i,child}(\vec{\tau}_1, \vec{\tau}_2, \ldots, \vec{\tau}_\rho) = P_{i,child}(\vec{\tau}_1, \vec{\tau}_2, \ldots, \vec{\tau}_{\frac{\rho}{2}}) +$ mutation
          $(P_{i,child}(\vec{\tau}_1, \vec{\tau}_2, \ldots, \vec{\tau}_{\frac{\rho}{2}}))$.
        $F_{i,child}(\vec{f}_1, \vec{f}_2, \ldots, \vec{f}_\rho)$ = evaluate objective $(P_{i,child}(\vec{\tau}_1, \vec{\tau}_2, \ldots, \vec{\tau}_\rho))$.
        $P_i(\vec{\tau}_1, \vec{\tau}_2, \ldots, \vec{\tau}_\rho)$ = sort $(P_{i,parent} + P_{i,child})$ w.r.t. $F_i$ and select optimal $\rho$
          solutions.
    end
    $\mathbf{Ten}_u = P$.
    sort $\mathcal{S}\_Op$ w.r.t. $\mathbf{Ten}_u$.
end
```

A) Network selection based on the fuzzy-logic approach: In this work, the proposed federation controller deploys the fuzzy-logic-based NSGA-II for optimal network selection for the tenants, as shown in algorithm 8. The forecasted demand of the uth tenant, denoted as $E[\vec{\tau}]$, on set $\mathcal{N}$ and service-specific network characteristics (i.e. available bandwidth, packet loss, required data rate, latency, and cost), are provided to the FLC as inputs for network selection. The lowest and highest guaranteed resource bounds, symbolised as $\{B_{(\gamma)}, B_{(h)}\}$, respectively, of the corresponding service network from $\mathcal{S}\_Op$ are selected for the tenant in view of forecasted demand. The tenant-acceptable resource demand among $\{B_{(\gamma)}, B_{(h)}\}$ are represented as a genome of size ρ for generation of the initial population, which is denoted as P_0. In the selection criteria, fitness of the resource demand characteristics of the initial population is evaluated with respect to objective functions such as tenant-desired QoE and maximum throughput. The selection process goes through many iterations or generations (denoted as gen), until convergence to a global optimum. After crossover of the parent and mutation of the child population, the most suitable statistics among the service-guaranteed network resource bounds are selected with respect to their fitness according to the defined objectives. The selected resource demand statistics $(\vec{\tau}_\gamma, \vec{\tau}_h)$ are placed in the tenant-forecasted demand matrix, symbolised as $\mathbf{Ten}_{u\,(\rho\times k)}$, where $k = 2|\mathcal{N}|$. $\mathbf{Ten}_{u\,(\rho\times k)}$ is arranged in descending order of the tenant $\mathcal{Q}$ and η for provisioning of the service from the selected network from $\mathcal{S}\_Op$. This is to present the corresponding selected network resources with guaranteed QoS to the tenant for customised network configuration.

Now, the uth tenant is admitted to a particular network from $\mathcal{S}\_Op$, subject to resource availability and provisioning of guaranteed QoS. However, the simultaneous access of various heterogeneous tenants to the network creates competition that leads to congestion in the network. Therefore, a priority-based admission queue has been generated in this approach for efficient admission control and resource allocation. Accordingly, a service allocation priority factor of the tenant, denoted as φ, can be acquired as:

$$\varphi_u = f(\kappa_u, \psi_u, \lambda_u, h_u)\,, \tag{4.24}$$

where, κ_u determines the requested service type across default classification, ψ_u represents generated revenue, λ_u determines the frequency of uth tenant requests, and h_u determines the uth tenant nth resource utilisation history, where $n \in \mathcal{N}$ and $|\mathcal{N}| = 1$. The priority list for tenant admission to the network is arranged in descending order of allocation factor. The tenant with the highest allocation factor is served first from among all tenants, belonging to set $\mathcal{U}$, by the network.

B) QoE-based admission control: Admission requests are processed in terms of QoE constraints in each network. Higher tenant satisfaction level from the network represents efficient admission control, and better network utilisation and revenue maximisation. The uth tenant-acquired QoE, symbolised as $\mathcal{Q}_{(R)}$, should not go beyond the expected QoE bounds, $\{\mathcal{Q}_{(\gamma)}, \mathcal{Q}_{(h)}\}$, where $R \leq \tau_h \leq \tau$. Thus, the uth tenant highest-demanded QoE ($\mathcal{Q}_{(h)}$) on the forecasted demand τ and the acquired QoE ($\mathcal{Q}_{(R)}$) on the acquired resources R are defined as:

$$\mathcal{Q}_{(h)} = f(\tau_h, \beta_\tau, \iota_\tau, \varphi_u)\,, \tag{4.25}$$

$$\mathcal{Q}_{(R)} = f\big(R, \beta_R, \iota_R, \varphi_u\big)\,, \tag{4.26}$$

where, β, ι, and φ are the acceptable user-application-specific packet loss, latency sensitivity, and priority, respectively. For simplicity, these measures are normalised for summation in $f(\cdot)$. Likewise, the tenant served with the least-expected QoE, denoted as $\mathcal{Q}_{(\gamma)}$, during peak hours due to limited resources availability in the network and to minimise the rejection. This occurs due to availability of softness in tenant QoE demand, such that $\gamma == R$. Thus, $\mathcal{Q}_{(\gamma)}$ can be determined as

$$\mathcal{Q}_{(\gamma)} = f(\tau_\gamma, \beta_\gamma, \iota_\gamma, \varphi_u)\,. \tag{4.27}$$

The uth tenant service request arrives in order with respect to φ from the prioritised admission queue, symbolised as

$$\mathcal{A}\_List = \{u_1(\mathcal{Q}_{(\gamma)}, \mathcal{Q}_{(h)}), u_2(\mathcal{Q}_{(\gamma)}, \mathcal{Q}_{(h)}), ...\}\,. \tag{4.28}$$

Upon arrival of each tenant request, the optimal networks are accessed from the list built by the service operator, represented as $\mathcal{S}\_Op$. For the customised network configuration, $\mathcal{S}\_Op$ is arranged according to tenant preferences in descending order. The NSCF assesses each tenant's desired $\mathcal{Q}$ across the network guaranteed QoS bounds, denoted as $\mathcal{S}_{\mathcal{Q}_\gamma}$ and $\mathcal{S}_{\mathcal{Q}_h}$, as shown in Algorithm 9. $\mathcal{S}_{\mathcal{Q}_\gamma}$ and $\mathcal{S}_{\mathcal{Q}_h}$ represents the lower and upper QoS bounds of the network, respectively. Next, the tenant τ resource demand (either guaranteed or demanded) is checked against network capacity for resource allocation. The tenant's resource demand should be less than the serving network resource capacity. Thus, the tenant-acquired QoE, resource utilisation, and overall network throughput are obtained to compute the fairness of resource allocation on set $\mathcal{U}$. Resource allocation fairness (i.e. $\mathcal{F}_\eta$) of that particular service operator network (i.e. $\mathcal{S}_v$) has been stored in the network service profile (i.e. Υ_List), along with the tenant-achieved QoE (i.e. $\mathcal{Q}_{(R)}$). This is for the federation controller to examine the fairness of resource allocation and user satisfaction level of the serving network. In the case of selected network resource unavailability or unsatisfied QoS bounds, the next network from $\mathcal{S}\_Op$ will be examined by the NSCF for admission control. After admission, the tenant-acquired $\mathcal{Q}$ is monitored to ensure efficient network performance. In the case of a violation of QoS/QoE bounds, the user will be dropped from the serving network and reassessed with higher priority by the NSCF. To summarise, by optimising the forecasted demand and service network statistics, a customised network is selected, and resources are allocated with guaranteed QoS bounds to ensure efficient resource utilisation and tenant-acquired QoE.

Algorithm 9: QoE-based admission control

```
Input: Service operator list ($\mathcal{S}\_Op$), tenant-forecasted demand ($\{\tau_\gamma, \tau_h\}$),
 and $\{\tau_\gamma, \tau_h\} \in \mathbf{Ten}_u$, admission queue ($\mathcal{A}\_List$).
Output: $\Upsilon\_List = \{u_1(\mathcal{Q}_{(R)}, \mathcal{F}_\eta, \mathcal{S}_v), \ldots\}$.
for $i = 1 \rightarrow \mathcal{A}\_List.length$ do
    for $(j = 1 \rightarrow \mathcal{S}\_Op.length)$ do
        Select $\{\mathcal{S}_{\mathcal{Q}_\gamma}, \mathcal{S}_{\mathcal{Q}_h}\}$ bounds of $\mathcal{S}\_Op(j)$.
        if $(\mathcal{S}_{\mathcal{Q}_\gamma} < \mathcal{Q}(\tau_h(i)) \leq \mathcal{S}_{\mathcal{Q}_h})\&\&(\mathcal{S}_{\mathcal{Q}_\gamma} \leq \mathcal{Q}(\tau_\gamma(i)) < \mathcal{S}_{\mathcal{Q}_h})$ then
            $\mathcal{R}\_Op$ = assign $\mathcal{S}\_Op(j)$ operator resources.
            if $(\tau_h(i) \leq \mathcal{R}\_Op)||(\tau_\gamma(i) > \mathcal{R}\_Op)$ then
                Allocate resources via $\mathcal{R}\_Op = \mathcal{R}\_Op - R(i)$.
                Obtain tenant-acquired QoE via
                $\mathcal{Q}_{(R)} = \frac{R(i)}{\tau_h(i)}$.
                Obtain resource utilisation via $\mathbf{U}_i(R(i))$.
                Compute throughput via $\eta_i = p_{R(i)} p_{\iota(i)}$.
                Update $\mathcal{F}_\eta$ by including $i$th tenant.
                $\mathcal{S}_v$ = save $\mathcal{S}\_Op(j)$.
                $\Upsilon\_List = u_i(\mathcal{Q}_{(R)}, \mathcal{F}_\eta, \mathcal{S}_v)$.
            else
                Check $j + 1 \in \mathcal{S}\_Op$ for tenant resource allocation.
            end
        else
            Check $j + 1 \in \mathcal{S}\_Op$ against tenant QoE demand.
        end
    end
end
```

4.3.3.3 Service and flow monitoring

After network selection for the tenant via fuzzy-logic optimisation, the network is configured with guaranteed QoS for the tenant for respective service provisioning. After network configuration and establishment of a connection, E2E service flow should also be monitored to ensure that the tenant's acquired QoE does not degrade, and that traffic flow is proportional to network capacity. Thus, proposed a QoS/QoE and traffic flow monitoring system (QTFM) with the following goals: (1) monitor flow to ensure the tenant's acquired QoE is within guaranteed bounds, whereby $\mathcal{Q}_{(\gamma)}$ represents the least-expected QoE, and $\mathcal{Q}_{(h)}$ represents the highest achievable QoE, (2) provide feedback to the analyser for modification of the forecasted demand in proportion to the actual demand and utilisation, as described in detail in the following subsections.

A) QoS/QoE monitoring: Continuous service monitoring is essential for network operators to ensure that network QoS and tenant-acquired QoE remain above the agreed least-guaranteed bounds, where violation in provisioning of agreed QoE and QoS can occur. Therefore, the proposed QTEM continuously monitors network QoS during the duration of service to ensure the tenant's acquired QoE is within expected bounds, as shown in Algorithm 10. In the case of violation of the agreed QoS/QoE bounds, the tenant is dropped from Ψ_List and added to the admission queue, $\mathcal{A}\_List$ with higher priority as compensation. Now, the tenant will be reassessed by the NSCF for network selection with the change in QoE statistics. The QTFM will also trigger the forecasting model to modify the demand to improve the overall network QoS and tenant-acquired QoE. The network service profile, Ψ_List, would also be updated by the federation controller to maintain the network service inventory.

B) Forecasted service demand monitoring: Inefficiency in the forecasting process might over-/under-utilise network resources, leading to inappropriate tenant admission to the network. This would result in a violation of the agreed QoS/QoE, due to poor network QoS and tenant QoE (Sciancalepore et al., 2017). Taking into account the above-mentioned issue, a monitoring procedure is designed to consistently monitor the forecasted and actual demand. This keeps track of the number of violations, such as inefficient resource utilisation, huge forecasting error or variance,

Algorithm 10: Service and flow monitoring

```
Input: Tenant QoE ($\mathcal{Q}_{R_u}$), serving network QoS ($\mathcal{S}_{\mathcal{Q}_{(h,u)}}$), service operator list
    ($\mathcal{S}\_Op$), service profile ($\Psi\_List$).
Output: Updated $\Psi\_List$.
if ($\mathcal{Q}_{R_u} > \mathcal{S}_{\mathcal{Q}_{(h,u)}}$) then
    $\Psi\_List = \Psi\_List - u(\mathcal{Q}_{R_u}, \mathcal{F}_{\eta}, \mathcal{S}_v)$.
    $\varphi_u$ = increase $\varphi_u$.
    $\mathcal{A}\_List = \mathcal{A}\_List + u(\mathcal{Q}_{\gamma}, \mathcal{Q}_h)$.
    Compute Algo. 9.
    Update $\mathcal{P}$ via (4.29).
    Update $\tau$ via (4.30).
else
    $u \in \Psi\_List$
end
```

agreement violation and poor QoE. For future forecasting optimisation, the QTFM provides feedback to the forecasting model, DCA, to update forecasted estimates using the penalty history function. This is symbolised as $\mathcal{P}$ and obtained on t_w as

$$\mathcal{P}_{(n)} = \exp\left(\frac{p_{(n)}}{\sum_{u\in\mathcal{U}} a_{(u,n)}}\right), \tag{4.29}$$

where, $n \in \mathcal{N}$, $p_{(n)} = 1$ indicates the penalty due to QoE violation on resource demand n of the uth tenant, otherwise zero. The admission indicator is $a = 1$ for the uth tenant due to the acquired resource n from the subscribed operator, respectively. On the given number of penalties for resource demand n, the forecasted demand will be updated for future services. Eq. (4.20) is updated by the forecasting modifier, (denoted as $\mathcal{P}\epsilon$), and defined as

$$f(\hat{\hat{\tau}}_{n,t}) = \frac{\hat{\tau}_{t-1}}{2} + \frac{25\,\hat{\tau}_t}{1+\hat{\tau}_t^2} + 8\cos\left(1.2\,(t-1)\right) + \mathcal{P}_{(n)}\,\epsilon_{(n)}\,, \tag{4.30}$$

where, $\frac{\hat{\tau}_{t-1}}{2} + \frac{25\hat{\tau}_t}{1+\hat{\tau}_t^2} + 8\cos\left(1.2(t-1)\right)$ gives the estimates from the posterior probability distribution function (Walters and Ludwig, 1994). ϵ_n is the noise covariance in the system for adjusting estimates according to actual demand. Unlike the conservative setting of the existing forecasting techniques (Holt–Winters, Bayesian, and Monte Carlo), the penalty function dynamically updates the system, such that no agreed QoE and QoS violation could occur. This is due to the adaptability of the service and flow monitoring feature, which obtains the effective demand from the forecasted information to release inefficient resources for better utilisation, thus permitting the network operator to accommodate more users.

4.3.4 Performance Evaluation Measures

To evaluate the proposed model, performance evaluation parameters are defined in this section. The chosen performance metrics include assessment of resources and network utilisation, resource allocation fairness among tenants, and user satisfaction. For comparison, the performance evaluation model has been aligned with relevant work found in existing literature.

4.3.4.1 Resource utilisation: To observe tenant admission by the vth network operator at a given time, t, a_u is introduced as an admission indicator that takes the value 1 upon tenant admission to the network, subject to availability of resources and services, otherwise zero. After successful admission, the nth resource assigned to tenant u from the service operator resource pool is defined as

$$R_u = a_u \, R_{u,n} \, , \tag{4.31}$$

where $R_u \leq \tau_h$. Now the aggregate resources assigned to the tenant set $\mathcal{U}$ is obtained from

$$\sum_{u \in \mathcal{U}} a_u \, R_{u,n} \leq \mathcal{R}\_Op \, . \tag{4.32}$$

Aggregate resources should not exceed network capacity. Similar to (Kuo and Liao, 2008), the uth tenant utility w.r.t. R_u, i.e. symbolised as $\mathbf{U}_u(R_u)$ in the proposed model, is computed as

$$\mathbf{U}_u(R_u) = \alpha \, e^{\omega q} \, , \tag{4.33}$$

whereby ω is the difference between achieved and desired resources, and q and α determine the slope and curve of the slope of the utility function, respectively. Now, the vth operator network utility can be computed as

$$\mathbf{U}_v = \sum_{u \in \mathcal{U}} \mathbf{U}_u \left(R_u \right) . \tag{4.34}$$

Subsequently, based on $\mathbf{U}_v$, the mean network utility, represented as $\mathbf{U}_{Net}$ on set $\mathcal{V}$ and $\mathcal{M}$ can be obtained as:

$$\mathbf{U}_{Net} = \frac{1}{MV} \sum_{i=1}^{M} \sum_{j=1}^{V} \mathbf{U}_{v_{ij}} . \tag{4.35}$$

4.3.4.2 Resource allocation fairness: Maximum resource utilisation and tenant-acquired throughput are crucial in the network to derive maximum revenue. More throughput determines higher resource allocation fairness and better tenant QoE in the network (Jiang, Condoluci, and Mahmoodi, 2016). The acquired fairness in admission control, represented as $\mathcal{F}_{\mathcal{A}}$, by Jain's fairness equation (Jain, Durresi, and Babic, 1999), is obtained as:

$$\mathcal{F}_{\mathcal{A}} = \frac{\left(\sum_{u \in \mathcal{U}} a_u \right)^2}{U \times \sum_{u \in \mathcal{U}} \left(a_u \right)^2} , \tag{4.36}$$

whereby, $a \in \{0, 1\}$, subject to the availability of resources and services from the subscribed operator v network. Likewise, at the time of user admission to the network, resource allocation is also a key factor to be considered. Resource allocation determines the acquired throughput on the probability of resource utilisation (p_R) within the given latency constraints (p_l) at massive tenant demand. Accordingly, this can be obtained as:

$$\eta_u = p_R \, p_l . \tag{4.37}$$

Significantly, the fairness factor in resource allocation, by Jain's fairness equation (Jain, Durresi, and Babic, 1999), can be achieved as follows:

$$\mathcal{F}_{\eta} = \frac{\left(\sum_{u \in \mathcal{U}} \eta_u \right)^2}{U \times \sum_{u \in \mathcal{U}} (\eta_u)^2} . \tag{4.38}$$

4.3.5 Performance Analysis and Results

For performance evaluation of the proposed model, a simulation environment is developed in MATLAB software. In this environment, a virtual network is constructed with different system parameters (as given in GSMA, 2019) to support four heterogeneous services belonging to $\mathcal{S}$. Tenants associated with this network are considered to be within the range $U = [5, 330]$. The average number of users associated with the tenant u is $E[u_i] = 100$, as considered in (Sciancalepore et al., 2017) and (Sciancalepore, Costa-Perez, and Banchs, 2019). Significantly, $\varphi = [1, 5]$, $\iota = [10, 200]$ ms, $\beta = [10^{-2}, 10^{-7}]$, $\mathcal{R}\_Op = 500$, and $\tau_h = [10, 100]$ MHz are the considered ranges of priority, latency sensitivity, tenant-service-specific packet loss, available operator resources, and desired resource demand for each service belong to $\mathcal{S}$, respectively. Overall demand is normalised for simplicity.

Figure 4.11 to Figure 4.15 determine the performance of the proposed FAC model in terms of user satisfaction, fairness of resource allocation, and utilisation gain. The achieved results are compared with Mobile Traffic Forecasting (MTF) (Sciancalepore et al., 2017), Reinforcement Learning (RL-NSB) (Sciancalepore, Costa-Perez, and Banchs, 2019), Online Auction (O-RAN) and Greedy Algorithm (Liang et al., 2019), and Bankruptcy Game (BG) (Jia et al., 2018), based resource allocation and admission control models found in the literature and summarised in Table 4.4.

4.3.5.1 Impact of forecasting

As discussed earlier, demand forecasting is essential for network operators for the sake of efficient network management and traffic engineering. More precise demand forecasting helps the operator in network planning and resource allocation to the tenants to ensure better network QoS and tenant QoE. The proposed model also admits the tenants to a corresponding network based on their forecasted demand. Thus, how the forecasting impacts tenant-acquired QoE, resource allocation, user satisfaction, and load distribution has been illustrated in Figure 4.11, 4.12 and 4.13 in the following subsection.

A) Tenant's QoE and fairness: Figure 4.11 illustrates tenant-perceived QoE and fairness in admission control on forecasted guaranteed resource bounds ($\tau_\gamma = 0.8$ and $\tau_h = 1$). The achieved trend in admission control fairness by (4.36) is more than 97%

on the entire range of U with respect to τ_γ and τ_h demand. The fairness achieved begins to rise with an increase in the number of tenants on the network. The rise in fairness is due to resource allocation to the tenants on their actual demand on the forecasted guaranteed resource bounds. From guaranteed and desired demand, the actual demand is acquired by the convergence of the forecasting modifier on the defined objectives. Hence, the relative gain in acquired admission control fairness by τ_γ over τ_h is 0.5% at $U = 150$. In a fully loaded network, this change in fairness gain is noticeable in keeping tenant admission rejection from the network as low as possible. The proposed model gives the operator precise future-demand estimates, due to self-healing of the forecasts via continuous monitoring of network QoS and tenant QoE. Thus, efficient demand forecasting and monitoring result in more appropriate network selection and admission control for tenants.

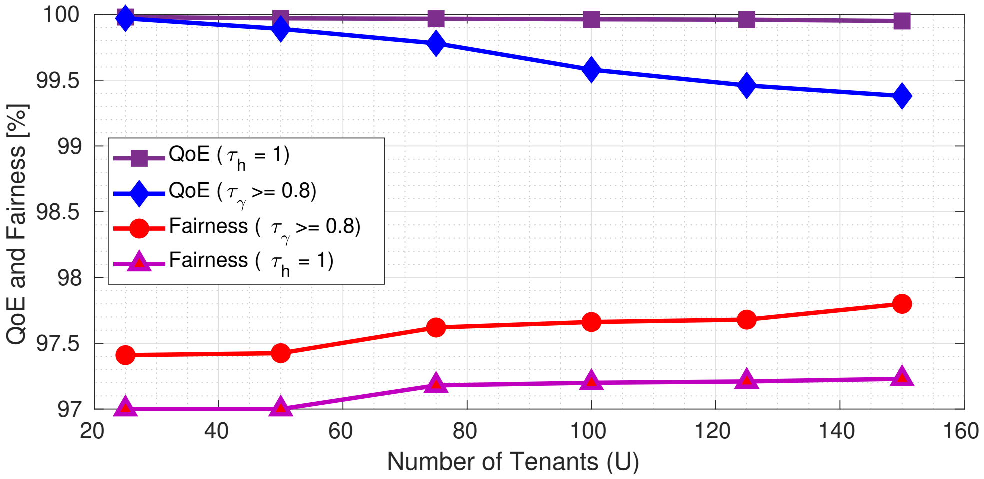


FIGURE 4.11: Computation of QoE and Fairness on $U = 150$ number of tenants with $\tau_\gamma \geq 0.8$ and $\tau_h = 1$ forecasted demand

Average QoE by (4.26) is high at the beginning of the acquired result. The reason is that tenants are acquiring resources on their forecasted demand ($\tau_h = 1$), which might be greater than the actual demand obtained after modification by the forecasting modifier. The achieved QoE begins to decline with an increase in the number of tenants, because tenants are acquiring resources on their actual demand to reduce rejection and improve fairness among tenants. The relative loss in QoE by τ_γ over τ_h is 0.6% at $U = 150$, which is noticeably low in a fully loaded network to keep tenant admission rejection from the network as low as possible. However, the achieved QoE is over 99% over the entire range of U. Therefore, because of the adaptability of the convergence to actual demand, the proposed model is significantly better at managing heavy demand on the network with better acquired QoE and resource allocation fairness.

B) User satisfaction level on forecasting: Consistent monitoring of various parameters during service provisioning is an added feature of the proposed model that can impact user satisfaction. Figure 4.12 shows the satisfaction level of 300 tenants with approximately 100 users each. In this work, the satisfaction level is obtained by accepted users acquired QoE from guaranteed resource bounds ($\tau_\gamma = 0.8$ and $\tau_h = 1$) on the desired QoE and the average of the total number of requests received. User satisfaction reflects the proportion of accepted users at their desired QoE for service provisioning. The performance of the proposed model is superior compared to its counterparts (i.e., O-RAN and Greedy algorithms) (Liang et al., 2019). At $U = 50$, the relative gain in user satisfaction by the proposed FAC model is 8% on O-RAN and 20% on the greedy algorithm, respectively. The variance in gain w.r.t user satisfaction increases with an increase in the number of tenants and their associated users. Similarly, at $U = 300$, the relative gain in user satisfaction by the proposed FAC model is 25% on O-RAN and 58% on the greedy algorithm.

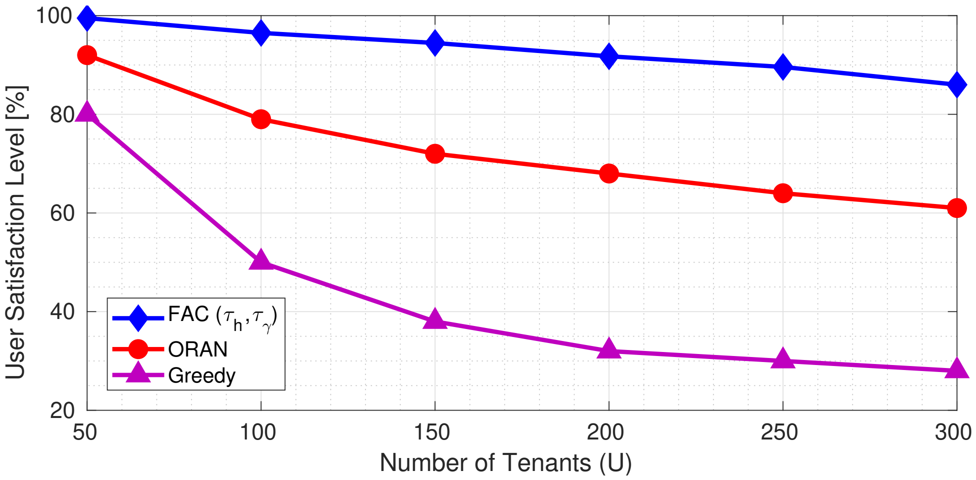


FIGURE 4.12: Comparison of user satisfaction at various demands; number of tenants $U = [50, 300]$ with respect to $|\mathcal{S}| = 4$ and $|\mathcal{N}| = 1$

First, acquired user satisfaction will be greater with the arrival of fewer tenants and associated users. This is because each tenant's user has access to its desired demand. However, congestion occurs with more tenants and associated users arriving on the network. This situation can cause the network to become inefficiently saturated, which leads to an increasing number of users being backed off from the service or rejected, as can be seen at $U = 300$ in the results of the O-RAN and greedy approaches in (Liang et al., 2019). The proposed FAC model reduces the number of rejections by providing services from the optimal network on the tenant's actual demand due to its QoS/QoE monitoring feature. The monitoring feature helps the efficient distribution of traffic flow among the services of set $\mathcal{S}$ to assure efficient admission control and resource allocation. In contrast, existing schemes have higher rejection rates, due to competition among users for limited desired resources and the adoption of the greedy approach. This deficiency in the existing scheme results in degradation of user satisfaction and network resource utilisation.

C) Traffic/load distribution across heterogeneous services: A detailed analysis of the traffic distribution for heterogeneous services provisioning from set $\mathcal{S}$ is presented here. Figure 4.13 shows average network utilisation with and without the proposed FAC model. Results are acquired on a fully loaded network; for instance, if 300 tenants arrive on the network. On admission without forecasting and with 100% utilisation, the 4G and 5G networks are inefficiently saturated. This saturation results in tenant QoE dropping due to congestion and more tenants being rejected or backed off from the network. Similarly, resources are underutilised with 62%, and 66% utilisation in the 2G and 3G networks. These circumstances become costly for an operator in an O-RAN-enabled network, because over/under resources utilisation in the respective service network not only increase operational cost but also reduces overall network performance and tenant-acquired QoE.

The network utilisation achieved by (4.35) in FAC is superior (i.e. more than 95%) in heterogeneous service provisioning from set $\mathcal{S}$ compared to the legacy approach. The proposed model minimises the drawbacks of the legacy approach by dynamically forecasting traffic demand for optimal tenant admission through fuzzy-logic-based network selection. The fuzzy-logic approach encourages efficient traffic load distribution based on demanded services and available capacity of various heterogeneous networks. The self-organisation feature of the proposed model ensures that the networks do not saturate in the case of 100% load. In congestion, the proposed model permits the tenant to accept resources over guaranteed bounds close to network capacity. In this way, each service will accept only relevant demand to accommodate more tenants at the agreed QoE. In contrast, without forecasting, traffic is randomly admitted by the network at desired demand and sensitivity, which leads to congestion and over/under network utilisation.

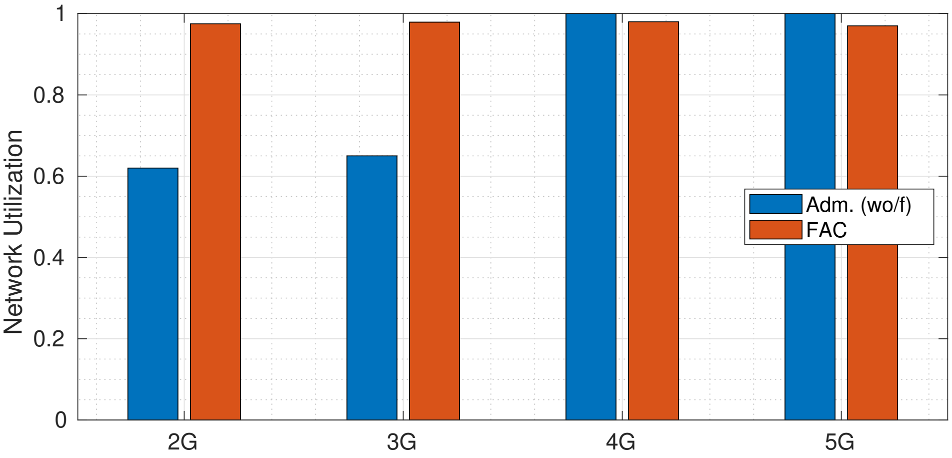


FIGURE 4.13: Average network utilisation of the fully loaded wireless network, with and without tenant demand forecasting and admission control on $|\mathcal{S}| = 4$ and $|\mathcal{N}| = 1$

4.3.5.2 Impact of optimisation and service monitoring

Tenant demand forecasting is crucial to run the optimisation. However, a higher degree of uncertainty in that demand can lead to inefficient admission. When a tenant accesses the network, the proposed model fetches the information from the tenant history to improve the efficiency of the demand forecasting and admission control mechanism. This enhanced mechanism reduces network saturation through efficient distribution of load among various networks and increases network utilisation, as shown in Figure 4.14 and 4.15 in the following subsections.

A) Priority-based supervised admission with optimisation: Figure 4.14 shows the performance of FAC with respect to bandwidth utilisation at various numbers of tenants. The results obtained from the proposed model are compared with those of existing models found in the literature. The utilisation gain is estimated by averaging the bandwidth utilisation of the tenants on legacy and forecasted demand as in (Sciancalepore, Costa-Perez, and Banchs, 2019). The relative gain achieved in bandwidth utilisation at $U = 5$ by FAC are 81.43% and 72.22% on RL-NSB (Sciancalepore, Costa-Perez, and Banchs, 2019) and 94.69% and 92% on MTF (Sciancalepore et al., 2017). The reason behind this achieved utilisation gain is the proposed admission priority factor (φ) by (4.24) and forecasting modifier ($\mathcal{P}\epsilon$) by (4.30). After reviewing tenant history upon arrival, φ prioritises tenant admission to the network to earn more revenue by efficient resource allocation. $\mathcal{P}\epsilon$ optimises the demand estimates through NSCF and QTFM to enhance admission control and network utilisation. However, existing schemes admit users upon their arrival on the network according to their resource demand forecasting.

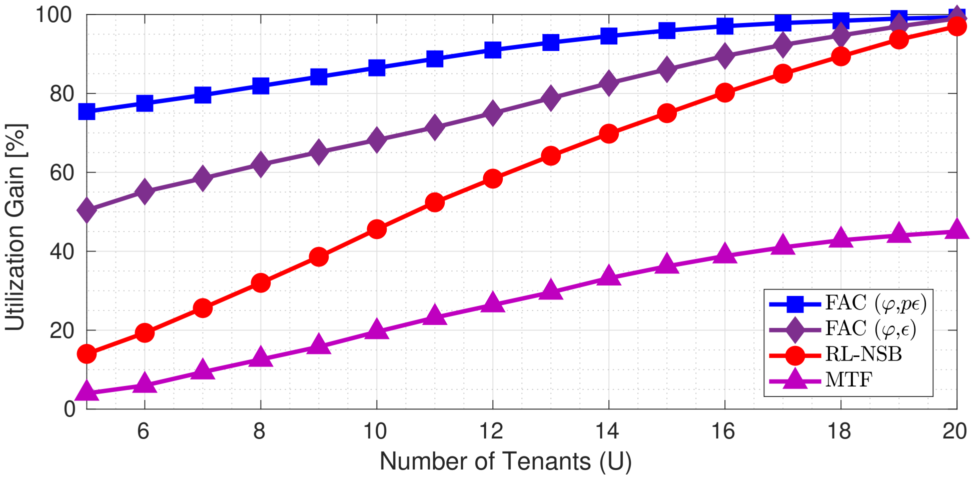


FIGURE 4.14: Computation of network utilisation gain by forecasting and legacy approach across $U = 20$ number of tenants along with $E[u_i] = 100$ user each, and $\tau_h = [10, 100]$ MHz

Resource utilisation continuously increases with an increase in the number of tenants. It can be seen that resource utilisation by FAC is above 95% at $U = 20$. The relative utilisation gain achieved at $U = 20$ by FAC($\mathcal{P}\epsilon$) and FAC (ϵ) are 3% and 2% on RL-NSB, and 55% and 54% with MTF, respectively. Existing models admit users on their forecasted demand and take more time to converge on an optimal solution to improve the admission process. Therefore, these models show less utilisation at the beginning and converge to higher utilisation with an increase in the number of tenants and processing time. In the case of fewer tenants arriving on the network, resources are allocated to tenants at their expected QoE bounds in the proposed approach. The remaining resources are returned to the pool for other operators to serve their associated tenants if needed. This adaptive mechanism provides the operator with an incentive to lease available resources and earn more revenue from resources utilisation. Tenants are accommodated at their guaranteed demand on negotiation in case of congestion on the network. This is to minimise tenant rejection or their being backed off from the network. Hence, the FAC model is efficient at demand forecasting due to the priority factor and self-organised forecasting mechanism. The proposed model yields better performance in terms of tenant-acquired QoE and resource utilisation.

B) Resource allocation with optimisation: The fitness function and its relationship with the demand are the key parameters to be considered for optimal network selection. This determines the appropriate network for the tenant, as well as fair resource distribution among tenants in the network. Figure 4.15 explains the fairness of resource allocation by (4.38) on $U = [315, 330]$ tenants, with 100 users each, across various approaches. The results obtained from the proposed model are also compared with relevant work found in existing literature (Jia et al., 2018). In the proposed model, fairness is obtained by the tenants' acquired average throughput at a given load. Each user belonging to same tenant shares the same proportion with regards to acquired QoE and resource utilisation. The result shows that the proposed QoE-based admission control attain efficient resource allocation with a fairness index of approximately 1 compared to the bankruptcy game allocation scheme with its fairness index floating around 0.99. The reason is that users randomly form groups for network admission and resource allocation in the bankruptcy game model. This results in more users being rejected at the edge of the service network due to resource scarcity. This also creates congestion within the network due to inefficient admission control and competition for limited resources. The relative gain in fairness of FAC is 0.6% and 0.65% at $U = 315$ and $U = 330$. It can be observed that the rise in the fairness gain is relatively low but noticeable on U. This achieved gain in fairness is because of the availability of optimal network solutions and multi-variate priority features for tenant admission and resource allocation, which is obtained through fuzzy-logic-based network selection in the proposed model. It helps to serve as many tenants as possible with guaranteed resource allocation and fewer tenants rejected or backed off from the network. Thus, efficient admission and resource allocation lead to maximised network utilisation and encourage fairness among tenants, as summarised in Table 4.4.

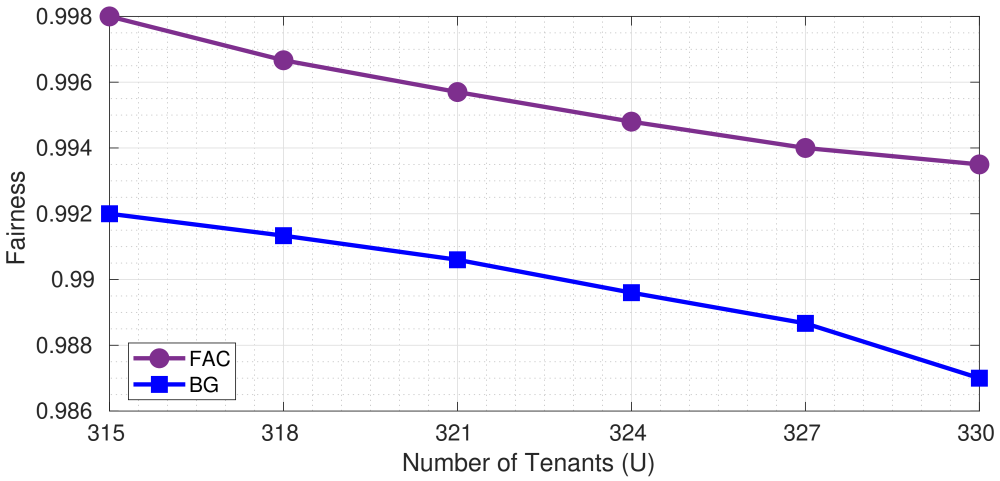


FIGURE 4.15: Resource allocation fairness on varying demand; number of tenants $U = [315, 330]$ with $E[u_i] = 100$ users each on $|\mathcal{S}| = 4$ and $|\mathcal{N}| = 1$

TABLE 4.4: Summary of comparisons of average efficiency between the proposed work and existing methods

Evaluation Parameters	**Approaches**	**Efficiency**
User satisfaction level	The forecasting model and availability of multiple heterogeneous networks (O-RAN) ensure optimal resource allocation to the tenants in the proposed work.	Average efficiency 93% from $U = 50$ to $U = 300$.
	Online auction on available resources and greedy approaches are applied for resource allocation in (Liang et al., 2019).	Average efficiency approximately 73%, and 43% from $U = 50$ to $U = 300$.
Network resources utilisation	In the proposed work, the forecasting modifier ($\mathcal{P}\epsilon$) and multi-variate priority factor (φ) ensure optimum admission control.	Approximately 90% average efficiency from $U = 5$ to $U = 20$.
	Mobile traffic forecasting (Sciancalepore et al., 2017) and reinforcement learning (Sciancalepore, Costa-Perez, and Banchs, 2019) approaches to tenant admission control are applied.	Average efficiency approximately 27%, and 59% from $U = 5$ to $U = 20$.
Resources allocation fairness	In the proposed work, fuzzy-logic-based network selection, QoE-based admission control and resource allocation approaches are applied.	Average efficiency approximately 99.5% from $U = 315$ and $U = 330$.
	Bankruptcy game approach is applied for admission control and resource allocation in (Jia et al., 2018).	Average efficiency approximately 99% from $U = 315$ and $U = 330$.

4.4 Slice Congestion and Admission Control (SCAC) Model

The growing demand for traffic heterogeneity support creates numerous challenges for wireless communication systems, such as bottleneck congestion and inefficient admission control (Mudassir et al., 2019; Gupta and Jha, 2015), which degrade network QoS and user-perceived QoE. Due to the increasing complexity of networks, the 3GPP consortium has proposed a novel and flexible architecture, built on the Network Slicing concept, to segment the network into various capabilities. In this architecture, a single physical network is logically split into multiple virtual networks by a dedicated or shared set of end-to-end network functions (aka network instances) (3GPP, 2018c). The virtual networks are then customised to balance the heterogeneous demand of emerging use cases to achieve maximum utility (Alliance, 2016; Kaloxylos, 2018).

Based on user-perceived QoE, 3GPP categorises the demand into three distinct 5G slice types (use cases or service types): ultra-reliable low latency communication (URLLC), enhanced mobile broadband (eMBB), and massive machine-type communication (mMTC) (Choi and Park, 2017; Shin et al., 2017). These service-specific slices can be classified into hard-QoE, guaranteed soft-QoE, and best-effort QoE traffic demand (Kuo and Liao, 2008). Depending on the service QoE requirements and network load, efficient selection and deployment of network slice instances are also essential for admission control and resource management (allocation and utilisation) (Mei, Wang, and Zheng, 2020; Ojijo and Falowo, 2020). When a 5G network is overloaded, requests are queued for slice admission. In the queue, the requests are arranged with respect to guaranteed soft and best-effort QoE traffic demand. Slice queue capacity is limited to avoid long delays in admission. When the number of slice requests exceeds queue capacity, bottleneck congestion occurs, due to no space being available in the waiting queue, which results in non-queued slice requests being instantly dropped (Han et al., 2018a). Similarly, inefficient admission into slices can lead to under- or over-utilisation of resources and can also create congestion through contention for shared resources. This, in turn, leads to increased rejection of queued slice requests (Haile et al., 2021).

For slice admission and congestion control, 3GPP has proposed a network slice management and orchestration function in 5G referenced architecture (3GPP, 2018b; 3GPP, 2018c). However, the 3GPP-proposed framework provides only the design principles and brief guidelines on service and reference interfaces. Based on the 3GPP network slice reference architecture, Vamshi et al. proposed a mobile virtual network operator slice resource reallocation architecture (MSRAA) in (Buyakar et al., 2020). In this work, resources are relocated from low-priority slices to high-priority slices to reduce the high-priority slice request rejection rate in overloaded networks. This approach causes resource starvation in the admitted low-priority slices and violation of the agreed QoE. The rejection rate of queued low-priority slice requests also increases. When the rejection rate crosses a certain threshold, more resource reallocation is prohibited, increasing the rejection of both high- and low-priority slice requests. Moreover, for each slice reconfiguration, this process goes through several rounds of signalling among core network functions, which creates signalling overheads and congestion in the core network (Han et al., 2018a; Najm et al., 2019; Dandachi et al., 2019). The Experiential Networked Intelligence (ENI) group under European Telecommunication Standard Institute (ETSI) is investigating the use of machine learning (ML) techniques in network slice management and orchestration (ETSI, 2017). However, conventional ML techniques (Supervised Learning, Unsupervised Learning, and Reinforcement Learning) have various shortcomings in solving emerging issues of wireless communication. These techniques require a sufficient amount of training data for resource optimisation in a particular scenario. Many wireless devices are also unable to run higher-complexity tasks, due to their limited computational capacity and power. Thus, the training data, which contains both high (AR/VR data or 4K videos)- and low-quality data (signals or audios), is sent to the central cloud node for training and processing, which might become costly in terms of extended training time and processing. A large amount of unnecessary data transmission into the cloud also creates communication overheads and congestion in the core network. One such problem is addressed by the authors Dandachi et al. in (Dandachi et al., 2019). Dandachi et al. proposed an ML-enabled slice deployment and management model for cross-slice congestion control in 5G networks. The authors evaluate the similarities between slice requests with respect to dedicated or shared network slice instance (NSI) demands by implementing Jaccard similarity-based assignment. According to their work, if the slice request requires a dedicated NSI or no existing NSI can serve the request, then a new NSI will be deployed and configured. Otherwise, the existing NSI is reconfigured with new slice requests and additional resources if needed. An alternative approach is spectral clustering with a computational complexity of $\mathcal{O}(n^3)$, which has also been applied to reduce slice request rejection rate. In this approach, the running slices are re-clustered based on similarity into a newly configured NSI. However, this approach has more competition for resource allocation and greater computational complexity. Each slice acquires the resources even partially. If more slice requests are added to the cluster, resource starvation may occur.

Similarly, advanced ML techniques, i.e. Deep learning, Federated Learning, and Deep Reinforcement Learning (Zhang, Patras, and Haddadi, 2019; Lim et al., 2020; Luong et al., 2019), are also trained on a large amount of data, which takes a long time to acquire an optimal solution and are also computationally complex. All these factors make these approaches impractical to apply to latency-sensitive and emerging applications from the core cloud network (Nguyen et al., 2021). A significant amount of research is available addressing this issue with the help of edge and fog computing (Zhang et al., 2019b; Santos et al., 2021). In this context, transfer learning (TL) has recently emerged as an effective solution for addressing the emerging problems of wireless communication. In this technique, a sufficient amount of only high-quality training data and knowledge transfer improves and speeds up the learning process. Prohibiting sending large amounts of unnecessary data helps to protect data privacy and also reduces communication overheads and congestion (Zhuang et al., 2020). The emphasis of this research is on ML-based slice congestion and admission control with the goal of achieving a minimum slice rejection ratio and maximum resource utilisation. To acquire the mentioned goal, the following are the major contributions of the proposed research work:

- A machine learning-based slice congestion and admission control model is proposed in this work to minimise the slice rejection ratio occurring due to bottlenecks and intra-slice congestion. Unsupervised learning algorithms, (i.e.

Ranking and K-mean clustering), along with optimisation and transfer learning, have been employed for slice request queuing.

- Derived a unified cost estimation function for slice selection to ensure fairness among slice requests. A reinforcement learning-based admission control policy is developed for taking appropriate action for the admission of guaranteed soft and best-effort slice requests in view of instantaneous network circumstances and load.
- A set of optimisation algorithms for Intra-slice and inter-slice resource allocation, along with adaptability of slice elasticity, are also proposed to maximise the slice acceptance ratio and resource utilisation. Robustness of the proposed model and algorithms are analysed by obtaining rejection ratio, bottleneck congestion, and fairness of resource allocation and utilisation at various traffic loads of mMTC and eMBB.

4.4.1 Related work

RL has been frequently used in wireless networks for network admission control, resource allocation and management (Jiang et al., 2016a; Gündüz et al., 2019; Chen et al., 2019). For example, Tong et al. proposed an RL-based call admission control model for wireless communication. In this work, the authors incorporated state-dependent and past-dependent constraints of QoS to maximise network revenue (Tong and Brown, 2000). However, the provided formulation is quite generic and not a good fit for solving recent wireless issues such as bottleneck and intraslice congestion in the dense environment. The authors, Mao et al. investigated Deep RL-based state-of-the-art techniques for resource management of large-scale wireless systems (Mao et al., 2016). Similarly, a Deep RL-based resource management model for 5G network slicing is proposed by the authors in (Li et al., 2018). In this work, demand-aware resource allocation is employed in two different slicing scenarios. However, higher-dimensional data generated by the devices in the scenario makes this approach computationally complex. Higher-dimensional data from various devices also contain redundancy that creates overhead on the core network function. Raza et al. proposed a slice selection and admission policy based on RL for 5G RAN (Raza et al., 2018). The authors investigated their proposed approach to high- and low-priority service demand to maximise the operator's revenue. Similarly, a dynamic reservation and DRL-based resource slicing model for virtual RAN is proposed by the authors in (Sun et al., 2019b). The authors employed Q function approximation for resource allocation in their work. Zhang et al. proposed a mode selection and resource allocation model for cellular networks (Zhang et al., 2019b). The authors applied the Markov decision process, as well as DRL algorithms, to solve the capacity problem. Bega et al. proposed a deep learning-based model known as DeepCog for cognitive network management in sliced 5G networks (Bega et al., 2019). The authors in (Tang, Zhou, and Kato, 2020) proposed a deep reinforcement learning-based model for dynamic uplink/downlink resource allocation in high mobility 5G HetNet. By utilising the deep reinforcement learning approach, another intelligent resource slicing model for URLLC and eMBB traffic in the 5G and beyond network is proposed in (Alsenwi et al., 2021). An end-to-end network slicing model based on deep Q-learning for a 5G network is proposed by the authors in (Li, Zhu, and Liu, 2020). However, if the available data is highly correlated, and the Q-function is estimated from a nonlinear function approximator, then DRL can diverge to unsuitability.

Recent studies have revealed that conventional ML approaches have shortcomings in solving future network problems, especially in emergency and mission-critical applications. As higher-dimensional data is required by conventional ML approaches, which require more time for processing and are computationally costly, they are not acceptable for latency-sensitive applications. Moreover, today's smart wireless devices are not capable of processing this higher-dimensional raw data, and such data needs to be processed on the cloud, which creates an extra burden on the network and also creates congestion. Recently, TL has emerged as an effective solution, where knowledge is transferred from one optimised task to solve another, related or similar task (Cook, Feuz, and Krishnan, 2013). TL has various advantages over conventional ML approaches. For example, the learning process in TL is faster due to the use of pre-trained models or policies, and knowledge sharing between tasks. Compared to traditional approaches, knowledge transfer in TL reduces the computing demand and congestion created in the network due to the huge amount of data. Just enhanced quality and quantity of training data is used in TL, which also provides data privacy protection (Zhuang et al., 2020). For example, the learning process in TL is faster due to the use of pre-trained models or policies, and knowledge sharing between tasks. Compared to traditional approaches, knowledge transfer in TL reduces the computing demand and congestion created in the network due to the huge amount of data. Just enhanced quality and quantity of training data is used in TL, which also provides data privacy protection (Zhuang et al., 2020). A significant amount of research into the applications of TL in wireless networks is available in the literature. For example, a novel TL-based paradigm for dynamic spectrum allocation and topology management of radio networks is proposed by authors Zhao et al. (Zhao et al., 2013). The knowledge learned through spectrum allocation is converted through their proposed priority algorithm and applied to topology management. During their research, Zhao et al. investigated the use of the K-means clustering approach for optimal spectrum and load management of mobile broadband networks (Zhao et al., 2015), whereby coefficients acquired from Q-parameters after demand clustering were transferred from spectrum allocation to broadband load management. Parera et al. proposed a transfer-based model for resource utilisation in wireless networks (Parera et al., 2020). The authors exploited deep learning and TL algorithms for dynamic resource allocation and efficient network control. Wagle et al. proposed three transfer learning algorithms for radio frequency allocation in wireless cellular networks (Wagle and Frew, 2012). The objective of their proposed TL algorithms is to identify the similarities in demand from the original data set to extract pertinent information, which was used in the target data set to achieve efficient radio frequency allocation. Zeng et al. proposed a deep TL-based traffic prediction model for wireless cellular networks (Zeng et al., 2020). The authors proposed a spatial-temporal cross-domain neural network model (STC-N) in this work. STC-N model uses cross-domain data along with a regional fusion TL strategy to improve the accuracy of future traffic prediction. TL-and DRL-based mode selection and resource management models for fog RAN, V2V communication and 5G networks are proposed in (Sun, Peng, and Mao, 2018; Zhang et al., 2019b; Dong et al., 2020). Similarly, Parera et al. proposed a TL model for channel quality prediction of a given frequency carrier in wireless networks (Parera et al., 2019). In this work, convolutional neural networks and long short-term memory networks have been considered as TL tasks.

Based on best of the knowledge, the existing research did not provide a TL-based solution for bottlenecks or intra-slice congestion problems to ensure efficient admission control in future networks. Dynamic slice congestion and admission control using advanced ML approaches is proposed in this work. The goal of this approach is to manage the demand proportionally with available capacity using two unsupervised learning (i.e. ranking-and k-mean-based clustering) and optimisation approaches for congestion control. In view of the eMBB network's complexity, knowledge learned by implementing optimisation of mMTC traffic load for clustering is implemented to eMBB traffic load to reduce bottleneck congestion. RL-based admission control and resource management have also been proposed using intra-slice and inter-slice resource allocation, along with adaptability of slice elasticity, to maximise admission gain and resource utilisation by reducing the slice request rejection ratio.

4.4.2 SCAC System Model

Managing a large amount of heterogeneous traffic flow proportionately with slice capacity at a tolerable latency in future networks is still an open issue Da Xu, He, and Li, 2014; Ojijo and Falowo, 2020. An ML-based dynamic slice congestion and admission control (SCAC) model is proposed for 5G and beyond networks. This model has been developed using an architecture similar to 3GPP release 15 3GPP, 2018c and *Next Generation Mobile Networks* (NGMN) slice architecture Alliance, 2016. The proposed SCAC model is composed of three major entities: a slice demand analysis and classification (SDAC) system, a demand clustering and queuing (DCQ) system, and an admission and resource management (ARM) controller, as shown in Fig. 4.16.

A network slice selection is performed at the time of user registration on the core network. When the user equipment (UE) is powered on, it sends a `Service Request and Registration` along with the user ID and service type (i.e., one of the slice types from the 3GPP defined categories) to the accessed *Next generation NodeB* (gNodeB/gNB) of the radio access network (RAN). The demand analyser analyses

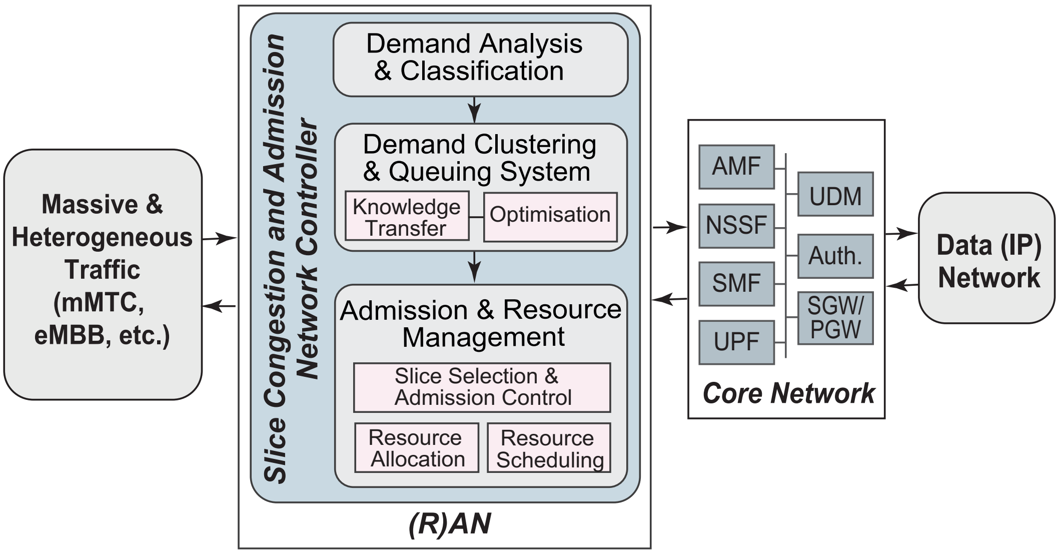


FIGURE 4.16: ML-based SCAC architecture for communication in future wireless networks

the requested slice QoE for the given service type. This analysis isolates guaranteed soft-QoE slice requests from best-effort QoE requests through a classification mask. The demand processing system or DCQ clusters requests using ML and optimisation techniques based on similar service types and QoE demand for queuing in a slice admission queue. After verification, the gNodeB sends the clustered slice requests from the slice queue to the access and mobility management function (AMF) of the default slice for admission control and resource management. The default AMF request for each UE's information from the unified data management (UDM). The UDM sends the subscribed user data to the default AMF and confirms if the user is authorised to be served from the core. After authentication for service provisioning, a slice ID is selected by the network slice selection function (NSSF). The slice ID contains NF instances, which are shared among users of clustered slice requests. The default slice stores all this information in an unstructured data storage function (UDSF) and finally forwards the clustered slice request to the session management function (SMF) and the serving gateway/packet gateway (SGW/PGW) for the establishment of a connection and its management for data transmission (Choi and Park, 2017; 3GPP, 2018b; 3GPP, 2014). Hence, in this work, the admission control is imposed by the signalling and the exchange of information among RAN, AMF, and NFFS, which optimises learning by accessing the current network situation on both sides for congestion control and implements the intra-slice, inter-slice, or slice elasticity approach for efficient resource allocation.

Network setup: Consider a 5G/6G cellular network with a set of slices denoted as $\mathcal{S} = \{1, 2, \ldots, S\}$. In this network, a set of M and N number of mMTC and eMBB devices (or users) are considered with best-effort and guaranteed soft-QoE demand, denoted as $\mathbf{U}_{MTC} = \{\mathbf{u}^{p_1}_{best}, \mathbf{u}^{p_1}_{soft}\}$, and $\mathbf{U}_{MBB} = \{\mathbf{u}^{p_2}_{best}, \mathbf{u}^{p_2}_{soft}\}$, whereby the users belong to mMTC and eMBB are denoted by p_1 and p_2 respectively. Its assumed that, $\mathbf{u}^{p_1}_{best} = \{1, 2, \ldots, \kappa\}$ and $\mathbf{u}^{p_1}_{soft} = \{\kappa + 1, \kappa + 2, \ldots, M\}$, and $\mathbf{u}^{p_1}_{best} \cap \mathbf{u}^{p_1}_{soft} = \emptyset$. Similarly, $\mathbf{u}^{p_2}_{best} = \{1, 2, \ldots, \iota\}$ and $\mathbf{u}^{p_2}_{soft} = \{\iota + 1, \iota + 2, \ldots, N\}$, and $\mathbf{u}^{p_1}_{best} \cap \mathbf{u}^{p_2}_{soft} = \emptyset$. Its also assumed that the best-effort and guaranteed soft-QoE demand have various characteristics $\mathcal{J} = \{1, 2, 3, \ldots, J\}$ with varied distributions. The characteristics are predetermined and quantified within a range of values between j_{min} and j_{max}, where $j \in \mathcal{J}$, stored in the AMF repository, as in (Perveen, Patwary, and Aneiba, 2019). Based on these characteristics the slice service request is classified as either best-effort or guaranteed soft-QoE demand. Each device can be connected to K number of heterogeneous application-specific service slices, denoted as $\Lambda = \{1, 2, 3, \ldots, K\}$, concurrently from set $\mathcal{S}$. However, for simplicity, K is assumed to be equal to 1. Key symbols used in this work are listed and described briefly in Table 4.5. A systematic diagram of the proposed SCAC model on this network setup is shown in Fig. 4.17 and discussed in detail in the following subsections.

4.4.3 Slice Demand Analysis and Classification

When user $u \in \mathbf{U}_{MTC}$ attempts to access the sth slice with desired QoE demand, it issues a request $\mathbf{a}_u = [a_{(u,1)}, a_{(u,2)}, \ldots, a_{(u,J)}]$, which is placed into the respective demand matrix, represented by $\mathbf{A}$ in (4.39), within the repository of the 5G slice controller in RAN. The vector $\mathbf{a}_u$ consists of the required user-application-specific statistical parameters, such as bandwidth, required data rate, latency, and packet

TABLE 4.5: SCAC model key symbols and definitions

Symbols	Definitions
$\mathbf{U}_{MTC}$	Set of users belonging to mMTC
$\mathbf{U}_{MBB}$	Set of users belonging to eMBB
$\mathbf{u}_{best}^{p_1}, \mathbf{u}_{best}^{p_2}$	Set of users belonging to best-effort demand of mMTC and eMBB, respectively.
$\mathbf{u}_{soft}^{p_2}, \mathbf{u}_{soft}^{p_2}$	Set of users belonging to guaranteed soft demand of mMTC and eMBB, respectively.
$\mathcal{S}$	Set of slices in the network
$\mathbf{A}_{MTC}$	Demand matrix of mMTC
$\mathbf{A}_{MBB}$	Demand matrix of eMBB
$\mathcal{M}_c$	Demand classification masks
$\mathcal{M}_R$	Demand ranking masks
C_{que}^s	Slice queuing capacity
C_{req}^s	Requests required capacity
$D_{(x)}$	Queue waiting time of cluster x
$\mathcal{D}_{(x)}$	Queue threshold time for each cluster
$d_{(x_u)}$	uth request waiting time from cluster x
$\mathbf{v}$	Cost estimation function for slice selection
$\mathbf{w}$	network weights for slice selection
$p_{(a_b,a_g)}$	RL-based admission control policy function
B_l^s	Lower slice configuration bounds
B_u^s	Upper slice configuration bounds
$\mathcal{Q}$	Number of rejected requests
w_b	Acquired reward on best-effort demand admission
w_g	Acquired reward on guaranteed soft demand admission
$U(R_x)$	xth cluster utility
U^s	sth slice utility
U	Network utility on set $\mathcal{S}$

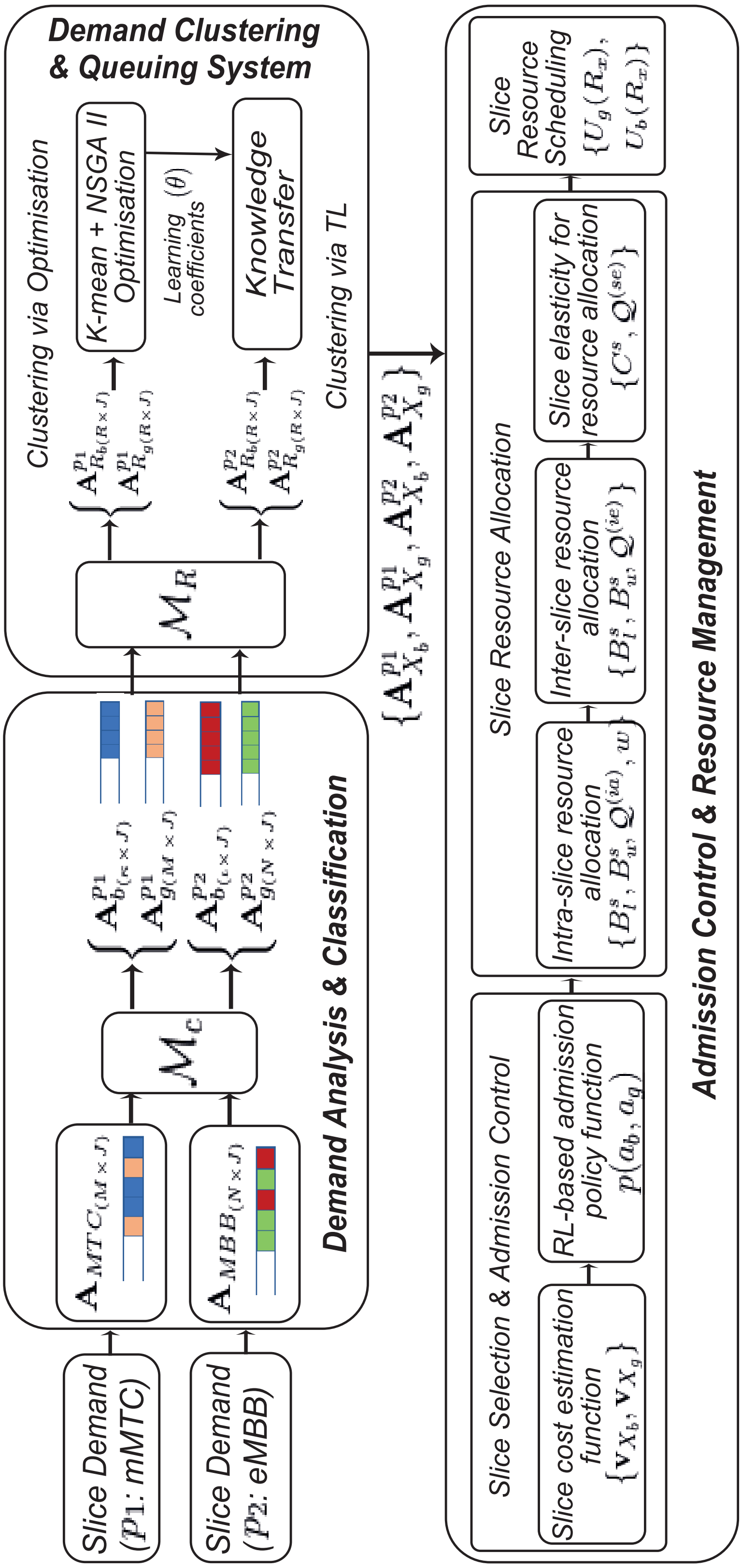

Slice Demand ($P1$: mMTC)
Slice Demand ($P2$: eMBB)
Demand Analysis & Classification
$\mathbf{A}_{MTC(M\times J)}$
$\mathbf{A}_{MBB(N\times J)}$
$\mathcal{M}_c$
$\{\mathbf{A}^{P1}_{b(\kappa\times J)}, \mathbf{A}^{P1}_{g(M\times J)}\}$
$\{\mathbf{A}^{P2}_{b(\iota\times J)}, \mathbf{A}^{P2}_{g(N\times J)}\}$
Demand Clustering & Queuing System
$\mathcal{M}_R$
$\{\mathbf{A}^{P1}_{R_b(R\times J)}, \mathbf{A}^{P1}_{R_g(R\times J)}\}$
$\{\mathbf{A}^{P2}_{R_b(R\times J)}, \mathbf{A}^{P2}_{R_g(R\times J)}\}$
Clustering via Optimisation
K-mean + NSGA II Optimisation
Learning coefficients (θ)
Knowledge Transfer
Clustering via TL
$\{\mathbf{A}^{P1}_{X_b}, \mathbf{A}^{P1}_{X_g}, \mathbf{A}^{P2}_{X_b}, \mathbf{A}^{P2}_{X_g}\}$
Admission Control & Resource Management
Slice Selection & Admission Control
Slice cost estimation function $\{\mathbf{v}_{X_b}, \mathbf{v}_{X_g}\}$
RL-based admission policy function $p(a_b, a_g)$
Slice Resource Allocation
Intra-slice resource allocation $\{B^s_l, B^s_u, \mathcal{Q}^{(ia)}, w\}$
Inter-slice resource allocation $\{B^s_l, B^s_u, \mathcal{Q}^{(ie)}\}$
Slice elasticity for resource allocation $\{C^s, \mathcal{Q}^{(se)}\}$
Slice Resource Scheduling $\{U_g(R_x), U_b(R_x)\}$

FIGURE 4.17: Systematic diagram of the proposed SCAC model for capacity and delay optimisation

loss ratio.

$$\mathbf{A}_{MTC} = \begin{bmatrix} a_{(1,1)} & a_{(1,2)} & a_{(1,3)} & \cdots & a_{(1,J)} \\ a_{(2,1)} & a_{(2,2)} & a_{(2,3)} & \cdots & a_{(2,J)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{(M,1)} & a_{(M,2)} & a_{(M,3)} & \cdots & a_{(M,J)} \end{bmatrix}. \tag{4.39}$$

A similar demand matrix, denoted as $\mathbf{A}_{MBB(N \times J)}$, is constructed for mMBB requests. Now, in the SDAC system, the request $\mathbf{a}_u$, either belongs to mMTC or eMBB, passes through the classification mask denoted as $\mathcal{M}_c$. This mask assists demand classification among the best-effort and guaranteed soft-QoE traffic to maximise the slice request acceptance ratio through clustering.

$$\mathcal{M}_c = \begin{cases} c_b = 1, & \text{if } \mathbf{a}_u \in \mathbf{u}_{best} \\ c_g = 1, & \text{if } \mathbf{a}_u \in \mathbf{u}_{soft} \end{cases}, \tag{4.40}$$

whereby, c_b and c_g represent the best-effort and guaranteed soft-QoE demand classifier, respectively. When a slice request for a particular service type arrives, the slice controller assesses each characteristic value of the request. After analysis, the controller classifies and places the request into the user-specific row of the best-effort (denoted as $\mathbf{A}_b^{p_1}$) or guaranteed soft-QoE demand matrix (denoted as $\mathbf{A}_g^{p_1}$) of the respective p_1, as shown below:

$$\mathbf{A}_b^{p_1} = \begin{bmatrix} b_{(1,1)} & b_{(1,2)} & b_{(1,3)} & \cdots & b_{(1,J)} \\ b_{2,1} & b_{(2,2)} & b_{(2,3)} & \cdots & b_{(2,J)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ b_{(\kappa,1)} & b_{(\kappa,2)} & b_{(\kappa,3)} & \cdots & b_{(\kappa,J)} \end{bmatrix}, \tag{4.41}$$

or

$$\mathbf{A}_g^{p_1} = \begin{bmatrix} g_{(\kappa+1,1)} & g_{(\kappa+1,2)} & g_{(\kappa+1,3)} & \cdots & g_{(\kappa+1,J)} \\ g_{(\kappa+2,1)} & g_{(\kappa+2,2)} & g_{(\kappa+2,3)} & \cdots & g_{(\kappa+2,J)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ g_{(M,1)} & g_{M,2} & g_{(M,3)} & \cdots & g_{(M,J)} \end{bmatrix}. \tag{4.42}$$

Similarly, the best-effort and guaranteed soft-QoE-based demand matrices, represented as $\mathbf{A}^{p_2}_{b(\iota \times J)}$ and $\mathbf{A}^{p_2}_{g(N \times J)}$, are constructed for users belong to eMBB. Now, the

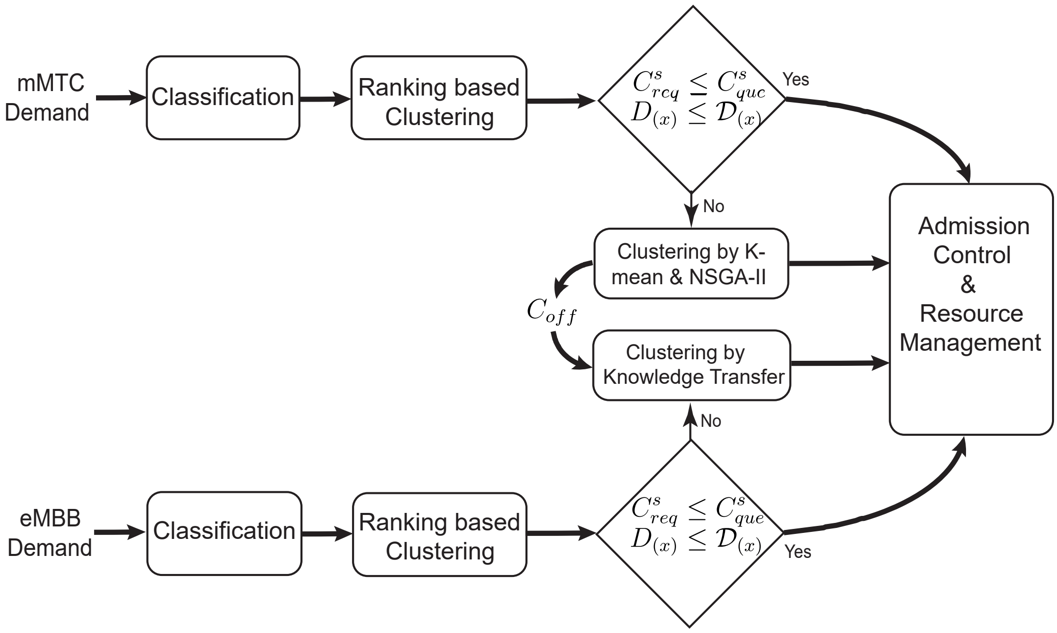

mMTC Demand
Classification
Ranking based Clustering
$C^s_{req} \leq C^s_{quc}$
$D_{(x)} \leq \mathcal{D}_{(x)}$
Yes
No
Clustering by K-mean & NSGA-II
C_{off}
Clustering by Knowledge Transfer
Admission Control & Resource Management
No
eMBB Demand
Classification
Ranking based Clustering
$C^s_{req} \leq C^s_{que}$
$D_{(x)} \leq \mathcal{D}_{(x)}$
Yes

FIGURE 4.18: Systematic diagram for clustering by optimisation and knowledge transfer

isolated QoE-based slice service demands are passed to the DCQ system for processing to reduce redundancy in the service request signalling, which can cause congestion due to resource starvation in the network.

4.4.4 Demand Clustering and Queuing System

In this section, the operations of the DCQ are explained in detail for slice request clustering, as shown in Figure 4.18. Two popular unsupervised learning techniques (i.e. k-mean and ranking) are implemented for clustering, along with optimisation and knowledge transfer, to maximise the slice request acceptance ratio by redundancy reduction in the requests, as explained in detail in the following subsections.

4.4.4.1 Demand Clustering for Capacity Optimisation: In the network, the slice queuing capacity, denoted as C^s_{que}, is limited to avoid long delays being encountered by slice requests in the queue. In the normal network situation, the capacity required by the incoming services requests, denoted as C^s_{req}, in the queue is proportional to the slice queuing capacity, due to the short waiting time in the queue. However, the required services request capacity increases exponentially with an increase in demand for services on the network (Gupta and Jha, 2015; Han et al., 2018a). Thus, in the case of a massive number of service requests, limited queuing capacity can cause bottleneck congestion at the network edge, which, in turn, increases the slice service request rejection ratio. This congestion at the edge also causes operator revenue and network QoS to drop due to inefficient resource utilisation in the core network. This problem can be modelled as an optimisation problem, whereby the objective is to manage the service requests, either guaranteed soft or/and best-effort QoE requests, belonging to MTC or p_1 users, in such a way that minimises the rejection rate at the edge, denoted as α, due to bottleneck congestion by efficiently utilising the slice queuing capacity. Mathematically, this can be described as follows:

$$\begin{aligned} \min \quad & \textstyle\sum_{i=b_{(1)}}^{b_{(\kappa)}} \alpha_{(i)} + \sum_{i=g_{(\kappa+1)}}^{g_{(M)}} \alpha_{(i)}, \\ \text{s.t.} \quad & \textstyle\sum_{i=b_{(1)}}^{b_{(\kappa)}} C^s_{req}(i) + \sum_{i=g_{(\kappa+1)}}^{g_{(M)}} C^s_{req}(i) \le C^s_{que}\,, \\ & \sum_{u=1}^{M} \beta_{(u)} \le 1\,. \end{aligned} \tag{4.43}$$

The aggregate capacity acquired by slice service requests, C^s_{req}, should not exceed the overall reserved slice queuing capacity over M number of slice requests for mMTC. An admission index β of 1 indicates that the request of the uth user is admitted to the queue, otherwise zero. All requests from set $\mathbf{U}_{MTC}$ should be admitted for queuing by the RAN controller.

In this work, ranking-based clustering techniques have been applied to reduce the overall rejection ratio in slice requests. Ranking-based clustering is a simple yet powerful approach to computing the similarity index within a cluster (Saxena et al., 2017). When a request is received by the SQC system, it is compared with the existing request for clustering based on homogeneous demand characteristics. Thus, the request, belongs either to the best-effort or soft-QoE demand matrix of p_1, passes through the ranking-based clustering mask, denoted as $\mathcal{M}_R$. After passing through the mask, the QoE-based demand matrices ($\mathbf{A}_b^{p_1}$ and $\mathbf{A}_g^{p_1}$) converted into the QoE- and Ranked-based demand matrices ($\mathbf{A}_{R_b}^{p_1}$ and $\mathbf{A}_{R_g}^{p_1}$), where $b_R \le b_\kappa$ and $g_R \le g_M$. Similarly, requests from p_2 users are passed through the mask $\mathcal{M}_R$, and $\mathbf{A}_{R_b}^{p_2}$ and $\mathbf{A}_{R_g}^{p_2}$ are constructed, where $b_R \le b_\iota$ and $g_R \le g_N$. Now, C^s_{req} acquired by the clustered requests would not exceed the overall reserved slice queuing capacity over a set of users from either mMTC,

$$\sum_{i=1}^{b_R} C^s_{req}(i) + \sum_{i=\kappa+1}^{g_R} C^s_{req}(i) \leq C^s_{que} \,. \tag{4.44}$$

or eMBB,

$$\sum_{i=1}^{b_R} C^s_{req}(i) + \sum_{i=\iota+1}^{g_R} C^s_{req}(i) \leq C^s_{que} \,. \tag{4.45}$$

4.4.4.2 Demand Clustering for Delay Optimisation: As discussed earlier, requests should be clustered in such a way that the delay (or waiting time) experienced by the requests in a cluster should not exceed the threshold waiting time, which would result in a violation of the provisioning of the agreed QoS to the users (Morgado et al., 2018; Khan et al., 2020b). Delay minimisation within clusters can be modelled as an optimisation problem. The objective is to cluster requests in a way that minimises delay, denoted as $D_{(x)}$, in the slice queue, and reduces the overall rejection ratio on the acquired capacity. This can be described mathematically as follows:

$$\begin{aligned} \min \quad & D_{(x)}, \\ \text{s.t.} \quad & \sum_{x_u=1}^{L} d_{(x_u)} \leq \mathcal{D}_{(x)} \,, \\ & \sum_{u=1}^{M} \sum_{x=1}^{X} \beta_{(u,x)} \leq |\mathbf{U}_{MTC}| \,. \end{aligned} \tag{4.46}$$

The aggregate waiting time of service requests belonging to cluster x should not exceed the threshold time, $\mathcal{D}_{(x)}$, of the particular cluster x in the queue. $\beta_{(u,x)} = 1$ only if the slice service request of user u belongs to a cluster x and is admitted, otherwise 0. All admitted requests from users of set $\mathbf{U}_{MTC}$ should belong to a particular cluster based on homogeneous demand.

Reducing the request rejection ratio at the network edge that occurs due to long delays in the queue and limited capacity is an NP-hard problem. Therefore, optimisation and ML-based approaches are investigated in this work to simultaneously obtain an optimal solution for (4.43) and (4.46). Based on homogeneous-slice service demand, user requests are distributed into R number of clusters by applying a ranking-based approach. *K-mean* (Likas, Vlassis, and Verbeek, 2003) and *Nondominated Sorting Genetic Algorithm II* (NSGA-II) (Deb et al., 2002) have been applied to obtain an optimal number of clusters, $\mathcal{X}$, and their associated requests L over the defined objectives. Due to the massive number of connectivity requests from users, an optimisation approach is applied to requests belonging to p_1. The knowledge gained from the p_1 requests during optimisation is transferred to requests of p_2 in form of coefficients to reduce the time spent on optimisation and speed up the process of admission control, as shown in Figure 4.18. In this optimisation approach, the crucial step is to define an appropriate genetic representation of the requests from set either $\mathbf{u}^{p_1}_{best}$ or $\mathbf{u}^{p_1}_{soft}$. The objective is to gather an optimal number of requests, L, within a cluster x such that the aggregate delay (or waiting time) $D_{(x)}$ of L requests should not exceed the threshold time $\mathcal{D}_{(x)}$. This can be obtained by efficiently scheduling each request u of p_1 in the cluster x based on the minimum aggregate waiting time:

$$D_{(x)} = \sum_{x_u=1}^{L} d_{(x_u)} \leq \mathcal{D}_{(x)} \ , \tag{4.47}$$

where, $d_{(x_u)}$ is the delay induced by request u in cluster x. Thus, by M/M/1 queuing theory (Schwarz et al., 2006), $d_{(x_u)}$ in the proposed model can be obtained as follows:

$$d_{(x_u)} = \frac{1}{(\mu - L)} - \frac{1}{\mu} , \tag{4.48}$$

where, μ represents the mean rate of the request execution from clusters, and L is the request arrival rate within a cluster x. After optimisation, the demand clustering metric, either $\mathbf{A}^{p_1}_{X_b}$ or $\mathbf{A}^{p_1}_{X_g}$, where $b_X \geq b_R$ and $g_X \geq g_R$, passes to the admission and resource management controller for resource allocation.

4.4.4.3 Demand Clustering through Knowledge Transfer: The performance of the optimisation techniques relies on the availability of a huge amount of training data that might include both high-quality and raw data. Many current wireless devices, especially smart devices, are unable to run highly complex tasks, due to their limited computation capacity (Nguyen et al., 2021). So, the data needs to be sent to the central network to acquire an optimum solution to a particular problem (e.g. admission control and resource scheduling of the requests belong to eMBB). This data not only consumes time in training and processing on the network but also creates communication overheads and congestion in the access and core network. TL provides a

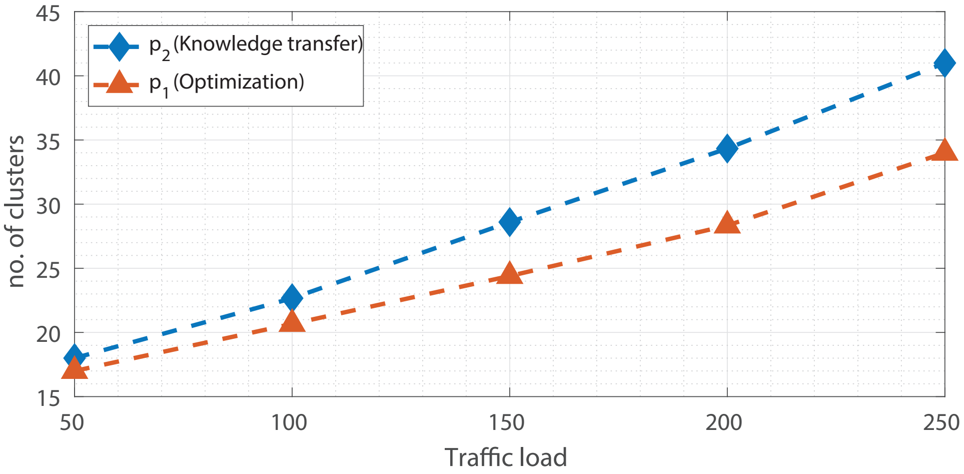


FIGURE 4.19: Slice request clustering using optimisation and machine learning approaches on load $[50, 250]$ and $C^s_{que} = 30$

highly effective solution to this sort of problem (Zhuang et al., 2020; Niu et al., 2020).

To reduce communication overheads or congestion that could occur due to eMBB requests, the concepts of transfer learning have been applied in this work. Upon arrival, the requests of p_2 initially belonging to eMBB pass through the $\mathcal{M}_c$ and $\mathcal{M}_R$ for QoE-based classification and clustering based on homogeneous demand characteristics. The constructed masks for processing the requests of p_1 are applied to the requests of p_2 to save the time spent on classification and clustering. This also blocks the entry of unnecessary data into the network. Next, the objectives, as given in (4.43) and (4.46), are analysed on the clustered requests belonging to $\mathbf{A}^{p_2}_{R_b}$ and $\mathbf{A}^{p_2}_{R_g}$, obtained from mask $\mathcal{M}_R$, as shown in Figure 4.18. In the case of a violation of the objectives, the coefficients, denoted as C_{off}, obtained from curve-fitting to a set of the optimum solutions of p_1 users requests (or mMTC traffic demand) are applied to the requests of p_2 (or mMBB traffic demand). The set of optimum solutions is obtained after running a number of experiments on the requests of p_1. The optimal L number of requests for each cluster of p_2 users obtained from C_{off} can be represented as follows:

$$L(p_2) = \sum_{i=0}^{\eta} C_{off}(i)(|\mathbf{U}_{MBB}|)^i , \tag{4.49}$$

where η is the degree of curve-fitting. Hence, the optimal L number of requests of p_2 are clustered in each cluster x based on knowledge gained from the requests of p_1 in the form of coefficients. Thereafter, $\mathcal{X}$ number of clusters would be acquired over the load of p_2, with L number of requests in each cluster. Such as, Figure 4.19 illustrates the no of clusters obtained by employing the optimisation and knowledge transfer in the proposed model on load [50,250]. Now, the demand clustering metric, either $\mathbf{A}_{X_b}^{p_2}$ or $\mathbf{A}_{X_g}^{p_2}$, where $b_X \geq b_R$ and $g_X \geq g_R$, passes to the admission and resource management controller for resource allocation.

4.4.5 Admission Control and Resource Management

In this section, a dynamically adaptive admission control and resource management scheme of the SCAC model is proposed. The proposed scheme reduces the dropping probability of slice requests occurring due to intra-slice congestion using RL-based admission control, intra/inter-slice and slice elasticity-based resource allocation approaches, as shown in Figure 4.20. Moreover, this scheme also leads to enhanced network resource utilisation through efficient resource allocation and scheduling. More details on this scheme of the proposed SCAC model are explained in the following subsections.

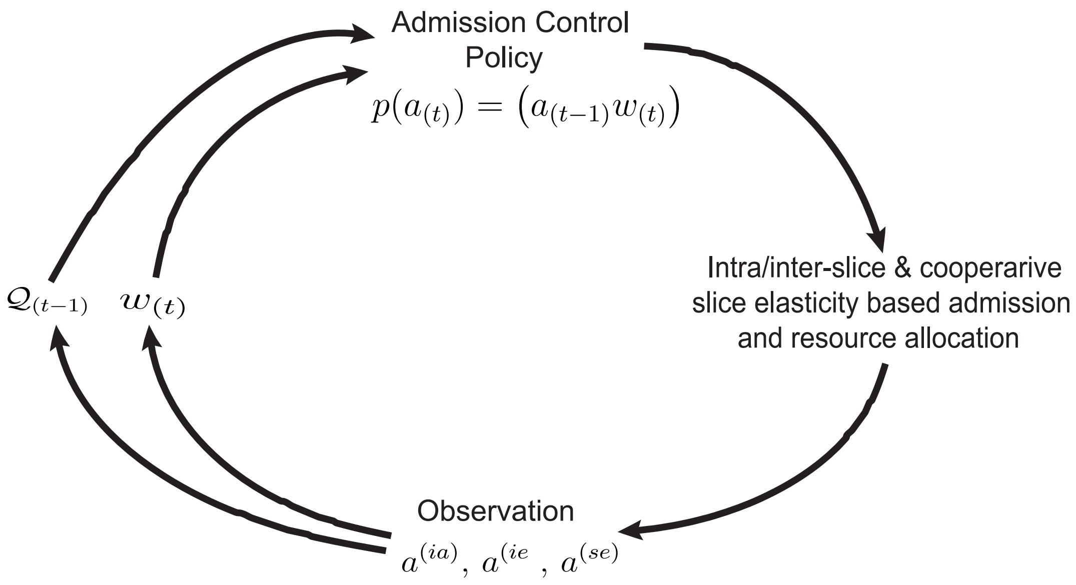

Admission Control
Policy
$p(a_{(t)}) = \big(a_{(t-1)}w_{(t)}\big)$
Intra/inter-slice & cooperarive
slice elasticity based admission
and resource allocation
Observation
$a^{(ia)}, a^{(ie}, a^{(se)}$
$\mathcal{Q}_{(t-1)}$
$w_{(t)}$

FIGURE 4.20: Systematic diagram of RL-based admission control, intra/inter-slice and slice elasticity based resource allocation

4.4.5.1 Slice Selection and Admission control: Future networks are expected to be reconfigured dynamically. The network slicing feature in 5G networks has the ability to provide heterogeneity of massive traffic with the capacity of adaptive and dynamic resource allocation within a slice (Gupta and Jha, 2015; Kaloxylos, 2018).

In this section, an ML-based admission control function is proposed that applies the aforementioned reinforcement learning technique to obtain optimal slice selection and admission control. When the network receives a `Service Request` and `Registration request` from a device with guaranteed soft (or best-effort) QoE demand attempting to access the network for kth application service, where $k \in \Lambda$, the AMF function accesses that request and verify from the respective repository $\mathbf{A}_{X_g}$ (or $\mathbf{A}_{X_b}$) to grant admission to a suitable slice. To achieve this, a cumulative soft-decision technique using cost function is proposed, which is derived from the demand matrix ($\mathbf{A}_{X_g}$ or $\mathbf{A}_{X_b}$) and network weights $\mathbf{w}$. Accordingly, the cost function, denoted as $\mathbf{v}_{X_g}$ and $\mathbf{v}_{X_b}$, can be represented as

$$\mathbf{v}_{X_g} = \mathbf{A}_{X_g}\mathbf{w} = \left[v_{\kappa+1}, v_{\kappa+2}, \ldots, v_{g_X}\right]^{\mathsf{T}} ,$$

and

$$\mathbf{v}_{X_b} = \mathbf{A}_{X_b}\mathbf{w} = \left[v_1, v_2, \ldots, v_{b_X}\right]^{\mathsf{T}} . \tag{4.50}$$

To achieve dynamic uniform slice allocation among requests, a set of network characteristic learning weights, denoted as $\mathbf{w} = \{\omega_1, \omega_2, \ldots, \omega_J\}$ is defined, which indicates the current network load status, resource availability, and other parameters. The learning weights can be computed using the normal equation for multivariate linear regression, as:

$$\mathbf{w} = (\mathbf{A}_{X_g})^{-1}\mathbf{I}, \tag{4.51}$$

where $\mathbf{I}$ is the identity matrix with respect to uniform slice distribution among all clustered requests and $(\mathbf{A}_{X_g})^{-1}$ is the Moore–Penrose inverse of $\mathbf{A}_{X_g}$. The resultant weighting factors are dynamic in nature and might be modified systematically or non-systematically by a change in network parameter values, with respect to time and network load. The learning weights can also be obtained using gradient descent linear regression techniques. However, such techniques are designed for scenarios in which a large amount of data (J >1000) is available, and they have high computational complexity to converge due to their iterative nature (Hospedales et al., 2020).

The estimated cost value v_x for xth clustered request is obtained as

$$v_{x_g} = \sum_{j=1}^{J} \mathbf{g}_{xj}\mathbf{w}_j, \ \forall \ \mathbf{g}_{xj} \in \mathbf{A}_{X_g} \text{ and } \mathbf{w}_j \in \mathbf{w}, \tag{4.52}$$

where, $\mathbf{g}_{xj}$ is the xth clustered request jth resource demand. Similarly, v_{x_b} obtained for the requests belong to $\mathbf{A}_{X_b}$. Now, v_{x_g} and v_{x_b} are placed in the respective queue for action taken for admission control. Action on the clustered slice requests is based on an admission control policy $p_{(a_b,a_g)}$ built through an RL algorithm. The policy aims to reduce the rejection ratio by taking appropriate action on instantaneous system rewards on guaranteed soft or best-effort request admission and resource allocation. The policy at time t can be expressed as follows:

$$p(a_{b_{(t)}}, a_{g_{(t)}}) = \left(a_{b_{(t-1)}} w_{b_{(t)}}, a_{g_{(t-1)}} w_{g_{(t)}}\right), \tag{4.53}$$

where, a_g, a_b, w_g, and w_b determine the number of accepted guaranteed soft (g_X) and best-effort (b_X) clustered requests, and their associated rewards on previous action, respectively. Initially, the action on the admission of the clustered requests (either guaranteed or best-effort) to the network is based on their ratio from overall demand with $w_b = w_g = 1$. For example, the ratio of g_X to b_X is $1 : 1$, available network resources are allocated evenly (or 50%) to requests belonging to g_X and (or) b_X. Following successful clustered request admission to the network, the proposed model applies the strategies of intra-slice and inter-slice resource allocation, as well as adaptability of slice elasticity for their service provisioning, as described in the following section.

4.4.5.2 Slice Resource Allocation: However, in a dense environment, where millions of devices with varied heterogeneous demands are deployed, blocking probabilities of slice requests may increase and reach the undesirable territory. In addition, if the host server of the slice is compromised or becomes unavailable, the slice operation would also be affected, causing outages (Sattar and Matrawy, 2019; Ojijo and Falowo, 2020). In this section, this problem is addressed by the dynamic reconfiguration of slice bounds. Three adaptive resource allocation techniques: intra-slice, inter-slice, and cooperative slice elasticity, have also been developed in this model, as discussed in detail in the following subsection:

A) Intra-slice resource allocation: 5G network slicing accommodates traffic heterogeneity with an adaptive and dynamic resource allocation mechanism within a slice through a slice resource pool (3GPP, 2018c; 3GPP, 2018b). Assumed that, slices are associated with re-configurable slice bounds, denoted as B_l^s and B_u^s, where B_l^s and B_u^s represent the slice lower and upper bounds, respectively, and is obtained by

$$B_l^s = \frac{1}{|\mathcal{S}|}(i) \text{ and } B_u^s = \frac{1}{|\mathcal{S}|}(i+1), \tag{4.54}$$

where i $=\{0, 1, 2, \ldots, |\mathcal{S}|\}$. Algorithm 11 represents clustered request admission and

Algorithm 11: Cluster request admission by intra-slice resource allocation

```
Input: $\mathbf{A}_X \neq 0$, $R_A^s \neq 0$, where, $x \in \mathcal{X}$ and $s \in S$, calculate $v_x$.
Output: $U > 0$, $\mathcal{Q}$, and $w$.
begin
    if $(B_l^s < v_x \leq B_u^s)$ then
        if $(R_A^s \geq R_x)$ then
            Admit $x$th cluster request and assign resources.
            Update available resources $R_A^s$ in slice $s$ resource pool.
            Calculate resource utilisation $U$.
            Compute respective values of $\mathcal{Q}$ and $w$.
        else
            if (Inter-slice resource allocation) then
                Update slice learning bounds ($B_l^s$, $B_u^s$) via (4.58) and (4.59)
                Compute Algo. 12
            else
                Apply slice elasticity approach via (4.62)
                Compute Algo. 13
            end
        end
    else
        Reassess $v_x$
        Compute Algo. 11
    end
end
```

intra-slice resource allocation. If the cost values are within the slice configurable bounds, clustered request admission can be guaranteed, subject to the availability of resources in the slice, denoted as R_A^s. Respectively, on overall admissions, the number of rejections, denoted as $\mathcal{Q}$ and rewards w are computed. Subsequently, if resources are unavailable, the possibility of inter-slice resource allocation or cooperative slice elasticity is assessed. Otherwise, the requests might be placed back in the respective matrix $\mathbf{A}_g$ (or $\mathbf{A}_b$) and reassessed according to the change in circumstances and cost values. Such consideration may include the possibility of back-off with the time-shift nature of application or assessment of QoS. Clustered requests belonging to best-effort QoE will go through a similar process. Now, rejections or $\mathcal{Q}$ value after admission of clustered requests can be obtained as:

$$\mathcal{Q}^{(ia)}(a_{g_{(t-1)}}) = g_X - a_g^{(ia)}\,, \tag{4.55}$$

and

$$\mathcal{Q}^{(ia)}(a_{b_{(t-1)}}) = b_X - a_b^{(ia)}\,, \tag{4.56}$$

where $a^{(ia)}$ determines the admission of the clustered requests using intra-slice resource allocation. Now, based on the previous action, rewards are obtained as

$$w = \begin{cases} w_{g_{(t)}} = 2w_{g_{(t-1)}}, w_{b_{(t)}} = w_{b_{(t-1)}} & \mathcal{Q}(a_{g_{(t-1)}}) > \mathcal{Q}(a_{b_{(t-1)}})\,, \\ w_{g_{(t)}} = w_{g_{(t-1)}}, w_{b_{(t)}} = w_{b_{(t-1)}} & \mathcal{Q}(a_{g_{(t-1)}}) = \mathcal{Q}(a_{b_{(t-1)}})\,, \\ \textit{Inter-slice admission} & \text{otherwise}\,. \end{cases} \tag{4.57}$$

In the case of $\mathcal{Q}(a_{g_{(t-1)}}) > \mathcal{Q}(a_{b_{(t-1)}})$ from intra-slice admission and resource allocation, the reward w_g will be updated, due to the higher priority of guaranteed soft-QoE demand over best-effort demand. The updated w_g reduces rejections of clustered requests from the g_R queue; however, it may increase the rejection ratio of requests belonging to b_R, due to resource scarcity in the network. In such circumstances, inter-slice admission control is applied for resource allocation, as explained in the following subsection.

B) Inter-slice resource allocation: Inter-slice admission control is another key feature for resource allocation. It is defined in recent 3GPP items and developed based on roaming techniques (3GPP, 2018c). In this setting, the devices belonging to a particular clustered request are configured with two slices: (1) the primary slice (aka serving slice) and (2) the neighbouring slice, denoted as $(s+1)$ and $(s-1)$ (used for fall-back during instances of primary slice unavailability). To access the neighbouring slices, the slice s bounds (B_l^s and B_u^s) will be updated (via (4.58) and (4.59)) by a certain bound index denoted as δ for the xth clustered request only. If handover to the neighbouring slice is completed, the neighbouring slice will have full control over the admitted clustered request. The bounds are mathematically expressed as

$$B_l^s = B_l^s - \delta_{(ie)}(B_u^{(s-1)} - B_l^{(s-1)})\,, \tag{4.58}$$

and

$$B_u^s = B_u^s + \delta_{(ie)}(B_u^{(s+1)} - B_l^{(s+1)}), \tag{4.59}$$

where $\delta_{(ie)} = [0, 0.5]$, according to the central limit theorem. Algorithm 12 represents the inter-slice resource allocation strategies from slice $(s-1)$ and $(s+1)$, respectively. Assumed that the capital expenditure (CAPEX) is proportional to the slice index, where $CAPEX_{(s-1)} < CAPEX_{(s+1)}$. The clustered user requests are guaranteed admission subject to the availability of resources, denoted as R_A from either $(s-1)$ or $(s+1)$ slice. Subsequently, rejection rate $\mathcal{Q}$, and resource utilisation are computed for the next action taken by the admission policy. The clustered requests belonging to best-effort QoE will go through a similar process if needed. Thus, the rejection ratio or $\mathcal{Q}^{(ie)}$ value after intra-slice and inter-slice admission and resource allocation will be as follows:

$$\mathcal{Q}^{(ie)}(a_{g_{(t-1)}}) = g_X - a_g^{(ia)} - a_g^{(ie)} = g_X - a_g^{(ia,ie)}\,, \tag{4.60}$$

and

$$\mathcal{Q}^{(ie)}(a_{b_{(t-1)}}) = b_X - a_b^{(ia)} - a_b^{(ie)} = b_X - a_b^{(ia,ie)}\,, \tag{4.61}$$

where, $a^{(ia,ie)}$ determines the admission of the clustered requests using intra-slice and inter-slice resource allocation, as given in (Perveen, Patwary, and Aneiba, 2019).

C) Cooperative slice elasticity for resource allocation: In the case of insufficient

Algorithm 12: Cluster request admission by inter-slice resource allocation

```
Input: $(s-1) \in S, \quad (s+1) \in S, \quad R^s_A = 0$ or $R^s_A < R_x$, Update slice
  learning bounds $(B^s_l, B^s_u)$ via (4.58) and (4.59).
begin
    if $(B^s_l < v_x \leq B^s_u)$ then
        if $(R^{(s-1)}_A \geq R_x) \vee (R^{(s+1)}_A \geq R_x)$ then
            Admit $x$th request and assign the resources.
            Update $R^{(s-1)}_A$ or $R^{(s+1)}_A$.
            Calculate resource utilisation $U$.
            Compute respective $\mathcal{Q}$ value
        end
    else
        Reassess $v_x$
        Compute Algo. 11
    end
end
```

primary slice capacity C^s and privacy constraints, cooperative slice elasticity is proposed to accommodate demand. Through this feature, the primary slice capacity extends to an absolute value by a certain elasticity index, denoted as $\delta_{(se)}$. The neighbouring slice capacity ($C^{(s+1)}$ or $C^{(s-1)}$) will be temporally allocated to the primary slice s for a defined time period, as expressed below.

$$C^s = C^s + \delta_{(se)}(C^{(s+1)} + C^{(s-1)})\,. \tag{4.62}$$

Thus, the rejection ratio or $\mathcal{Q}^{(se)}$ value after intra-slice, inter-slice, and cooperative slice elasticity admission and resource allocation will be as follows:

$$\mathcal{Q}^{(se)}(a_{b_{(t-1)}}) = b_X - a^{(ia)}_b - a^{(ie)}_b - a^{(se)}_b = b_X - a^{(ia,ie,se)}_b\,, \tag{4.63}$$

and

$$\mathcal{Q}^{(se)}(a_{g_{(t-1)}}) = g_X - a^{(ia)}_g - a^{(ie)}_g - a^{(se)}_g = g_X - a^{(ia,ie,se)}_g\,. \tag{4.64}$$

Algorithm 13 presents the proposed resource allocation approach with the cooperative slice elasticity. This contingency solution is one of the key features of the proposed work. It implements capacity elasticity among the slices and captures the slice policies, which are defined by the mobile network operator.

Algorithm 13: Admission control with cooperative slice elasticity

```
Input: $s \in S$, $R_x \neq 0$, $v_x$, $C^s \neq 0$.
if $(B^s_l < v_x \leq B^s_u)$ then
    if $(C^s \geq R_x)$ then
        Admit $x$th request and assign the resources.
        Update $R^s_A$.
        Calculate resource utilisation $U$.
        Compute respective $\mathcal{Q}$ value
    else
        Update slice capacity bounds $C^s$ via (4.62)
        Compute Algo. 13
    end
else
    Reassess $v_x$
    Compute Algo. 11
end
```

4.4.5.3 Slice Resource Scheduling: Along with optimised admission control, resource scheduling also has a significant impact on network QoS. An adequate slice scheduling guarantees support for diverse QoS requirements of different use cases, as identified by the International Telecommunication Union (ITU) (Schmidt, Chang, and Nikaein, 2019). One of the important measures to quantify network QoS is the resource utility estimation on efficient scheduling. Additionally, the shape of the utility function varies in line with the device application, as well as with network characteristics (Kuo and Liao, 2008; Ojijo and Falowo, 2020). In the proposed work, traffic demand is categorised into two types: (1) best-effort QoS slice traffic demand, and (2) guaranteed Soft-QoS slice traffic demand. Thus, the goal of the proposed multi-slice resource scheduling is efficient resource allocation among clustered slice requests to maximise resource utilisation and overall throughput. The utility function is the projection of slice request demand on allocated and desired resources (Tan et al., 2015; Han et al., 2019). Accordingly, the uth request utility $U(R_{x_u})$ is obtained by the following equation:

$$U(R_{x_u}) = \begin{cases} \varphi\, e^{pq}, & R_{x_u} < R_d \\ (1-\varphi)\, e^{-pq} - 1 & R_{x_u} \geq R_d, \end{cases} \tag{4.65}$$

where R_{x_u} and R_d represent the achieved and desired resources of the uth slice request, p is the difference between the achieved and desired resources, q and φ represents the utility function slope and the utility function curve slope (as in Tan et al., 2015). The achieved resources, R_{x_u}, can be obtained as

$$R_{x_u} = \frac{\nu_{x_u}}{\sum_{x_u=1}^{L} \nu_{x_u}} r_{x_u}, \tag{4.66}$$

where ν as a channel condition is the non-negative resource share of the slice request among clustered requests. r_{x_u} is the peak rate or maximum achievable rate of the uth request from the cluster x. The aggregate resource allocation to all clustered requests should be equivalent to or less than the total slice capacity C^s. The γ_u, minimum guaranteed rate requirement of the uth soft QoS traffic device is non-negative and non-zero (i.e., $R_{x_u} \geq \gamma_u > 0$). In the case of best-effort traffic, γ_u can be zero such that $\gamma_u = R_d = 0$. Thus, (4.65) can be rewritten as

$$U_b(R_{x_u}) = (1-\varphi)\, e^{-qR_{x_u}} - 1. \tag{4.67}$$

The marginal utilities, denoted as $u(R_{x_u})$, of the achieved resource can be computed by taking the derivative of (4.65) and can be expressed as

$$u(R_{x_u}) = \frac{dU(R_{x_u})}{d(R_{x_u})} = \begin{cases} \varphi q e^{pq}, & R_{x_u} < R_d \\ (1-\varphi)\, q e^{-pq} & R_{x_u} \geq R_d. \end{cases} \tag{4.68}$$

By utility $U(R_{x_u})$, the xth cluster utility $U(R_x)$ is the sum of individual utilities, as in (Caballero et al., 2018), and can be obtained as

$$U(R_x) = \sum_{x_u=1}^{L} U(R_{x_u}), \tag{4.69}$$

where, $U(R_x)$ is computed with regards to best-effort and/or guaranteed soft cluster demand, which are represented as $U_b(R_x)$ and $U_g(R_x)$, respectively. Now, slice utility can be obtained as

$$U^s = \sum_{x=1}^{X} \alpha_x U(R_x)\,, \tag{4.70}$$

whereby, α_x determines that the cluster x is admitted to the slice s. Accordingly, the overall network utility U over the slices from set $\mathcal{S}$ is derived as

$$U = \sum_{s=1}^{S} U^s. \tag{4.71}$$

Network utility maximisation is key to optimal resource scheduling and allocation. Thus, resource allocation problems to the clustered best-effort and soft QoS traffic are formulated in terms of the maximisation of the utility function, as proved mathematically in the following:

Lemma 1. *Assumed that slice s is serving massive heterogeneous traffic with the best-effort QoS cluster demand only. To ensure maximum slice utilisation, the maximum aggregate resource allocation to the X number of clusters belonging to best-effort QoS demand from the slice s will be equivalent to or less than the total capacity of the slice C^s. This can be expressed mathematically as*

$$\max : \sum_{x=1}^{X} \alpha_x U_b(R_x) = U(R_{(b,x)}) \leq 1\,, \tag{4.72}$$

and

$$C^s \geq R_b^s = \sum_{x=1}^{X} R_{(b,x)}, \;\; s.t. \;\; R_{(b,x)} > 0\,, \tag{4.73}$$

where $R_{(b,x)}$ determines the resources assigned to the best-effort cluster x on request, R_b^s is the slice assigned aggregate resources to X number of admitted clusters, and $U_b(R_x)$ is the x cluster acquired utility from the slice and $U(R_{(b,x)})$ is the overall slice utility with respect to best-effort demand.

Proof. To prove the statement accordingly and achieve an optimal solution, the Lagrange function has been considered.

$$L = U(R_{(b,x)}) + \lambda(C^s - R_b^s). \tag{4.74}$$

Its assume the gradient of L (denoted as ∇L), according to $R_{(b,x)}$ and λ, is equal to zero, and simplified as follows.

$$u(R_{(b,x)}) = \lambda \frac{d(R_b^s)}{d(R_{(b,x)})}, \tag{4.75}$$

$$\nabla L_\lambda = C^s - R_b^s,$$

$$C^s = R_b^s, \tag{4.76}$$

where C^s is the slice capacity as in (4.73) and $u(R_{(b,x)})$ is the best-effort cluster request marginal utility that gives us the slope of the utility curve φ, and $\frac{R_b^s}{R_{(b,x)}}$ will gives the slope of utility function q, where utilisation will be at a maximum. □

Lemma 2. *Suppose slice s is serving massive heterogeneous traffic with soft QoS demand only. To ensure maximum slice utilisation, the maximum aggregate resource allocation to the X number of clusters belonging to soft QoS demand from the slice s will be equivalent to or less than the total capacity of the slice C^s. This can be expressed mathematically as*

$$\max : \sum_{x=1}^{X} \alpha_x U_g(R_x) = U(R_{(g,x)}) \leq 1, \tag{4.77}$$

and

$$C^s \geq R_g^s = \sum_{x=1}^{X} R_{(g,x)}, \ \ s.t. \ \ R_{(g,x)} > 0, \tag{4.78}$$

where $R_{(g,x)}$ determines the resources assigned to the guaranteed soft cluster x on request, R_g^s is the slice assigned aggregate resources to X number of clusters, and $U_g(R_x)$ is the x cluster acquired utility from the slice and $U(R_{(g,x)})$ is the overall slice utility with respect to soft demand.

Proof. Let's prove the statement using the Lagrange function.

$$L = U(R_{(g,x)}) + \lambda\big(C^s - R_g^s\big). \tag{4.79}$$

Let's consider the gradient of L (denoted as ∇L) according to $R_{(g,x)}$ and λ, equal to zero and simplify it.

Let's consider the gradient of L (denoted as ∇L) according to $R_{(g,x)}$ and λ, is equal to zero and simplified as follows.

$$u(R_{(g,x)}) = \lambda \frac{d\big(R_g^s\big)}{d\big(R_{(g,x)}\big)}, \tag{4.80}$$

$$\nabla L_\lambda = C^s - R_g^s,$$

$$C^s = R_g^s, \tag{4.81}$$

where C^s is the slice capacity as in (4.78) and $u(R_{(g,x)})$ is the soft cluster request marginal utility that gives the slope of the utility curve φ, and $\frac{R_g^s}{R_{(g,x)}}$ will gives the slope of utility function q, where the utilisation will be at a maximum. □

Lemma 3. *Let's assume slice s is serving incoming heterogeneous traffic with both best-effort and soft QoS demand. To ensure maximum slice utilisation, the maximum aggregate resource allocation to the number of clusters belonging to best-effort and guaranteed soft demand from the slice s will be equivalent to or less than the total capacity of the slice C^s. Mathematically*

$$\max : \sum_{x=1}^{X_b} \alpha_x U_b(R_x) + \sum_{x=1}^{X_g} \alpha_x U_g(R_x) = U(R_{(b,x)}) + U(R_{(g,x)}) \leq 1\,, \tag{4.82}$$

and

$$C^s \geq \sum_{x=1}^{X_g} R_{(b,x)} + \sum_{x=1}^{X_g} R_{(g,x)} = R_b^s + R_g^s\,, \;\; s.t. \;\; R_{(g,x)} > 0, R_{(b,x)} > 0\,. \tag{4.83}$$

Proof. Let's prove the statement using the Lagrange function.

$$L = U(R_{(b,x)}, R_{(g,x)}) + \lambda\big(C^s - (R_b^s + R_g^s\big)\,. \tag{4.84}$$

Let's consider the gradient of L (denoted as ∇L) according to R_b^s, R_g^s as in (4.75) and (4.80), and λ equal to zero to simplify it.

$$u(R_{(b,x)}) = \lambda \frac{d(R_b^s)}{d(R_{(b,x)})}, \quad \text{and} \quad u(R_{(g,x)}) = \lambda \frac{d(R_g^s)}{d(R_{(g,x)})},$$

$$\nabla L_\lambda = C^s - (R_b^s + R_g^s),$$

$$C^s = R_b^s + R_g^s. \tag{4.85}$$

Subsequently, $\frac{u(R_{(b,x)})}{u(R_{(g,x)})}$ will gives the slope of the utility curve φ, and $\frac{R_b^s}{R_g^s}$ will gives the slope of utility function q, where the utilisation will be the maximum. □

Requests within a cluster are scheduled in descending order based on their maximum utility and waiting time in the queue for resource allocation. Resources are allocated to the clustered requests according to the scheduling order. In the case of extra resources, the remaining resources are returned to the pool for successful operations of inter-slice resource allocation or slice elasticity if needed.

4.4.6 Performance Analysis and Results

For performance evaluation of the proposed model, an analytical model is developed in MATLAB. In this model, a virtual network is established with different system parameters, as given in (GSMA, 2019), to support mMTC and eMBB demand. Traffic load associated with mMTC and eMBB are considered to be in the range of 50 to 250 user requests. Considering the number of supporting slices, queue capacity and threshold waiting time are $S = 5$, $C_{que}^s = 30$, $\mathcal{D}_{(x)} = 0.2$ ms, respectively. $\mathcal{J}(1) = [1,5]$, $\mathcal{J}(2) = [10,80]$ ms, $\mathcal{J}(3) = [10^{-2}, 10^{-7}]$, and $\mathcal{J}(4) = [10,100]$ MHz are the considered ranges of priority, latency sensitivity, user-service-specific packet loss, and desired resource demand from the slice belonging to $\mathcal{S}$, respectively. The overall demand is normalised for simplicity. Through these considered parameters, outcomes of our proposed SCAC model are shown in Table 4.6 to Table 4.11 and also discussed in detail in the following:

4.4.6.1 Impacts of Optimisation and Knowledge Transfer on Bottleneck Congestion Reduction: Table 4.6 illustrates the ratio of bottleneck congestion at various loads. It can be seen that the acquired congestion from the proposed approach is significantly lower compared to the ground truth, whereby the results are obtained without using any approach. The achieved gain with regards to bottleneck congestion control from the proposed approach on p_1 (or mMTC traffic) and p_2 (or eMBB traffic) at load 50 is 40% on the ground truth values. The gain with regards to bottleneck congestion control from the proposed approach increases significantly with increasing load. For example, at load 250, the gains of p_1 and p_2 over ground truth on bottleneck congestion are 91% and 74%, respectively. The lower bottleneck congestion is due to clustering of the requests by using optimisation and machine learning approaches in proportional to the slice queue capacity. Moreover, bottleneck congestion occurring among p_1 requests is lower compared to that among p_2 requests. The increasing gain of p_1 requests on bottleneck congestion is due to the implementation of ranking-based and K-mean clustering with optimisation on capacity, as well as delay minimisation. The resource demand from mMTC requests is lower than that of eMBB requests. Thus, within the range of defined objectives, the proposed approach results in more users being accommodated from a cluster belongs to the mMTC traffic load. Ranking-based clustering is implemented on p_2 to efficiently utilise queue capacity. The knowledge gained by applying the optimisation approach to p_1 is implemented on p_2 to acquire an optimal number of requests in each cluster for delay minimisation. The capacity demand from p_2 requests is greater; hence, the number of clusters on eMBB load would be more with a fewer number of requests in each to obey the defined objectives. In bottleneck congestion, the proposed optimisation and knowledge transfer show superiority over the ground truth approach with 2.8% and 7.8% mean value and 2.8% and 8.2% standard deviation respectively.

TABLE 4.6: Computation of slice bottleneck congestion using optimisation and machine learning approaches on load $[50, 250]$ and $C^s_{que} = 30$

	Traffic load				
Approaches	50	100	150	200	250
Ground truth	40%	63%	72%	84%	87%
	Mean: 69.2% Std.: 16.94 %				
Knowledge transfer on p_2	0%	1.8%	5%	9%	23%
	Mean: 7.8% Std.: 8.2%				
Optimisation on p_1	0%	0%	2.2%	4.1%	7.5%
	Mean: 2.8% Std.: 2.8%				

4.4.6.2 Impacts of Proposed Resource Allocation Approaches on Intra-Slice Congestion Reduction: Table 4.7 illustrates the behaviour of intra-slice, inter-slice, and cooperative slice elasticity-based resource allocation with varied loads. It can be seen that the achieved request rejection ratio in the case of intra-slice resource allocation is greater than that achieved using cooperative and inter-slice resource allocation. Clustered requests are admitted to the particular slices based on their cost estimation value to attain fair admission. After slice allocation, resources are allocated to the clustered requests in order from the admission queue. When the rejection ratio begins to increase due to resource scarcity within the slices, the clustered requests from the admission queue are diverted to neighbouring slices, along with a change in their cost bounds. Now, the diverted clustered requests are admitted to the neighbouring slices to reduce the rejection ratio that occurs due to intra-slice congestion. The rejection ratio in cooperative slice elasticity is greater than that of the inter-slice approach but less than that of intra-slice recourse allocation. This significant difference from the other two approaches is due to the slice scalability factor. Based on bounds, each slice is allowed to scale its capacity up to δ value of the available capacity of the neighbouring slice. This results in more requests being admitted compared to the intra-slice approach but less than inter-slice resource allocation.

TABLE 4.7: Comparison of p_1 and p_2 requests rejection due to intra-slice congestion over intra-slice, inter-slice, and cooperative slice elasticity-based resource allocation at a load of $[50, 250]$

Resource allocation	p_1 traffic load					p_2 traffic load				
	50	100	150	200	250	50	100	150	200	250
Intra-slice	0%	18%	44%	52%	56%	30%	45%	63%	74%	78%
	Mean: 34% Std.: 21.54%					Mean: 58% Std.: 18%				
Cooperative	0%	0%	28%	29%	31%	15%	26%	33%	40%	53%
	Mean: 17.6% Std.: 14.4%					Mean: 33.4% Std.: 12.8%				
Inter-slice	0%	0.80%	2.00%	3.17%	5.02%	0%	12%	18%	23%	27%
	Mean: 2.2% Std.: 1.8%					Mean: 16% Std.: 9.4%				

In Table 4.7, at the beginning of the results, the requests belonging to p_1 have a lower rejection ratio in all three cases, due to their resource demand. The lower resources demand from mMTC leads to smart accommodation of requests and reduction of congestion using the proposed approaches. It can be seen that the overall request rejection ratio increases with increased load. However, the achieved gain of the inter-slice resource allocation approach with load 250 is 84% and 91% on cooperative and intra-slice resource allocation approaches, respectively. A similar trend can be seen in the case of p_2 requests. Moreover, mean rejection ratio of the clustered requests of p_2 in all three cases is greater than that on p_1 for loads between 50 and 250. The greater rejection ratio on eMBB load is due to the aggregate resource demand, which creates more competition for resource allocation among clusters. The achieved gain of the inter-slice resource allocation approach at the load of 50 is 15% and 30% on the cooperative and intra-slice resource allocation approaches, respectively. With an increase in load, a similar trend of increase in rejection ratio can be seen in Table 4.7.

4.4.6.3 Impacts of RL-based Admission Policy on Guaranteed soft and Best-effort Traffic Load: This section illustrates the performance of the proposed policy-based admission control approach in terms of rejection ratio of guaranteed soft ($p_{1(g)}$ or $p_{2(g)}$) and best-effort ($p_{1(b)}$ or $p_{2(b)}$) requests at various loads. The aim is to minimise the intra-slice rejection ratio of the guaranteed soft requests belonging to either mMTC or eMBB. Therefore, an RL-based admission control policy has been proposed in this work. In the case of congestion, guaranteed soft requests belonging either to p_1 or p_2 would have higher priority on best-effort requests for admission and resource allocation from the slice. The proposed approach is evaluated through intra-slice (see Table 4.8), inter-slice (see Table 4.9), and cooperative slice elasticity (see Table 4.10), based resource allocation approaches.

TABLE 4.8: Comparison of guaranteed soft and best-effort request rejection on RL-based admission policy and intra-slice resource allocation

	p_1 traffic load					p_2 traffic load				
Requests	50	100	150	200	250	50	100	150	200	250
Best-effort	0%	23%	38%	40%	43%	19%	33%	40%	46%	54%
	Mean: 28.8% Std.: 15.96%					Mean: 38.4% Std.: 12%				
Guaranteed soft	0%	6%	11%	19%	27%	14%	25%	31%	37%	49%
	Mean: 12.6% Std.: 9.5%					Mean: % 31.2 Std.: 11.7%				

As explained earlier, the overall request rejection ratio from intra-slice admission and resource allocation is greater than that of the other two approaches, as shown in Table 4.8. At the beginning, the rejection ratio in the case of $p_{1(g)}$ and $p_{1(b)}$ admissions is the same. But as demand increases, rejection ratio increases such that at load 250, $p_{1(b)}$ have approximately 37% more rejections as compare to $p_{1(g)}$. The lower rejection ratio in $p_{1(g)}$ is achieved as a result of the proposed admission policy, whereby a number of requests are admitted to a particular slice with regards to their cost value and rewards. Therefore, clustered requests are queued in the admission queue in order with respect to their rewards. Rewards are set based on the rejection ratio of the requests in the previous action, as seen in Figure 4.20. Based on the preferences mentioned in the policy, the acquired reward of the guaranteed soft requests would be more than the best-effort request on rejection in each action. Therefore, the rejection percentage of the guaranteed soft is less than best-effort requests in intra-slice admission and resource allocation. A similar trend can be seen in the case of $p_{2(g)}$ and $p_{2(b)}$. However, the greater number of rejections on $p_{2(g)}$ and $p_{2(b)}$ compared to $p_{1(g)}$ and $p_{1(b)}$ are due to implementation of knowledge transfer, which give solutions approximately closer to the optimal solutions, as obtained in case of p_1.

TABLE 4.9: Comparison of guaranteed soft and best-effort request rejection on RL-based admission policy and inter-slice resource allocation

	p_1 traffic load					p_2 traffic load				
Requests	50	100	150	200	250	50	100	150	200	250
Best-effort	0%	3%	12%	20%	37%	0%	22%	29%	37%	43%
	Mean: 14.4% Std.: 13.3%					Mean: 26.2% Std.: 14.9%				
Guaranteed soft	0%	0%	0%	0%	0%	0%	7%	14%	21%	25%
	Mean: 0% Std.: 0%					Mean: 13.4% Std.: 9%				

When the rejected requests from intra-slice congestion are diverted to neighbouring slices for admission, rewards will be updated. Accordingly, the guaranteed soft request would have more preferences on best-effort for admission and resource allocation due to the change in the rewards, as shown in Table 4.9. The inter-slice admission and resource allocation attempts to accommodate more guaranteed soft requests to lower their rejection ratio. Therefore, the rejection ratio of best-effort requests is significantly more than that of guaranteed soft requests. However, the overall rejection ratio of guaranteed soft and best-effort requests is significantly less than that of intra-slice and cooperative slice elasticity-based admission and resource allocation. It is also noticed that the mean rejection ratio in the case of eMBB traffic load is more compared to mMTC, as shown in Table 4.9.

A similar trend can be seen in the case of the cooperative slice elasticity-based admission and resource allocation approach, as shown in Table 4.10, where the rejection ratio of guaranteed soft requests is also less than that of best-effort requests. Moreover, the overall rejection ratio of cooperative slice elasticity is lower than that of intra-slice but more than that of inter-slice-based admission and resource allocation. This is due to the scalability feature of the cooperative scheme, where each slice has access to δ value of the available resources of the neighbouring slices to admit the rejected requests. In this way, the primary slice capacity extends to an absolute value with regards to δ for a particular period, that leads to enhance network performance and user-acquired QoE. It is also noticed that the mean rejection ratio in the case of eMBB traffic load is more compared to mMTC, as shown in Table 4.10. This is due to more resource demand from eMBB compared to mMTC traffic load.

TABLE 4.10: Comparison of guaranteed soft and best-effort request rejection on RL-based admission policy and cooperative slice elasticity-based resource allocation

	p_1 traffic load					p_2 traffic load				
Requests	50	100	150	200	250	50	100	150	200	250
Best-effort	0%	8%	18%	27%	40%	16%	29%	36%	43%	49%
	Mean: 18.6% Std.: 14%					Mean: 34.6% Std.: 11.5%				
Guaranteed soft	0%	0%	2.3%	4%	10%	14%	24%	30%	37%	44%
	Mean: 3.26% Std.: 3.7%					Mean: 29.8% Std.: 10.36%				

4.4.6.4 Impact of Resource Scheduling on Network Utilisation: Table 4.11 illustrates the network utilisation by the guaranteed soft and best-effort requests belonging to mMTC (or p_1) and eMBB (or p_2), respectively. The achieved resource utilisation is evaluated using intra-slice, inter-slice, and cooperative slice elasticity-based resource allocation. In the case of p_1 requests, resource utilisation is low at the beginning, due to the lower demand of the mMTC traffic, as shown in Table 4.11. However, with an increase in demand, utilisation increases. Greater utilisation can be observed on inter-slice admission compared to intra-slice and cooperative schemes. For example, at a load of 250, the achieved utilisation gain of inter-slice over cooperative and intra-slice schemes is 28% and 60%, respectively. As in intra-slice admission, the requests are bound to the specific slices only, which leads to low utilisation of the slices. In cooperative resource allocation, due to the scalability factor, utilisation is higher than that of intra-slice but lower than that of inter-slice.

A similar trend can be seen in the case of p_2. However, in the case of p_2, due to higher demand, resource utilisation is significantly more with a gain of almost 8% in the intra-slice scheme compared to p_1, as shown in Table 4.11. With an increase in the demand of both p_1 and p_2 requests, utilisation also increases. However, at a greater load, overall utilisation in the case of p_1 load is more than that of p_2, due to a lower rejection ratio. To summarise, the proposed RL-based admission and resource allocation approach not only reduces Intra-slice congestion but also improves network utilisation. Mean resource utilisation in the case of eMBB traffic load is more compared to mMTC, as shown in the table below.

TABLE 4.11: Computation of resource utilisation through intra-slice, inter-slice, and cooperative slice elasticity-based resource allocation to guaranteed soft and best-effort traffic load [50, 250]

Resource	p_1 traffic load					p_2 traffic load				
allocation	50	100	150	200	250	50	100	150	200	250
Inter-slice	35%	63%	75%	93%	97%	53%	65%	74%	86%	90%
	Mean: 73% Std.: 22.5%					Mean: 74% Std.: 13.57%				
Cooperative	35%	50%	58%	66%	70%	48%	54%	59%	62%	69%
	Mean: 56% Std.: 12.4%					Mean: 58.4% Std.: 7%				
Intra-slice	26%	30%	33%	37%	38%	28%	32%	35%	40%	47%
	Mean: 32.8% Std.: 4.4%					Mean: 36.34% Std.: 6.7%				

4.5 Summary

In this chapter, three optimised admission control models have been presented for future wireless networks. In Section 4.2, a novel edge architecture for future core networks is presented, whereby, a clustering-based signalling optimisation and admission control model has been presented to derive benefits from the proposed architecture. The proposed model is a three-stage approach: Demand analysis and categorisation, Demand processing system, and Admission control and resource allocation. Moreover, two popular unsupervised learning algorithms, k-mean and ranking-based clustering, have been employed in this model to reduce communication overheads on the edge by reducing signalling redundancy, providing low latency and efficient resource utilisation. The proposed clustering mechanism reduces the complexity from $\mathcal{O}(U)$ to $\mathcal{O}(R)$ for service signalling and $\mathcal{O}(N)$ for resource signalling. This represents a significant saving in the uplink control plane signalling and link capacity compared to the results found in the existing literature.

In Section 4.3, a dynamic traffic forecasting and admission control (FAC) model has been presented for a federated O-RAN in this thesis work. FAC consists of three stages: Demand and capacity analyser, Network selection and configuration, and QoS/QoE and traffic flow management. The role of FAC is to predict future traffic demand to select an optimal network from multiple heterogeneous service networks and for efficient resource management to assure better tenant-acquired QoE and resource utilisation. In this model, a fuzzy-logic-based network selection scheme has been introduced with a multi-variate admission priority feature for optimal admission control, and service as well as resource allocation to tenants. A QoS/QoE-based service monitoring scheme is also presented to update the demand estimates with the support of a forecasting modifier. The provided service monitoring feature helps resource allocation to tenants, approximately closer to the actual demand of the tenants, to improve tenant-acquired QoE and overall network performance. FAC outperforms existing legacy approaches in terms of efficient network utilisation, enhanced tenant QoE, and fairness of resource allocation, as well as better user satisfaction in the provisioning of various heterogeneous services in O-RAN networks.

In Section 4.4, a dynamic slice congestion and admission control model is presented to minimise the slice rejection ratio that occurs due to bottlenecks and intra-slice congestion. This model consists of a slice demand analysis and classification system, a demand clustering and queuing system, and an admission and resource management controller. Two popular unsupervised learning algorithms, known as Ranking and K-mean clustering algorithms, along with optimisation and transfer learning, have been employed for slice request queuing. A unified cost estimation function is also derived for slice selection to ensure fairness among slice requests. In view of instantaneous network circumstances and load, an RL-based admission control policy is also established for taking appropriate action on guaranteed soft and best-effort slice requests admissions. Intra-slice, as well as inter-slice resource allocation, along with the adaptability of slice elasticity, are also proposed for maximising slice acceptance ratio and resource utilisation. The proposed SCAC model and algorithms are analysed by obtaining the rejection ratio, bottleneck congestion, and utilisation at various traffic loads of mMTC and eMBB. Accordingly, a summary of the analysis of the presented E-RMAC, FAC, and SCAC models is also presented in Table 4.12.

TABLE 4.12: Summary of E-RMAC, FAC, and SCAC models analysis

Analysis Measures	**E-RMAC Model**	**FAC Model**	**SCAC Model**
Admission objective	Objective of the E-RMAC model is link capacity and latency optimisation via core control signalling redundancy minimisation and efficient admission control.	Objective of the FAC model is to enhance overall network throughput through efficient resource allocation and utilisation within a multi-operator environment or O-RAN.	Objective of the SCAC model is the bottleneck (as well as interslice) congestion and admission control in 5G and beyond networks to support mMTC and eMBB traffic demand.
Slice elasticity	Slice elasticity not supportive	Support tenant-aware slice reconfiguration from optimised service network list	Slice reconfigurable resource bounds for inter-slice and cooperative slice elasticity based admission and resource allocation
Tenancy	Multi-tenant support	Multi-tenant support	Multi-tenant support
Slicing domain	E2E slice management and orchestration support in Edge network	E2E slice management and orchestration support in O-RAN network	E2E slice management and orchestration support in 5G and beyond network
Admission strategy	Multi-objective optimisation	Multi-objective optimisation	Multi-objective optimisation
Optimisation Algorithm	Unsupervised learning and NSGA-II	SMC based particle filtering and NSGA-II	Unsupervised learning, NSGA-II, reinforcement and transfer learning
Admission domain	Intra and inter-slice admission and resource allocation	Intra and inter-slice admission and resource allocation	Intra/inter-slice and cooperative slice elasticity based admission and resource allocation
Admission efficiency	The average admission efficiency of the E-RMAC model over a fully loaded edge network is 98% due to lower signalling redundancy in the access and core of the edge network.	The average admission efficiency of the FAC model over a fully loaded O-RAN network is 94.12% due to the availability of various heterogeneous RATs and optimised admission control.	The average admission efficiency of the SCAC model over a fully loaded 5G network is 95% due to the proposed clustering and optimised admission control.

Chapter 5

Conclusion and Future Work

5.1 Conclusion

Dynamically optimised admission control is considered a promising solution for efficient resources management in future wireless networks. This approach exploits the nature of user demand and network resource statistics for more appropriate network selection and management of heterogeneous traffic flow in future networks (i.e. 5G and beyond). This occurs in such a way that it improves overall user-acquired QoE and network QoS. By applying this approach, a number of admission control models were considered in this research. Based on user preferences from demand analysis, available network opportunities were identified for admission control and resource allocation in a dense network. However, dynamically selecting an optimal network with regard to numerous heterogeneous user-demand characteristics at an agreed SLA becomes a challenging task. This thesis has contributed to the research by proposing dynamically reconfigurable admission control models using optimisation and advanced machine learning approaches (i.e. unsupervised learning, reinforcement learning and transfer learning) for network QoS and user QoE enhancement.

In Chapter 3 of this thesis, a novel dynamical slice allocation and admission control (DSAAC) model has been presented for 5G and beyond networks, in which a unified cost estimation function is proposed for dynamic optimising slice allocation and admission control. This model considers varied user demands, as well as multiple real-time network resource characteristics. These characteristics include user and slice required bandwidth, data rate, priority, latency sensitivity, and cost revenue. Moreover, to maximise network utility, adjustable minimum and maximum slice resource bounds have also been derived. In the case of user blocking from the primary slice due to congestion or resource scarcity, a set of admission control and resource allocation algorithms has been derived, such as inter-slice admission control and resource allocation algorithm, and adaptability of slice elasticity for resource allocation algorithm. These algorithms ensure efficient utilisation of network slices on optimal admission control. For the access network control signaling redundancy minimisation, a novel optimised signaling and admission control (SAC) model is also presented for 5G and beyond networks in Chapter 3. In this model, a three-stage approach including pre-clustering analysis, usage-specific clustering, and a signalling optimisation and admission mechanism has been introduced. This model deals with the usage and user-device-specific heterogeneity in a single-layer approach instead of a two-layer approach. By reducing redundancy in demand, the unsupervised ML-based clustering approach implemented reduces the additional burden on the network in terms of unnecessary resource utilisation and computational time. Thus, signalling redundancy reduction from the cluster decreases the massive amount of unnecessary control messages flowing into the network. Eventually, the proposed models are evaluated in terms of GoS, network utility, mean delay, throughput, uplink signalling load, and admission gain. The results obtained are also compared with those of relevant strategies from the literature, suggesting that the proposed optimised admission control models outperform their existing counterparts. From the comparative results, it is observed that a flexible but efficient decision metric can be obtained through the accumulation of user demand and network resource characteristics. The proposed models provide explicit definitions of the requirements of network slice characteristics, which leads to better admission control and resource utilisation to ensure enhanced network QoS and user-acquired QoE.

A novel learning-based optimised edge redundancy minimisation and admission control (E-RMAC) model for 5G/6G edge networks has been presented in Chapter 4. The proposed model is a three-stage approach including demand analysis and categorisation, demand processing system, and admission control and resource allocation. Moreover, two popular unsupervised learning algorithms, K-mean and Ranking-based clustering, were employed in this model to reduce communication overheads at the edge by reducing signalling redundancy, and providing low latency and efficient resource utilisation. The proposed clustering mechanism reduces the complexity from $\mathcal{O}(U)$ to $\mathcal{O}(R)$ for service signalling, and $\mathcal{O}(N)$ for resource signalling. This represents a significant saving in the uplink control plane signalling and link capacity compared to results found in the existing literature. Moreover, a set of optimisation algorithms are also established in this model for efficient resource allocation and admission control, whereby K-mean is employed in combination with NSGA-II. In Chapter 4, a dynamic traffic forecasting and admission control (FAC) model for a federated O-RAN (also called FORAN in this thesis) is presented. FAC also consists of three-stages: the demand and capacity analyser, network selection and configuration, and QoS/QoE and traffic flow management. The role of FAC is to predict future traffic demand for optimal network selection from amongst multiple service networks and resource management to assure better tenant-acquired QoE and network utilisation. A fuzzy-logic-based network selection scheme with a multi-variate admission priority feature is introduced in this model for optimal admission control and service allocation to tenants. Moreover, a QoS/QoE-based service monitoring approach is also presented to update demand via a forecasting modifier. This is to allocate resources approximately closer to the actual demand of tenants to improve tenant-acquired QoE and overall network QoS. The proposed models outperform existing legacy approaches in terms of control signalling redundancy reduction, admission gain, more efficient network utilisation, enhanced tenant-acquired QoE and resource allocation fairness, as well as better user satisfaction levels in the provisioning of various heterogeneous services in edge and O-RAN networks, respectively.

A novel slice congestion and admission control (SCAC) model is presented in Chapter 4 to minimise the slice rejection ratio that occurred due to a bottleneck, as well as intra-slice congestion. This model consists of a slice demand analysis and classification system, a demand clustering and queuing system, and an admission and resource management controller. The demand analyser analyses the requested slice QoE from the given service type, which isolates the guaranteed soft-QoE slice requests from best-effort QoE requests through a classification mask. The demand processing system clusters the requests using two popular unsupervised learning algorithms known as Ranking and K-mean clustering algorithms, along with multi-objective optimisation and TL techniques for slice request queuing that reduces bottleneck congestion. Next, a unified cost estimation function for slice selection was derived to ensure fairness among slice requests. In view of instantaneous network circumstances and load, an RL-based admission control policy was established to take appropriate action on the arrival of guaranteed soft and best-effort slice requests for admission into the network. Intra-slice and inter-slice resource allocation, along with the adaptability of slice elasticity, are proposed for maximising slice acceptance ratio and resource utilisation. The proposed model and algorithms are analysed by obtaining the ratio of bottleneck congestion that occurred at varying loads of mMTC and eMBB after applying clustering along with the multi-objective optimisation and transfer learning approaches. The requests' rejection ratio and resource utilisation are also acquired for analysis of the proposed intra-slice, inter-slice, and cooperative slice-elasticity-based admission and resource allocation. The results obtained suggest that the proposed optimised clustering approach has a significantly greater impact in terms of congestion control on the massive volume of network load than the conventional approaches. Moreover, the proposed RL-based admission policy ensures fairness among requests on admission and a lower rejection ratio due to the proposed resource allocation schemes. It can also be observed that the difference in outcomes achieved from implementing optimisation and transfer learning for clustering is not significant. Precise solutions to a particular problem would be acquired from the optimisation; however, they are computationally complex and costly in terms of resources. Instead, knowledge sharing from an already optimised problem to another relevant problem can speed up the process and preserve resources. Thus, among ML techniques, transfer learning is considered an effective approach for slice congestion and admission control in future networks.

5.2 Research Objectives Achievement

The achievements regarding the research objectives mentioned in Chapter 1 are as follows:

1. How to enhance Grade-of-Service (GoS) in future networks with limited or without scaling up network capacity to support a massive amount of heterogeneous traffic demand?

   - Efficient slice allocation and admission control has been achieved within 5G and beyond networks via dynamically reconfigurable slice resource bounds, inter-slice and intra-slice resource allocation, and adaptation of the slice elasticity approach, as presented in Section 3.2 and disseminated internationally via the publication ***C01*** (i.e. 2019 IEEE 89th Vehicular Technology Conference).

2. Can signalling redundancy minimisation in access and core networks enhance admission and resource utilisation without degrading network QoS and user-desired QoE demand within a dense environment?

   - For the first time access and core network control signalling redundancy minimisation is achieved via two unsupervised learning approaches, along with optimisation for efficient admission control and resources utilisation within future networks, as presented in Sections 3.3 and 4.2, and disseminated internationally via the publications ***C02*** (2020 IEEE International Conference on Communications) and ***C03*** (2021 IEEE 5G World Forum).

3. Can the coexistence of the various heterogeneous cellular technologies (2G, 3G,4G, and 5G) and their integration help enhance overall network throughput via efficient resource allocation fairness among users and resources utilisation within a multi-operator environment?

   - In the O-RAN environment, resource allocation fairness among users and efficient network resources utilisation has been achieved via demand forecasting and fuzzy-logic-based optimal admission control and resource allocation, as presented in Section 4.3 and disseminated internationally via the publication ***J01*** (2020 Springer Journal).

4. How can the techniques of ML and optimisation help in bottleneck and inter-slice congestion control along with efficient admission control in 5G and beyond networks to support mMTC and eMBB traffic demand?

   - Optimisation and ML approaches have been implemented for the first time for bottlenecks, as well as intra-slice congestion control. Moreover, reinforcement learning-based admission control policy, intra/inter-slice, and cooperative slice elasticity based resource allocation approaches have been proposed to enhance admission gain and achieve efficient network utilisation, as presented in Section 4.4.

5.3 Research Limitations

The proposed optimised admission control models have novel contributions to the research. However, there are some limitations associated with the presented research. For example, the DSAAC model is not supported in a multi-slice connectivity environment. By the network slicing concept given in (3GPP, 2018c; Alliance, 2016), a user can be connected to multiple slices simultaneously. However, the user was supposed to be connected to one slice only for the service provisioning of a particular application in the DSAAC model. This is because when the user will be connected to multiple slices for different applications, signalling redundancy would occur that would overwhelm the network and create congestion. Moreover, due to mobility admission control on handover would be complicated. If high mobility users are to be considered in DSAAC, slices can lose their ability of uniform utilisation. Similarly, in the SAC model redundancy minimisation in data transmission was not been considered. For data redundancy minimisation in the access network, advanced multiplexing techniques are required to be investigated in depth, which was out of the scope of this research.

The E-RMAC model is not supported in a multi-edge connectivity environment. As, edge-to-edge configuration can be different, which can induce complexity in the network operations. Users profiles can be stored on various edges due to mobility. Some of the edges would not have much capacity to store users profiles. Moreover, each edge has to update user's profile consistently with changes in user application-specific characteristics. In the FAC model, due to changes in network circumstances or tenant preference, a tenant might be redirected to the next available network slice from the optimised network operator list to ensure the provisioning of agreed QoE. In this case, the tenant's slice reconfiguration induces latency in the communication. As the fuzzy logic-based technique has been applied in the FAC model for optimal network selection on the forecasting demand. To apply optimisation on actual demand is costly in terms of computation time and memory. Therefore, for the tenant's slice reconfiguration more optimised policies are required. The FAC model is evaluated in the simulated environment in MATLAB. However, for a more robust evaluation and in-depth analysis real experimental environment is required. Similarly, the optimised clusters and their associated number of requests can be acquired by applying optimisation techniques in the SCAC model but optimisation itself takes time to converge to an optimal solution, which is not acceptable in communication. Moreover, the acquired knowledge in the SCAC model is from a simulated environment, which gives performance on the target task less than the optimisation task. Therefore, a real environment is mandatory to acquire a more robust and well-trained model from optimisation and knowledge transfer.

5.4 Future Work

This research contributes to the theme of dynamic admission control by using ML and optimisation algorithms. In this section, I am going to discuss possible directions of research work into future wireless networks. There are several possible directions for dynamically reconfigurable slice allocation and admission control, as suggested in Chapter 3. The most obvious is user connectivity to multiple slices simultaneously, which can generate several issues. For example, simultaneous access to multiple access and core NFs creates a burden on the network. Moreover, mobility issues such as handover would be complex, as the user is connected to multiple slices. Little mobility can be confused with the handover, which should be managed dynamically during slice setup and maintenance. Hence, there is a need for a more-optimised design for decision-making that can efficiently manage user multi-slice connectivity and mobility, and create less burden via reducing redundant connectivity to NFs. Redundancy in requests of numerous user devices for connectivity can be eliminated with the proposed SAC model. However, redundancy in non-critical data transmission also needs to be reduced to achieve efficient resource utilisation. Multiple devices can send similar data, such as in IoT and MTC. Redundancy in their data needs to be filtered and compressed via optimisation, along with help of learning approaches. Redundancy reduction in non-critical data transmission would prohibit the entry of raw or low-quality data into the network to improve overall network QoS and user-acquired QoE.

In Chapter 4, the design of a novel edge architecture to support the visions of advanced technology is still an open topic. This emerging topic requires further research that would cover other issues. For example, multiple-edge self-configuration, management and synchronisation, and the need for network slicing support with efficient resource sharing strategies among slices of various edges (inter-slice) or within slices (intra-slice) of a particular edge. Multiple configured edges could have a single, shared edge core to cope with security issues and signalling redundancy. Another emerging topic is support for O-RAN in future networks and its integration with existing technologies and physical infrastructure. O-RAN is a few years old only, so it needs a significant amount of research in dynamic admission control and resource management via optimised network policies and standards. Moreover, learning and model training using optimisation, AI, and ML approaches can help O-RAN dynamically support various heterogeneous future use cases with reduced latency and higher reliability. As, 5G applications have an extremely diverse set of requirements such as high-definition real-time streaming, automated vehicular systems, and remote operations using robotic hands. These applications require a higher data rate with low latency and higher accuracy in transmission. Any significant disturbance in the network can have a catastrophic impact on critical applications such as automated vehicle systems. Therefore, the networks must be monitored regularly to address any performance decline before it leads to the failure of any of these applications due to congestion in the network. However, it takes a lot of time and money to regularly test the network strength of an entire city for optimising admission control and resource management that can minimise the congestion in the network. As discussed in Chapter 4, TL has proved efficient in various domains of image processing and computer vision. Therefore, a key topic is currently the role of TL in cellular networks. Now, how TL can address admission and congestion control issues of a 5G and beyond network needs a significant amount of research. Inefficient admission control can occur in a network due to data scarcity. Moreover, the huge amount of raw and multidimensional data creates congestion and induces additional latency in a network. Training and optimisation are costly for operators in terms of resources and time. Hence, the availability of well-trained and open-source TL-based models is essential for wireless communication. In my future work, the plan is to enhance the models presented in Chapter 4 and develop a well-trained model in a real multi-edge computing and O-RAN environment for admission control and resource management. For this project, data would be collected from various locations of the Birmingham City Council United Kingdom via an android application. The collected data would be analysed and tested on the enhanced model to provide an optimal solution for admission control and resource management of an entire city network without human resources.

Bibliography

3GPP (2012). “GPP system to wireless local area network (WLAN) interworking; system description”. In: *3GPP TS 23.234, V11.0.0, 3*.

— (2013a). “Architecture enhancements for non-3GPP accesses”. In: *3GPP TS 23.402 Release 12*.

— (2013b). “GPRS enhancements for E-UTRAN access”. In: *3GPP TS 23.401 Release 12*.

— (2014). “Architecture enhancements for dedicated core networks; Stage 2”. In: *3GPP TS 23.707 Release 13*.

— (2016). “Enhancements of Dedicated Core Networks selection mechanism”. In: *3GPP TS 23.711 Release 14*.

— (2017). “Technical Specification Group Radio Access Network; Study on New Radio Access Technology;Radio Interface Protocol Aspects”. In: *3GPP TR 38.804 V14.0.0*.

— (2018a). “Radio Frequency (RF) requirements for Multicarrier and Multiple Radio Access Technology (Multi-RAT) Base Station (BS)”. In: *3GPP TR 37.900 V15.0.0*.

— (2018b). “System Architecture for the 5G System”. In: *3GPP TS 23.501 Release 15*.

— (2020). “Technical Specification Group Core Network and Terminals; 5G System; Unified Data Management Services”. In: *3GPP TS 29.503 Release 17*.

3GPP, Third Generation Partnership project (2018c). “Telecommunication management;Study on management and orchestration of network slicing for next generation network”. In: *3GPP TS 28.801 Release 15*.

Abbas, Nasir et al. (2017). “Mobile edge computing: A survey”. In: *IEEE Internet of Things Journal* 5.1, pp. 450–465.

Adou, Yves, Ekaterina Markova, and Irina Gudkova (2018). “Performance measures analysis of admission control scheme model for wireless network, described by a queuing system operating in random environment”. In: *2018 10th International*

Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT). IEEE, pp. 1–5.

Agiwal, Mamta et al. (2021). "A Survey on 4G-5G Dual Connectivity: Road to 5G Implementation". In: *IEEE Access* 9, pp. 16193–16210.

Aguwamba, CB (2020). "User Plane Optimization in a 5G Radio Access Network". In.

Ahmad, Wan Siti Halimatul Munirah Wan et al. (2020). "5G technology: Towards dynamic spectrum sharing using cognitive radio networks". In: *IEEE Access* 8, pp. 14460–14488.

Al-Fuqaha, Ala et al. (2015). "Internet of things: A survey on enabling technologies, protocols, and applications". In: *IEEE communications surveys & tutorials* 17.4, pp. 2347–2376.

Allen, David M (1971). "Mean square error of prediction as a criterion for selecting variables". In: *Technometrics* 13.3, pp. 469–475.

Alliance, NGMN (2015). "5G white paper". In: *Next generation mobile networks, white paper* 1.

— (2016). "Description of network slicing concept". In: *NGMN 5G P* 1.

Alliance, O.R (2018). "O-RAN: Towards an Open and Smart RAN". In: *White paper*.

Alsenwi, Madyan et al. (2021). "Intelligent resource slicing for eMBB and URLLC coexistence in 5G and beyond: A deep reinforcement learning based approach". In: *IEEE Transactions on Wireless Communications*.

Altman, Eitan (2000). "Applications of markov decision processes in communication networks: A survey". PhD thesis. INRIA.

Amjad, Muhammad, Mubashir Husain Rehmani, and Shiwen Mao (2018). "Wireless multimedia cognitive radio networks: A comprehensive survey". In: *IEEE Communications Surveys & Tutorials* 20.2, pp. 1056–1103.

Anand, S and Ananthnarayanan Chockalingam (2003). "Performance analysis of voice/data cellular CDMA with SIR-based admission control". In: *IEEE Journal on selected areas in communications* 21.10, pp. 1674–1684.

Andrews, Jeffrey G et al. (2014). "What will 5G be?" In: *IEEE Journal on selected areas in communications* 32.6, pp. 1065–1082.

Antevski, Kiril et al. (2020). “A Q-learning strategy for federation of 5G services”. In: *ICC 2020-2020 IEEE International Conference on Communications (ICC)*. IEEE, pp. 1–6.

Arjoune, Youness and Naima Kaabouch (2019). “A comprehensive survey on spectrum sensing in cognitive radio networks: Recent advances, new challenges, and future research directions”. In: *Sensors* 19.1, p. 126.

Aryafar, Ehsan et al. (2013). “RAT selection games in HetNets”. In: *2013 Proceedings IEEE INFOCOM*. IEEE, pp. 998–1006.

Banchs, Albert et al. (2015). “A Novel Radio Multiservice adaptive network Architecture for 5G networks”. In: *Vehicular Technology Conference (VTC Spring), 2015 IEEE 81st*. IEEE, pp. 1–5.

Barakabitze, Alcardo Alex et al. (2020). “5G network slicing using SDN and NFV: A survey of taxonomy, architectures and future challenges”. In: *Computer Networks* 167, p. 106984.

Bega, Dario et al. (2017). “Optimising 5G infrastructure markets: The business of network slicing”. In: *IEEE INFOCOM 2017-IEEE Conference on Computer Communications*. IEEE, pp. 1–9.

Bega, Dario et al. (2019). “DeepCog: Cognitive network management in sliced 5G networks with deep learning”. In: *IEEE INFOCOM 2019-IEEE Conference on Computer Communications*. IEEE, pp. 280–288.

Behrad, Shanay et al. (2020). “A new scalable authentication and access control mechanism for 5G-based IoT”. In: *Future Generation Computer Systems* 108, pp. 46–61.

Bhandari, Sabin, Shree Krishna Sharma, and Xianbin Wang (2018). “Device Grouping for Fast and Efficient Channel Access in IEEE 802.11 ah based IoT Networks”. In: *2018 IEEE International Conference on Communications Workshops (ICC Workshops)*. IEEE, pp. 1–6.

Bhatia, Nishita, Piyush Chauhan, and Hitesh Yadav (2021). “Applications of Hybrid Particle Swarm Optimization Algorithm: A Survey”. In: *Proceedings of the Second International Conference on Information Management and Machine Intelligence*. Springer, pp. 291–297.

Bouali, Faouzi, Klaus Moessner, and Michael Fitch (2016). “A context-aware user-driven framework for network selection in 5G multi-RAT environments”. In: *2016 IEEE 84th Vehicular Technology Conference (VTC-Fall)*. IEEE, pp. 1–7.

Brichet, Francois and Alain Simonian (1998). “Conservative Gaussian models applied to measurement-based admission control”. In: *Quality of Service, 1998.(IWQoS 98) 1998 Sixth International Workshop on*. IEEE, pp. 68–71.

Brown, Gabriel (2012). “The evolution of the signaling challenge in 3G & 4G networks”. In: *White Paper, jun*.

Busari, Sherif Adeshina et al. (2017). “Millimeter-wave massive MIMO communication for future wireless systems: A survey”. In: *IEEE Communications Surveys & Tutorials* 20.2, pp. 836–869.

Buyakar, Tulja Vamshi Kiran et al. (2020). “Resource allocation with admission control for GBR and delay QoS in 5G network slices”. In: *2020 International Conference on COMmunication Systems & NETworkS (COMSNETS)*. IEEE, pp. 213–220.

Caballero, Pablo et al. (2018). “Network slicing for guaranteed rate services: Admission control and resource allocation games”. In: *IEEE Transactions on Wireless Communications* 17.10, pp. 6419–6432.

Camps Mur, Daniel et al. (2020). “5G-CLARITY: Integrating 5GNR, WiFi and LiFi in private networks with slicing support”. In: *2020 European Conference on Networks and Communications (EuCNC): 15-18 June 2020, Dubrovnik, Croatia: poster 1*. European Conference on Networks and Communications (EuCNC), pp. 1–2.

Cao, Jin et al. (2019). “Anti-quantum fast authentication and data transmission scheme for massive devices in 5G NB-IoT system”. In: *IEEE Internet of Things Journal* 6.6, pp. 9794–9805.

Challa, Rajesh et al. (2019). “Network slice admission model: Tradeoff between monetization and rejections”. In: *IEEE Systems Journal* 14.1, pp. 657–660.

Challita, Ursula, Henrik Ryden, and Hugo Tullberg (2020). “When machine learning meets wireless cellular networks: Deployment, challenges, and applications”. In: *IEEE Communications Magazine* 58.6, pp. 12–18.

Chen, He et al. (2018). “Ultra-reliable low latency cellular networks: Use cases, challenges and approaches”. In: *IEEE Communications Magazine* 56.12, pp. 119–125.

Chen, Mingzhe et al. (2019). "Artificial neural networks-based machine learning for wireless networks: A tutorial". In: *IEEE Communications Surveys & Tutorials* 21.4, pp. 3039–3071.

Choi, Young-il and Noik Park (2017). "Slice architecture for 5G core network". In: *Ubiquitous and Future Networks (ICUFN), 2017 Ninth International Conference on*. IEEE, pp. 571–575.

Chowdhury, Mostafa Zaman, Yeong Min Jang, and Zygmunt J Haas (2013). "Call admission control based on adaptive bandwidth allocation for wireless networks". In: *Journal of Communications and Networks* 15.1, pp. 15–24.

Cook, Diane, Kyle D Feuz, and Narayanan C Krishnan (2013). "Transfer learning for activity recognition: A survey". In: *Knowledge and information systems* 36.3, pp. 537–556.

Da Xu, Li, Wu He, and Shancang Li (2014). "Internet of things in industries: A survey". In: *IEEE Transactions on industrial informatics* 10.4, pp. 2233–2243.

Dandachi, Ghina et al. (2019). "An artificial intelligence framework for slice deployment and orchestration in 5G networks". In: *IEEE Transactions on Cognitive Communications and Networking* 6.2, pp. 858–871.

Deb, K. et al. (2002). "A fast and elitist multiobjective genetic algorithm: NSGA-II". In: *IEEE Transactions on Evolutionary Computation* 6.2, pp. 182–197.

Deb, Kalyanmoy et al. (2002). "A fast and elitist multiobjective genetic algorithm: NSGA-II". In: *IEEE transactions on evolutionary computation* 6.2, pp. 182–197.

Dong, Rui et al. (2020). "Deep learning for radio resource allocation with diverse quality-of-service requirements in 5G". In: *IEEE Transactions on Wireless Communications* 20.4, pp. 2309–2324.

Dorigo, Marco and Gianni Di Caro (1999). "Ant colony optimization: a new metaheuristic". In: *Proceedings of the 1999 congress on evolutionary computation-CEC99 (Cat. No. 99TH8406)*. Vol. 2. IEEE, pp. 1470–1477.

Du, Jianbo et al. (2016). "Enhanced PSO based energy-efficient resource allocation and CQI based MCS selection in LTE-A heterogeneous system". In: *China Communications* 13.11, pp. 197–204.

Dudek, Grzegorz (2016). "Neural networks for pattern-based short-term load forecasting: A comparative study". In: *Neurocomputing* 205, pp. 64–74.

Dudek, Grzegorz (2019). "Multilayer perceptron for short-term load forecasting: from global to local approach". In: *Neural Computing and Applications*, pp. 1–13.

El-Saleh, Ayman A et al. (2021). "Multi-objective optimization of joint power and admission control in cognitive radio networks using enhanced swarm intelligence". In: *Electronics* 10.2, p. 189.

Elayoubi, Salah Eddine et al. (2019). "5G RAN slicing for verticals: Enablers and challenges". In: *IEEE Communications Magazine* 57.1, pp. 28–34.

Emara, Mustafa, Miltiades C Filippou, and Dario Sabella (2018). "MEC-assisted end-to-end latency evaluations for C-V2X communications". In: *2018 European Conference on Networks and Communications (EuCNC)*. IEEE, pp. 1–9.

ETSI (2020). "5G; NR; Medium Access Control (MAC) protocol specification". In: *3GPP TS 38.321 version 16.1.0 Release 16*.

ETSI, ESIISG (2017). "Improved operator experience through experiential networked intelligence (ENI)". In: *White Paper, Oct*.

Ewert, Jörg, Lennart Norell, and Soner Yamen (2012). "Diameter Signaling Controller in next-generation signaling networks". In: *Ericsson Review* 284, pp. 23–31761.

Faris, Hossam et al. (2019). "An intelligent system for spam detection and identification of the most relevant features based on evolutionary random weight networks". In: *Information Fusion* 48, pp. 67–83.

Farooq, Umar and Ghulam Mohammad Rather (2019). "Millimeter wave (MMW) communications for fifth generation (5G) mobile networks". In: *Progress in Advanced Computing and Intelligent Engineering*. Springer, pp. 97–106.

Fischer, Hans (2010). *A history of the central limit theorem: from classical to modern probability theory*. Springer Science & Business Media.

Forum, xRAN (2016). "The mobile access network, beyond connectivity". In: *xRAN Forum*.

Foukas, Xenofon et al. (2017). "Network slicing in 5G: Survey and challenges". In: *IEEE Communications Magazine* 55.5, pp. 94–100.

Fourati, Hasna, Rihab Maaloul, and Lamia Chaari (2021). "A survey of 5G network systems: challenges and machine learning approaches". In: *International Journal of Machine Learning and Cybernetics* 12.2, pp. 385–431.

Gavrilovska, Liljana, Valentin Rakovic, and Daniel Denkovski (2018). "Aspects of resource scaling in 5G-MEC: technologies and opportunities". In: *2018 IEEE Globecom Workshops (GC Wkshps)*. IEEE, pp. 1–6.

— (2020). "From Cloud RAN to Open RAN". In: *Wireless Personal Communications*, pp. 1–17.

Ge, Fei and Liansheng Tan (2014). "Network utility maximization in two-way flow scenario". In: *ACM SIGCOMM Computer Communication Review* 44.2, pp. 13–19.

Gharghan, Sadik K et al. (2015). "Accurate wireless sensor localization technique based on hybrid PSO-ANN algorithm for indoor and outdoor track cycling". In: *IEEE Sensors Journal* 16.2, pp. 529–541.

Ghosh, Sumit et al. (1998). "A survey of recent advances in fuzzy logic in telecommunications networks and new challenges". In: *IEEE Transactions on Fuzzy Systems* 6.3, pp. 443–447.

Gohil, Asvin, Hardik Modi, and Shobhit K Patel (2013). "5G technology of mobile communication: A survey". In: *2013 international conference on intelligent systems and signal processing (ISSP)*. IEEE, pp. 288–292.

Gong, Wenrong et al. (2019). "PSO-based resource allocation in software-defined heterogeneous cellular networks". In: *KSII Transactions on Internet and Information Systems (TIIS)* 13.5, pp. 2243–2257.

Goudarzi, Shidrokh et al. (2019). "A hybrid intelligent model for network selection in the industrial Internet of Things". In: *Applied Soft Computing* 74, pp. 529–546.

Gozalvez, Javier (2017). "5G worldwide developments [mobile radio]". In: *IEEE Vehicular Technology Magazine* 12.1, pp. 4–11.

Gözüpek, Didem and Fatih Alagöz (2011). "Genetic algorithm-based scheduling in cognitive radio networks under interference temperature constraints". In: *International Journal of Communication Systems* 24.2, pp. 239–257.

GSMA (2019). *5G Implementation Guidelines*. URL: https://www.gsma.com/futurenetworks/wp-content/uploads/2019/03/5G-Implementation-Guideline-v2.0-July-2019.pdf (visited on 11/11/2019).

Gu, Jaheon et al. (2015). "Heuristic algorithm for proportional fair scheduling in D2D-cellular systems". In: *IEEE Transactions on Wireless Communications* 15.1, pp. 769–780.

Gündüz, Deniz et al. (2019). "Machine learning in the air". In: *IEEE Journal on Selected Areas in Communications* 37.10, pp. 2184–2199.

Gupta, Akhil and Rakesh Kumar Jha (2015). "A survey of 5G network: Architecture and emerging technologies". In: *IEEE access* 3, pp. 1206–1232.

Gutierrez-Estevez, David M et al. (2018). "The path towards resource elasticity for 5G network architecture". In: *2018 IEEE wireless communications and networking conference workshops (WCNCW)*. IEEE, pp. 214–219.

Gutierrez-Estevez, David M et al. (2019). "Artificial intelligence for elastic management and orchestration of 5G networks". In: *IEEE Wireless Communications* 26.5, pp. 134–141.

Habibi, Mohammad Asif et al. (2019). "A comprehensive survey of RAN architectures toward 5G mobile communication system". In: *IEEE Access* 7, pp. 70371–70421.

Haile, Habtegebreil et al. (2021). "End-to-end congestion control approaches for high throughput and low delay in 4G/5G cellular networks". In: *Computer Networks* 186, p. 107692.

Hamdi, Monia and Mourad Zaied (2019). "Resource allocation based on hybrid genetic algorithm and particle swarm optimization for D2D multicast communications". In: *Applied Soft Computing* 83, p. 105605.

Han, Bin, Di Feng, and Hans D Schotten (2018). "A Markov model of slice admission control". In: *IEEE Networking Letters* 1.1, pp. 2–5.

Han, Bin et al. (2018a). "Admission and congestion control for 5G network slicing". In: *2018 IEEE Conference on Standards for Communications and Networking (CSCN)*. IEEE, pp. 1–6.

Han, Bin et al. (2019). "A utility-driven multi-queue admission control solution for network slicing". In: *IEEE INFOCOM 2019-IEEE Conference on Computer Communications*. IEEE, pp. 55–63.

Han, Ren et al. (2018b). "An effective multi-objective optimization algorithm for spectrum allocations in the cognitive-radio-based Internet of Things". In: *IEEE Access* 6, pp. 12858–12867.

Hicham, Magri, Noreddine Abghour, and Mohammed Ouzzif (2014). "4G System: Network Architecture and Performance". In: *International Journal of Innovative Research in Advanced Engineering* 4.2, pp. 215–220.

Hippert, Henrique Steinherz, Carlos Eduardo Pedreira, and Reinaldo Castro Souza (2001). "Neural networks for short-term load forecasting: A review and evaluation". In: *IEEE Transactions on power systems* 16.1, pp. 44–55.

Hospedales, Timothy et al. (2020). "Meta-learning in neural networks: A survey". In: *arXiv preprint arXiv:2004.05439*.

Hu, Yun Chao et al. (2015). "Mobile edge computing—A key technology towards 5G". In: *ETSI white paper* 11.11, pp. 1–16.

Hung, Chi-Hsiang, Yao-Chou Hsieh, and Li-Chun Wang (2017). "Control plane latency reduction for service chaining in mobile edge computing system". In: *2017 13th International Conference on Network and Service Management (CNSM)*. IEEE, pp. 1–5.

Inaba, Takaaki et al. (2015). "A secure-aware call admission control scheme for wireless cellular networks using fuzzy logic and its performance evaluation". In: *Journal of Mobile Multimedia*, pp. 213–222.

Jain, Madhu and Ragini Mittal (2016). "Adaptive call admission control and resource allocation in multi server wireless/cellular network". In: *Journal of Industrial Engineering International* 12.1, pp. 71–80.

Jain, Raj, Arjan Durresi, and Gojko Babic (1999). "Throughput fairness index: An explanation". In: *ATM Forum contribution*. Vol. 99. 45.

Jia, Yang et al. (2018). "Bankruptcy game based resource allocation algorithm for 5G Cloud-RAN slicing". In: *2018 IEEE Wireless Communications and Networking Conference (WCNC)*. IEEE, pp. 1–6.

Jiang, Chunxiao et al. (2016a). "Machine learning paradigms for next-generation wireless networks". In: *IEEE Wireless Communications* 24.2, pp. 98–105.

Jiang, Menglan, Massimo Condoluci, and Toktam Mahmoodi (2016). "Network slicing management & prioritization in 5G mobile systems". In: *European Wireless 2016; 22th European Wireless Conference*. VDE, pp. 1–6.

Jiang, Menglan et al. (2016b). "Economics of 5G network slicing: optimal and revenue-based allocation of radio and core resources in 5G". In: *Kings collections London*.

Johnson, Oliver (2004). *Information theory and the central limit theorem*. World Scientific.

Kakalou, Ioanna et al. (2017). "Cognitive radio network and network service chaining toward 5G: Challenges and requirements". In: *IEEE communications Magazine* 55.11, pp. 145–151.

Kaloxylos, Alexandros (2018). "A Survey and an Analysis of Network Slicing in 5G Networks". In: *IEEE Communications Standards Magazine* 2.1, pp. 60–65.

Kaloxylos, Alexandros et al. (2014). "An efficient RAT selection mechanism for 5G cellular networks". In: *2014 International Wireless Communications and Mobile Computing Conference (IWCMC)*. IEEE, pp. 942–947.

Kammoun, Amal et al. (2018). "Admission control algorithm for network slicing management in SDN-NFV environment". In: *2018 6th international conference on multimedia computing and systems (ICMCS)*. IEEE, pp. 1–6.

Kato, Nei et al. (2020). "Ten challenges in advancing machine learning technologies toward 6G". In: *IEEE Wireless Communications* 27.3, pp. 96–103.

Kennedy, James and Russell Eberhart (1995). "Particle swarm optimization". In: *Proceedings of ICNN'95-international conference on neural networks*. Vol. 4. IEEE, pp. 1942–1948.

Khan, Ammara Anjum et al. (2019). "A hybrid-fuzzy logic guided genetic algorithm (H-FLGA) approach for resource optimization in 5G VANETs". In: *IEEE Transactions on Vehicular Technology* 68.7, pp. 6964–6974.

Khan, Humayun Zubair et al. (2021). "Joint admission control, cell association, power allocation and throughput maximization in decoupled 5G heterogeneous networks". In: *Telecommunication Systems* 76.1, pp. 115–128.

Khan, Latif U et al. (2020a). "Network slicing: Recent advances, taxonomy, requirements, and open research challenges". In: *IEEE Access* 8, pp. 36009–36028.

Khan, Muhammad Fahad et al. (2020b). "Survey and taxonomy of clustering algorithms in 5G". In: *Journal of Network and Computer Applications* 154, p. 102539.

Kuo, Wen-Hsing and Wanjiun Liao (2008). "Utility-based radio resource allocation for QoS traffic in wireless networks". In: *IEEE Transactions on Wireless Communications* 7.7, pp. 2714–2722.

Le, Luong-Vy et al. (2018). "SDN/NFV, machine learning, and big data driven network slicing for 5G". In: *2018 IEEE 5G World Forum (5GWF)*. IEEE, pp. 20–25.

Le, Nam Tuan et al. (2016). "Survey of promising technologies for 5G networks". In: *Mobile Information Systems* 2016.

Lee, Ying Loong et al. (2018). "Dynamic Network Slicing for Multitenant Heterogeneous Cloud Radio Access Networks". In: *IEEE Transactions on Wireless Communications* 17.4, pp. 2146–2161.

Li, Rongpeng et al. (2018). "Deep reinforcement learning for resource management in network slicing". In: *IEEE Access* 6, pp. 74429–74441.

Li, Taihui, Xiaorong Zhu, and Xu Liu (2020). "An end-to-end network slicing algorithm based on deep Q-learning for 5G network". In: *IEEE Access* 8, pp. 122229–122240.

Li, Xujie et al. (2017). "Allocation Optimization Based on Multi-population Genetic Algorithm for D2D Communications in Multi-services Scenario". In: *International Conference on Machine Learning and Intelligent Communications*. Springer, pp. 23–32.

Liang, Liang et al. (2019). "Online Auction-Based Resource Allocation for Service-Oriented Network Slicing". In: *IEEE Transactions on Vehicular Technology* 68.8, pp. 8063–8074.

Likas, Aristidis, Nikos Vlassis, and Jakob J Verbeek (2003). "The global k-means clustering algorithm". In: *Pattern recognition* 36.2, pp. 451–461.

Lim, Wei Yang Bryan et al. (2020). "Federated learning in mobile edge networks: A comprehensive survey". In: *IEEE Communications Surveys & Tutorials* 22.3, pp. 2031–2063.

Liu, Jianhui and Qi Zhang (2018). "Offloading schemes in mobile edge computing for ultra-reliable low latency communications". In: *Ieee Access* 6, pp. 12825–12837.

Luong, Nguyen Cong et al. (2019). "Applications of deep reinforcement learning in communications and networking: A survey". In: *IEEE Communications Surveys & Tutorials* 21.4, pp. 3133–3174.

Madan, Rishabh and Partha Sarathi Mangipudi (2018). "Predicting computer network traffic: a time series forecasting approach using DWT, ARIMA and RNN".

In: *2018 Eleventh International Conference on Contemporary Computing (IC3)*. IEEE, pp. 1–5.

Mahmood, Nurul Huda et al. (2019). "Six key enablers for machine type communication in 6G". In: *arXiv preprint arXiv:1903.05406*.

Maksymyuk, Taras et al. (2018). "Deep learning based massive MIMO beamforming for 5G mobile network". In: *2018 IEEE 4th International Symposium on Wireless Systems within the International Conferences on Intelligent Data Acquisition and Advanced Computing Systems (IDAACS-SWS)*. IEEE, pp. 241–244.

Mao, Hongzi et al. (2016). "Resource management with deep reinforcement learning". In: *Proceedings of the 15th ACM workshop on hot topics in networks*, pp. 50–56.

Mei, Jie, Xianbin Wang, and Kan Zheng (2020). "An intelligent self-sustained RAN slicing framework for diverse service provisioning in 5G-beyond and 6G networks". In: *Intelligent and Converged Networks* 1.3, pp. 281–294.

Memisoglu, Ebubekir et al. (2019). "Guard band reduction for 5G and beyond multiple numerologies". In: *IEEE Communications Letters* 24.3, pp. 644–647.

Miao, Dandan et al. (2016). "MSFS: multiple spatio-temporal scales traffic forecasting in mobile cellular network". In: *2016 IEEE 14th Intl Conf on Dependable, Autonomic and Secure Computing, 14th Intl Conf on Pervasive Intelligence and Computing, 2nd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress (DASC/PiCom/DataCom/CyberSciTech)*. IEEE, pp. 787–794.

Mirjalili, Seyedali (2016). "Dragonfly algorithm: a new meta-heuristic optimization technique for solving single-objective, discrete, and multi-objective problems". In: *Neural Computing and Applications* 27.4, pp. 1053–1073.

— (2019). "Genetic algorithm". In: *Evolutionary algorithms and neural networks*. Springer, pp. 43–55.

Mirjalili, Seyedali, Seyed Mohammad Mirjalili, and Andrew Lewis (2014). "Grey wolf optimizer". In: *Advances in engineering software* 69, pp. 46–61.

Mirjalili, Seyedali et al. (2017). "Salp Swarm Algorithm: A bio-inspired optimizer for engineering design problems". In: *Advances in Engineering Software* 114, pp. 163–191.

Monteil, Jean-Baptiste et al. (2020). "Resource reservation within sliced 5g networks: A cost-reduction strategy for service providers". In: *2020 IEEE International Conference on Communications Workshops (ICC Workshops)*. IEEE, pp. 1–6.

Moradi, Mehrdad et al. (2018). "SoftBox: A Customizable, Low-Latency, and Scalable 5G Core Network Architecture". In: *IEEE Journal on Selected Areas in Communications* 36.3, pp. 438–456.

Morgado, António et al. (2018). "A survey of 5G technologies: regulatory, standardization and industrial perspectives". In: *Digital Communications and Networks* 4.2, pp. 87–97.

Mudassir, Ahmad et al. (2019). "Game theoretic efficient radio resource allocation in 5G resilient networks: A data driven approach". In: *Transactions on Emerging Telecommunications Technologies*, e3582.

Najm, Ihab Ahmed et al. (2019). "Machine learning prediction approach to enhance congestion control in 5G IoT environment". In: *Electronics* 8.6, p. 607.

Narmanlioglu, Omer et al. (2017). "Prediction of active UE number with Bayesian neural networks for self-organizing LTE networks". In: *2017 8th International Conference on the Network of the Future (NOF)*. IEEE, pp. 73–78.

Nguyen, Cong T et al. (2021). "Transfer learning for future wireless networks: A comprehensive survey". In: *arXiv preprint arXiv:2102.07572*.

Nguyen, Dinh C et al. (2020). "Enabling AI in future wireless networks: a data life cycle perspective". In: *IEEE Communications Surveys & Tutorials* 23.1, pp. 553–595.

Niknam, Solmaz et al. (2020). "Intelligent O-RAN for beyond 5G and 6G wireless networks". In: *arXiv preprint arXiv:2005.08374*.

Niu, Shuteng et al. (2020). "A decade survey of transfer learning (2010–2020)". In: *IEEE Transactions on Artificial Intelligence* 1.2, pp. 151–166.

Nokia (2020). "What is Open RAN and why is it important?" In: *Nokia*.

Ofcom, UK (2017). ""Update on 5G spectrum in the UK". In.

Ojijo, Mourice O and Olabisi E Falowo (2020). "A survey on slice admission control strategies and optimization schemes in 5G network". In: *IEEE Access* 8, pp. 14977–14990.

Ordonez-Lucena, Jose et al. (2017). “Network slicing for 5G with SDN/NFV: Concepts, architectures, and challenges”. In: *IEEE Communications Magazine* 55.5, pp. 80–87.

Panwar, Nisha, Shantanu Sharma, and Awadhesh Kumar Singh (2016). “A survey on 5G: The next generation of mobile communication”. In: *Physical Communication* 18, pp. 64–84.

Parera, Claudia et al. (2019). “Transfer learning for channel quality prediction”. In: *2019 IEEE International Symposium on Measurements & Networking (M&N)*. IEEE, pp. 1–6.

Parera, Claudia et al. (2020). “Transfer learning for multi-step resource utilization prediction”. In: *2020 IEEE 31st Annual International Symposium on Personal, Indoor and Mobile Radio Communications*. IEEE, pp. 1–6.

Parvez, Imtiaz et al. (2018). “A survey on low latency towards 5G: RAN, core network and caching solutions”. In: *IEEE Communications Surveys & Tutorials* 20.4, pp. 3098–3130.

Perveen, Abida, Mohammad Patwary, and Adel Aneiba (2019). “Dynamically reconfigurable slice allocation and admission control within 5G wireless networks”. In: *2019 IEEE 89th Vehicular Technology Conference (VTC2019-Spring)*. IEEE, pp. 1–7.

— (2020). “End-use Aware Optimized Control Signaling for User Admission within 5G and Beyond Networks”. In: *ICC 2020-2020 IEEE International Conference on Communications (ICC)*. IEEE, pp. 1–7.

Perveen, Abida et al. (2021a). “Clustering-based Redundancy Minimization for Edge Computing in Future Core Networks”. In: *2021 IEEE 4th 5G World Forum (5GWF)*. IEEE, pp. 453–458.

Perveen, Abida et al. (2021b). “Dynamic traffic forecasting and fuzzy-based optimized admission control in federated 5G-open RAN networks”. In: *Neural Computing and Applications*, pp. 1–19.

Pham, Quoc-Viet et al. (2020). “A survey of multi-access edge computing in 5G and beyond: Fundamentals, technology integration, and state-of-the-art”. In: *IEEE Access* 8, pp. 116974–117017.

Pouyanfar, Samira et al. (2018). “A survey on deep learning: Algorithms, techniques, and applications”. In: *ACM Computing Surveys (CSUR)* 51.5, pp. 1–36.

Pratap, Ajay and Sajal K Das (2021). “Stable Matching based Resource Allocation for Service Provider’s Revenue Maximization in 5G Networks”. In: *IEEE Transactions on Mobile Computing*.

Qiao, Jian et al. (2015). “Enabling device-to-device communications in millimeter-wave 5G cellular networks”. In: *IEEE Communications Magazine* 53.1, pp. 209–215.

Qu, Yuben et al. (2020). “Empowering the Edge Intelligence by Air-Ground Integrated Federated Learning in 6G Networks”. In: *arXiv preprint arXiv:2007.13054*.

Raikwar, Aditya R et al. (2017). “Long-term and short-term traffic forecasting using holt-winters method: A comparability approach with comparable data in multiple seasons”. In: *International Journal of Synthetic Emotions (IJSE)* 8.2, pp. 38–50.

Rappaport, Theodore S et al. (2013). “Millimeter wave mobile communications for 5G cellular: It will work!” In: *IEEE access* 1, pp. 335–349.

Raza, Muhammad Qamar and Abbas Khosravi (2015). “A review on artificial intelligence based load demand forecasting techniques for smart grid and buildings”. In: *Renewable and Sustainable Energy Reviews* 50, pp. 1352–1372.

Raza, Muhammad Rehan et al. (2018). “A slice admission policy based on reinforcement learning for a 5G flexible RAN”. In: *2018 European Conference on Optical Communication (ECOC)*. IEEE, pp. 1–3.

Richart, M. et al. (2016). “Resource Slicing in Virtual Wireless Networks: A Survey”. In: *IEEE Transactions on Network and Service Management* 13.3, pp. 462–476.

Roh, Wonil et al. (2014). “Millimeter-wave beamforming as an enabling technology for 5G cellular communications: Theoretical feasibility and prototype results”. In: *IEEE communications magazine* 52.2, pp. 106–113.

Roman, Rodrigo, Javier Lopez, and Masahiro Mambo (2018). “Mobile edge computing, fog et al.: A survey and analysis of security threats and challenges”. In: *Future Generation Computer Systems* 78, pp. 680–698.

Rosenblatt, Murray (1956). “A central limit theorem and a strong mixing condition”. In: *Proceedings of the National Academy of Sciences of the United States of America* 42.1, p. 43.

Russell, Stuart J and Peter Norvig (2010). *Artificial Intelligence-A Modern Approach, Third International Edition.*

Santos, Guto Leoni et al. (2020). “When 5G meets deep learning: a systematic review”. In: *Algorithms* 13.9, p. 208.

Santos, José et al. (2021). “Resource provisioning in fog computing through deep reinforcement learning”. In.

Sattar, Danish and Ashraf Matrawy (2019). “Optimal slice allocation in 5G core networks”. In: *IEEE Networking Letters* 1.2, pp. 48–51.

Saxena, Amit et al. (2017). “A review of clustering techniques and developments”. In: *Neurocomputing* 267, pp. 664–681.

Schmidt, Robert, Chia-Yu Chang, and Navid Nikaein (2019). “Slice scheduling with QoS-guarantee towards 5G”. In: *2019 IEEE Global Communications Conference (GLOBECOM)*. IEEE, pp. 1–7.

Schwarz, Maike et al. (2006). “M/M/1 queueing systems with inventory”. In: *Queueing Systems* 54.1, pp. 55–78.

Sciancalepore, Vincenzo, Xavier Costa-Perez, and Albert Banchs (2019). “RL-NSB: Reinforcement Learning-Based 5G Network Slice Broker”. In: *IEEE/ACM Transactions on Networking* 27.4, pp. 1543–1557.

Sciancalepore, Vincenzo et al. (2017). “Mobile traffic forecasting for maximizing 5G network slicing resource utilization”. In: *IEEE INFOCOM 2017-IEEE Conference on Computer Communications*. IEEE, pp. 1–9.

Shami, Tareq M, Ayman A El-Saleh, and Aymen M Kareem (2014). “On the detection performance of cooperative spectrum sensing using particle swarm optimization algorithms”. In: *2014 IEEE 2nd International Symposium on Telecommunication Technologies (ISTT)*. IEEE, pp. 110–114.

Shin, Myung-Ki et al. (2017). “A way forward for accommodating NFV in 3GPP 5G systems”. In: *Information and Communication Technology Convergence (ICTC), 2017 International Conference on*. IEEE, pp. 114–116.

Shrestha, Ajay and Ausif Mahmood (2019). “Review of deep learning algorithms and architectures”. In: *IEEE Access* 7, pp. 53040–53065.

Shrimali, Bela, Harshad Bhadka, and Hiren Patel (2018). “A fuzzy-based approach to evaluate multi-objective optimization for resource allocation in cloud”. In: *International Journal of Advanced Technology and Engineering Exploration* 5.43, pp. 140–150.

Siami-Namini, Sima, Neda Tavakoli, and Akbar Siami Namin (2018). "A comparison of ARIMA and LSTM in forecasting time series". In: *2018 17th IEEE international conference on machine learning and applications (ICMLA)*. IEEE, pp. 1394–1401.

Silva, Ketyllen C et al. (2018). "Self-tuning handover algorithm based on fuzzy logic in mobile networks with dense small cells". In: *2018 IEEE Wireless Communications and Networking Conference (WCNC)*. IEEE, pp. 1–6.

Soliman, Hazem M and Alberto Leon-Garcia (2016). "QoS-aware frequency-space network slicing and admission control for virtual wireless networks". In: *2016 IEEE Global Communications Conference (GLOBECOM)*. IEEE, pp. 1–6.

Su, Ruoyu et al. (2019). "Resource allocation for network slicing in 5G telecommunication networks: A survey of principles and models". In: *IEEE Network* 33.6, pp. 172–179.

Sun, Guolin et al. (2019a). "Autonomous resource provisioning and resource customization for mixed traffics in virtualized radio access network". In: *IEEE Systems Journal* 13.3, pp. 2454–2465.

Sun, Guolin et al. (2019b). "Dynamic reservation and deep reinforcement learning based autonomous resource slicing for virtualized radio access networks". In: *Ieee Access* 7, pp. 45758–45772.

Sun, Mengying et al. (2020). "Large-Scale User-Assisted Multi-Task Online Offloading for Latency Reduction in D2D-Enabled Heterogeneous Networks". In: *IEEE Transactions on Network Science and Engineering*.

Sun, Yan, Fuhong Lin, and Haitao Xu (2018). "Multi-objective optimization of resource scheduling in Fog computing using an improved NSGA-II". In: *Wireless Personal Communications* 102.2, pp. 1369–1385.

Sun, Yaohua, Mugen Peng, and Shiwen Mao (2018). "Deep reinforcement learning-based mode selection and resource management for green fog radio access networks". In: *IEEE Internet of Things Journal* 6.2, pp. 1960–1971.

Suresh, K and N Kumaratharan (2021). "Call Admission Control Decision Maker Based on Optimized Fuzzy Inference System for 5G Cloud Radio Access Networks". In: *Wireless Personal Communications*, pp. 1–21.

Tan, Chuanqi et al. (2018). "A survey on deep transfer learning". In: *International conference on artificial neural networks*. Springer, pp. 270–279.

Tan, Liansheng et al. (2015). “Utility maximization resource allocation in wireless networks: Methods and algorithms”. In: *IEEE Transactions on systems, man, and cybernetics: systems* 45.7, pp. 1018–1034.

Tang, Fengxiao, Yibo Zhou, and Nei Kato (2020). “Deep reinforcement learning for dynamic uplink/downlink resource allocation in high mobility 5G HetNet”. In: *IEEE Journal on Selected Areas in Communications* 38.12, pp. 2773–2782.

Tang, Jianhua, Byonghyo Shim, and Tony QS Quek (2019). “Service multiplexing and revenue maximization in sliced C-RAN incorporated with URLLC and multicast eMBB”. In: *IEEE Journal on Selected Areas in Communications* 37.4, pp. 881–895.

Techplayon (2019). “Open RAN: (O-RAN) Reference Architecture”. In: *Techplayon O-RAN Alliance*.

Tong, Hui and Timothy X Brown (2000). “Adaptive call admission control under quality of service constraints: a reinforcement learning solution”. In: *IEEE Journal on selected Areas in Communications* 18.2, pp. 209–221.

Trivisonno, Riccardo et al. (2018). “mIoT Slice for 5G Systems: Design and Performance Evaluation”. In: *Sensors* 18.2, p. 635.

Tseliou, Georgia et al. (2016). “A capacity broker architecture and framework for multi-tenant support in LTE-A networks”. In: *2016 IEEE International Conference on Communications (ICC)*. IEEE, pp. 1–6.

Ullah, Hanif et al. (2019). “5G communication: an overview of vehicle-to-everything, drones, and healthcare use-cases”. In: *IEEE Access* 7, pp. 37251–37268.

Uwaechia, Anthony Ngozichukwuka and Nor Muzlifah Mahyuddin (2020). “A comprehensive survey on millimeter wave communications for fifth-generation wireless networks: Feasibility and challenges”. In: *IEEE Access* 8, pp. 62367–62414.

Vashi, Shivangi et al. (2017). “Internet of Things (IoT): A vision, architectural elements, and security issues”. In: *I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC), 2017 International Conference on*. IEEE, pp. 492–496.

Wagle, Neeti and Eric W Frew (2012). “Transfer learning for dynamic RF environments”. In: *2012 American Control Conference (ACC)*. IEEE, pp. 1406–1411.

Walingo, Tom Mmbasu and Fambirai Takawira (2014). “Performance analysis of a connection admission scheme for future networks”. In: *IEEE Transactions on wireless communications* 14.4, pp. 1994–2006.

Walters, Carl and Donald Ludwig (1994). “Calculation of Bayes posterior probability distributions for key population parameters”. In: *Canadian Journal of Fisheries and Aquatic Sciences* 51.3, pp. 713–722.

Wang, Cheng-Xiang et al. (2020). “Artificial intelligence enabled wireless networking for 5G and beyond: Recent advances and future challenges”. In: *IEEE Wireless Communications* 27.1, pp. 16–23.

Wang, Dan et al. (2019). “Intelligent Cognitive Radio in 5G: AI-Based Hierarchical Cognitive Cellular Networks”. In: *IEEE Wireless Communications* 26.3, pp. 54–61.

Wang, J, H Roy, and C Kelly (2019). “OpenRAN: The next generation of radio access networks”. In: *Telecom Infra Project*.

Wang, Xiaoqian, Xin Su, and Bei Liu (2019). “A novel network selection approach in 5G heterogeneous networks using Q-learning”. In: *2019 26th International Conference on Telecommunications (ICT)*. IEEE, pp. 309–313.

Wang, Xiong et al. (2018). “Millimeter wave communication: A comprehensive survey”. In: *IEEE Communications Surveys & Tutorials* 20.3, pp. 1616–1653.

Wang, Zhi and Yigang Cai (2019). “Management Optimization of Mobile Edge Computing (MEC) in 5G Networks”. In: *2019 IEEE International Conference on Communications Workshops (ICC Workshops)*. IEEE, pp. 1–6.

Wireless, Parallel (2020). “5G 4G 3G 2G WI-FI OPENRAN CONTROLLER”. In: *Parallel Wireless*.

Wu, Dapeng et al. (2018). “Biologically Inspired Resource Allocation for Network Slices in 5G-Enabled Internet of Things”. In: *IEEE Internet of Things Journal*.

Xiang, Bin et al. (2019). “Joint network slicing and mobile edge computing in 5G networks”. In: *ICC 2019-2019 IEEE International Conference on Communications (ICC)*. IEEE, pp. 1–7.

Xiong, Zehui et al. (2019). “Deep reinforcement learning for mobile 5G and beyond: Fundamentals, applications, and challenges”. In: *IEEE Vehicular Technology Magazine* 14.2, pp. 44–52.

Xu, Yanfang et al. (2012). “Interference-aware channel allocation for device-to-device communication underlaying cellular networks”. In: *2012 1st IEEE International Conference on Communications in China (ICCC)*. IEEE, pp. 422–427.

Ye, Hao, Geoffrey Ye Li, and Biing-Hwang Fred Juang (2019). "Deep reinforcement learning based resource allocation for V2V communications". In: *IEEE Transactions on Vehicular Technology* 68.4, pp. 3163–3173.

Yi, Bo, Xingwei Wang, and Min Huang (2018). "Optimised approach for VNF embedding in NFV". In: *IET Communications* 12.20, pp. 2630–2638.

Yu, Qi-Yue, Hong-Chi Lin, and Hsiao-Hwa Chen (2019). "Intelligent Radio for Next Generation Wireless Communications: An Overview". In: *IEEE Wireless Communications* 26.4, pp. 94–101.

Zeng, Haiyong et al. (2019). "A green coordinated multi-cell NOMA system with fuzzy logic based multi-criterion user mode selection and resource allocation". In: *IEEE Journal of Selected Topics in Signal Processing* 13.3, pp. 480–495.

Zeng, Qingtian et al. (2020). "Traffic prediction of wireless cellular networks based on deep transfer learning and cross-domain data". In: *IEEE Access* 8, pp. 172387–172397.

Zhang, Chaoyun, Paul Patras, and Hamed Haddadi (2019). "Deep learning in mobile and wireless networking: A survey". In: *IEEE Communications surveys & tutorials* 21.3, pp. 2224–2287.

Zhang, Chiya et al. (2019a). "Research Challenges and Opportunities of UAV Millimeter-Wave Communications". In: *IEEE Wireless Communications* 26.1, pp. 58–62.

ZHANG, Ping, Yun-zheng TAO, and Zhi ZHANG (2016). "Survey of several key technologies for 5G". In: *Journal on Communications* 37.7, p. 15.

Zhang, Xinran et al. (2019b). "Deep-reinforcement-learning-based mode selection and resource allocation for cellular V2X communications". In: *IEEE Internet of Things Journal* 7.7, pp. 6380–6391.

Zhang, Zhen et al. (2017). "A traffic prediction algorithm based on Bayesian spatio-temporal model in cellular network". In: *2017 International Symposium on Wireless Communication Systems (ISWCS)*. IEEE, pp. 43–48.

Zhao, Qiyang et al. (2013). "Transfer learning: A paradigm for dynamic spectrum and topology management in flexible architectures". In: *2013 IEEE 78th Vehicular Technology Conference (VTC Fall)*. IEEE, pp. 1–5.

Zhao, Qiyang et al. (2015). "Using k-means clustering with transfer and Q learning for spectrum, load and energy optimization in opportunistic mobile broadband networks". In: *2015 International Symposium on Wireless Communication Systems (ISWCS)*. IEEE, pp. 116–120.

Zhao, Zhijin et al. (2009). "Cognitive radio adaptation using particle swarm optimization". In: *Wireless Communications and Mobile Computing* 9.7, pp. 875–881.

Zheng, Jiaxiao et al. (2018). "Statistical multiplexing and traffic shaping games for network slicing". In: *IEEE/ACM Transactions on Networking*.

Zhuang, Fuzhen et al. (2020). "A comprehensive survey on transfer learning". In: *Proceedings of the IEEE* 109.1, pp. 43–76.