Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

School of Computer Science and Statistics

On Multi-Radio Multi-Server Powered Multi-Access Edge Computing

Asad Ali

May 4, 2023

A dissertation submitted in partial fulfilment
of the requirements for the degree of
MAI (Computer Science)

Declaration

I hereby declare that this dissertation is entirely my own work and that it has not been submitted as an exercise for a degree at this or any other university.

I have read and I understand the plagiarism provisions in the General Regulations of the University Calendar for

I have also completed the Online Tutorial on avoiding plagiarism 'Ready Steady Write', located at

Signed: _____ Date: _____

Abstract

Highly intelligent, automated and ubiquitous digital world will be hallmark of the coming decade. To achieve this, we need high-speed, highly-reliable connectivity between physical, digital and biological world. In terms of cloud systems, Multi-access Edge Computing (MEC) has been playing a key role in enabling mobile devices to have swift connectivity to resource-rich cloud servers. However, the current state-of-art may be unable to meet the full connectivity and processing demands of the future compute- and bandwidth-hungry applications transpiring the envisioned digital society. To make up for the capacity, 5G and upcoming 6G extend the channel bandwidth. This exacerbates the already daunting spectrum resource scarcity and adds to the cost of the network. To minimize the cost of the network and delay, a solution recently proposed in the literature is the concept of parallel offloading to multiple servers over multiple radios access technologies (RATs) that a mobile device comes equipped with such as Wi-Fi Direct, Wi-Fi and macro-cellular technology such as 5G.

Using multi-radio multi-server powered MEC, we work on minimizing network delay as well as jointly minimizing network and computation delay. To minimize the network delay, we measure the performance on different radios. Using the obtained performance, we optimally utilize the joint capacities of the radios and schedule the traffic in such a way that packet order at the source and destination is maintained thereby completely avoiding packet reordering delay to keep the throughput intact. We develop a Continuous Non-Linear Program (CNLP) that vary the load on the radio access technologies according to their performances. The proposed CNLP is solved through Lagrange's Multiplier theorem for several constraints. Furthermore, to ensure smooth relay of the MEC traffic, capacity distribution at the relay node is optimized according to the arrival of the MEC traffic. Numerical results show significant improvement in terms of throughput, delay and QoS compared with other techniques using multiple radios for computation offloading

To jointly minimize network and computation delay, we develop a technique that chooses the most optimal servers. Further, to minimize server migration and to achieve a convergence point in the algorithm, we formulated a max-min based non-linear lexicographic minimization problem. To solve the formulated problem in polynomial time, we transform the non-linear objective function to a linear one and solve it through the simplex algorithm. Based on the obtained network performance and computation delay, we formulate a multi-server multi-radio load distribution problem to optimally utilize the available capacities of the radios. This problem is solved using techniques from algorithmic game theory. Illustrative numerical results show that proposed technique significantly minimizes computational and network delay.

Contents

List of Figures

List of Tables

Nomenclature

MEC	Multi-Access Edge Computing
RAT	Radio Access Technology
Wi-Fi D	Wi-Fi Direct
λ_U	Traffic Load on a particular link u
d_u	Delay of the link u
D_r	End-to-End Delay on RAT r
ζ_i	Capacity of RAT i
μ	Packet length
R_u	Data rate of link u
λ	Packet Arrival Rate
L_i	Load on link i
CNLP	Continuous Non-Linear Program
QoS	Quality of Service
MIMO	Multiple Input Multiple Output
MSMR	Multi-Server Multi-Radio
TCP	Transmission Control Protocol
IoT	Internet of Things
VEC	Vehicular Edge Networking
LTE	Long Term Evolution
UE	User Equipment
D_l	Delay when computation performed locally
D_v''	Delay of the slowest radio when task are offloaded
Δ	Packet re-ordering delay
δ	Packet retransmission delay
Θ	Propagation Delay
BER	Bit Error Rate
LDPC	Low Density Parity Check
h_r	Load share ratio of an arbitrarily chosen radio r
T_{sd}	Traffic from node s to node d
τ_{ij}	Delay of the link u
Γ	System delay
K	Lagrange Multiplier

1 Introduction

1.1 Background

Mobile devices with their small sizes come handy but with the limited processing power and small storage, cannot run computationally intensive applications such as high-processing gaming, demanding scientific algorithms and computations that require large data storage. Because of this fundamental limitation, mobile devices are coupled with servers located at the edge of the networks to assist the execution of demanding tasks such as online gaming, e-Health, virtual assistance, Internet of things, augmented/virtual reality [2]. The coexistence of mobile devices with the edge servers has lead to a new category of cloud computing known a Multi-access Edge Computing (MEC) [3]. In traditional cloud computing model, a cloud server is placed on the core network. Consequently, a data packet traversing different nodes and travelling all the way to the server and back to the user will cause packets of real-time applications to miss their deadlines, thus eventually get discarded by the application [4]. This results in drastic degradation of the service delivery which we define as the time taken by the application to produce output of given request [5]. Therefore, MEC came into existence to make up for the computationally-intensive tasks that are characterized by their ultra-low latency and high throughput requirement. In MEC, a compute and resource-intensive task performed by the application that is too demanding for device hosting it, may be delegated to the edge server. The edge server in return processes the task and feedback the output to the application. This scheme makes the mobile device a mere input-output terminal while the actual processing is performed by the server whereas the end-user is oblivious to the actual processing flow. This way, network traffic is reduced, congestion and latency are minimized while overall application performance is improved significantly.

Multi-Access Edge Computing (MEC) has been paramount to research for about a decade. It's a fabulous way to make mobile devices appear to execute high-processing gaming, demanding scientific algorithms and computations despite their limited processing power, storage, and battery size. There have been a significant number of groundbreaking works in this area. However, as we are forging forward to a fully con-

nected digital world, the current state-of-art may fall short to accommodate the explosive growth in the number of wireless devices and real-time bandwidth hungry applications such a virtual/augment reality, holographic telepresence, Internet-of-Everything, smart grid 2.0, Industry 5.0, robotics for their stringent requirements [6]. For example, requirement of these applications in terms of ultra-high data rates, real-time access to powerful computing resources, ultra-low latency, and extremely high reliability and availability already surpass the network capabilities promised by current infrastructure [7]. To support the aforementioned services and applications thereby materializing the envisioned digital world of the next decade, both in 5G and the upcoming 6G, communication capacity is improved by extending the channel bandwidth. This exacerbates the already daunting spectrum resource scarcity and adds to the cost of the network. Finally, with the advent of Massive IoT, 6G, and MEC itself, there will be massive densification of devices. We shall notice several disruptive changes in the networks and applications. For these reason, the existing techniques are insufficient to cope up with the requirements of future networks.

On the processing side, in existing techniques, a single server is chosen ignoring its available capacity, existing load on it and its intermittent unavailability. The idea behind MEC is to bring large number of servers close to users, so that each server serves a small number of users to expedite the processing. However, if large number of users congregate around a single MEC server and overload it, the available capacity can no longer be disregarded. In worst cases, the server may even become unavailable. Therefore, we need to manage the numbers of users and load on the servers, as to minimize service delay, we need to minimize both network delay as well as computation related delays at the server.

Motivated by these concerns, we propose multi-server multi-radio (MSMR) powered MEC where offloading occurs over multiple radios and depending upon the conditions, the radios may be connected to different servers. Seemingly straightforward, adopting MSMR however, has many inherent problems such as difficulty in multi-radio packet scheduling and optimizing packet distribution due to difference in the capacity and delay of radio access technologies, optimizing the allocation of resources, maintaining packet order after arriving at the destination, server selection.

Another motivation behind this work is the number of shortcomings that subsists in the existing techniques. For example, ignoring or assuming constant values for several important parameters such as SNR, bitrate, coding rate, available server capacity etc. Moreover, in an MEC system, performance is measured in terms of transmission capacity of the RAT and then offloading decisions are taken. However, transmission capacity between the relay node and the MEC server is ignored or taken constant. This issue will further exacerbate when there are multiple RATs involved because with difference

in performance measures, there will be packet reordering issues. Therefore, we need to consider and manage the transmission capacity at the relay node to ensure smooth relay of the MEC packets at the relay node.

1.2 Research Gap

- The communication and processing capacity of the current infrastructure is insufficient to accommodate the future bandwidth and compute-hungry applications. Both in 5G and 6G, communication capacity is improved by extending the channel bandwidths. This exacerbates the already daunting spectrum resource scarcity and adds to the cost of the network. On processing side, available capacity and the intermittent unavailability of the server hampers the processing speed.

- Packet scheduling over multiple RATs has a lot of inefficiencies. The existing techniques do not utilize available capacities of the RATs optimally. A parameter, ratio of residual capacity to total capacity, that shows how optimal packet distribution over multiple RATs is, is sub-optimal for the existing techniques. The scheduling and load distribution over multiple radio access technologies gives rise to packet arriving out-of-order which is another factor that plummets the throughput and drastically increases the end-to-end delay.

- Resource sharing in current technique are sub-optimal ignoring different important parameters such as post-relay node capacity of the links, server unavailability and channel related parameters such as SNR, bitrate, coding rate.

1.3 Structure

The transfer report is structured as follows.

- In Chapter 2 we give a general overview of Multi-Radio Multi-Server MEC. We discuss relevant problems of Multi-RAT based task offloading and present a state-of-the-art.

- In Chapter 3 we show our network delay optimizing models in a Multi-Radio based MEC. We show parallel offloading over multiple radios to a single edge server.

- Chapter 4 is on joint optimization of network delay as well as processing delay. We show parallel offloading over multiple Radios to multiple server.

- We conclude the thesis in Chapter 5.

1.4 Key Contributions

A brief list of major contributions is as follows.

- We present a multi-server multi-radio based architecture for MEC based computation offloading. We exploit the concept of the parallel offloading over multi-radio end-nodes. We develop a Continuous Non-Linear Program (CNLP) to minimize network delay. The developed CNLP, solved through Lagrange's Multiplier Theorem for several constraints, avoid packet re-ordering delay by ensuring packets arrive at the destination simultaneously by equalling delays of the radios. To minimize network delay, we also optimize load distribution and packet scheduling to optimally utilize available capacities of the radios.
- We jointly minimize network and computation delay for the multi-server architecture. We develop a technique that chooses the optimal server for computation and avoid server migration in real time using max-min based non-linear lexicographic minimization problem. With the computation and transmission capacity of the servers and radios computed, we formulate a multi-server multi-radio load distribution problem to optimally utilize the available capacities. The load distribution technique schedules the packets in such a way that packet order at the destination is maintained.

2 Multi-Radio MEC

In this chapter, we describe Multi-Radio based MEC. We highlight challenges with such a system and provide a provide a brief overview of the state-of-the-art.

2.1 Introduction

In MEC, quality of service is measured in terms of how quickly the user gets the results of their generated request, commonly known as service delay whereas service delay is the sum of network delay and processing delay at the edge server. The principal requirement of MEC is to have indiscernible service delay which, despite of prioritized processing, is impeded by factors such as traffic congestion in the network and link's limited capacity [8]. Operators are finding increasingly difficult to provide adequate QoS to emergence of real-time, bandwidth hungry applications. When combined with the increasing popularity of wireless technologies beyond smartphones, additional home appliances requiring wireless connectivity and plunging cost of electronic equipment, our next generation wireless networks are expected to have enormous capacity to accommodate the ever increasing demands from bandwidth-intensive applications [9]. Figure (2.1) and Table (2.1) shows some recent statistics about the amount of data generated in different times to give us the idea about the capacity our data networks need to have. The compound annual growth rate of internet traffic from 2017 to 2022 is 47% [1] which implies 1000 times more capacity relative to the current level of capacity in next few decades.

Different techniques such as massive multiple-input multiple-output (MIMO) and beam-forming [10], spatial multiplexing [11], multi-band transmission [12] [13], channel bonding and bandwidth aggregation [14], prioritized processing [15] have been devised to keep up with the rapidly growing real-time bandwidth hungry applications. However, due to new applications' stringent performance requirement, MEC communication still struggles to provide adequate connectivity [16].

Motivated by these concerns, we propose multi-server multi-radio (MSMR) powered MEC where offloading occurs simultaneously on all the radios that a mobile phone comes equipped with. A smartphone today comes equipped with different radios supporting

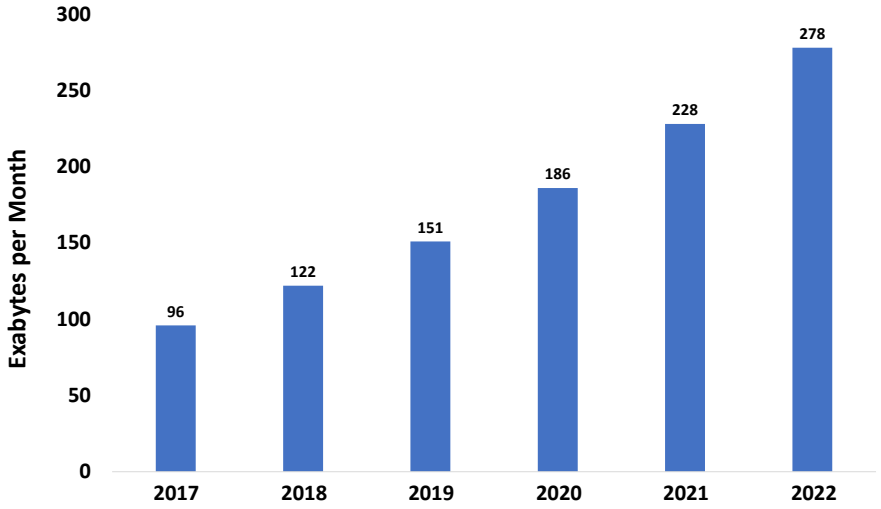

Figure 2.1: Annual mobile traffic growth [1].

Table 2.1: Historical internet context [1]

Year	Global Internet Traffic
1992	100 GB per day
1997	100 GB per hour
2002	100 GB per second
2007	2,000 GB per second
2016	26,600 GB per second
2022	105,800 GB per second

different technologies such as WiFi-Direct, WiFi and macro-cellular technology such as 5G, and depending upon the conditions, the radios may be connected to different servers. Starting with 4G/5G, our smartphones are connected to macro-base station to provide broad coverage and medium speed connectivity. On the other hand Wi-Fi, based on IEEE 802.11 standard, provides ultra fast connectivity, but confined to a local area networks. Similarly, Wi-Fi Direct is a peer-to-peer WiFi standard for device-to-device communication without involving intermediary central access point or router [17]. Wi-Fi Direct has been shown to be a successful avenue for task offloading [18]. A simple illustration of a smartphone with multiple radio access technologies is shown in Figure 2.2.

Seemingly straightforward, adopting MSMR however, has many inherent problems. For example, the radios working on different technologies and standards have different capacities and different response to physical conditions such as distance, interference, barriers and the environment in which they operate, hence, different capacity and network delays. This inherent nature of a multi-radio system makes traffic scheduling a daunting task. Unequal delay of radios have detrimental effect on system throughput. For example, MEC server cannot process data transmitted by a faster radio, as it will have to wait to receive the data from the slower radio to combine the two chunks to

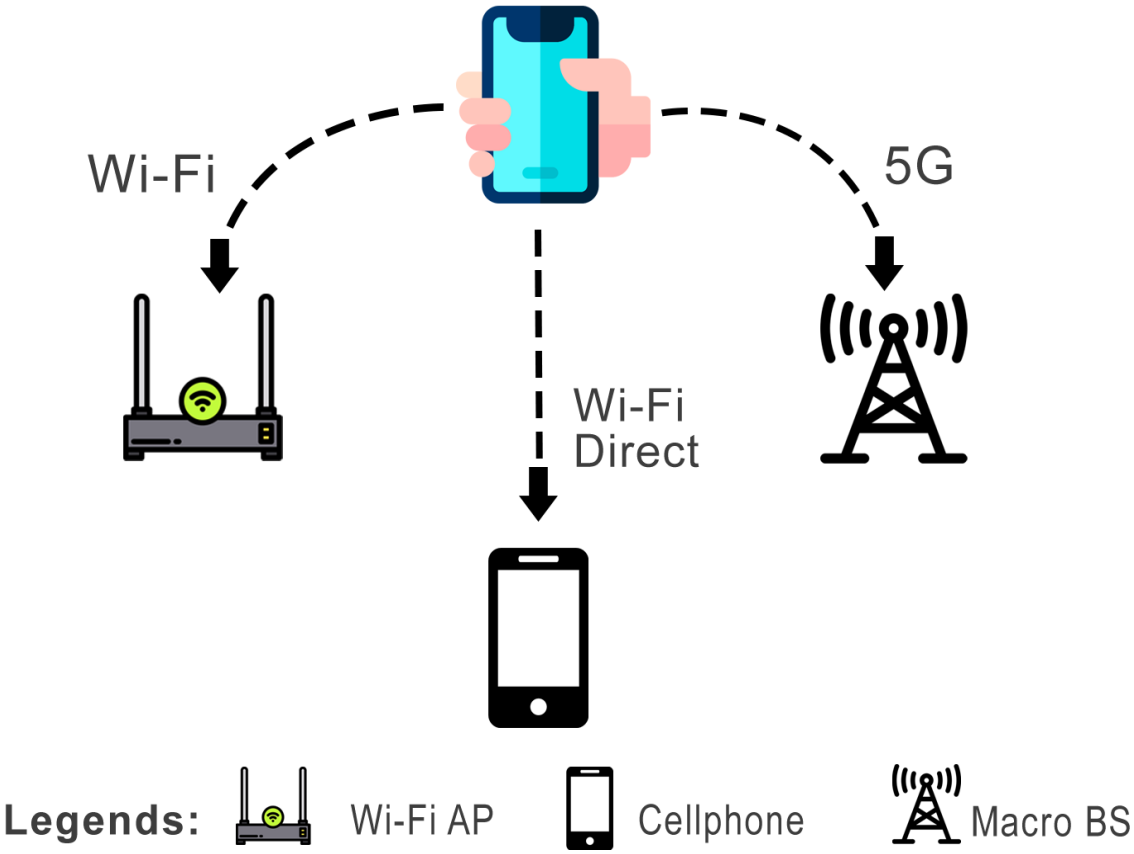

Figure 2.2: A Simple Illustration of a smartphone with Multi-RAT capabilities.

make one whole. Consequently, processing is hampered by the slower radio. This under-utilization of capacity is caused at many level due to inefficient traffic scheduling as a radio with larger capacity having transmitted its load would sit idle until all other radios are done with their transmission before next transmission session begins. This idle time of the radio could have been used for more data transmission. Moreover, because of the disproportionate or equally shared load, data packets are likely to arrive out-of-order. Factors such as edge server waiting to receive packets from the slower radio and putting received packet back in order increase end-to-end system delay, thus causing some of the packets of real-time applications to miss their respective deadlines and get discarded. This packet reordering issue is further disseminated to Transmission Control Protocol (TCP) layer as packets arrived out-of-order will either be kept in the buffer or discarded depending on the magnitude of the out-of-order packets as Transmission Control Protocol (TCP) can allow packet reordering by a maximum of two positions, which are corrected through inherent re-sequencing mechanism [22]. However, packet reordering beyond two positions is taken as a loss, thus leading to TCP reducing its transmission window size. Consequently, the aggregated capacity will be underutilized, and the application throughput may drop drastically. An efficient MSMR solution is, therefore, one that includes mechanisms to minimize packet reordering to alleviate its effects.

2.2 Challenges of Using Multi-Radio Multi-Server System

In the following, we shall summarize the problems associated with multi-radio multi-server based MEC system.

2.2.1 Traffic Scheduling

An efficient traffic scheduling algorithm is a paramount of importance for giving high Quality of Service (QoS). It is one of the oldest topic of research in computer networks [19]. Traffic scheduling becomes even more challenging when there are multiple radios and multiple MEC servers involved.

2.2.2 Unequal Delay of Radios

The different radios have different capacities, hence, different delays. This inherent nature of a multi-radio system makes traffic scheduling a daunting task. Unequal delay of radio have detrimental effect on system throughput. For example, for real-time applications, MEC server cannot process data transmitted by a faster radio, as it will have to wait to receive the data from the slower radio. Processing in this case the MEC server is hampered by the slow radios. Therefore, it is very important to exploit the physical characteristics of the radios to make the end-to-end delay of all the RATs equal.

2.2.3 Under-Utilization of Resources

A disproportionate or equally shared load will lead to under-utilization of capacity on many level. To begin with, a radio with larger capacity having transmitted its load would sit idle until all other radios are done with their transmission before next transmission session. This is the first case of capacity under-utilization where the idle time could have been used for more data transmission. Moreover, the MEC server cannot act upon the transmitted data as it will be waiting to receive the remaining data from the sender.

2.2.4 Maintaining Packet Order at Receiver

Because of the difference in capacity and delay, and inefficient traffic scheduling, data packets are likely to arrive out-of-order. Packets arrived out-of-order will either be kept in the buffer or discarded depending on the magnitude of the out-of-order packets as TCP can allow packet re-ordering by a maximum of two positions only [22]. Overwhelming the slower radio and putting the total data in correct order increases the overall end-to-end delay. Bringing data packets back in order, consumes a significant amount of time,

which causes packets of real-time application to miss their respective deadlines and get discarded. Therefore, it is necessary for the packets to be received in order and to avoid packet re-ordering delay.

2.2.5 Server Selection

The idea behind Multi-access edge computing is to bring large number of processing servers to the proximity of the networks, so that each MEC server serves a small number of users to expedite the processing. However, MEC servers being considerably less powerful, processing is drastically impacted when large number of users request the same server for their services. This situation is exacerbated when offloading algorithms ignore the factors like load on the server, computational requirement of the application, user mobility and unavailability of servers. Therefore, it is very important to discreetly choose a server for ultimate user experience [20].

2.3 Literature Survey

Computation offloading is one of the oldest topics of computing and probably the main motivation behind computer networks, as can be seen in the memo shared by J. Licklider in 1963 [21]. More recently, the increasing popularity of bandwidth hungry applications in conjunction with mobile devices brought this issue into limelight. Computation offloading to remote central cloud servers is often unsuitable for real-time applications, as the transmission distance and number of hops required to reach a central computing node typically incur latency of several tens of millisecond, with comparably high jitter. Multi-access edge computing (MEC), on the contrary, outdoes traditional cloud computing by significantly enhancing the capabilities of capacity-limited mobile devices thereby remarkably reducing the service delay[4]. It is for this reason that MEC manifests itself as promising technology for extending the computation and storage capabilities of mobile devices.

We acknowledge that over the last few years there have been a large number of studies focusing on the technical aspects of the MEC [5], [8]. Most of the solutions are single radio based and are inadequate to incorporate several key characteristics and are often too simple to reflect real world scenarios. In the following discussion, we shall divide our literature review into two parts where we shall discuss the motivation behind this work by reviewing the shortcomings in the existing offloading techniques in general and then in the second part, we shall review the work done in the context of multi-radio systems.

2.3.1 Computation Offloading in General

A comprehensive survey on Multi-access edge computing is given in [5], [8] where the authors provide a detailed insight into the problem of computation offloading and resource allocation. Most of the techniques highlighted are inadequate to incorporate several key characteristics and are often too simple to reflect real world scenarios. In the following sections, we attempt to classify the literature according to the assumptions and scenarios assumed.

Binary Offloading Models

In binary offloading models, a task is either offloaded wholly or not at all. MEC systems are essentially multi-tiered in nature. A task can be executed locally by the mobile device as well as servers at the edge or cloud servers at the core network. Most of the existing work ignore the processing power of the mobile device and wholly offload the task [22], [23]. These algorithms are simple and easy to implement, however, they are inefficient and cannot make full use of resources [24].

Constant Values for Important Parameters

Most of the existing works have assumed constant values for several important parameters such as SNR, Bitrate, received signal power and path loss etc. [25], [26]. Similarly, [27] has considered constant values for transmission and processing delay. Assuming constant values for these important parameters is not realistic and leave little room for improving the performance.

Capacity of MEC Server

Most existing techniques also ignore capacity of or load on an MEC server. Capacity of MEC server is either assumed to be constant [28], [29] or assumed to be capable of always processing the task [30]. Processing delay of a task depends upon the capacity of the MEC server. Assuming fix capacity makes the processing delay fix which is clearly not practical and leaves little room for improving performance and server efficiency particularly in a scenario where end-user is mobile and often switch between servers.

Static Offloading

Several techniques assume static offloading where network haphazardry (i-e the fact that networks are dynamic in nature) and spatio-temporal variation in the network is ignored [31], [32], [33]. With user mobility and nature of applications combined with variation in wireless channels, MEC based wireless networks are highly dynamic. Therefore, using deterministic optimization models fall short in real-life scenarios.

Massive IoT and Introduction of 6G

Once Internet of Things (IoT) is wholly deployed, there will be massive increase in the number of active devices in the internet [34]. Similarly, with the advent of 6G, there will be several changes in the network architecture such as introduction of Terahertz frequency, NOMA, beamforming, massive number of base stations and servers are anticipated few at this stage [35] [36]. With large number of clients and requests, servers and their variables, and the new technology, the existing solutions for MEC will no longer work [35].

In some recent works, to make efficient use of resources, [37] considers local computation by partially offloading the tasks to MEC servers. The authors considers both single as well multi-users scenarios of MEC resource allocation for computation offloading, which are solved by branch-and-bound and iterative based heuristics, respectively. Schemes like partial offloading require perfect user-MEC server-remote cloud coordination that leads to high signalling overhead. Moreover, these schemes assume that a task can be arbitrarily divided into subtasks which is an unreal assumption. To further ameliorate the resource allocation, authors in [38] investigated the efficiency of deep reinforcement learning and developed solutions for joint resource allocation and energy minimization based on Deep Q-Networks (DQN). The authors developed techniques based on DQN, convex optimization and traditional Q-learning. However, offloading learning policy is for fixed topology and given the efficiency of DQN, they are not suitable for edge video processing. The goal of minimizing energy consumption and processing delay is carried forward in [39], where authors have developed an evolutionary algorithm that jointly optimizes energy consumption and processing delay and attempts to find pareto-optimal point between energy consumption and processing delay. Computation offloading is also investigated in vehicular edge networks (VEC) in [40], where authors have worked on selecting least congested edge server with an aim to minimize cellular hand-offs to avoid obstruction in computation.

To summarize, given the dynamic nature of MEC applications and wireless networks, the assumption taken in most of the existing solutions are not at par with real world. The networking and processing models have several flaws such as taking fixed delays, load and capacities. Moreover, the fact that base-station (BS) serves as a relay node and that the transmission capacity and communication related delays post- and pre-BS can be different, is ignored. Furthermore, the existing solutions are not scalable enough to cope up with massive IoT and service requirement of future applications. Therefore, we need a solution that is scalable, flexible and completely represents the actual networking and processing operation of the real world.

2.3.2 Parallel Offloading over Multi-Radio

The co-existence of Wi-Fi and macro-cell networks such as LTE has been a widely studied research area [41]. However, in WiFi-LTE integrated networks only, a portion of the capacity of the WiFi AP is used and data is offloaded to Wi-Fi with the aim to improve the cost and throughput of a cell. Similarly, most of the studies investigate downlink performance [42]. We, on the contrary, investigate the synergy of WiFi, WiFi Direct and cell network, and offload the data to a remote server and use any portion of capacity of any radio depending upon the channel condition. Leveraging multi-radio access technologies (RAT) in the context of MEC offloading has been carried out in [43], [44] and [45]. Computation offloading techniques and protocols differ in purpose and how they model the computation offloading process. A detailed review of computation offloading modeling is given in [2]. Within this frame of reference, authors in [43] offload data on the basis of the tasks. For instance, one task is sent over one radio while another task is sent over another RAT. Distributing data on the basis of the computational tasks can lead to packet reordering delay as they can be of different size. Moreover, performance is measured on the basis of the transmission delay and ($Load/Bandwidth$) metric. Different important parameters such as queuing delay, processing delay at the node and congestion are ignored. [44] requires the end-node to send all the information to the relay node such as required latency, data rate, average packet length, average packet arrival rate, required computing power and so on. We believe, sending so much information for real-time applications will be cumbersome and will defy the real purpose of task offloading in real-world. Moreover, the radio-access technologies are not used simultaneously, rather the choice is made for best radio-edge pair. Finally, a detailed analytical framework of the presented work is also missing. Similarly, [45] distributes data flows on the basis of tasks which is subject to packet reordering delay.

We argue that load, when offloading, must be distributed among radios according to the data size rather than tasks be sent over different radios. This distribution must be done according to the channel health and performance of the link which must be duly computed. In addition, the system should maintain the order of data packets.

2.4 Summary

MEC is a plausible mechanism to prop the less powerful mobile devices. For MEC to become a reality, there are several challenges that must be addressed before. A lot of work has been done in this area to resolve the many challenges of MEC. However, the current solutions do not model the networking and processing elements of MEC correctly. Moreover, MEC applications and wireless networks in general are dynamic in nature. There has been high degree of spatio-temporal variations. However, on the

other hand, the current solutions take several important parameters as constant and the overal network is assumed to be static. Finally, the introduction of 6G and massive deployment of IoT, the existing solutions will fail to meet the expectation. Therefore, we need a solution that that truly optimizes the performance of the system in terms of communication and processing delay and a solution that completely models the true networking and processing elements of the system.

3 Multi-RATs Single Server

To minimize the cost of network delay, in this chapter, we describe our multi-radio single server architecture. We describe our network delay formulation, problem formulation and our proposed solution. Multiple radios have been combined to improve throughput and capacity has been carried out in past [46], [47]. In this Chapter, we provide an analytical model to optimize communication related delays by optimization scheduling, and capacity utilization and distribution.

3.1 Computational Offloading over Multi-Radios

In this section, we first introduce the multi-radio computation offloading MEC model considered in our work, followed by a description of our proposed computation offloading. After that, we will formulate the problem for multi-radio simultaneous computation offloading.

3.1.1 Assumed System Model

Our assumed system model is summarized in Figure 3.1. In order to differentiate between the radios, we assume the three radios be WiFi, Wifi-Direct for short range connectivity and 5G for macro-cellular network. Furthermore, for WiFi we assume IEEE 802.11ax standard and the model defined in the standard [48] and in [49], [50] are borrowed for performance estimation. Similarly, for 5G performance estimation is carried out using model described in standard [51]. A detailed description on performance estimation is given in Section 4.5.1.

Starting from the end-node, we have a smartphone as user equipment (UE) that acts as a source node. The UE is equipped with multiple radio access technologies (RATs) such as 5G Transceiver, Wi-Fi and Wi-Fi Direct. We assume that it is in range of a 5G base-station, Wi-Fi access point and occasionally, a peer device comes in its transmission range. Therefore, it can transmit through 5G, Wi-Fi and Wi-Fi Direct simultaneously. In the figure, a peer is any device that serves as a relay node and has same features as the end-node itself that is, any device capable of transmitting over 5G, Wi-Fi, Wi-Fi Direct.

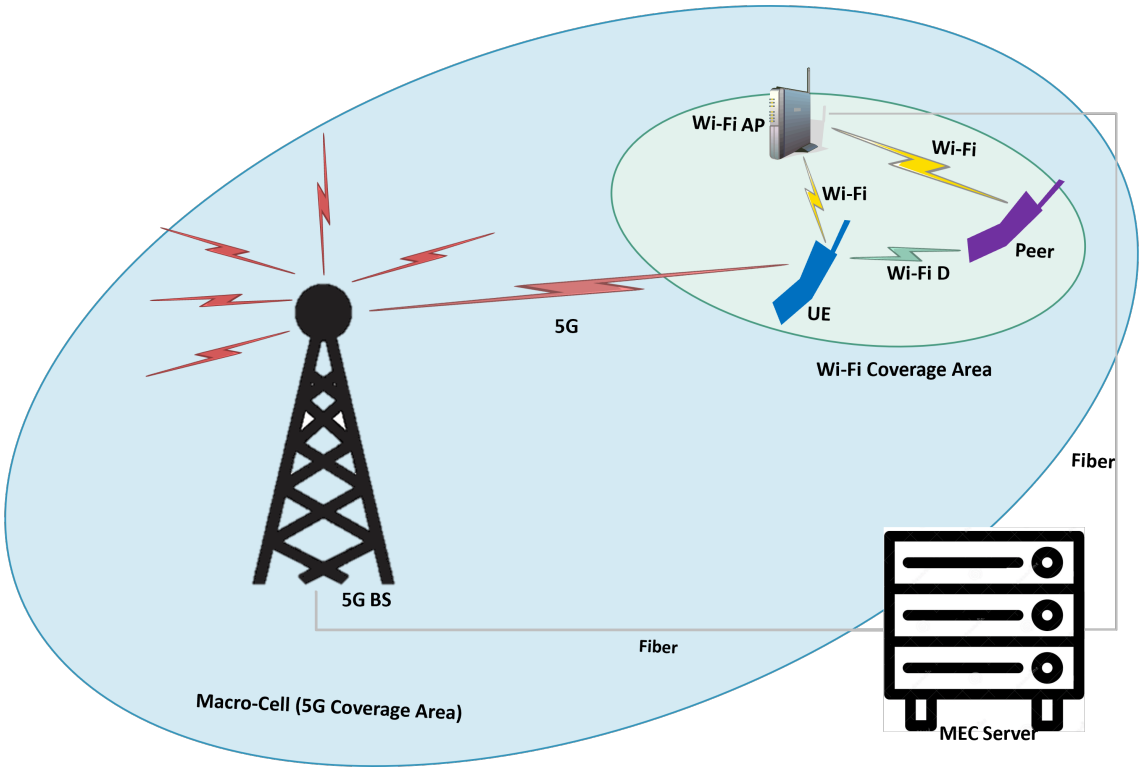

Figure 3.1: A Simple Illustration of Multi-Radio Access Technologies Computation Offloading.

In the figure, the end-node is mobile in nature whereas the Wi-Fi access point and 5G base-station are fixed. We also assume that Wi-Fi access point and 5G base-station can serve multiple users simultaneously. Both the Wi-Fi AP and 5G base-station are connected to a single MEC server by optical fiber connection.

Suppose a task has been generated by the application, the end-node has two options, either execute the task locally or offload the computation to MEC server. Thus, service delay can be mathematically expressed as follows;

$$D = \begin{cases} 1. & D_l \\ 2. & D_v'' + \Delta + \delta \end{cases} \tag{1}$$

Here D_l is the s delay (including computation and queuing time) if a task is performed locally, D_v'' is the service delay of the slowest radio (in this case also including the radio transmission time) when the task is offloaded over v RATs where upper bound of v depends upon the number of radio available.

We assume not all radios are available all the time. Δ is packet reordering delay when data packets arrives at the receiver node which happens to be the MEC server in this case. Finally, δ is the packet retransmission delay when because of the gap between the received packets, receiver is unable to order the packets as per the sequence number

and asks the sender to retransmit the packets. Further details about packet reordering and retransmission are given in 3.1.3.

If a task is executed locally, the service delay will be the sum of processing time and the time it waits in the queue to get the processor. Assuming Poisson processes with rate λ, if ζ is the processing capacity of the device, processing time will be $1/\mu\zeta$, where μ is the size of the process. Similarly, the queuing time will be $\lambda/\left((\mu\zeta)(\mu\zeta-\lambda)\right)$. Therefore, the service delay if the task is executed locally will be:

$$D_l = \frac{1}{\mu\zeta} + \frac{\lambda}{\mu\zeta\left(\mu\zeta-\lambda\right)} \tag{2}$$

We will formulate the service delay for the offloaded task in the next section. Next, suppose the task is computationally intensive and cannot be executed locally or local execution time is very large from what would have been if the task was executed locally, that is:

$$D_l >> D_v^{''} + \Delta + \delta \tag{3}$$

Therefore, the end-node must offload the task to the MEC server to speed up the processing. Now the questions arise how much traffic load should each radio get, and how to schedule the traffic among the radios to avoid packet re-ordering delay at the receiver's end. Therefore, once the system decides to offload the task, the goal is to minimize the delay while keeping in view these considerations. We would also like to mention that following the general notations trend, μ will be used as the packet length and λ will be used as data load.

3.1.2 Delay When Computation Offloaded

In this section, we provide the mathematical model for computation offloading and formulate the objective function for our proposed continuous non-linear program (CNLP).

When computation offloading is decided, other than the processing delay and queuing delay mentioned above, we will have transmission delay and slot-synchronization delay for wireless network data transmission. The data in wireless networks is governed mainly by four different types of delays namely queuing delay, slot synchronization delay, transmission delay and propagation delay [52]. When a data packet arrives at certain node, it is kept in the queue before it gets its turn for processing or transmission. This is the time the packet spends in routing queues and is called as queuing delay. Queuing delay depends upon the capacity of the transmitter and packet arrival rate. Denote μ as the packet length, ζ_u as the capacity of the link (u) and λ_u as the load on link (u): the average queuing delay for a single link (u) can be obtained as $\lambda_u/(\mu\zeta_u)(\mu\zeta_u-\lambda_u)$

[53]. Similarly, assuming a time division multiple access (TDMA) based transmission where synchronization among the nodes is important, slot-synchronization delay will incur when the node has to synchronize its operation with the neighboring wireless nodes. The packet will wait for getting its designated time-slot before it is transmitted. Average slot-synchronization delay can be obtained as $1/2\mu\zeta_u$ [52]

After getting its designated slot, the packet is transmitted into the link. The associated delay is given by $1/\mu\zeta_u$. Finally, time taken by the signals to propagate from source to destination is referred to as propagation delay, which depends upon the propagation distance of the signal [54]. We can see that these delays will keep adding as the packet traverse relay nodes. Combining the four quantities, we get packet delay of the link (u) as follows:

$$d_u = \frac{1}{2\mu\zeta_u} + \frac{1}{\mu\zeta_u - \lambda_u} + \Theta \tag{4}$$

Here Θ is the propagation delay. Let the total number of hops from source to destination be m; for any arbitrarily chosen radio access technology r, our goal is to minimize the following:

$$D_r = \sum_{u=1}^{m} \left(\frac{1}{2\mu\zeta_u} + \frac{1}{\mu\zeta_u - \lambda_u} \right) + \sum_{u=1}^{m} \Theta_u \tag{5}$$

Where ζ_u and λ_u respectively are the capacity and load of link u. For the same packet size, Equation (5) shows that delay is a function of the capacity of the link and the load. Therefore, in order to minimize the delay, we must optimize the load on the radio access technologies in order to optimally utilize the obtained capacities. Since the propagation delay is independent of capacity and load, and only dependent on distance, its value is added at the end of the computation.

3.1.3 Continuous Non-Linear Program Formulation

We assume the source node is mobile, and its transmission capacity is driven by its SNR which is primarily a function of its distance from the relay node. Assuming both WiFi and cellular networks are equipped with scheduled access [50], Non-orthogonal multiple access (NOMA) and beamforming capabilities [9],[55], both the technologies have interference suppression. Therefore, we ignore the interference and take the SNR obtained as a result of distance and operating environment only, for computing the transmission capacity of both the RATs. Moreover, we develop a technique to handle any change in channel condition due to SNR or other factors in Section 3.3. Furthermore, for computation purpose, instead of relying on transmission capacity, we take a more practical approach

by considering the number of bits per second received successfully which is essentially synonymous to system throughput. If BER_u is the Bit error rate and R_u is the data rate of link u, we can write the capacity as follows.

$$\zeta_u = (1 - BER_u)^l \times R_u \tag{6}$$

Where l is the packet length. The bit error rate (BER_u) of the channel taken here is after applying a low-density parity check ($LDPC$) correction code. To optimally utilize the available capacity, the load can be contrived in such a way that makes maximum use of the available capacity as disproportionate or equally shared load will lead to under-utilization of capacity on many levels. To begin with, the radio with larger capacity, having transmitted its load, would sit idle until all other radios are done with their transmission before next transmission session begins. This is the first case of capacity under-utilization where the idle time of the faster radio could have been used for more data transmission. Here it may be noted that with equally distributed load, the channel and the time-slot of the faster radio, once done with transmission of its load share, can be employed by other nodes in the network but the radio of the UE remains idle despite the fact that there exist data load which the UE has allocated to other radios.

Moreover, the MEC server cannot take action on the transmitted data as it is waiting to receive the remaining data. Therefore, the processing at the MEC server is hampered by the slow RATs. Additionally, if the order of the packets at the receiver is different from the order of the same packets at the sender, the processing will be further hindered by packet reordering. In case of out-of-order reception, packets are cached in the receiver's buffer and reordered according to the sender's sequence number. Consequently, the transmission window is reduced, as these losses are attributed to unfavourable channel conditions. As a consequence, the sender drops the transmission rate e.g. using a lower order modulation, in order to make up for the change in channel condition [56].

As a result, we see a sharp decline in the system throughput. The decrease in transmission rate as a result of reduction in transmission is clearly under-utilization of the available capacity. This situation can be made up for if the delays of all the radio access technologies are equal. Therefore, the first objective of our system is to make the delays of all the RATs equal, that is;

$$D_r = D_t = D_v \tag{7}$$

In (7) D_r, D_t and D_v are the delays of the radio r, t and v. For (7) to hold, it is necessary for the participating radios to always assume some load during the transmission.

$$h_r > 0 \; ; \quad \forall \; \lambda > 0 \tag{8}$$

Here h_r is load share ratio of an arbitrarily chosen radio r and λ is the total load. It follows that sum of load share ratios of all three RATs cannot exceed 1, that is;

$$\sum_{r=0}^{v} h_r = 1 \tag{9}$$

(9) ensures that sum of loads on individual loads cannot exceed total incoming load, that is;

$$\sum_{r=1}^{v} \lambda_r = \lambda \tag{10}$$

In (10), λ_r is the load on radio r while λ is total load generated by the device. Moroever, load on the RATs cannot be negative. Therefore, we have to make sure that load share ratios of all the RATs are always positive.

$$h_r \geq 0 \tag{11}$$

Finally, the load on a radio cannot exceed its capacity.

$$\sum_{r=1}^{v} \lambda_r \leq \zeta_r \tag{12}$$

Equation (7)$-$(12) ensure optimal capacity utilization and in-order delivery of packets and destination. Based on the discussion above, we formulate a continuous non-linear program (CNLP) where our objective functions is as follows.

$$\begin{gathered} \text{minimize } D_r \\ \text{s.t. (3), (6)}-\text{(12)} \end{gathered}$$

Where D_r is the delay of arbitrarily chosen radio r. Minimizing delay of one radio will ensure delay of all the RATs are minimized as given in (7).

3.1.4 Proof of Convexity

Considering that the objective function for delay minimization problem is non-linear, we need to verify that any solution we find is a correct global minimum solution. Therefore, in this section we attempt to proof convexity of the delay minimization problem to

confirm that the local optima is also the global optima.

Theorem 1. The delay minimization function given in Equation (5) is a convex function.

Corollary 1. If $f(x)$, where $x \in R$ is a convex function, $f(x) + w$ is also a convex function, where w is a positive real number.
Corollary 2. If $d_u = \left(\frac{1}{2\mu\zeta_u} + \frac{1}{\mu\zeta_u - \lambda_u} + \Theta\right)$ is true for one link u, it is true for all $u = 1$ to n links.

Proof. Following Corollary 1, we ignore propagation delay and draw Hessian matrix for all links, all RATs.

$$H(f) = \begin{pmatrix} \frac{\partial^2 f}{\partial D_1^2} & \frac{\partial^2 f}{\partial D_1 \partial D_2} & \cdots & \frac{\partial^2 f}{\partial D_1 \partial D_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial D_n \partial D_1} & \frac{\partial^2 f}{\partial D_n \partial D_2} & \cdots & \frac{\partial^2 f}{\partial D_n^2} \end{pmatrix} \tag{13}$$

To make the computation simple, let us draw Hessian matrix for single link only, without the loss of generality, as allowed by Corollary 2. The resultant matrix is given in (14). We began with calculating Eigenvalues of the matrix and then putting the smallest and largest possible values for all the variable in the Eigenvalues. The results were positive for both the minimum and maximum values, indicating the function being convex. However, we duly prove its convexity through principle minor technique. The resultant matrix given in (14) is a 3×3 matrix which implies that there will be three orders of principal minors where first order leading principal minor P_1 is obtained by deleting the last two rows and columns of the matrix, that is;

$$H(d_u) = \begin{pmatrix} \frac{1}{\mu^3\zeta} + \frac{2\zeta^2}{(\mu\zeta-\lambda)^3} & \frac{1}{2\mu^2\zeta^2} + \frac{2\mu\zeta}{(\mu\zeta-\lambda)^3} - \frac{1}{(\mu\zeta-\lambda)^2} & -\frac{2\zeta}{(\mu\zeta-\lambda)^3} \\ \frac{1}{2\mu^2\zeta^2} + \frac{2\mu\zeta}{(\mu\zeta-\lambda)^3} - \frac{1}{(\mu\zeta-\lambda)^2} & \frac{2\mu^2}{(\mu\zeta-\lambda)^3} + \frac{1}{\mu\zeta^3} & -\frac{2\mu}{(\mu\zeta-\lambda)^3} \\ -\frac{2\zeta}{(\mu\zeta-\lambda)^3} & -\frac{2\mu}{(\mu\zeta-\lambda)^3} & \frac{2}{(\mu\zeta-\lambda)^3} \end{pmatrix} \tag{14}$$

$$P_1 = \frac{1}{\mu^3\zeta} + \frac{2\zeta^2}{(\mu\zeta - \lambda)^3} \tag{15}$$

Examining (15), we see that none of the terms is negative here. Therefore, the P_1 is greater than 0. Please note that $(\mu\zeta - \lambda)$ is a very large positive number. Similarly, we find second order leading principal minor P_2. P_2 will be the determinant of matrix

obtained by deleting last row and column from $H(d_u)$.

$$P_2 = -\frac{1}{4\mu^4\zeta^4} + \frac{1}{\mu^4\zeta^4} - \frac{4\mu\zeta^2}{(\mu\zeta-\lambda)^6} + \frac{4\mu^2\zeta^2}{(\mu\zeta-\lambda)^6} + \frac{4\mu\zeta}{(\mu\zeta-\lambda)^5} - \frac{1}{(\mu\zeta-\lambda)^4} - \frac{2\mu\zeta}{\mu^2\zeta^2(\mu\zeta-\lambda)^3} + \frac{2}{\mu\zeta(\mu\zeta-\lambda)^3} + \frac{2\zeta^2}{\mu\zeta^3(\mu\zeta-\lambda)^3} + \frac{1}{\mu^2\zeta^2(\mu\zeta-\lambda)^2} \tag{16}$$

Examining 16, there are 10 terms. The result of the first two terms will be positive as second terms is greater than first one. The result of the 3^{rd} and 4^{th} terms is also positive as 4^{th} terms is greater than 3^{rd}. The result of 5^{th} and 6^{th} term will be again positive as 5^{th} term is greater than 6^{th}. Finally, 7^{th} terms is smaller than $8^{th}+9^{th}+10^{th}$. Therefore, the net result of these four terms will be positive which implies that overall P_2 is positive. Next, we move to third order principal minor P_3 which is the determinant of the Hessian Matrix itself and is given by;

$$P_3 = \frac{\frac{3(\lambda-\mu\zeta)^4}{\mu^4\zeta^4} + \frac{4(\lambda-\mu\zeta)^2}{\mu^2\zeta^2} - 4}{2\left(\mu\zeta-\lambda\right)^7} \tag{17}$$

Again, in (17), the only negative term here is 4. However, the first two terms in numerator are larger than 4 due to which net result of numerator will be positive. Therefore, it is safe to say P_3 is also positive. From the net results of P_1, P_2 and P_3, we can say that first, second and third order leading principal minors are all positive. Therefore, we can say that the resultant Hessian matrix of the function is positive definite. From this, we conclude that the delay minimization function is convex.

3.2 Capacity Optimization

In this section, we solve our formulated CNLP to optimize capacity utilization at source node. After that, we optimize the capacity at the relay node where it is shared among multiple receivers connected to it.

3.2.1 Optimizing Capacity Utilization at Source Node

In this section, we develop a solution for the proposed CNLP. We use Lagrange's Multiplier theorem for several constraints. The goal here is to find the optimal loads share λ_i for all the radio access technologies for which the delay is minimum. Using Lagrange Multiplier Theorem, we re-write our problem as follows.

$$G = \left(\frac{1}{2\mu\zeta_u} + \frac{1}{\mu\zeta_u-\lambda_u}\right) - K_1\left(\lambda - \sum_{r=1}^{n}\lambda_r\right) - K_2\left(C_1\right) - K_3\left(C_2\right) - \cdots \tag{18}$$

In (18), K_1, K_2, $\cdots$ are Lagrange multipliers, λ is the total load, λ_r is the load radio r will get and C_1, C_2, $\cdots$ are the constraints defined in (3), (6)$-$(12). Taking partial derivative of (18) with respect to every variable and equalling to 0, we get;

$$\frac{2}{\mu\zeta_r-\lambda_r} = -\frac{1}{\mu\zeta_r}\left(\frac{-\frac{2}{\zeta_r}+\frac{1}{\zeta_t}+\frac{1}{\zeta_v}}{2\mu} + \frac{1}{\mu\zeta_t-\lambda_t} + \frac{1}{\mu\zeta_v-\lambda+\lambda_r+\lambda_t}\right) \tag{19}$$

From (7), we have;

$$\frac{1}{2\mu\zeta_r} + \frac{1}{\mu\zeta_r-\lambda_r} = \frac{1}{2\mu\zeta_t} + \frac{1}{\mu\zeta_t-\lambda_t} \tag{20}$$

Solving (20) for λ_t, we get;

$$\lambda_t = \frac{\mu\zeta_t\left(\lambda_r\left(3\zeta_r-\zeta_t\right)-3\mu\zeta_r\left(\zeta_r-\zeta_t\right)\right)}{\lambda_r\left(\zeta_r-\zeta_t\right)-\mu\zeta_r\left(\zeta_r-3\zeta_t\right)} \tag{21}$$

Finally, λ_v for radio v can be obtained by subtracting λ_r and λ_t from total load that is, $\lambda_v = \lambda - \lambda_r - \lambda_t$. Here it may be worth mentioning that channel condition is time varying. Change in physical conditions lead to change in channel quality. In such a case, these obtained loads do not hold any longer. Therefore, we have to incorporate the change in channel condition to the obtained load. The procedure to incorporate such a change is given in Section 3.3. Moreoever, the procedure to assign the load shares to radios is given in the following sub-section.

Next, we prove λ_r, λ_t and λ_t to be optimal loads that utilize available capacity optimally.

Theorem 2. λ_r, λ_t and λ_t are the optimal load shares.

Proof. Using proof by contradiction, let us assume λ_r, λ_t and λ_v are not optimum and instead x, y and z are the optimal load shares. Therefore, we attempt to optimize these quantities by extending Nash Bargaining theorem [57] to three players as follows;

$$\text{maximize } J = \left(\lambda_r - x\right)\left(\lambda_t - y\right)\left(\lambda_v - z\right) \tag{22}$$

Taking $\frac{\partial J}{\partial \lambda_i}$ with respect to $\lambda_i = \lambda_r$, λ_t and λ_v and equalling to 0, we get;

$$0 = \lambda_t\lambda_v - z\lambda_t - y\lambda_v + yz \tag{23}$$

$$0 = \lambda_r\lambda_v - z\lambda_r - x\lambda_v + xz \tag{24}$$

$$0 = \lambda_r\lambda_t - y\lambda_r - x\lambda_t + xy \tag{25}$$

Solving (23), (24) and (25) for λ_r, λ_t and λ_v, the quantities remain unchanged, substantiating the fact that the quantities are optimum. This contradicts our assumption and hence prove the theorem.

3.2.2 Optimizing Capacity Distribution at Relay Nodes

According to our system model, the peer node, the Wi-Fi access point and the 5G base station are serving as relay nodes. Capacity Utilization at relay node is different from that at source node. Unlike source node, capacity at relay node is shared among multiple receivers connected to it. If ζ_t is the total capacity of the relay node and k users are connected to it, mathematically we can write;

$$\zeta_t = \sum_{u=1}^{k} \zeta_u \tag{26}$$

Relay nodes will be a major bottleneck if packets are not relayed smoothly as a result of dwindling capacity. We overcome this situation by optimizing the distribution of the total capacity ζ_t, such that ζ_u for link u is optimal according to the load λ_u on it.

Suppose a packet travels from source s to destination d. let the traffic from source to destination be T_{sd} and the traffic in other direction be T_{ds}. Also, let there be N sources and M destinations in the network. Therefore, total traffic (T) in the network will be;

$$T = \sum_{s}^{N} \sum_{d}^{M} (T_{sd} + T_{ds}) \tag{27}$$

Next, consider two nodes i and j. Let the link between i and j be u and the load on the link u be λ_u. Also, let T_n be traffic load of another node n passing through link u. If there are N nodes in the network whose traffic load passes through link u, total load on link u is given by;

$$\lambda_u = \sum_{n=1}^{N} T_n \tag{28}$$

We know that each link carry a fraction of total traffic load of the network. Assuming number of links from source to destination is essentially the number of hops and if $\bar{n}$ is the average number of hops that data take from source to destination, mathematically we can express the fraction of traffic load per hop as follows;

$$\bar{n} = \frac{\sum_{n=1}^{N} T_n}{\sum_{s}^{N} \sum_{d}^{M} (T_{sd} + T_{ds})} \tag{29}$$

Let τ_{ij} be delay of the link u. We can exploit Little's Law to get system delay (Γ) as follows.

$$\Gamma = \frac{\sum_{n=1}^{N} T_n}{\sum_{s}^{N} \sum_{d}^{M} (T_{sd} + T_{ds})} \cdot \sum^{N} \tau_{ij} \tag{30}$$

Assuming M/M/1 queuing system with capacity ζ_{ij} and Poisson arrival with an average of λ_{ij} packets and average service time of $1/\mu\zeta_{ij}$, τ_{ij} can be obtained as follows [53];

$$\tau_{ij} = \frac{1}{\mu\zeta_{ij} - \lambda_{ij}} \tag{31}$$

Here μ is the average packet length. Using value of τ_{ij} in Equation 30, we get;

$$\Gamma = \frac{\sum_{n=1}^{N} T_n}{\sum_{s}^{N} \sum_{d}^{M} (T_{sd} + T_{ds})} \cdot \sum^{N} \frac{1}{\mu\zeta_{ij} - \lambda_{ij}} \tag{32}$$

Using Equation (27) and (28), and replacing ij with u, we can re-write (32) as follows;

$$\Gamma = \frac{1}{T} \cdot \sum^{n} \frac{\lambda_u}{\mu\zeta_u - \lambda_u} \tag{33}$$

Equation (33) shows the significance of capacity for system delay. In order to minimize the system delay, we must optimize the capacity. We again use Lagrange multiplier theorem [58] and re-write our capacity optimization problem as follows.

$$W = \frac{1}{T} \sum_{u=1}^{k} \frac{\lambda_u}{\mu\zeta_u - \lambda_u} - K \left(\sum_{u=1}^{k} \zeta_u - \zeta_t \right) \tag{34}$$

Here K is the Lagrange multiplier and $(\sum_{u=1}^{k} \zeta_u - \zeta_t)$ is the capacity conservation constraint as shown in (26) . Taking $\frac{\partial W}{\partial \zeta_u}$ and equalling to 0, we get;

$$\zeta_u = \frac{\lambda_u}{\mu} + \frac{\left(\zeta_t - \sum_{u=1}^{k} \frac{\lambda_u}{\mu} \right) \cdot \sqrt{\lambda_u}}{\sum_{k=1}^{n} \sqrt{\lambda_u}} \tag{35}$$

Equation (35) shows the capacity that a link u will get according to its load λ_u. Having obtained optimal load shares and capacity at relay nodes optimized, we briefly highlight our proposed MEC offloading technique in Algorithm 1.

While it is true that service delay of local processing is larger than offloading the tasks to MEC server, the proposed technique will compute end-to-end delay of all the RATs as desccribed in Section 3.1.2. Based on performances of RATs, optimal load shares are computed as shown in Section 3.2.1. Next, we allocate the obtained load to RATs. Load allocation mechanism and the choice of relay nodes is described in Section 3.2.3. Having transmitted the traffic at source node, we make sure that MEC traffic is relayed smoothly at relay node. Therefore, capacity is optimized at relay node as explained in Section 3.2.2.

Algorithm 1 Multi-Radio Traffic Offloading

Input: λ and ζ_r, ζ_t, ζ_u of the three RATs r, t, v
Output: Communication Delay
while *true* **do**
• **Radio Performance Computation**
1. Compute end-to-end performance of every radio using (5).
• **Optimal Capacity Utilization at Source Link**
1. Compute the three load shares using (19) and (20).
2. Assign the obtained load shares to the RATs in such a way that minimizes the delay.
• **Capacity Optimization at Relay Nodes**
1. Determine incoming MEC traffic and its outgoing link.
2. Assign the capacity on its outgoing link according to (35).
end while

Traditionally, macro-cellular technologies employees proportional fair scheduling that is, capacity is allocated according to weight of the traffic load while Wi-Fi employees a throughput-based fairness model that is, capacity is shared in way to give all the nodes equal throughput [59]. Therefore, throughput of macro-cell and Wi-Fi are respectively given by;

$$t_m = \frac{w_i \zeta_i}{\sum_n w_i} \tag{36}$$

$$t_{wf} = \frac{u}{\sum_n \frac{u w_i}{\zeta_i}} \tag{37}$$

Here w_i is the weight of the user i. We, on the contrary argue that capacity must be shared according to (35).

Theorem 3. Capacity distributed at relay node is optimal.

Proof. Here prove that the obtained capacity in (35) on the basis of Lagrange Theorem defined in (34), is optimal. We argue that ζ_u for link u is optimal for certain $\zeta_t = \sum_{u=1}^{n} \zeta_u$.

We used Lagrange Multiplier Theorem [58], on that account let the original function be $f(x, y)$ and let for the sake of simplicity $g(x, y) = \sum_{u=1}^{n} \zeta_u - \zeta_t$ and $g(x, y) = 0$, but $g \neq 0$, without loss of generality $\frac{\partial g}{\partial y} \neq 0$. Writing (34) in its standard form, we get;

$$W = f(x, y, k) = f(x, y) - k(g(x, y)) \tag{38}$$

Where $W = f(x, y, k)$ is the new function obtained as a result of incorporating multiplier k.

Lagrange Multiplier theorem is based on implicit function theorem (IFT). Therefore, by IFT we can assume that there is a function $y = y(x)$ such that $g(x, y(x)) = 0$ which follows that $f(x, y(x))$. Furthermore, using the same theorem, we have;

$$y'(x) = -\frac{g_x}{g_y} \tag{39}$$

Since $f(x, y(x))$ is assumed to be optimal, its derivative has to be 0. Using chain rule we have;

$$f_x + f_y \cdot y'(x) = 0 \tag{40}$$

Equation (40) shows an optimal value. Next, we have to show that this optimal value is equal to the value of original Equation (38).

Using (39), we get;

$$f_x - f_y \cdot \frac{g_x}{g_y} = 0 \tag{41}$$

Let $-k$ denote f_y/g_y;

$$f_y + kg_y = 0 \tag{42}$$

Using (41), we get;

$$f_x + kg_x = 0 \tag{43}$$

Equation (43) shows that gradient of $(f + kg)$ at points defined by constraint is 0. Also, following (40), Equation (43) shows that original value defined by function in (38) is optimal, hence our capacity is optimal.

3.2.3 Assigning Load Shares to RATs

We formulate an integer linear program to assign load shares to the *RATs*. Let delay for the three load shares λ_r, λ_t, λ_u over *RAT* r be $d_{r,r}$, $d_{t,r}$ and $d_{u,r}$, delay for the same load shares over *RAT* t be $d_{r,t}$, $d_{t,t}$ and $d_{u,t}$. Similarly, delay for these load share over *RAT* u be $d_{r,u}$, $d_{t,u}$ and $d_{u,u}$, as shown in Table 3.1.

Before formulating the integer linear program, let us define a binary variables $x_{i,j}$ such that;

$$x_{i,j} = \begin{cases} 1, & \text{if load } i \text{ is assigned to } RAT\ j \\ 0, & \text{otherwise} \end{cases}$$

Similarly for load share λ_r over *RAT* u, the assignment variable will be $x_{r,u}$ and its value will be 0 or 1 depending upon whether or not λ_r is assigned to u. Next, we formulate our integer linear program as follows;

$$minimize \begin{pmatrix} d_{r,r} \cdot x_{r,r} + d_{v,r} \cdot x_{r,t} + d_{t,r} \cdot x_{r,v} + \\ d_{r,t} \cdot x_{t,r} + d_{t,t} \cdot x_{t,t} + d_{v,t} \cdot x_{t,v} + \\ d_{r,v} \cdot x_{v,r} + d_{t,v} \cdot x_{v,t} + d_{v,v} \cdot x_{v,v} \end{pmatrix}$$

Subject to

$$x_{r,r} + x_{r,t} + x_{r,v} = 1 \tag{44}$$

$$x_{t,r} + x_{t,t} + x_{t,v} = 1 \tag{45}$$

$$x_{v,r} + x_{v,t} + x_{v,v} = 1 \tag{46}$$

$$x_{r,r} + x_{t,r} + x_{v,r} = 1 \tag{47}$$

$$x_{r,t} + x_{t,t} + x_{v,t} = 1 \tag{48}$$

Table 3.1: Delay of the obtained load shares on different RATs.

	RAT_r	RAT_t	RAT_u
λ_r	$d_{r,r}$	$d_{v,r}$	$d_{t,r}$
λ_t	$d_{r,t}$	$d_{t,t}$	$d_{v,t}$
λ_v	$d_{r,v}$	$d_{t,v}$	$d_{v,v}$

$$x_{r,v} + x_{t,v} + x_{v,v} = 1 \tag{49}$$

$$x_{i,j} \geq 0 \tag{50}$$

The objective functions says to minimize total delay of the three *RATs* when loads are assigned to them. The decision variable 0s indicate that if a load is not assigned, its value will be zero that is, the corresponding load will not go to the radio where the value is 0. The constraints (44) - (49) show that load are assigned to one single radio only and one radio will get one share of load only. No two loads can go to a single radio conversely, no radio can be assigned more than one load share. Finally, (50) is the positivity constraint. We solved the integer linear program with simplex method.

3.3 Managing Channel Variation

In Section 3.1.2, we computed radio performance in terms of delay from source node to destination and in Section 3.2, we showed optimal load distribution according to the obtained performance. However, as a result of change in channel condition, performance estimates obtained may become soon outdated and as a result, the load distribution and capacity optimization effectuated may not hold and the constraints may be violated. Confronted with such a situation, we have to allocate the load in such a way that the impact of the change in the performance is averted. Furthermore, we have to identify how frequent the radio performance must be updated in order to reap the correct performance.

3.3.1 Frequency of Radio Performance Update

Given the temporal variation in a wireless channel, it is important to identify a suitable interval and frequency of radio performance update that is, how frequent should the radio performance be updated to get the optimal performance? With larger interval between two consecutive performance updates, there is a possibility of decreasing performance due to stale information. Likewise, smaller intervals will result in sacrificing the bandwidth for network updates and making the task cumbersome. Performance of a radio is subject to

user mobility and network load in addition to the small scale channel fading. Optimizing radio update interval with respect to both instantaneous position and network load simultaneously is NP-hard and beyond the scope of this thesis. However, we strive to find a suitable interval that satisfies the performance of the network.

In this research, we propose a dynamic performance update interval. The interval between two consecutive performance updates varies according to the variation in the performance and will keep increasing as long as the variation (increase as well as decrease) in the performance is within the acceptable limit taken as a threshold. Moreover, for data transmission at a particular instance, performance of a radio is estimated by taking the weighted moving average (WMA) of performance of last n seconds from the time the performance was last updated. Whenever the variation in the performance of a radio is greater than a certain threshold, the current data is transmitted using technique shown in Section 3.3.2 and performance is updated immediately.

3.3.2 Managing Change in Radio Performance

Suppose there is a change in the performance of the radio access technologies. Such a situation will lead to a violation of the constraints defined above unless the performance of RATs are updated. For example, with the change in channel condition, the delay of RATs will be different, thus violating the constraint defined in (7). Here we attempt to temporarily reinstate the constraint before the performance of the RATs are updated in the next interval. This is carried out by adding certain amount of load to the faster RATs. Adding load will increase the delay of the faster RATs, thereby bringing them at par with slower RATs. This process is performed in three steps. In the first step, we determine how fast the faster RATS are with respect to the slowest RAT. Denote D_v as the delay of the fastest RAT, followed by D_t and D_r being the slowest among all the three. With this information given, the following holds true.

$$D_r = D_t - A = D_v - B \tag{51}$$

Assuming D_r is the slowest RAT, D_t is faster than D_r by A μs (hence, A μs subtracted from it) while D_v is faster than D_r by B μs. With simple manipulation of Equation (7), A and B can be obtained as follows.

$$A = \frac{1}{2\mu}\left(\frac{\zeta_t - \zeta_r}{\zeta_r \zeta_t}\right) + \left(\frac{\mu\zeta_t - \lambda_t - \mu\zeta_r + \lambda_r}{(\mu\zeta_t - \lambda_t)(\mu\zeta_r - \lambda_r)}\right) \tag{52}$$

$$B = \frac{1}{2\mu}\left(\frac{\zeta_v - \zeta_r}{\zeta_r \zeta_v}\right) + \left(\frac{\mu\zeta_v - \lambda_v - \mu\zeta_r + \lambda_r}{(\mu\zeta_v - \lambda_v)(\mu\zeta_r - \lambda_r)}\right) \tag{53}$$

In the second step, using (52) we derive Equation (54) to get the equivalent load of A μs that is λ_A;

$$\lambda_A = \frac{3\mu\zeta_A - 2A\mu^2\zeta_A^2}{1 - 2A\mu\zeta_A} \tag{54}$$

Here ζ_A is the capacity of the radio access technology used with A which happens to be t as per Equation (51). Similarly, we can derive expression for λ_B using Equation (53) or simply by replacing A with B and capacity of radio t with capacity of radio v in Equation (54).

Assuming traffic load is continuously being generated by the user, in the third step, we add λ_A and λ_B amount of load to their respective RATs, t and v respectively in this case, to make the delays equal. Adding loads λ_A and λ_B to radio t and v respectively, Equation (51) will become;

$$\begin{aligned} D_r &= D_t + \left\{\sum_{u=1}^{m}\left(\frac{1}{2\mu\zeta_u} + \frac{1}{\mu\zeta_u - \lambda_A}\right) + \sum_{u=1}^{m}\Theta_u\right\} \\ &= D_v + \left\{\sum_{u=1}^{m}\left(\frac{1}{2\mu\zeta_u} + \frac{1}{\mu\zeta_u - \lambda_B}\right) + \sum_{u=1}^{m}\Theta_u\right\} \end{aligned} \tag{55}$$

The new delays of the three RATs are now equal. The change in radio performance is incorporated.

3.4 Performance Evaluation

In this section, we provide mathematical comparative analysis results to show the performance of our proposed scheme. We call our proposed scheme "Multi-Radio Parallel Offloading (MPO)" and compare our performance with 5G, Wi-Fi, Wi-Fi Direct and schemes that distribute the load on the basis of the tasks such as [43], [45]. For elaboration purpose, we refer to these schemes as atomic load distribution schemes (ALD). We show how different RATs take different loads for their corresponding performance and compare their delay. We then compare the performance of our proposed scheme with Wi-Fi, 5G and ALD. We also consider impact on the services of a newly arrived traffic when the node is busy serving the existing traffic in its queue. Finally, we show data outage probability comparison to verify service-level agreement (SLA) of MPO with Wi-Fi and 5G.

3.4.1 Environment Setting and Parameters

We consider the scenario shown in Fig. 3.2 where an end-user is assumed to be based inside a multi-storey building. A Wi-Fi access point and a peer device is assumed to be

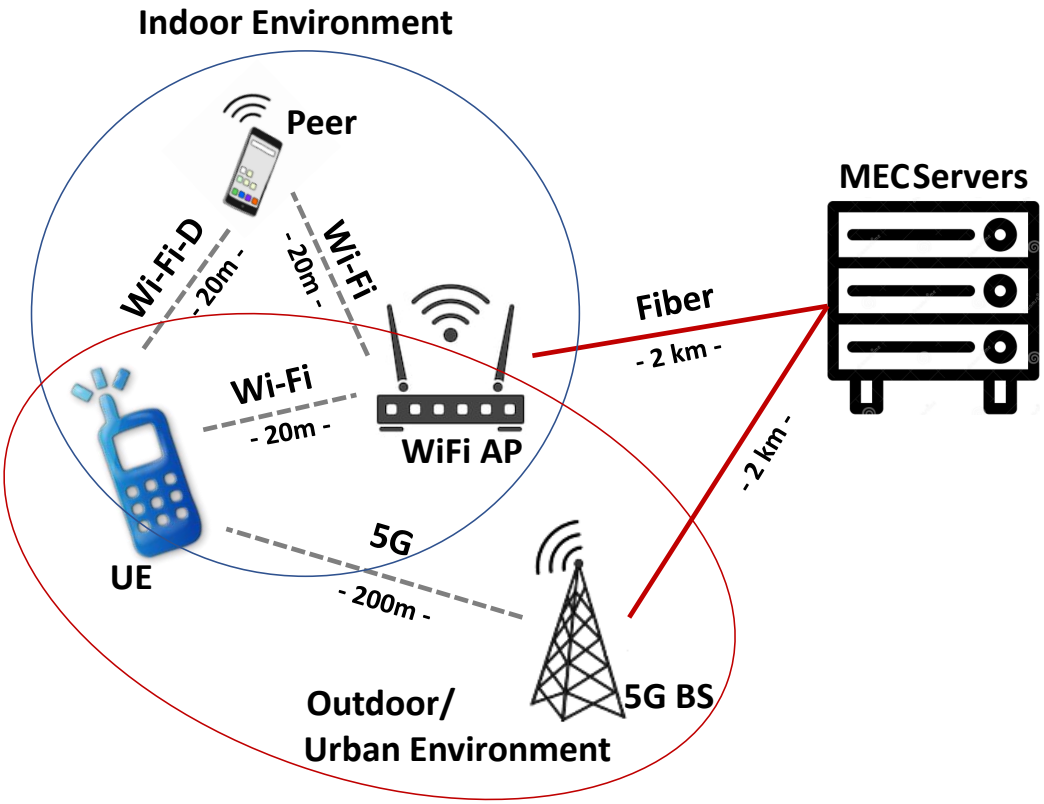

Figure 3.2: Assumed topology where an end-user inside a building is served by a peer device, Wi-Fi access point and 5G macro-cell technology.

inside the building while 5G macro-cell base-station is assumed to be at a distance of 200 m in an urban environment. The end-user is assumed to be simultaneously connected to a peer, Wi-Fi AP and 5G base-station. For Wi-Fi, we have used a frequency band of 5 GHz whereas for 5G, we have used 3.4 GHz band from Frequency Range 1 [60]. Similarly, EIRP for Wi-Fi is 30 dBm and 43 dBm for 5G. Next we describe how to compute different parameters in order to get performance measures of different RATs.

For WiFi, SNR is computed on the basis of indoor path loss model as described in [61]. For ease of reference, we write the path loss formula here.

$$PL_{wi} = PL_{wi}(d_o) + 10\alpha log\left(\frac{d_{wi}}{d_{wi,o}}\right) + \beta d \tag{56}$$

Where PL_{wi} is the indoor path loss for $Wi - Fi$ in dB, α is path-loss exponent, β is specific attenuation, $PL(d_{wi,o})$ is path loss at a reference distance which is taken to be 1 m. The values of both α and β is taken to be 2.

Similarly, for 5G, SNR is computed on the basis of the path loss model given in [62] where macro-cell path loss is divided into two parts, that is outdoor propagation loss and the building penetration loss. The outdoor propagation loss is given by

$$PL_{mo} = 54 + 40logd_{mo} - 30loghb + 21logf \tag{57}$$

Where PL_{mo} is the outdoor path loss for macro-cell in dB, d_{mo} is the distance of user from macro-cell base station in meters, hb is the height of base-station and f is the frequency. The corresponding building penetration loss is given by [62];

$$PL_{mi} = 0.6d_{mi} - 0.6h + 10 \tag{58}$$

Where PL_{mi} is the loss in dB when the signal from the macro-cell base-station penetrates the building, d_{mi} is the indoor distance of the user from the wall, h is the height of the floor. The BER obtained on the basis of computed SNR is considered after LDPC code correction.

Data rate for Wi-Fi $R(w)$ is calculated as follows.

$$R(w) = M \cdot S \cdot R_c \cdot \frac{1}{T_s} \tag{59}$$

Where M is the modulation scheme used, S is the number of subcarriers, R_c is the coding rate and T_s is the symbol duration for Wi-Fi. Similarly, 5G data rate computation

Table 3.2: Parameters Setting.

Technology	Wi-Fi (802.11ax)	5G
Distance	20 m	200 m
Bandwidth	80 MHz	100 MHz
Capacity	SNR Driven	
EIRP	30 dBm	43 dBm
Modulation	SNR Driven	
Code Rate	SNR Driven	
Frequency	5 GHz	3.4 GHz (FR-1)
α	2	-
β	2	-
Height of 5G Base Station	-	45 m
Height of Floor	-	10 m
Aggregated Carrier	1	1
Number of Streams	1	1
5G Numerology	-	1

is based on 3GPP TS 38.306 standard [51] and is given by (60);

$$R(m) = 10^{-6} \cdot \sum_{j=1}^{J} \left(v_L^{(j)} \cdot Q_m^{(j)} \cdot f^{(j)} \cdot R_{max} \cdot \frac{N_{PRB}^{BW(j),\psi} \cdot 12}{T_s^{\psi}} \cdot (1 - OH)^{(j)} \right) \tag{60}$$

Here, J is the aggregated carrier component. In our case, we have not used carrier aggregation, therefore, its value is 1. $v_L^{(j)}$ is number of streams. Again our computation is based on single-input single-output signal, therefore its value is 1. $Q_m^{(j)}$ is the modulation scheme used, $f^{(j)}$ is the scaling factor whose value we have taken to be 1. R_{max} is coding rate, ψ is the numerology which defines the guard interval. Its value is 0 to 4 that corresponds to 15kHz, 30kHz and so on up to 120kHz, respectively. We are using a bandwidth of 100 MHz for which the recommended guard interval is 30kHz therefore, its value will be 1 as 0 is not supported for 100 MHz according to the standard. T_s^{ψ} is the average symbol duration and is given by $\frac{10^{-3}}{14 \cdot 2^{\psi}}$. The data rates obtained here are fed to Equation (6) to get the capacities of the RATs. Based on the obtained capacities, loads distribution is carried out as described in Section 3.2.1.

The parameters used in computations are summarized in Table 3.2.

3.4.2 Results

All the results shown here are mathematically computed using Mathematica software.We begin with load distribution and system delay analysis of the proposed scheme where system delay is essentially network-wide packet delay. Fig. 3.3 shows the load each radio will get for different load generated by the end-user. 5G, for having higher bandwidth,

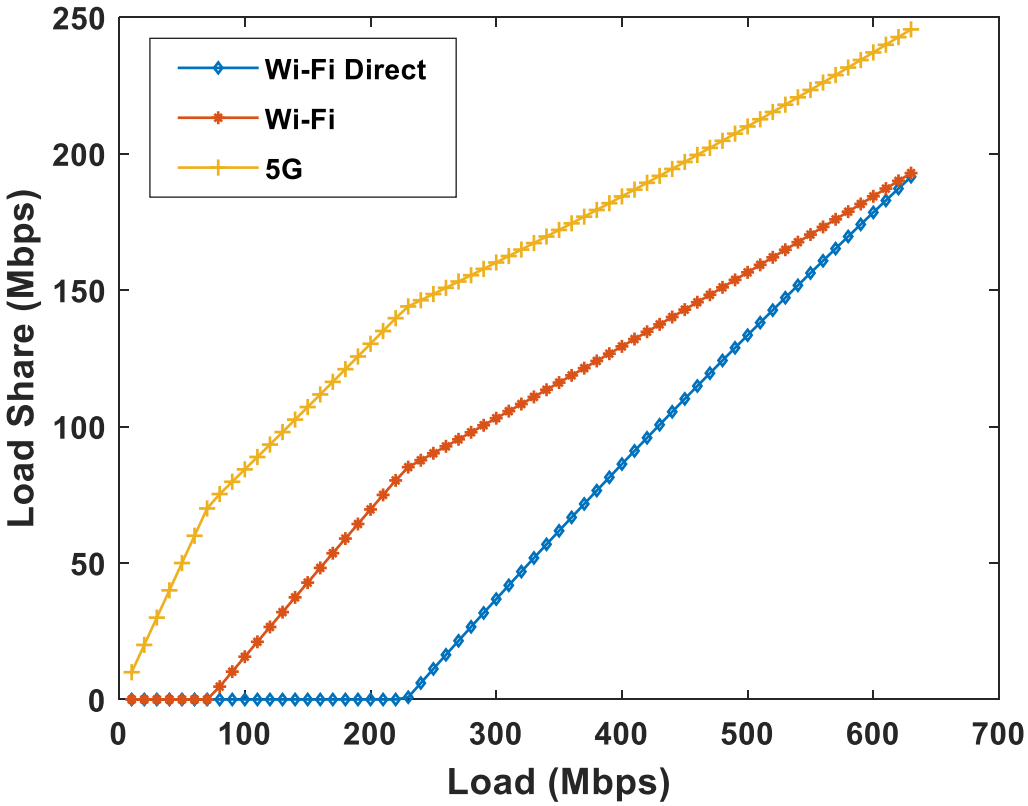

Figure 3.3: Load shares assumed by different RATs as a result of increase in the incoming load.

has highest capacity among all the RATs and as a result has least delay as per (5), therefore the load share taken by 5G is the highest. On the other hand, the increase in load share with the increase in generated traffic for Wi-Fi Direct is highest. This is because the more the load is taken by a RAT, the sooner it will reach its saturation point. Therefore, to avoid saturation, more traffic is transferred to the radio that has the lowest traffic load, which in this case happens to be Wi-Fi Direct.

We then analyze the delay for the corresponding load assumed by these RATs in Fig. 4.1. There are three curves in the figure which appear to be one single curve. The load shares assumed by different RATs are different as shown in Fig. 3.3, their delay, however, is equal. This is very important outcome of our proposal. We argued that packet re-ordering delay in multi-path multi-radio packet routing impedes the throughput significantly and is a major reason of real-time transmission missing the delay deadlines. However, with all the data packet reaching simultaneously, there will be no packet reordering delay. Another important outcome of the proposed scheme is the significantly high data that it can handle. The delay for up to 600 Mb is less than 0.1 ms, after which point it jumps to saturation point.

We also compare delay performance when data is offloaded through Wi-Fi Direct

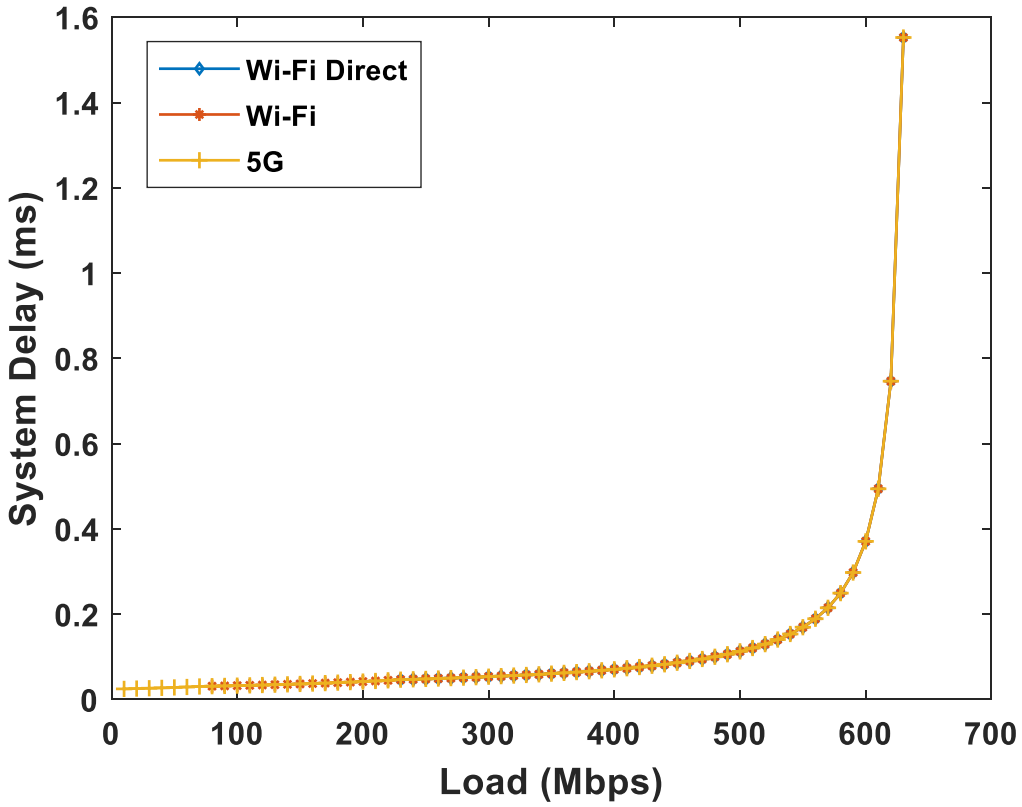

Figure 3.4: Delay for different RATs as a result of increase in the incoming load.

alone, Wi-Fi alone and 5G alone, with MPO. As can be seen in Fig. 3.5, MPO outperforms Wi-Fi and 5G offloading in terms the amount of data that they can carry. Both Wi-Fi Direct and Wi-Fi reach saturation before 200 Mbps, 5G reaches saturation at slightly beyond 200 Mbps whereas MMPO on the contrary, performs well all the way till 600 Mbps and the delay remains less than 0.1 ms for up to 600 Mbps. This is a gain of about 70% as compared to Wi-Fi Direct and Wi-Fi, and about 63% as compared to 5G.

Next, assuming MEC traffic is subject to prioritized processing [15] [63], the conventional traffic will be affected. Similarly, if a node's capacity is already heavily used, the services of the new incoming traffic will be impacted. Therefore, here we measure the impact on the quality of services of the newly arrived data when the nodes are processing the existing data in their queues. We thus measure the impact (I) on the new arrived data relative to the prior load on the node.

Let $\zeta_{r,cr}$ be the current capacity of a certain radio r. Suppose the newly arrived normalized load at time t requires the capacity $\zeta_{r,req}$ for time Δt seconds. The impact in terms of degradation in the quality of services of the new arrived data will be;

$$I = \sum_{r=1}^{v} \left(\frac{\int_t^{\Delta t} \zeta_{r,req} - \int_t^{\Delta t} \zeta_{r,cr}}{\int_t^{\Delta t} \zeta_{r,req}} \right) \tag{61}$$

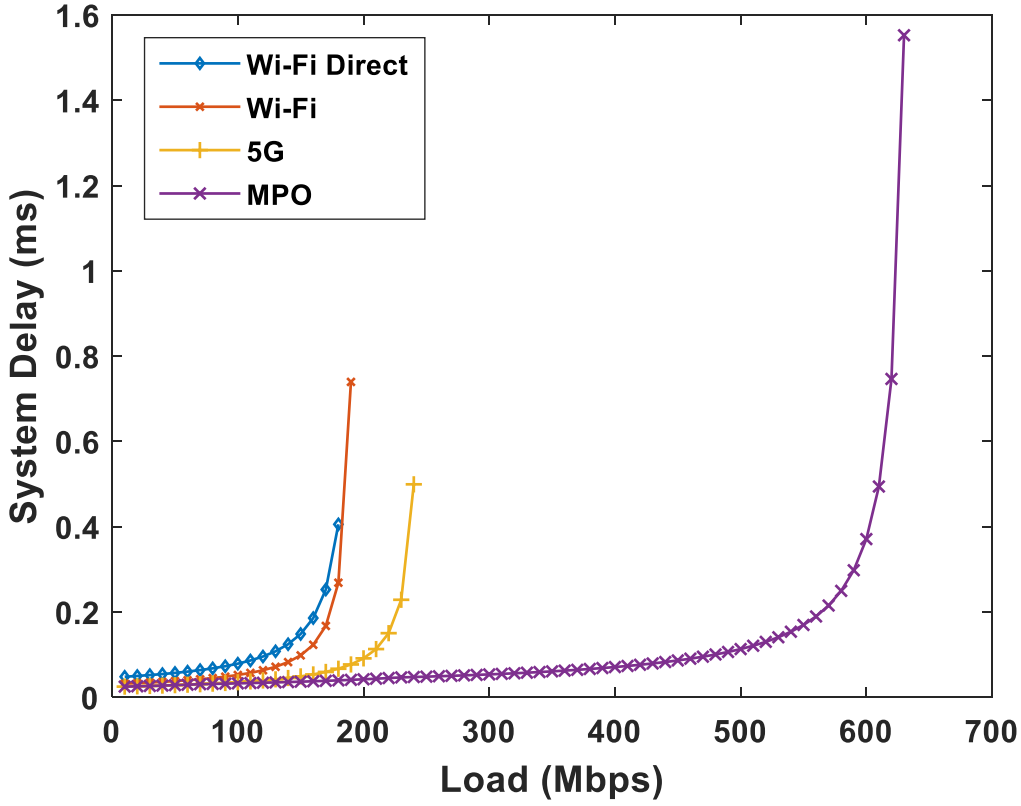

Figure 3.5: Delay comparison of the proposed MPO schemes with when data is offloaded through Wi-Fi alone, Wi-Fi Direct alone and 5G alone.

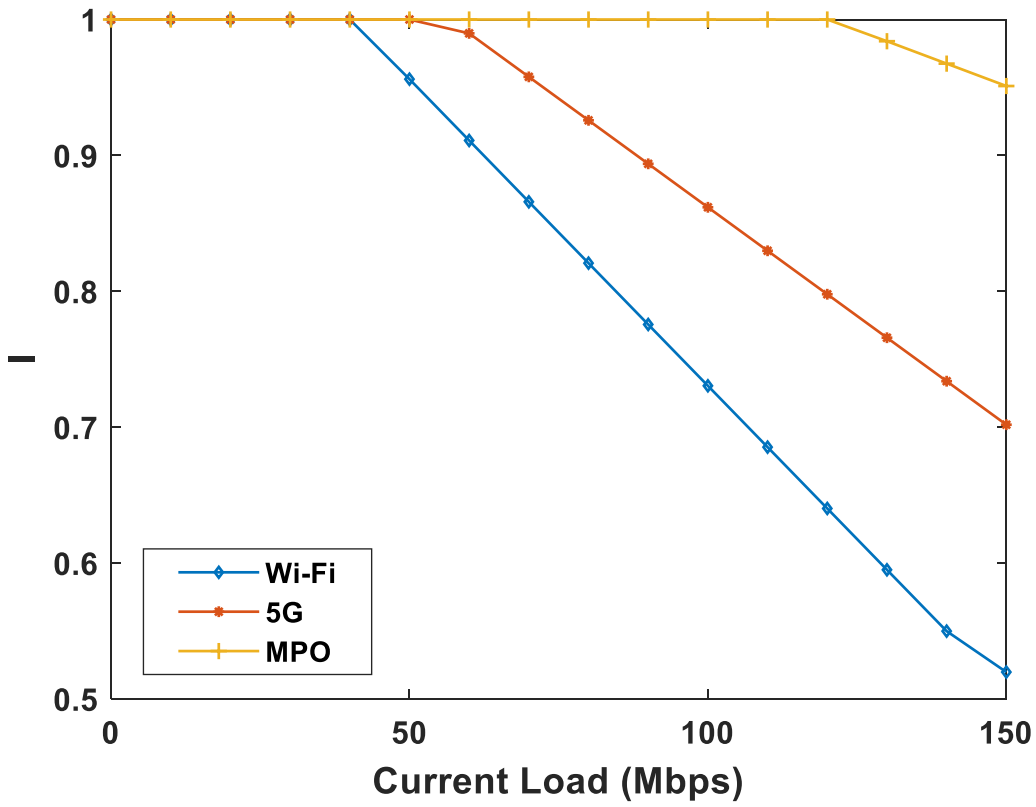

Figure 3.6: Impact on the Services of different RATs.

In Fig. 3.6, we have shown the relative impact on the services of the newly arrived data packet with respect to the data that is already present at the node. The results are for average 150 Mbps of incoming load, for 10 ms whereas x-axis shows the current or existing load at the node. Here 1 means no impact on the service at all and on the other extreme 0 means that new packets will not be serviced at all. Therefore, higher value implies lesser impact on the services. We can see that for the first 50 Mbps, no radio is affected. However, Wi-Fi begins to have impact on the newly arrived data after 50 Mbps of existing load and 5G is showing decline in service after 65 Mbps of existing data. MPO performs steadily till 130 Mbps of load after which there is a decline in the quality of service. It is a gain of 61% compared to Wi-Fi and 50 % compared to 5G. At a load of 150 Mbps, impact on service for Wi-Fi is around 55% which means the new traffic load will be 45% impacted, the impact on the services of 5G is 70% which commensurate to 30% decline the quality of service while MMPO is performing at a rate of more than 90%. There is only a minute impact on the quality of services of the newly arrived data for MPO.

We also compare the packet outage probability of the three schemes. The knowledge of packet outage probability is useful for verifying service-level agreement (SLA) compliance. Packet outage probability is linked to the probability of load getting greater than

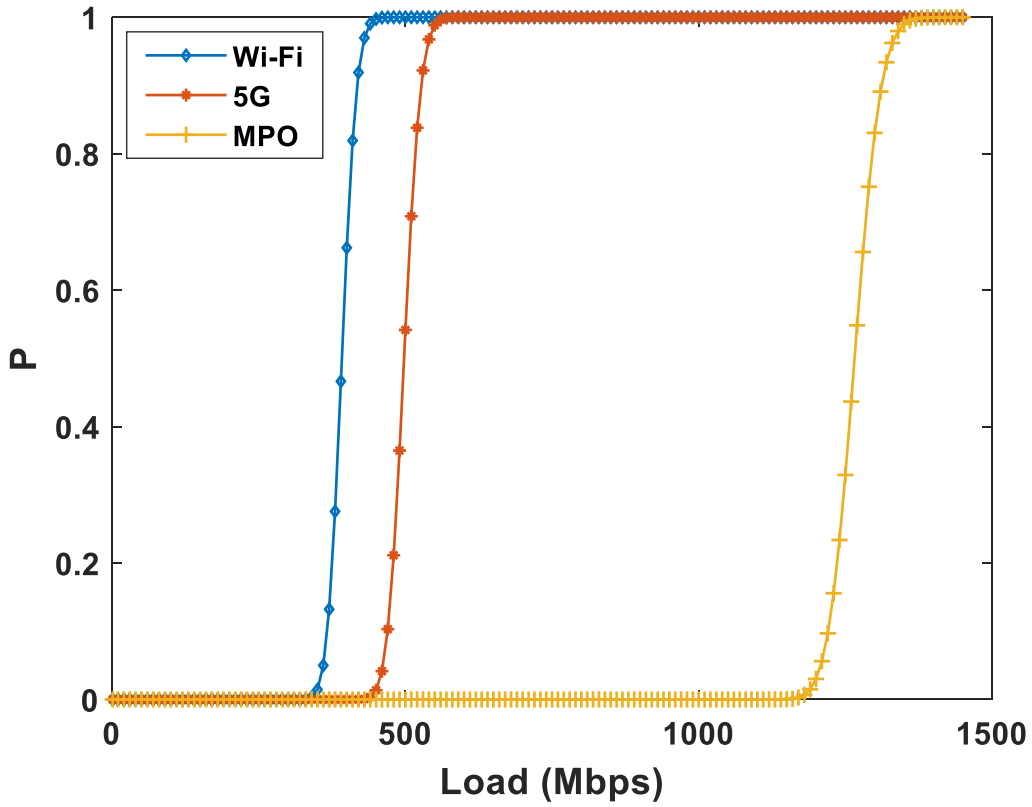

Figure 3.7: Outage probability comparison of different scheme to show SLA conformity

a given threshold, as for load, $\lambda > \zeta$, there will be outages in packets. When outages are greater than certain threshold, the SLA terms will be violated. For Poisson packet arrival, the probability of load getting greater than capacity is given by;

$$P(\lambda > \zeta) = 1 - \left(e^{-\rho} \sum_{i=0}^{k} \frac{\rho^i}{i!} \right) \tag{62}$$

Where $\left(e^{-\rho} \sum_{i=0}^{k} \frac{\rho^i}{i!} \right)$ is the cumulative distribution function (CDF) of the Poisson distribution. In Fig. 3.7, we compare the probability of load getting greater than capacity for Wi-Fi, 5G and MPO to analyze the packet outage probability. The three schemes have full SLA compliance, with no packet loss, until about 350 Mbps. The probability of Wi-Fi tends to fall at this point and reaches 0 at about 450 Mbps. For 5G this probability is impacted after 450 Mbps approximately and reaches 0 at 580 Mbps. Whereas, the MPO has a consistent probability of 1 until 1100 Mbps approximately. That is a gain of 67% as compared to Wi-Fi and 58% gain as compared to 5G.

So far we showed the gain in performance by using multiple radio resources. Next, we compare performance of our proposed MLO scheme with ALD that distributes the load on the basis of the task rather than the load itself. Considering a task size of

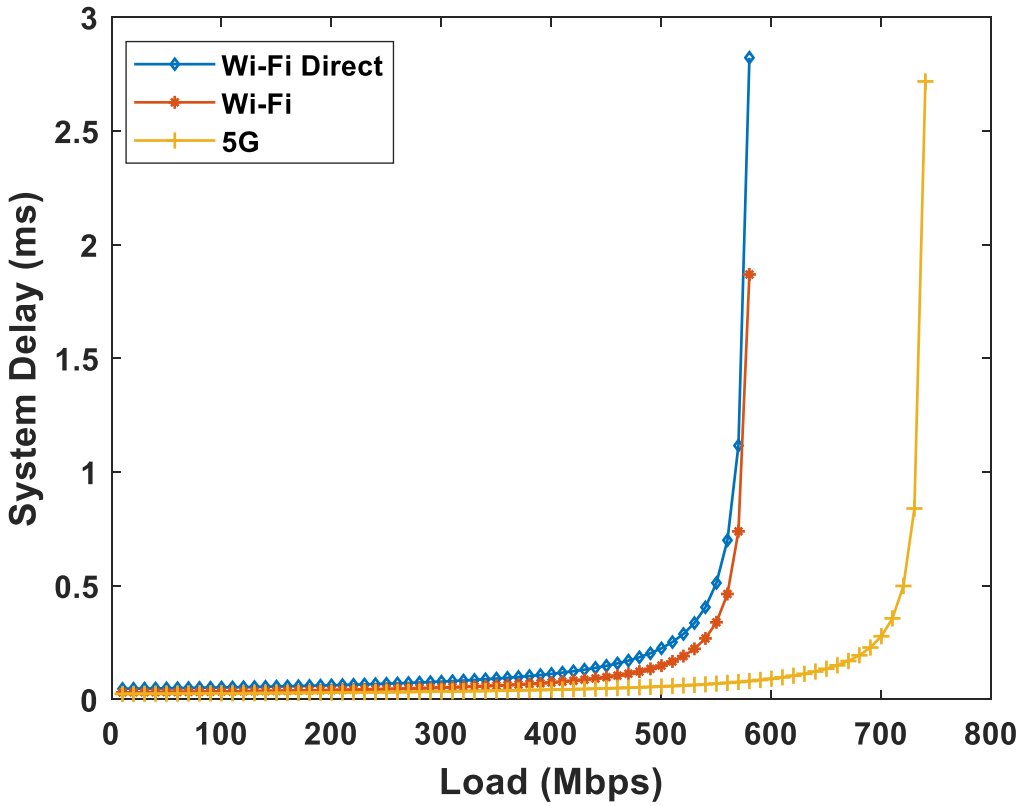

Figure 3.8: System delay of different RATs for ALD

800 KB, the delay of different RATs for ALD is shown in Fig. 3.8. As can be seen, different RATs have different delays. Now this is a major bottleneck as applications depends on the reception of all three tasks in order to provide seamless services to the end users. Therefore, all data packets must arrive at the transport protocol layer in sequence whereas data arrived out-of-order is either kept in buffer or totally discarded depending upon the magnitude of latency of slower RATs.

In Fig. 3.9, we thus compare the system delay of the proposed MPO with ALD. Here, we can see that the performance of ALD is limited by a slower radio whereas the proposed MPO scheme apportion loads according to the performance of the RATs by virtue of which a slower radio receives a lower load and thus its effect on performance are minimised. As can be seen, for the given scenario, the proposed MPO scheme carry approximately 80 Mbps more load in comparison with ALD. ALD is saturated at the offered load of about 550 Mbps whereas MPO can carry a load up to 630 Mbps. Similarly, MPO has consistently lesser system delay in comparison to ALD. The higher system delay of ALD is contributed by higher load share allocated to slower radio which happens to be Wi-Fi Direct in this case.

We also analyze the performance of our proposed capacity optimization technique at relay node. Until now, we have compared ALD and MPO with both schemes having

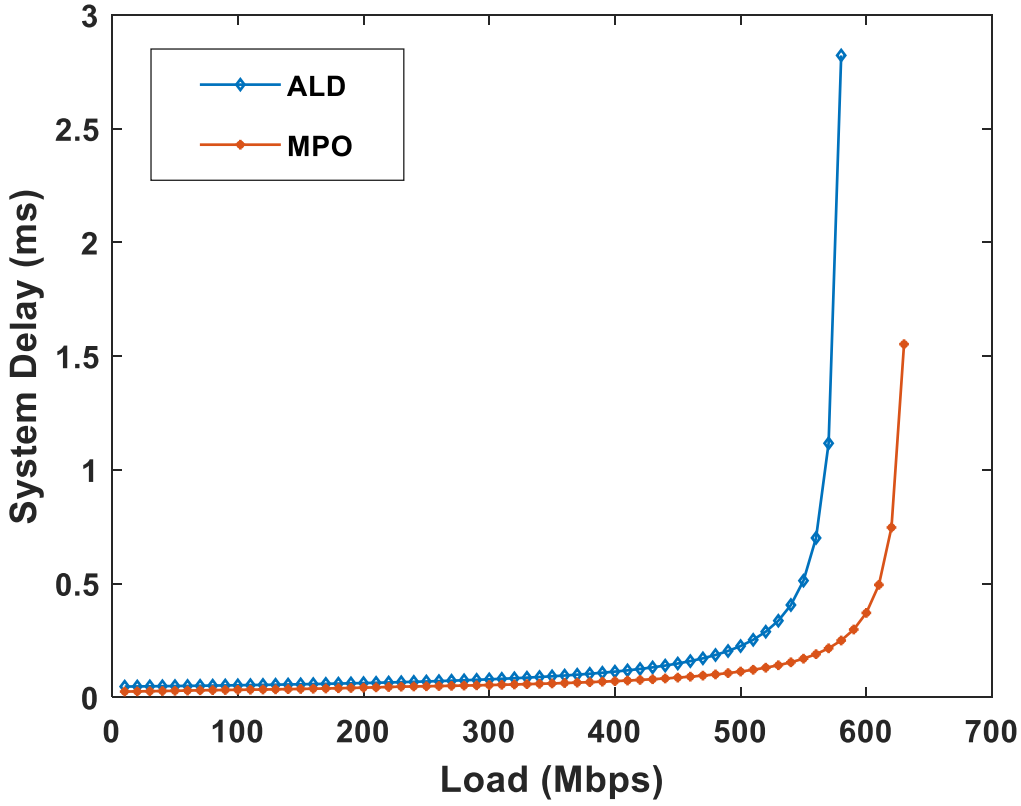

Figure 3.9: System delay comparison of MPO with ALD when capacity distribution at relay is optimized for both the schemes.

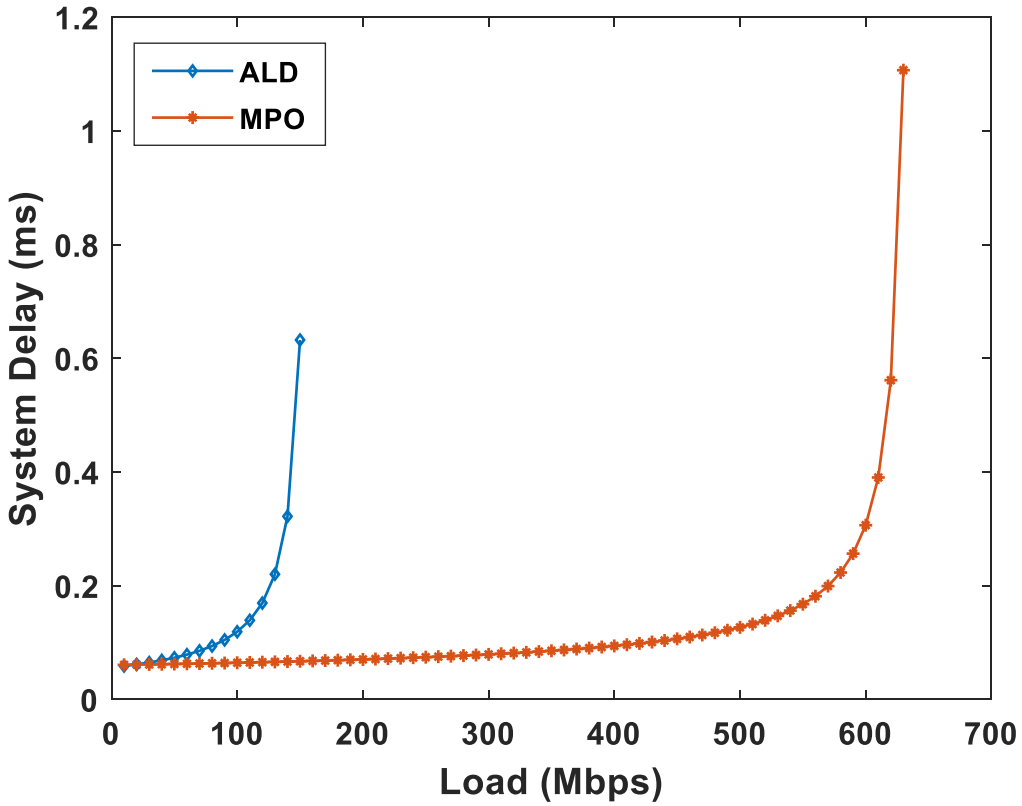

Figure 3.10: System delay comparison of MPO with ALD when capacity distribution at relay is optimized for MPO only.

capacity optimized according to the load, because the goal was to show system performance in terms of load distribution under the same parameters. Here we compare the performance of the proposed optimal capacity distribution scheme against conventional technique where data at relay node is forwarded with even capacity or non-optimal capacity distribution among the links.

In Fig. 3.10 we plot the system delay for 4 users with 40% MEC traffic. Thus, the load on the x-axis indicates 60% conventional and 40% MEC traffic for 4 users operating simultaneously. It is clear from the Fig. 3.10 that proposed MPO with optimized capacity distribution can support nearly 4 times the maximum load that ALD can while giving lower system delay. With the same total capacity, ALD reaches its saturation point at about 150 Mbps while MPO, intelligently distributing the capacity according to the load, maintains a stable delay until 600 Mbps.

The delay and the load shared by RATs against the incoming load is not linear. If there is existing load at the node, the incoming traffic will incur more delay and accordingly the load shares will be different as discussed in Section 3.1. Therefore, we next we show the impact of current or existing delay at node on delay and load share. In Fig. 3.11 and 3.12, we have compared load share and delay for different RATs for

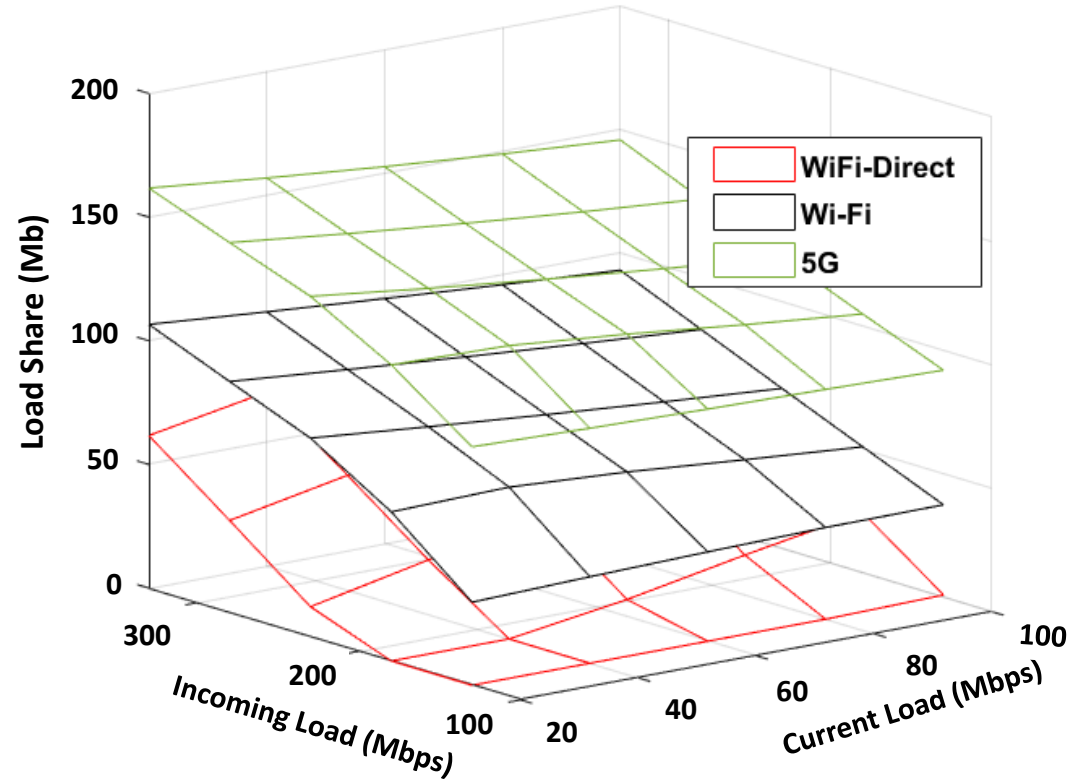

Figure 3.11: Impact of current load on Load Share.

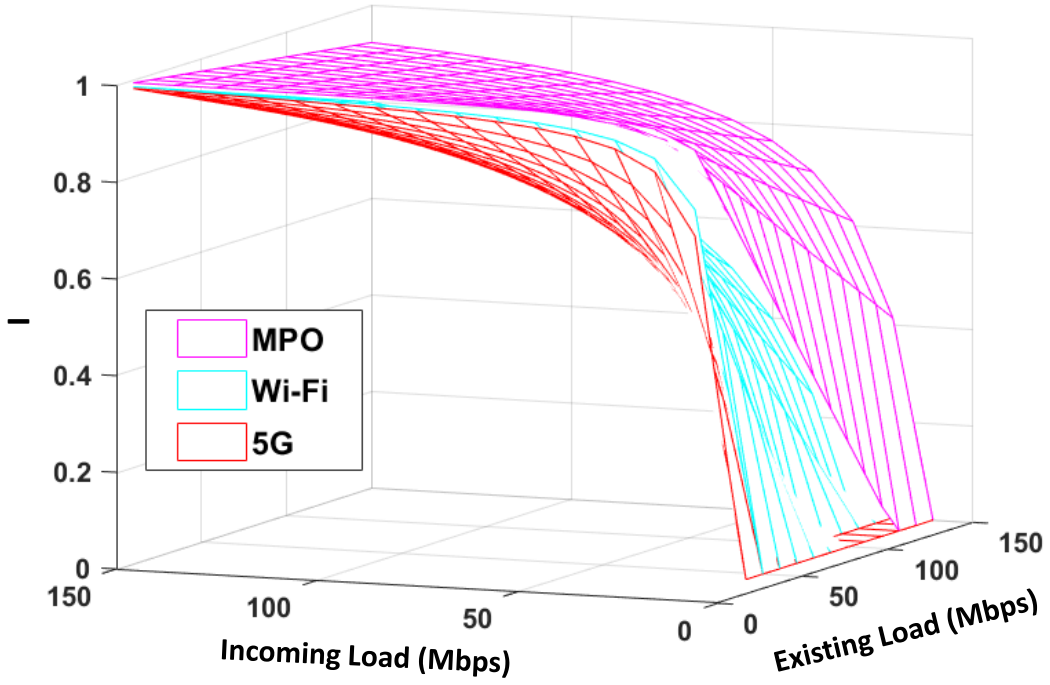

Figure 3.12: Impact of current load on Services on Incoming Data.

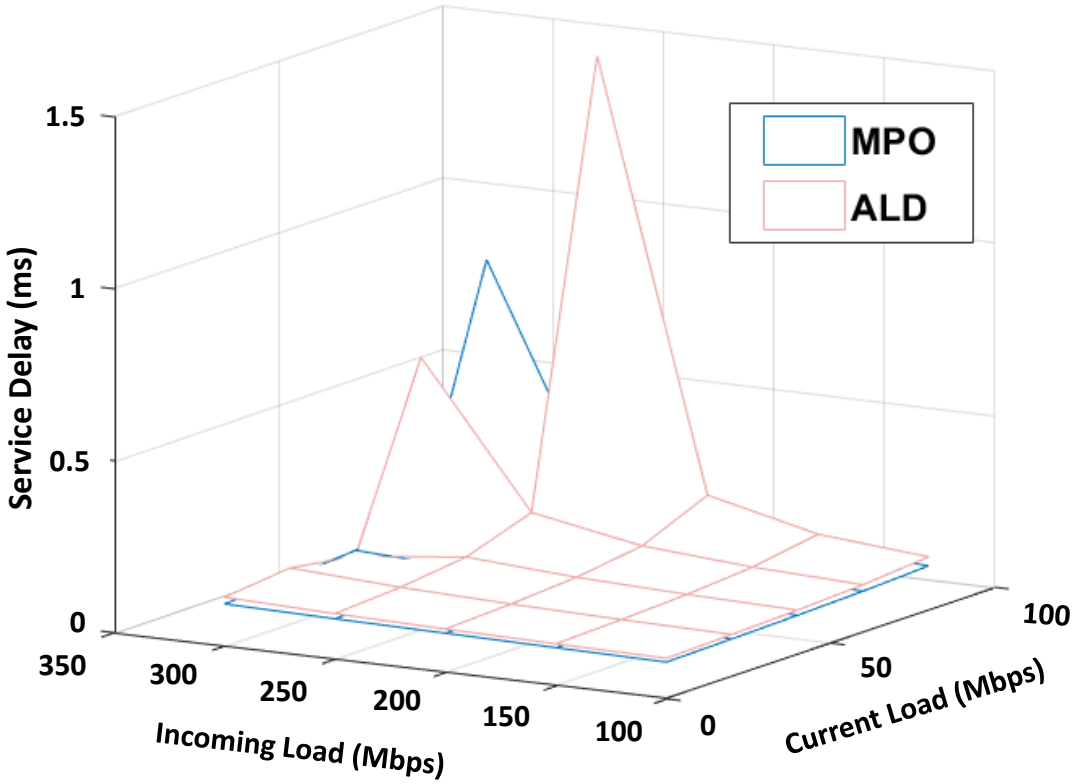

Figure 3.13: Impact of current load on Delay.

MPO. We have also compared the delay of the proposed MPO with ALD in Fig.3.13, that linearly distributes the load on the basis of the tasks. The figure shows that the proposed MPO has consistently lower delay than ALD.

3.5 Summary

We proposed simultaneous offloading over multiple radio access technologies available for a phone. We began with optimally utilizing available capacity at the source node and then optimized the capacities at outgoing links of relay node according to the incoming traffic, so that MEC traffic is relayed smoothly. We developed a non-linear continuous program that takes performance of all the RATs into account and accordingly optimally distribute the traffic load among the RATs in such a way that delay for all the RATS is equal, thereby avoiding the packet re-ordering delay at the destination node. We shrewdly used Lagrange Multiplier Theorem to solve our program and optimize the capacity at the relay nodes. As a proof-of-concept, we showed that to minimize service delay and maximize throughput, QoS and SLA compliance, we must optimize capacity utilization at the source node and capacity distribution on the outgoing links at relay nodes. Furthermore, our illustrative results showed that contemporary techniques for dealing with service delay are not favorable and to get optimal performance, traffic load must be distributed in a way

to avoid re-ordering delay. We believe that simultaneous offloading over multiple RATs will not only improve MEC performance for future applications but is also a plausible mechanism to make up for the debilitated telecom infrastructure in low- and middle-income countries.

4 Multi-RAT Multi-Server

Multiple RATs have been combined to improve throughput and capacity has been carried out in past [46], [47]. In this paper, we provide an analytical model to optimize communication delay by optimization scheduling, and capacity utilization and distribution.

The idea behind MEC is to bring large number of servers close to UEs, so that each server serves a small number of UEs to expedite the processing. However, if large number of users gather around a single MEC Server and overloading it, computation is hampered [1]. Therefore, we need to manage these users and load on the servers. Similarly, on communication side future applications such as extended reality, industry 5.0, smart grid 2.0, holographic telepresence, space and deep-sea tourism keep coming expeditiously. On the requirements side, these applications are characterized with ultra-high data rates, real-time access to powerful computing resources, ultra-low latency, and extremely high reliability and availability surpassing the network capabilities promised by existing infrastructure [2]. Therefore, to minimize service delay, we need to minimize both computation as well as communication related delays. To achieve this objective, we propose multi-server multi-RAT (MSMR) MEC systems. Using MSMR MEC systems, we minimize computation and communication delay by exploiting relevant parameters such as managing server selection, managing load distribution on servers and RATs, optimally utilizing RATs capacity and get rid of unnecessary delays such as packet reordering delay, server migration loops etc.

4.1 System Model

Assumed system model is shown in Figure 1 where we have a user (UE) connected to the network by two RATs simultaneously. The RATs are a Wireless LAN (WLAN) such as IEEE 802.11ax and a macro-cellular network such as 5G or 6G. The two RATs in return are connected to two different MEC Servers via a high-speed optical fiber. Both the technologies and servers are operational simultaneously. The two servers have same basic capacity. Moreover, since 80% of the communication takes place in indoor environment [4], we assume the user is indoor served by WiFi AP and an macro-cell BS from outside.

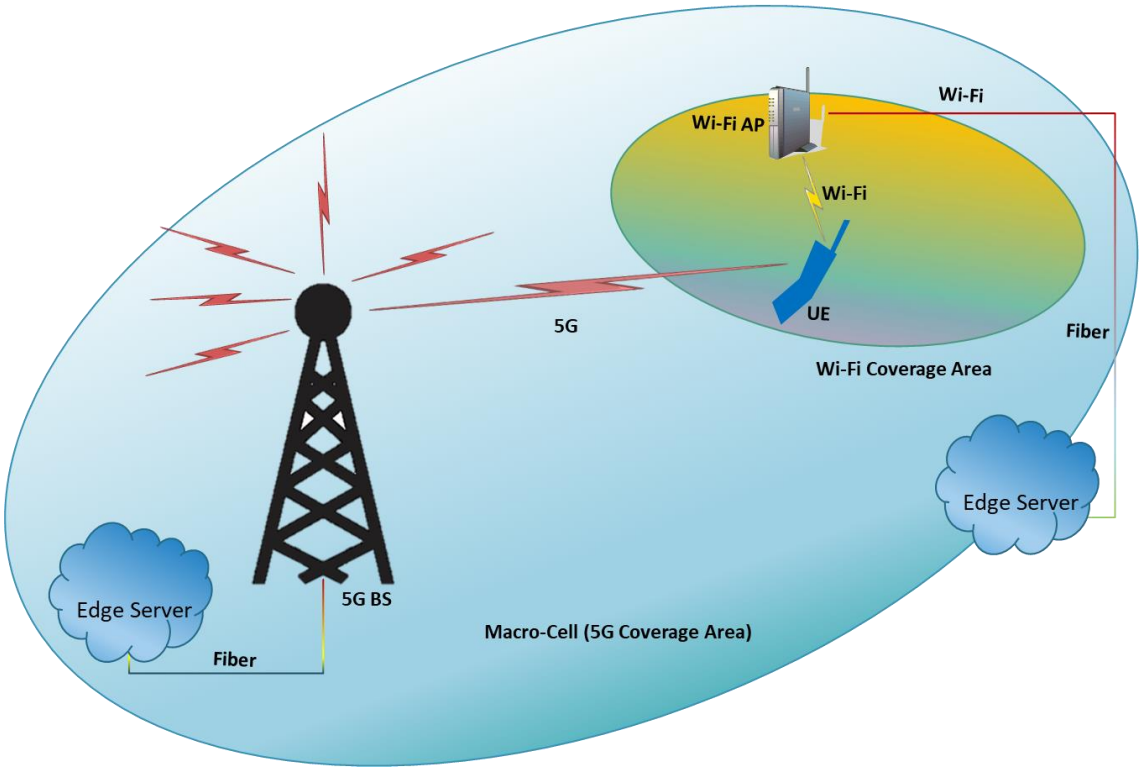

Figure 4.1: A Simple Illustration of Multi-RAT Multi-Server MEC System.

Next we show our computation and communication performance estimation model.

4.2 Computational Delay

Here we assume Multiple servers (c) are available to connect to for every RAT. Tasks arrival are random with rate λ. Assuming exponential distribution, let processing time for server i will be $1/\mu_i$. Server Occupation rate at server i is given by [4].

$$\rho_i = \frac{\lambda}{c_i \mu_i} \tag{1}$$

Where λ is the task arrival rate at server c_i and c is how many processors are there. Therefore, the probability that the task has to wait in the queue is given by:

$$\omega_i = \frac{(c \cdot \rho_i)^c}{c!} \left((1 - \rho_i) \sum_{n=0}^{c-1} \frac{c \cdot \rho_i}{n!} + \frac{c \cdot \rho_i}{c!} \right)^{-1} \tag{2}$$

Now we can calculate queueing delay as follows.

$$D_q = \omega_i \cdot (1 - \rho_i)^{-1} (c \cdot \mu_i)^{-1} \tag{3}$$

Similarly, processing delay is given by;

$$D_c = \omega_i \cdot (1 - \rho_i)^{-1} (\varsigma \cdot \mu_i)^{-1} + \frac{1}{\mu_i} \tag{4}$$

4.3 Communication Delay

We assume the scenario as explained in the previous chapter. Therefore, path losses, data rate and throughput will remain the same as in previous chapter. However, here we separate the uplink and downlink traffic for both WiFi and 5G.

4.3.1 Delay for WiFi Part

Uplink Delay for WiFi Part

The single biggest addition to IEEE 802.11ax is that of Orthogonal Frequency Division Multiple Access (OFDM) where transmission is organized on a per-frame basis. This means that a frame can carry information to and from multiple STAs [8]. In such a frame, physical resource, i.e.,spectrum, is divided into multiple orthogonal subchannels—referred to as a resource unit (RU) in the 802.11ax terminology. The number of RUs assigned to a particular user are driven by equation (4). Moreover, the RUs are distributed on the basis of two criteria that is, Scheduled Access (SA) and Random Access (RA). For associated users, the communication begins by AP transmitting Trigger Frame (TF). Upon receiving a TF, the associated users enter a scheduled access (SA) mode whereby only those clients can transmit or receive frames that are allocated RUs by the AP. This behaviour is in contrast to legacy 802.11 standards that use a contention-based mechanism for channel access. When the AP transmits the TF, users other than those that are assigned RUs will defer their transmissions for an interval specified by the TF's Network Allocation Vector (NAV). The non-associated users send Buffer Status Report (BSR) in the RUs allocated for RA.

Assuming RU_N are the total RU_s, RU_N=RU_{SA}+RU_{RA}, where RU_{SA} is RU_s allocated for scheduled access and RU_{RA} are RU_s allocated for random access. Sending BSR in RU_{RA} follows the legacy IEEE 802.11 technique that is, all the users will contend for access to channel and a user will transmit when its back-off counter reaches 0. Therefore, for BSR transmission in RU_{RA} borrow the model of Bianchi [9] where the author has modelled back-off process by a two-dimensional Markov chain. Based on the model, the probability that a user will transmit its BSR in RU_{RA} given by;

$$\varepsilon = \frac{2\,(1-p)}{(1-2p)\left(\frac{W}{RU_{RA}}\right) + p\frac{W}{RU_{RA}}\,(1-(2p)^m)} \tag{5}$$

Where p denotes probability that a transmitted packet collides. The probability that a transmitted packet results in a collision can be computed as,

$$p = 1 - \left(1 - \frac{\varepsilon}{RU_{RA}}\right)^{n-1} \tag{6}$$

Where n is the number of non-associated users contending for the channel. From the above two equations, now we can compute the probability that at least one user transmits in a considered RU_{RA} during the TF as follows;

$$P_{tr} = 1 - \left(1 - \frac{\varepsilon}{RU_{RA}}\right)^{n} \tag{7}$$

In case of multiple simultaneous transmissions, there will be collision. Therefore, the probability P_s that a transmission in an RU_{RA} successful is given by the probability of exactly one transmission given that there has been a transmission on the considered RU.

$$P_s = \frac{n\frac{\varepsilon}{RU_{RA}}\left(1 - \frac{\varepsilon}{RU_{RA}}\right)^{n-1}}{1 - \left(1 - \frac{\varepsilon}{RU_{RA}}\right)^{n}} \tag{8}$$

Similarly, the probability $RARU_s$ idle that all RA RUs are idle because none of the STAs were able to complete their backoff procedure is given as,

$$P_{idle} = (1 - P_{tr})^{N_{RA}} \tag{9}$$

Now, based on RU allocation, we have following cases.

1. RU_s are divided between RU_{RA} and RU_{SA} — in such a case, there will be scheduled access as well as random access. Therefore, delay will be;

$$D_w^{'} = \left(\frac{(RU_{SA} + RU_{RA}P_{tr}P_s)\cdot L}{T_1 \cdot I \cdot \tau}\right) \tag{10}$$

Where I is the packet size and L is the total load of a user. T_1 is the total time taken by relevant frames. In this case, we have TF, BSR and packet transmission and the corresponding acknowledgements. Therefore, T_1 will be;

$$T_1 = T_H + (T_{TF} + SIFS + T_\delta) + (T_{BSR} + SIFS + T_\delta) +$$

$$(T_{BSR\_ACK} + SIFS + T_\delta) + (T_P + SIFS + T_\delta) + (T_{ACK} + SIFS + T_\delta) \quad (11)$$

Here T_h is the frame headers bit and T_δ is the propagation delay.

2. All RU_s are allocated for SA — implies that there are no non-associated users, and all the RU_s are assigned to RU_{SA} associated users. In such a case, delay will be;

$$D_w^{'} = \left(\frac{N_{SA} \cdot L}{T_5 \cdot I \cdot \tau} \right) \quad (12)$$

Since all the *RUs* are allocated for Scheduled Access, there we will be TF, packet transmission and Its ACK only.

$$T_2 = T_H + (T_{TF} + SIFS + T_\delta) + (T_P + SIFS + T_\delta) + (T_{ACK} + SIFS + T_\delta) \quad (13)$$

3. All RUs are allocated for RA – implies that there are no associated users and all the RU_s are assigned to RU_{SA}. Therefore, delay will be;

$$D_w^{'} = \left(\frac{RU_{RA} P_{tr} P_s \cdot L}{(P_1 T_1 + P_{idle} T_4 + (1 - P_1 - P_{idle})\, T_3) \cdot I \cdot \tau} \right) \quad (14)$$

$P_1 T_1$ indicates the time taken when there is at least one BSR delivered to the AP. $P_{idle} T_4 + (1 - P_1 - P_{idle})\, T_3$ indicates that no BSRs reach the AP due to none of the STAs finishing their respective backoffs where T_4 and T_3 are given by;

$$T_3 = T_H + (T_{TF} + SIFS + T_\delta) + (T_{BSR} + SIFS + T_\delta) \quad (15)$$

$$T_4 = T_H + (T_{TF} + SIFS + T_\delta) \quad (16)$$

Downlink Delay for WiFi Part

The downlink delay is based on pure schedule-based transmissions [10]. In the DL, the AP has a global view of its associated users. The AP can assign parameters related to QoS requirements, fairness etc. while scheduling resources for the downlink. The important point to note is that as long as the AP's transmission queue is full, the downlink

throughput is deterministic. Also, in the downlink direction the entire channel is devoted to transmissions of the AP. Therefore, for downlink $RU_N = RU_{SA}$ and corresponding delay will be equal to;

$$D_w^{'} = \left(\frac{N_{SA}^{"} \cdot L}{T_5 \cdot I \cdot \tau} \right) \tag{17}$$

Where $N_{SA}^{"}$ is the RUs allocated for SA downlink transmission. Since all the *RUs* are allocated for Scheduled Access and there is no TF involved, therefore there will be time taken by packet transmission and its ACKs only.

$$T_5 = T_H + (T_P + SIFS + T_\delta) + (T_{ACK} + SIFS + T_\delta) \tag{18}$$

Backhaul communication, that is communication from the BS to MEC server is carried out through optical fibre connections. In an optical fibre medium, we can assume that bandwidth is abundant enough that the transmission rate is high and propagation delay dominates [11]. Having known the propagation distance, we can calculate propagation delay assuming propagation speed to be 2/3rd of speed of light.

Therefore, total delay on Wi-Fi will be;

$$D_w = D_w^{'} + D_w^{"} + 2\delta \tag{19}$$

Where δ is the backhaul communication the detailed computation of which are given in the paper.

4.3.2 Communication Delay of 5G

Downlink Communication Delay of 5G

For macro-cellular technology, 6G is still at the research phase, therefore, we mostly borrow the techniques defined in 3GPP 5G-NR release-15 standard [12] which defines one-way downlink delay as follows.

$$D_m^{"} = d_{bsp}^{"} + d_q^{"} + d_{fa}^{"} + d_{tti}^{"} + d_{uep}^{"} \tag{20}$$

Where $D_m^{"}$ is the downlink delay from macro-cell BS to UE $d_{bsp}^{"}$ is the BS processing delay, $d_q^{"}$ is the queuing delay, $d_{fa}^{"}$ is the frame alignment delay, $d_{tti}^{"}$ is the transmission delay (Transmission Time Interval) and $d_{uep}^{"}$ is the UE processing delay.

Uplink Communication Delay of 5G

For uplink, 5G NR defines two types of scheduling that is Dynamic Grant (DG) and Grant-Free (GF) Scheduling. In GF, 5G eliminates the need for UEs to request resources and wait until the network grants them. Grant-free scheduling reserves radio resources for dedicated UEs or for groups of UEs. While in case of DG, user first align to the first available transmission opportunity of the uplink control channel, in order to send the scheduling request (SR), and accordingly wait for the scheduling grant (SG) from the serving BS over the downlink control channel. Thus, uplink delay $D_m^{'}$ is given by;

$$d_m^{'} = d_{dg}^{'} + d_q^{'} + d_{fa}^{'} + d_{tti}^{'} + d_{bsp}{}^{'} \tag{21}$$

Where the right-hand side in the equation represent delay incurred by DG, queuing, frame alignment, transmission, and BS processing respectively. If the communication mode is GF, there will be no $d_{dg}^{'}$. For DG, $d_{dg}^{'}$ is given by;

$$d_{dg}^{'} = d_{uep}^{'} + d_{fa'}^{'} + d_{sr'}^{'} + d_{bsp}^{'} + d_{fa}^{'} + d_{sg'}^{'} \tag{22}$$

Where the right-hand side shows UE processing, SR frame alignment, SR TTI, SR BS Processing and delay incurred by SG transmission.

Having uplink and downlink delay for macro-cellular network, the total delay of will be;

$$D_m = D_m^{"} + D_m^{"} + 2\delta \tag{23}$$

And total communication delay will be maximum of Wifi and Macro-cellular delay, that is

$$D_t = Max\,(D_w, D_m) \tag{24}$$

Combining Equation 4 and 24, we have service delay as follows;

$$D_s = D_t + D_c \tag{25}$$

Equation (4) shows that service delay is dependent upon communication delay and computation delay. Therefore, in this research, we strive to minimize computation and communication delay. We shall investigate relationship of service delay with different parameters to find the optimal parameter setting that minimizes service delay.

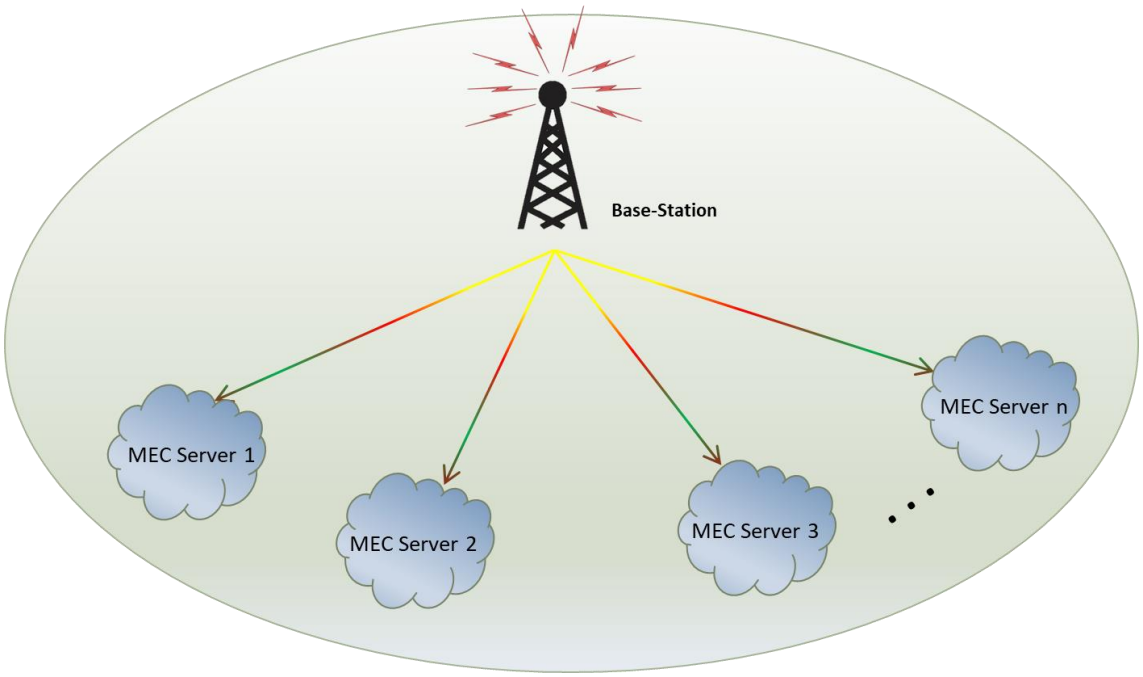

Figure 4.2: An Illustration of pool of servers available to choose from.

From Equation (4), it is clear that computation delay is inversely proportion to capacity of the chosen MEC server and directly proportion to the load on it. The computation delay will increase with the increase in the load on the server and decrease with the increase in the available capacity. Therefore, computation delay can be minimized by discreetly choosing the server and by managing the capacity and load on it.

Similarly, when two servers and two RATs are involved, load distribution and traffic scheduling take utmost significance. Disproportionate load distribution and traffic scheduling will result in significantly lower throughput [13]. Therefore, to minimize the service delay of MEC applications, we need to optimally manage the load distribution and optimally utilized the capacity obtained for the given SNR.

In the following sections, we show mathematical models to minimize computation delay and communication delay thereby managing server selection and load, and traffic scheduling and distribution.

4.4 Server Selection Model

We assume multiple servers are available to choose from as shown in Fig. 4.2. Following equation (2) and (3), let's assume a server with highest $(\varsigma_r/\varsigma_t)$ ratio is chosen, where c_r

the residual capacity of the server and c_t is the total processing capacity of the server. We define residual capacity as ($totalcapacity - currentloadontheserver$) that is, whatever capacity is left after the load occupy certain amount of capacity of the server and is the actual available capacity for processing. Having said that, higher $(\varsigma_r/\varsigma_t)$ ratio means lesser load and more available processing capacity to process our tasks.

It is worth mentioning here that we are working on two RATs simultaneously that is, a macro-cellular technology (R_m) and WiFi R_w. Therefore, one MEC servers will be chosen when all three of the following conditions are met and two servers in all other cases.

1. Computation demand of loads of both the RATs, $D^{'} + D^{"} <$ Residual capacity of MEC server $\varsigma_r^{'}$ with highest $(\varsigma_r/\varsigma_t)$ ratio among all.
2. Computation demand of the load over slower RAT $D^{"} >$ Residual capacity of MEC server $(\varsigma_r^{"})$ with second highest $(\varsigma_r/\varsigma_t)$ ratio among all.
3. Computation + Communication delay of slower RAT for $\varsigma_r^{"} >$ Computation + Communication delay of both RATS for $\varsigma_r^{'}$.

Let's assume that two servers S_i and S_j are chosen with total basic capacity ς_i and ς_j respectively. We also assume that ς_i and ς_j are equal. Therefore, the performance S_i and S_j only differ with the difference in load and under zero load, ς_i and ς_j. Therefore, a server with highest $(\varsigma_r/\varsigma_t)$ ratio is guaranteed to give minimum computation delay.

Also, under normal conditions, the capacity is equally distributed among the number of users. Let number of users connected to ς_1 be x_n and ς_2 be y_m. Therefore, capacity allocated to R_m and R_w is given by;

$$\varsigma_{R_m} = \frac{\varsigma_i}{x_n} \quad and \quad \varsigma_{R_w} = \frac{\varsigma_j}{y_n} \tag{26}$$

This was initial selection of MEC server and is guaranteed to give minimum computation delay. However, problem arises when we have MEC servers with better $(\varsigma_r/\varsigma_t)$ ratio than the current one. In such a case, the migration to server with higher (c_r/c_t) ratio is indispensable.

Let us define $(\varsigma_r/\varsigma_t)^t - (\varsigma_r/\varsigma_t)^{t-1}$ as change in utility ΔU, where $(\varsigma_r/\varsigma_t)^t$ is the ratio at time t while $(\varsigma_r/\varsigma_t)^{t-1}$ is the ratio at time $t-1$. Therefore, whenever ΔU is positive, RATs will change the servers. This entire process is summarized in algorithm 1 below.

Algorithm 1: Server Selection and Migration

Input: list of MEC servers with their current load, total capacity ς_t, RAT_m and RAT_w

Output: Selected Server and Migration to the new Server

1. With load and total capacity given, calculate $(\varsigma_r/\varsigma_t)$ for all servers, where $\varsigma_r = \varsigma_t - load$.
2. Choose the MEC servers with highest $(\varsigma_r/\varsigma_t)$ ratio.
3. If all the three conditions are not met, choose the second MEC server with second highest $(\varsigma_r/\varsigma_t)$ ratio.
4. Calculate ΔU at time t for every server where $\Delta U = (\varsigma_r/\varsigma_t)^t - (\varsigma_r/\varsigma_t)^{t-1}$
5. If for any server, $\Delta U > 0$,

   Change the server

   Else

   Continue current server

4.4.1 Server Migration Model

Here we shall model and predict MEC server migration. For the sake of simplicity, let's assume there are two servers only. We shall give a general case of n servers later.

Let's call the sum of the two allocated capacities as the total utility experienced by users u that is, using (26), we obtain;

$$U_u = \frac{\varsigma_i}{x_n} + \frac{\varsigma_j}{y_n} \tag{27}$$

However, sharing capacity on the basis of the number of users lead to inefficient allocation of capacity, as some user's application may not be very compute intensive. We propose capacity sharing on the basis of the computation load which is defined by the number of tasks that a server has to process rather than the number of the users.

Let's assume that capacity is distributed on the basis of the computational load that is, number of tasks. And $\sum x$ be the number of tasks run by server S_1 and $\sum y$ are

Table 4.1: All Possible Cases of Server Migration.

Case	Current Server	Candidate Server
1	1	0
2	M	0
3	1	M
4	M	M

the number of tasks run by server S_2 . $\sum x$ and $\sum y$ are total tasks of all the users accommodated by the servers. Since basic total capacity is same for both the servers, therefore, $\varsigma_1 = \varsigma_2 = \varsigma$. Let i and $j\epsilon$ users u, be the number of tasks transmitted via the two RATs that what we are using. Having said this, computational capacity that two RATs will get is given by;

$$\varsigma_{R_m} = \frac{i \cdot \varsigma}{\sum x} \quad and \quad \varsigma_{R_w} = \frac{j \cdot \varsigma}{\sum y} \tag{28}$$

Now, the total utility experienced by users u will be;

$$U_u = \frac{i \cdot \varsigma}{\sum x} + \varsigma_{R_w} = \frac{j \cdot \varsigma}{\sum y} \tag{29}$$

Next suppose, an arbitrary user changed his server from S_1 to S_2. In such a case, its corresponding load will go to S_2. Let its load be m. As a result, when load m moves to S_2, some room will be created in S_1 that will allow other users to occupy more space on server S_1. Let r amount of capacity is added to ς_{R_m} due to this change. Similarly, all users at S_2 will have to sacrifice certain amount of capacity to make room for load m of this new user. Again, let our RAT ς_{R_w} sacrifice r amount of capacity. Therefore, (29) will become;

$$\frac{(i + r) \cdot \varsigma}{\sum x - m} + \frac{(j - r) \cdot \varsigma}{\sum y + m} \tag{30}$$

Therefore, change in utility will be;

$$\Delta U_u = \left(\frac{(i + r) \cdot \varsigma}{\sum x - m} + \frac{(j - r) \cdot \varsigma}{\sum y + m} \right) - \left(\frac{i \cdot \varsigma}{\sum x} + \frac{j \cdot \varsigma}{\sum y} \right) \tag{31}$$

A user will switch to another server if its ΔU_u is positive for that server. When switching between the servers, there can be following possible cases.

Next, we shall discuss these cases to know if the value of u will be positive and negative.

- **Case 1:** Let's assume that S_1 is serving one user and S_2 is empty by which we mean no user is attached to it and its not processing anything. Therefore, before switching, $\sum y = 0$, $j = 0$ and also $m = 0$ because there is only one user, and all the load belongs to it which is represented by i. Equation (31) will become;

$$\Delta U_u = \left(0 + \frac{i \cdot \varsigma}{\sum y}\right) - \left(\frac{i \cdot \varsigma}{\sum x} + 0\right) = 0 \tag{32}$$

Because $\sum x = \sum y$, since the same user has switched the server. As a result, change in utility ΔU is 0. Therefore, there will be no migration to the new server.

- **Case 2:** There are multiple users on current server S_1 and no users attached to the candidate server S_2. Suppose one of the RAT moves to S_2. The ΔU_u will be calculated as follows. Also before migration, $\sum x = 0$, $j = 0$.

$$\Delta U_u = \left(\frac{(i + r) \cdot \varsigma}{\sum x - m} + \frac{(r) \cdot \varsigma}{m}\right) - \left(\frac{i \cdot \varsigma}{\sum x} + \frac{0 \cdot \varsigma}{0}\right) \tag{33}$$

Also, $r = m$ because all the load on S_2 belongs to the RAT of user u. Therefore,

$$= \varsigma + \left(\frac{(i + r) \cdot \varsigma}{\sum x - m}\right) - \left(\frac{i \cdot \varsigma}{\sum x} + 0\right) > 0 \tag{34}$$

ΔU_u is greater than 0. Therefore, migration to another server is justifiable. On the contrary, even if we assume that $r \neq m$, still $\frac{r \cdot \varsigma}{m} > \frac{i \cdot \varsigma}{\sum x}$. The ΔU_u is still positive and migration is again justifiable.

However, there is a different case when instead of the RAT of user u, another user switch the server that is, both of our RATs are still on S1. Therefore, $\sum y = 0$, $j = 0$, $r = 0$.

$$\Delta U_u = \left(\frac{i \cdot \varsigma}{\sum x - m} + \frac{0 \cdot \varsigma}{0 + m}\right) - \left(\frac{i \cdot \varsigma}{\sum x} + 0\right) \tag{35}$$

$$= \left(\frac{i \cdot \varsigma}{\sum x - m}\right) - \left(\frac{i \cdot \varsigma}{\sum x} + 0\right) \tag{36}$$

Since $(\sum x - m) < \sum x$, therefore, first term is greater than 0. This means, $\Delta U > 0$. Therefore, there will be again an increase in the total utility.

- **Case 3:** There are multiple users on S_2 and only one user on S1. Therefore, if this user switches to S_2, all it's processing data will migrate to S_2 and there will be nothing left on S_1. Also, before migration, $i = \sum x$ because there is only user.

$$\Delta U_u = \left(\frac{(0)\cdot\varsigma}{0} + \frac{(j+r)\cdot\varsigma}{\sum y + m}\right) - \left(\varsigma + \frac{j\cdot\varsigma}{\sum y}\right) \tag{37}$$

$$= -\left(\frac{\varsigma\,(i\cdot j) + \sum y^2}{\sum y\,(i + \sum y)}\right) < 0 \tag{38}$$

The ΔU_u is negative which shows there will be no migration.

For rest of the cases, we simplify equation (31).

$$\Delta U_u = \left(0 + \frac{i\cdot\varsigma}{\sum y}\right) - \left(\frac{i\cdot\varsigma}{\sum x} + 0\right) = 0 \tag{39}$$

Therefore, whenever the numerator is positive, ΔU_u will be positive and migration is justified.

$$\Delta U_u = \frac{\varsigma\,(jm^2\sum x - jm\sum x^2 + im^2\sum y + im\sum y^2 + 2mr\sum x\sum y + r\sum x\sum y^2 - r\sum x^2\sum}{\sum x\sum y\,(\sum x - m)\,(\sum y + m)} \tag{40}$$

$$\Delta U_u = \varsigma\left(jm^2\sum x - jm\sum x^2 + im^2\sum y + im\sum y^2 +\right.$$

$$2\text{mr}\sum x\sum y + r\sum x\sum y^2 - r\sum x^2\sum y \tag{41}$$

4.4.2 Convergence

The algorithm above is guaranteed to give the optimal computation delay. However, it has a serious drawback of convergence, particularly when there are three or more servers. The users will keep switching the servers upon finding a server that gives better utility. Therefore, to bring the algorithm to convergence, we manage the load on the server by minimizing and keeping the load in non-increasing order.

Let there be n servers in total and their load be $x_1, x_2, x_3 \cdots x_n$ Let this total load be L. Assuming load x_j are in random order, we define another collection of variables $y_1, y_2, y_3 \cdots y_n$ to order the x_j's. Now $y_1$is the largest in x_j and y_n is smallest. Furthermore, y_i is the *ith* largest of x_j.

Now there are two goals. In order to minimize the computation time, we need to minimize loads on the servers. Additionally, to achieve convergence of the server selection algorithm above, we need to keep the load in non-decreasing order. Mathematically, we can say;

$$Minimize\,(y_1, y_2, y_3, \cdots y_n,) \tag{42}$$

Subject to

$$\sum_{i=1}^{n} y_i = L \tag{43}$$

However, there are a few problems with this formulation. We can handle one server at a time. e.g., Minimize the worst loaded server that is, y_1. We cannot process the second or more than one server simultaneously. Similarly, we need to know where to shift the extra load once we minimized the load on the server. Finally, the above formulation is non-linear as loads are randomly distributed. Therefore, we assign weights to $y_1, y_2, y_3, \cdots y_n$ and formulate our problem as follow.

$$Minimize \quad w_1y_1 + w_2y_2 + w_3y_3 + \cdots w_ny_n + L \tag{44}$$

Subject to

$$y_j - x_i \leq KZ_j \tag{45}$$

$$\sum_{j=1}^{n} Z_j = n - j \tag{46}$$

$$\sum_{i=1}^{n} w_i = 1 \tag{47}$$

$$y_{i+1} \leq y_i \tag{48}$$

$$Z_j \epsilon \{0, 1\} \tag{49}$$

Now the problem formulation is in linear form. We can easily solve this using simple algorithm such as simplex. Moreover, this ILP above has two major advantages. Apart from bringing the algorithm to equilibrium point, the ILP has a characteristic of minimizing the load on the servers thus, minimizing the server occupation rate as per Equation (2) which in return again will minimize the computation load as given in (3). Therefore, with the help of algorithm above and the formulated ILP, we have the most

possible computation minimum delay.

Having obtained minimum computation delay, we next formulate our multi-RAT multi-server load distribution problem to minimize communication related delays.

Our goal is to minimize the service delay defined in (25). Mathematically;

$$Minimize \quad D_s \tag{50}$$

While subject to following constraints

$$D_w = D_w \tag{51}$$

$$L_\varsigma = \varsigma_i \tag{52}$$

$$L_{si} = R_i \tag{53}$$

$$v\frac{R_i - Mod\,(x_i, R_i)}{\varsigma_i} = \frac{R_j - Mod\,(x_j, R_j)}{\varsigma_j} \tag{54}$$

$$\sum L_{si} = L \tag{55}$$

Where constraint 51 states that service delay of the load transmitted over both the RATs should be same. This is to avoid reordering delay. Constraint 52 and 53 are capacity conservation constraints and ensure that load on a server cannot exceed its processing capacity and load assigned to a RAT for transmission cannot exceed its transmission capacity whereas constraint 54 ensures that capacity RATs are optimally utilized. It ensures fair shares of load by ensuring ratio f residual capacity to total capacity are equal. Finally, constraint 55 ensure traffic load conservation and ensures that sum of load shares distributed over both the RATs should be equal to total data load generated by the user.

We solve the ILP through a heuristic outlined in Algorithm 2. We define a few variables before explaining the heuristic. Let T be threshold delay which is the service delay of the faster RAT (or fastest if more than two RATs are involved) for entire data load L generated by the user. Let L_T be the load served in time T and L_F be the load left after transmitting L_T in T seconds which is obviously zero for the faster RAT. We define another auxiliary variable x to balance the load which is explained later.

Our heuristic works as follows. Having found T, we shuffle the RATs according to the number of RATs and give them different orders. For example, starting with C_w and putting it first in the order, we calculate L_T for C_w and move L_F to C_m. Similarly, in next step, we assume C_m to be the first RAT and calculate its L_T and move L_F to C_w. Subsequenty, we take sum of L_F and L_T for both the RATs and divide it by $N! + x$. Here $N!$ is the number of shuffles which is 2 in this case and x is the auxiliary variable that is used to balance the load. Its value is $-\left(\frac{B-A}{B}\right)$ for slower RAT and $\left(\frac{A(B-A)}{B^2}\right)$ for a faster RAT where A and B are transmission capacities of slower and faster RATS respectively. The service delay for $L_{s,i}$ obtained for the RAT i is equal and optimal for both the RATs.

Algorithm 2: Load Distribution

Input: Total traffic load (L) generated by the user, Capacity R_w and R_m for WiFi and Macro-cellular radio.

Output: Load shares L_{si} of a RAT i.

1. Find T where $T = Min\,(D_w^t, D_m^t)$
2. For RAT R_w do
   - Calculate throughput L_T in time T.
   - Move L_F to R_m.
3. For RAT R_m do
   - Calculate throughput L_T in time T.
   - Move L_F to C_w.
4. $L_{s,i} = \frac{\sum(L_{F,i}+L_{T,i})}{N!+x}$

In this particular case where there are only two RATs, mathematically, we can write the two obtained loads as follows.

$$Min\,(D_w^t, D_m^t) \tag{56}$$

$$\frac{A \cdot T}{2 + x_1} + \frac{2L - L_F}{2 + x_2} = L \tag{57}$$

Where the first expression is the load obtained by a slower RAT of the two and second expression is the load obtained by the faster RAT and $L_F = L - B \cdot T$. The 2 in the denominator shows the number of permutations. The sum of the two loads is equal to total load generated by the user which signifies conservation of the total load.

4.5 Performance Evaluation

In this section, we provide numerical results to show the performance of our proposed scheme. We compare the performance with WiFi, 5G and multi-RAT enabled cybertwin [64]. We show how different RATs take different loads for their corresponding performance and compare their service delay. We then compare the performance of our proposed scheme with Wi-Fi, 5G and cybertwin.

4.5.1 Environment Setting and Parameters

We consider the same environment setting and parameters as shown as in previous chapter except that the user has to choose a server among a pool of available servers to minimize processing delay. We consider the scenario shown in Fig. 4.3 where an end-user is assumed to be based inside a multi-storey building. A Wi-Fi access point is assumed to be inside the building while 5G macro-cell base-station is assumed to be at a distance of 200 m in an urban environment. The end-user is assumed to be simultaneously connected to Wi-Fi AP and 5G base-station. For Wi-Fi, we have used a frequency band of 5 GHz whereas for 5G, we have used 3.4 GHz band from Frequency Range 1 [60]. Similarly, EIRP for Wi-Fi is 30 dBm and 43 dBm for 5G. Next we describe how to compute different parameters in order to get performance measures of different RATs.

The parameters used in computations are summarized in Table 4.2.

4.5.2 Results

We begin with load distribution and service delay analysis of the proposed scheme where system delay is essentially the time between a user sending the request and the corresponding results, that is the output of its request. Under the same condition for all the servers, Fig. 4.4 shows the load each radio will get for different load generated by the end-user. 5G, having higher bandwidth, has higher capacity and lower network delay among the two radios, Therefore, the load share taken by 5G is the higher.

We then analyze the delay for the corresponding load assumed by the two radios in Fig. 4.5. There are two curves in the figure which appear to be one single curve. The load shares assumed by the two radios is different as shown in Fig. 4.4, their delay,

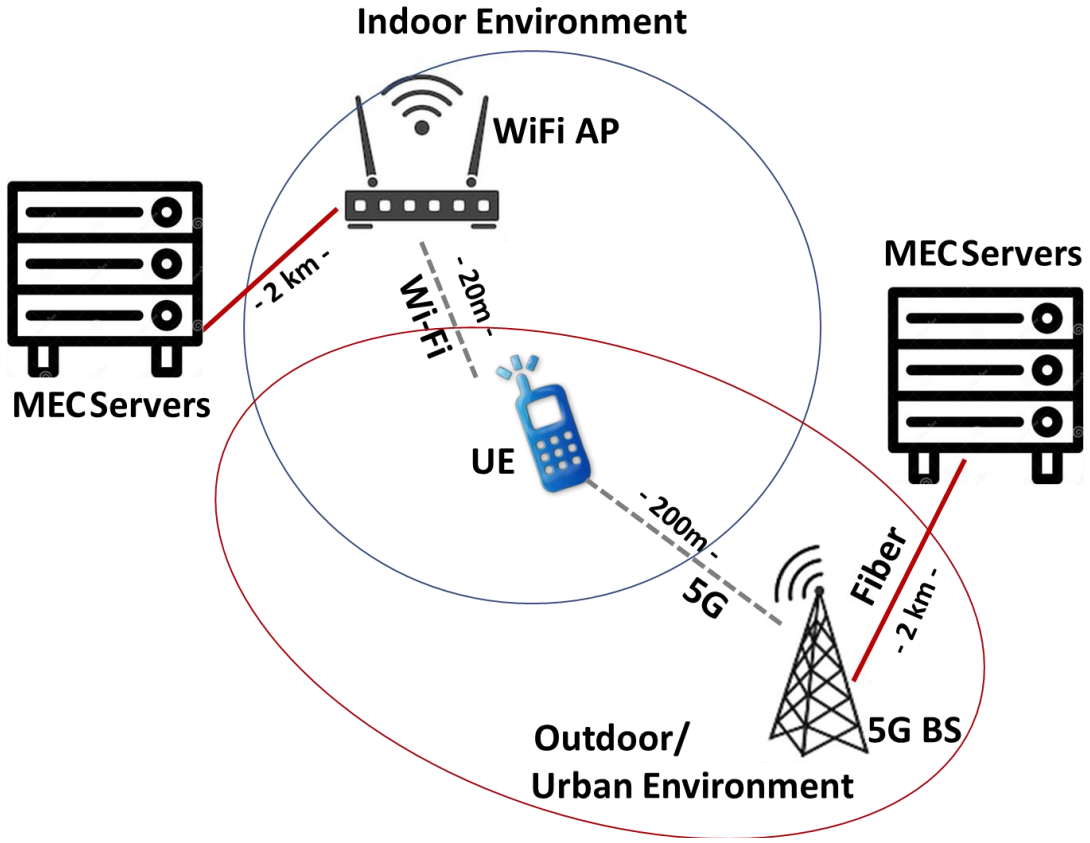

Figure 4.3: Assumed topology where an end-user inside a building is served by Wi-Fi access point and 5G.

Table 4.2: Parameters Setting.

Technology	Wi-Fi (802.11ax)	5G
Distance	20 m	200 m
Bandwidth	80 MHz	100 MHz
Capacity	SNR Driven	
EIRP	30 dBm	43 dBm
Modulation	SNR Driven	
Code Rate	SNR Driven	
Frequency	5 GHz	3.4 GHz (FR-1)
α	2	-
β	2	-
Number of Servers	8	
Height of 5G Base Station	-	45 m
Height of Floor	-	10 m
Aggregated Carrier	1	1
Number of Streams	1	1
5G Numerology	-	1

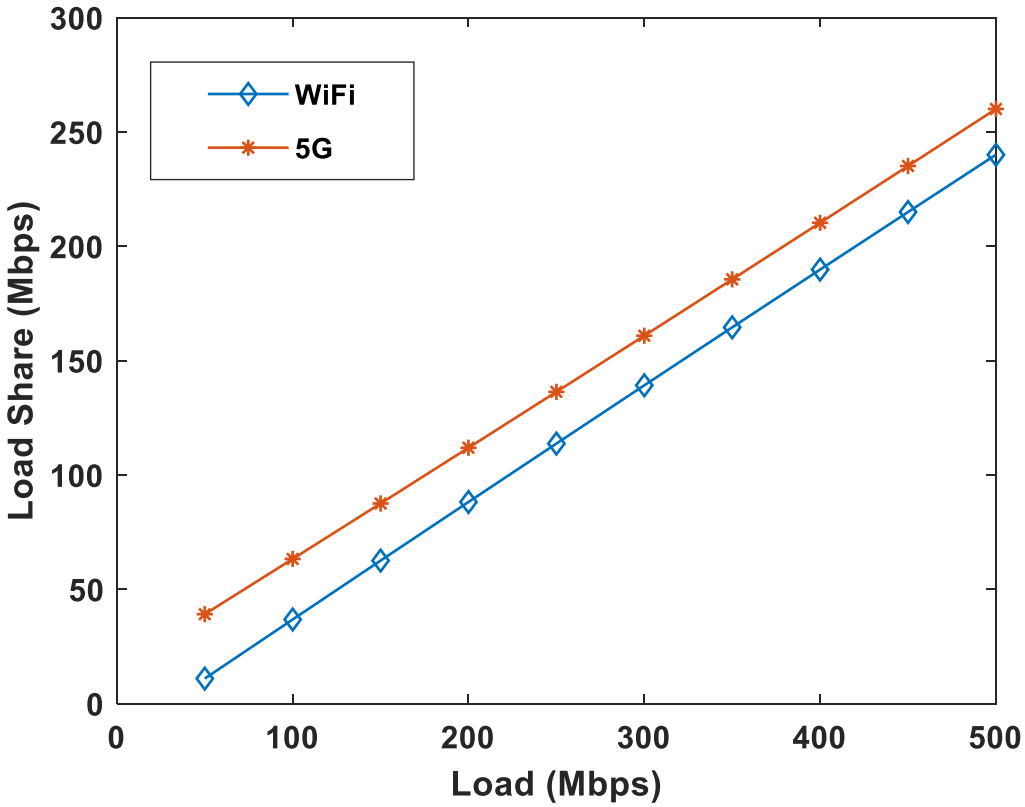

Figure 4.4: Load shares assumed by WiFi and 5G as a result of increase in the incoming load.

however, is equal. The proposed scheme is quite stable up to a load of 450 Mbps. The service delay has been consistently lower than 200 ms for the said load. Finally, with the data packet arriving simultaneously, there will be no packet reordering delay.

We also compare delay performance when radios are used data is offloaded through Wi-Fi alone and 5G alone with with our proposed scheme. As can be seen in Fig. 4.6, the proposed scheme outperforms Wi-Fi and 5G offloading in terms the amount of data that they can carry. Wi-Fi reach saturation at around 200 Mbps whereas 5G reach saturation at around 250 Mbps. The proposed scheme however, performs well untill 450 Mbps and the service delay remains less than 200 ms. This is a gain of more than 65% as compared to Wi-Fi and 5G. In addition to processing more load, the sevice delay for the proposed scheme has been consistenly lower than WiFi and 5G.

We also compare service delay for our proposed scheme with cybertwin technology. Cybertwin is a recent technology that has more coordinated information sharing process among the servers, however at the cost of more network traffic, storage occupation and an extra layer of cybertwin virtual servers. Cybertwin employs multiple server for task processing, albeit, one server is used at particular time choosing the most optimal server for processing. Figure 4.7 compare the service delay of the proposed scheme

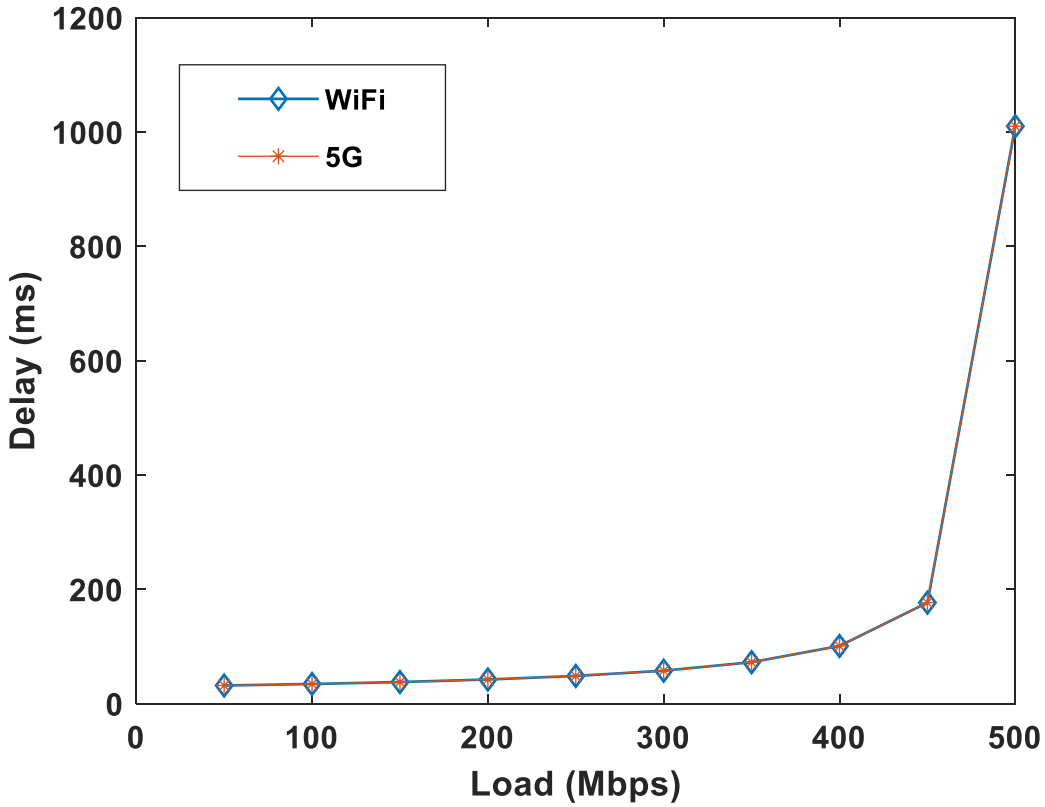

Figure 4.5: Delay for different RATs as a result of increase in the incoming load.

with cybertwin-based system. Both the proposed technology and the cybertwin, having option of choosing among multiple servers, have almost same saturation point, however, proposed scheme has slightly lower service delay at lower load of up to 400 Mbps which tends to get double at higher load of 500 Mbps and above. The higher delay of cybertwin is due to the fact that it uses single server for processing which reach saturation point earlier at higher load.

4.6 Summary

we jointly minimize networking and processing delay. To minimize computation delay, we developed a technique that chooses the most optimal servers. Further, to minimize server migration and to achieve a convergence point in the algorithm, we formulated a max-min based non-linear lexicographic minimization problem. To solve the formulated problem in polynomial time, I transformed the non-linear objective function to a linear one and solved it through the simplex algorithm. Based on the obtained network performance and computation delay, we formulated a multi-server multi-RAT load distribution problem to optimally utilize the available capacities of the radios.

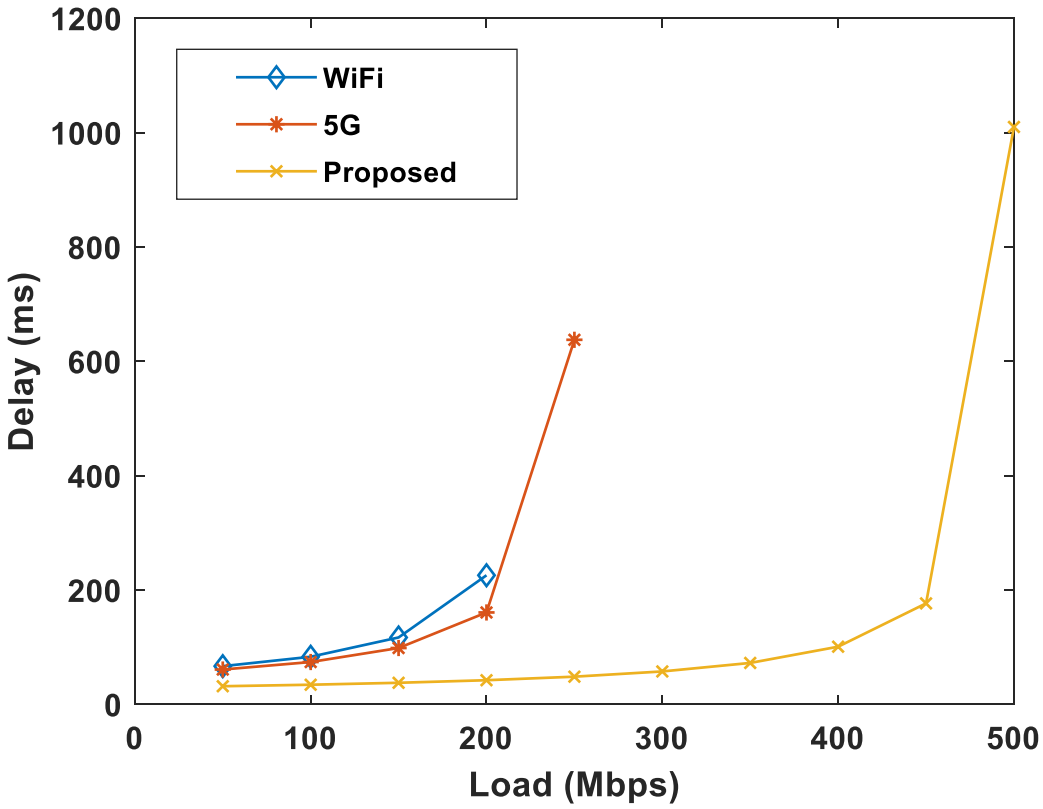

Figure 4.6: Service delay of individual radios when they are used as solo radios for offloading.

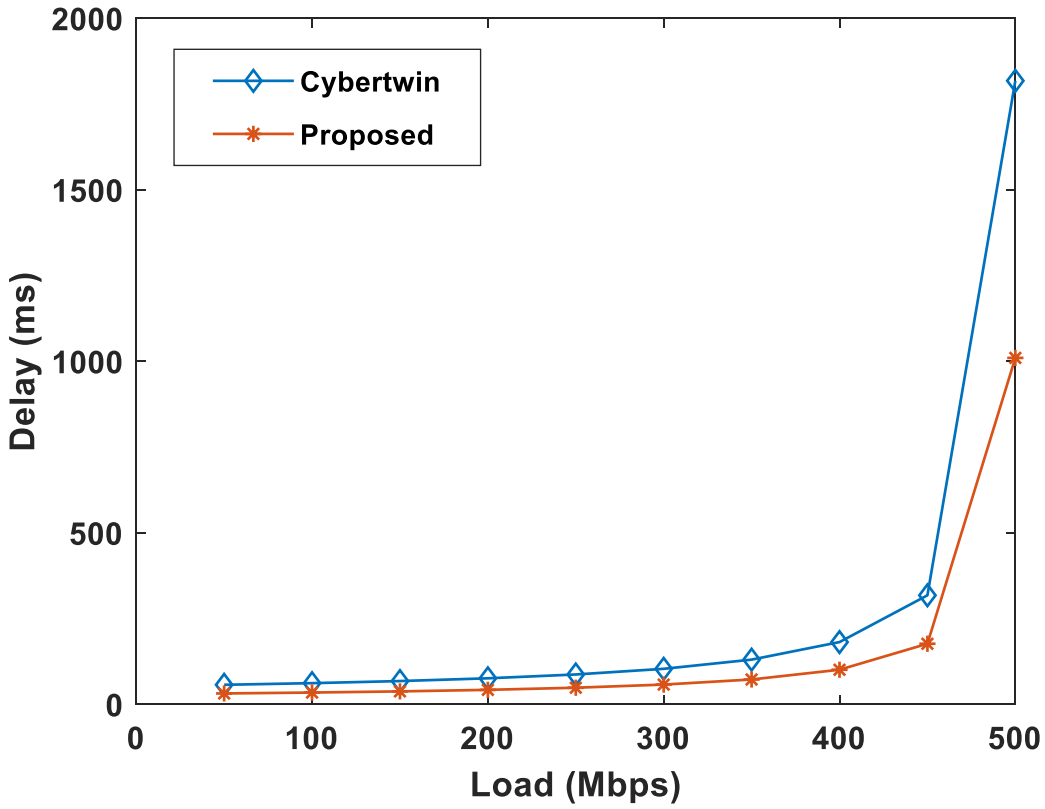

Figure 4.7: Service delay of individual radios when they are used as solo radios for offloading.

5 Conclusion

5.1 Conclusion

MEC-based networks are a complex systems with multiple parameters to be taken into consideration. Such systems have a wide array of resources organized across multiple tiers of the network and, if orchestrated properly, MEC can be a plausible platform for a wide variety applications. In this thesis, we worked on two important aspects of multi-radio multi-server MEC-based systems that is, minimizing network delay and service delay.

We developed a technique that optimally utilizes the capacity at source node and optimally distributes the available capacity among the links at relay node. We considered the performance of all the radios and distributed the traffic among the radios in such a way that delay for all the RATS is equalized, thereby avoiding the packet re-ordering delay at the destination node. As a proof-of-concept, we showed that to minimize system delay and maximize throughput, QoS and SLA compliance, we must optimize capacity utilization at the source node and capacity distribution on the outgoing links at relay nodes. Our numerical results demonstrated that our proposed technique fares better than contemporary techniques that distribute the data on the basis of the number of tasks.

We also worked on multi-server multi-RAT (MSMR) powered MEC where offloading occurs on both the radio access technologies (RAT) that a mobile phone comes equipped with that is, WiFi and macro-cellular technology such as 5G, and depending upon the conditions, both may be connected to different servers. We developed a technique that optimally utilize the available capacity, overcome packet re-ordering delay. The proposed technique minimizes the processing and networking delay. Numerical results showed equal delay for the two RATs for the different loads assumed by them. The ratio of residual capacity to total capacity was equal for the two RATs indicating optimal utilization of the available capacity.

5.2 Future Work

Availing the system of multiple radios and multiple servers simultaneously adds to its energy consumption. In future, work can done on incorporating energy efficiency in the system particular when system load is low. Similarly, we computed SLA compliance in terms of throughput. In addition, we plan to carry out a more in depth analysis of SLA compliance, including additional parameters such as service delay and QoS. Moreover, we used WMA for instantaneous capacity estimation between two consecutive performance updates. Work can be done on improving accuracy of instantaneous capacity estimation. Finally, optimizing update interval with respect to instantaneous position and network load simultaneously in polynomial time is a good direction for future.

In order to check the viability of the proposed scheme, we have been working on practical implementation of the developed algorithms. We have been planning to develop a SDN-driven controller that takes the relevant decisions related load sharing and traffic scheduling on run-time.

Bibliography

[1] Cisco. *Cisco Annual Internet Report (2018–2023) White Paper*, 2022 (accessed on August 18, 2022). URL `https://www.cisco.com/c/en/us/solutions/collateral/executive-perspectives/annual-internet-report/white-paper-c11-741490.html`.

[2] Hai Lin, Sherali Zeadally, Zhihong Chen, Houda Labiod, and Lusheng Wang. A survey on computation offloading modeling for edge computing. *Journal of Network and Computer Applications*, 169:1–25, 2020. ISSN 1084-8045. doi: https://doi.org/10.1016/j.jnca.2020.102781.

[3] Nasir Abbas, Yan Zhang, Amir Taherkordi, and Tor Skeie. Mobile edge computing: A survey. *IEEE Internet of Things Journal*, 5(1):450–465, 2018. doi: 10.1109/JIOT.2017.2750180.

[4] A. Mukhopadhyay and M. Ruffini. Learning automata for multi-access edge computing server allocation with minimal service migration. In *ICC 2020 - 2020 IEEE International Conference on Communications (ICC)*, pages 1–6, 2020. doi: 10.1109/ICC40277.2020.9148802.

[5] P. Porambage, J. Okwuibe, M. Liyanage, M. Ylianttila, and T. Taleb. Survey on multi-access edge computing for internet of things realization. *IEEE Communications Surveys Tutorials*, 20(4):2961–2991, 2018. doi: 10.1109/COMST.2018.2849509.

[6] Mostafa Zaman Chowdhury, Md. Shahjalal, Shakil Ahmed, and Yeong Min Jang. 6g wireless communication systems: Applications, requirements, technologies, challenges, and research directions. *IEEE Open Journal of the Communications Society*, 1:957–975, 2020. doi: 10.1109/OJCOMS.2020.3010270.

[7] Chamitha De Alwis, Anshuman Kalla, Quoc-Viet Pham, Pardeep Kumar, Kapal Dev, Won-Joo Hwang, and Madhusanka Liyanage. Survey on 6G frontiers: Trends, applications, requirements, technologies and future research. *IEEE Open Journal*

of the Communications Society, 2:836–886, 2021. doi: 10.1109/OJCOMS.2021.3071496.

[8] Q. Pham, F. Fang, V. N. Ha, M. J. Piran, M. Le, L. B. Le, W. Hwang, and Z. Ding. A survey of multi-access edge computing in 5G and beyond: Fundamentals, technology integration, and state-of-the-art. *IEEE Access*, 8:116974–117017, 2020. doi: 10.1109/ACCESS.2020.3001277.

[9] J. R. Bhat and S. A. Al-Qahtani. 6G ecosystem: Current status and future perspective. *IEEE Access*, pages 1–34, 2021. doi: 10.1109/ACCESS.2021.3054833.

[10] Nir Shlezinger, George C. Alexandropoulos, Mohammadreza F. Imani, Yonina C. Eldar, and David R. Smith. Dynamic metasurface antennas for 6G extreme massive mimo communications. *IEEE Wireless Communications*, 28(2):106–113, 2021. doi: 10.1109/MWC.001.2000267.

[11] Daniele Pinchera, Marco Donald Migliore, and Fulvio Schettino. Optimizing antenna arrays for spatial multiplexing: Towards 6G systems. *IEEE Access*, 9:53276–53291, 2021. doi: 10.1109/ACCESS.2021.3070198.

[12] Asad Ali and Faisal Ahmed Khan. Condition and location-aware channel switching scheme for multi-hop multi-band WLANs. *Computer Networks*, 168:107048, 2020. ISSN 1389-1286. doi: https://doi.org/10.1016/j.comnet.2019.107048.

[13] A. Ali, F. Hussain, R. Hussain, A. M. Khan, and A. Ferworn. Multi-band multi-hop WLANs for disaster relief and public safety applications. In *2020 IEEE Wireless Communications and Networking Conference Workshops (WCNCW)*, pages 1–6, 2020.

[14] Seung-seob Lee, TaeYoung Kim, SuKyoung Lee, Kyungsoo Kim, Yoon Hyuk Kim, and Nada Golmie. Dynamic channel bonding algorithm for densely deployed 802.11ac networks. *IEEE Transactions on Communications*, 67(12):8517–8531, 2019. doi: 10.1109/TCOMM.2019.2944382.

[15] M. Ruffini, A. Ahmad, S. Zeb, N. Afraz, and F. Slyne. Virtual DBA: virtualizing passive optical networks to enable multi-service operation in true multi-tenant environments. *IEEE/OSA Journal of Optical Communications and Networking*, 12(4): B63–B73, 2020. doi: 10.1364/JOCN.379894.

[16] J. Xia, L. Fan, N. Yang, Y. Deng, T. Q. Duong, G. K. Karagiannidis, and A. Nallanathan. Opportunistic access point selection for mobile edge computing networks. *IEEE Transactions on Wireless Communications*, 20(1):695–709, 2021. doi: 10.1109/TWC.2020.3028102.

[17] Gurjot Singh Gaba, Gulshan Kumar, Tai-Hoon Kim, Himanshu Monga, and Pardeep Kumar. Secure device-to-device communications for 5G enabled internet of things applications. *Computer Communications*, 169:114–128, 2021. ISSN 0140-3664. doi: https://doi.org/10.1016/j.comcom.2021.01.010. URL https://www.sciencedirect.com/science/article/pii/S0140366421000311.

[18] Haneul Ko and Sangheon Pack. Distributed device-to-device offloading system: Design and performance optimization. *IEEE Transactions on Mobile Computing*, early access:1–1, 2020. doi: 10.1109/TMC.2020.2994138.

[19] Mohammed Dighriri, Ali Saeed Dayem Alfoudi, Gyu Myoung Lee, Thar Baker, and Rubem Pereira. Comparison data traffic scheduling techniques for classifying qos over 5g mobile networks. In *2017 31st International Conference on Advanced Information Networking and Applications Workshops (WAINA)*, pages 492–497, 2017. doi: 10.1109/WAINA.2017.106.

[20] Tai Manh Ho and Kim-Khoa Nguyen. Joint server selection, cooperative offloading and handover in multi-access edge computing wireless network: A deep reinforcement learning approach. *IEEE Transactions on Mobile Computing*, pages 1–1, 2020. doi: 10.1109/TMC.2020.3043736.

[21] J. Licklider. *Memorandum for : Members and affiliates of the intergalactic computer network*, 1963, accessed on June 25, 2021). URL http://shannon.usu.edu.ru/Papers/Lick/.

[22] Shangguang Wang, Yan Guo, Ning Zhang, Peng Yang, Ao Zhou, and Xuemin Shen. Delay-aware microservice coordination in mobile edge computing: A reinforcement learning approach. *IEEE Transactions on Mobile Computing*, 20(3):939–951, 2021. doi: 10.1109/TMC.2019.2957804.

[23] Yaping Sun, Zhiyong Chen, Meixia Tao, and Hui Liu. Communication, computing and caching for mobile vr delivery: Modeling and trade-off. In *2018 IEEE International Conference on Communications (ICC)*, pages 1–6, 2018. doi: 10.1109/ICC.2018.8422519.

[24] Xing Chen and Guizhong Liu. Energy-efficient task offloading and resource allocation via deep reinforcement learning for augmented reality in mobile edge networks. *IEEE Internet of Things Journal*, 8(13):10843–10856, 2021. doi: 10.1109/JIOT.2021.3050804.

[25] Amit Samanta and Yong Li. Latency-oblivious incentive service offloading in mobile edge computing. In *2018 IEEE/ACM Symposium on Edge Computing (SEC)*, pages 351–353, 2018. doi: 10.1109/SEC.2018.00042.

[26] Dan Zhao, Tan Yang, Yuehui Jin, and Yue Xu. A service migration strategy based on multiple attribute decision in mobile edge computing. In *2017 IEEE 17th International Conference on Communication Technology (ICCT)*, pages 986–990, 2017. doi: 10.1109/ICCT.2017.8359782.

[27] Yu Ma, Weifa Liang, Jing Li, Xiaohua Jia, and Song Guo. Mobility-aware and delay-sensitive service provisioning in mobile edge-cloud networks. *IEEE Transactions on Mobile Computing*, pages 1–1, 2020. doi: 10.1109/TMC.2020.3006507.

[28] Mohammad Goudarzi, Huaming Wu, Marimuthu Palaniswami, and Rajkumar Buyya. An application placement technique for concurrent iot applications in edge and fog computing environments. *IEEE Transactions on Mobile Computing*, 20(4): 1298–1311, 2021. doi: 10.1109/TMC.2020.2967041.

[29] Shuyue Ma, Shudian Song, Jingmei Zhao, Linbo Zhai, and Feng Yang. Joint network selection and service placement based on particle swarm optimization for multi-access edge computing. *IEEE Access*, 8:160871–160881, 2020. doi: 10.1109/ACCESS.2020.3020935.

[30] Van Dat Tuong, Thanh Phung Truong, Anh-Tien Tran, Arooj Masood, Demeke Shumeye Lakew, Chunghyun Lee, Yunseong Lee, and Sungrae Cho. Delay-sensitive task offloading for internet of things in nonorthogonal multiple access mec networks. In *2020 International Conference on Information and Communication Technology Convergence (ICTC)*, pages 597–599, 2020. doi: 10.1109/ICTC49870.2020.9289406.

[31] Tuyen X. Tran and Dario Pompili. Joint task offloading and resource allocation for multi-server mobile-edge computing networks. *IEEE Transactions on Vehicular Technology*, 68(1):856–868, 2019. doi: 10.1109/TVT.2018.2881191.

[32] Zheming Yang, Bing Liang, and Wen Ji. An intelligent end-edge-cloud architecture for visual iot assisted healthcare systems. *IEEE Internet of Things Journal*, pages 1–1, 2021. doi: 10.1109/JIOT.2021.3052778.

[33] Xiaowen Cao, Feng Wang, Jie Xu, Rui Zhang, and Shuguang Cui. Joint computation and communication cooperation for energy-efficient mobile edge computing. *IEEE Internet of Things Journal*, 6(3):4188–4200, 2019. doi: 10.1109/JIOT.2018.2875246.

[34] Kalpit D Ballal, Lars Dittmann, Sarah Ruepp, and Martin Nordal Petersen. Iot devices reliability study: Multi-rat communication. In *2020 IEEE 6th World Forum on Internet of Things (WF-IoT)*, pages 1–2, 2020. doi: 10.1109/WF-IoT48130.2020.9221163.

[35] Tiago Koketsu Rodrigues, Katsuya Suto, and Nei Kato. Edge cloud server deployment with transmission power control through machine learning for 6G internet of things. *IEEE Transactions on Emerging Topics in Computing*, pages 1–1, 2019. doi: 10.1109/TETC.2019.2963091.

[36] Walid Saad, Mehdi Bennis, and Mingzhe Chen. A vision of 6G wireless systems: Applications, trends, technologies, and open research problems. *IEEE Network*, 34 (3):134–142, 2020. doi: 10.1109/MNET.001.1900287.

[37] Zhaolong Ning, Peiran Dong, Xiangjie Kong, and Feng Xia. A cooperative partial computation offloading scheme for mobile edge computing enabled internet of things. *IEEE Internet of Things Journal*, 6(3):4804–4814, 2019. doi: 10.1109/JIOT.2018.2868616.

[38] Yi-Chen Wu, Thinh Quang Dinh, Yaru Fu, Che Lin, and Tony Q. S. Quek. A hybrid dqn and optimization approach for strategy and resource allocation in mec networks. *IEEE Transactions on Wireless Communications*, 20(7):4282–4295, 2021. doi: 10.1109/TWC.2021.3057882.

[39] Arash Bozorgchenani, Farshad Mashhadi, Daniele Tarchi, and Sergio A. Salinas Monroy. Multi-objective computation sharing in energy and delay constrained mobile edge computing environments. *IEEE Transactions on Mobile Computing*, 20(10):2992–3005, 2021. doi: 10.1109/TMC.2020.2994232.

[40] Arash Bozorgchenani, Setareh Maghsudi, Daniele Tarchi, and Ekram Hossain. Computation offloading in heterogeneous vehicular edge networks: On-line and off-policy bandit solutions. *IEEE Transactions on Mobile Computing*, pages 1–1, 2021. doi: 10.1109/TMC.2021.3082927.

[41] Rojeena Bajracharya, Rakesh Shrestha, Yousaf Bin Zikria, and Sung Won Kim. LTE in the unlicensed spectrum: A survey. *IETE Technical Review*, 35(1):78–90, 2018. doi: 10.1080/02564602.2016.1251344.

[42] F. Tian, Y. Yu, X. Yuan, B. Lyu, and G. Gui. Predicted decoupling for coexistence between WIFI and LTE in unlicensed band. *IEEE Transactions on Vehicular Technology*, 69(4):4130–4141, 2020. doi: 10.1109/TVT.2020.2976939.

[43] T. BRAUD, P. ZHOU, J. KANGASHARJU, and P. HUI. Multipath computation offloading for mobile augmented reality. In *2020 IEEE International Conference on Pervasive Computing and Communications (PerCom)*, pages 1–10, 2020. doi: 10.1109/PerCom45495.2020.9127360.

[44] K. C. Lin, H. Wang, Y. Lai, and Y. Lin. Communication and computation offloading for multi-rat mobile edge computing. *IEEE Wireless Communications*, 26(6):180–186, 2019. doi: 10.1109/MWC.001.1800603.

[45] Z. Jing, Q. Yang, M. Qin, J. Li, and K. s. Kwak. Long-term max-min fairness guarantee mechanism for integrated multi-rat and MEC networks. *IEEE Transactions on Vehicular Technology*, pages 1–15, 2021 (Early access). doi: 10.1109/TVT.2021.3059944.

[46] Olga Galinina, Alexander Pyattaev, Sergey Andreev, Mischa Dohler, and Yevgeni Koucheryavy. 5g multi-rat lte-wifi ultra-dense small cells: Performance dynamics, architecture, and trends. *IEEE Journal on Selected Areas in Communications*, 33 (6):1224–1240, 2015. doi: 10.1109/JSAC.2015.2417016.

[47] Guanding Yu, Yuhuan Jiang, Lukai Xu, and Geoffrey Ye Li. Multi-objective energy-efficient resource allocation for multi-rat heterogeneous networks. *IEEE Journal on Selected Areas in Communications*, 33(10):2118–2127, 2015. doi: 10.1109/JSAC.2015.2435374.

[48] Ieee standard for information technology–telecommunications and information exchange between systems local and metropolitan area networks–specific requirements part 11: Wireless lan medium access control (mac) and physical layer (phy) specifications amendment 1: Enhancements for high-efficiency wlan. *IEEE Std 802.11ax-2021 (Amendment to IEEE Std 802.11-2020)*, pages 1–767, 2021. doi: 10.1109/IEEESTD.2021.9442429.

[49] Yousri Daldoul, Djamal-Eddine Meddour, and Adlen Ksentini. Performance evaluation of OFDMA and MU-MIMO in 802.11ax networks. *Computer Networks*, 182: 107477, 2020. ISSN 1389-1286.

[50] Evgeny Khorov, Anton Kiryanov, Andrey Lyakhov, and Giuseppe Bianchi. A tutorial on ieee 802.11ax high efficiency wlans. *IEEE Communications Surveys Tutorials*, 21(1):197–216, 2019. doi: 10.1109/COMST.2018.2871099.

[51] 3GPP TS 38.306. *Technical Specification Group Radio Access Network; User Equipment (UE) radio access capabilities*, December, 2017 (accessed on June 25, 2021). URL `https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3193`.

[52] Asad Ali, Kanza Ali, and Aftab Ahmad Shaikh. Energy and delay aware routing algorithm for fiber-wireless networks. *Wireless Networks*, 20(6):1313–1320, Aug 2014. ISSN 1572-8196. doi: 10.1007/s11276-013-0679-5.

[53] Leonard Kleinrock. *Queuing systems*. Wiley, New York, NY, 1975. URL `https://cds.cern.ch/record/103535`.

[54] A. Ali, I. Ullah, T. Taqueer, and S. M. H. Zaidi. Performance enhancement of WLANs. In *Proc. Intl. Conf. on High-capacity Optical Netw. and Emerg. Technol.*, pages 148–152, Dec 2011. doi: 10.1109/HONET.2011.6149806.

[55] Won-Jae Lee, Wonjae Shin, Joan A. Ruiz-de Azua, Lara Fernandez Capon, Hyuk Park, and Jae-Hyun Kim. Noma-based uplink ofdma collision reduction in 802.11ax networks. In *2021 International Conference on Information and Communication Technology Convergence (ICTC)*, pages 212–214, 2021. doi: 10.1109/ICTC52510.2021.9621014.

[56] S. Song, J. Jung, M. Choi, C. Lee, J. Sun, and J. Chung. Multipath based adaptive concurrent transfer for real-time video streaming over 5G multi-rat systems. *IEEE Access*, 7:146470–146479, 2019. doi: 10.1109/ACCESS.2019.2945357.

[57] John F Nash Jr. The bargaining problem. *Econometrica: Journal of the Econometric Society*, 18:155–162, April 1950.

[58] E. J. McShane. The lagrange multiplier rule. *The American Mathematical Monthly*, 80(8):922–925, 1973. doi: 10.1080/00029890.1973.11993409.

[59] Rajesh Mahindra, Hari Viswanathan, Karthik Sundaresan, Mustafa Y. Arslan, and Sampath Rangarajan. A practical traffic management system for integrated LTE-WIFI NETWORKS. In *Proceedings of the 20th Annual International Conference on Mobile Computing and Networking*, MobiCom '14, pages 189–200, New York, NY, USA, 2014. Association for Computing Machinery. ISBN 9781450327831. doi: 10.1145/2639108.2639120.

[60] Qualcomm. *Global Updates on Spectrum for 4G/5G*, 2020, accessed on June 25, 2021). URL `https://www.qualcomm.com/media/documents/files/spectrum-for-4g-and-5g.pdf`.

[61] V. Degli-Esposti, G. Falciasecca, F. Fuschini, and E. M. Vitucci. A meaningful indoor path-loss formula. *IEEE Antennas and Wireless Propagation Letters*, 12(1):872–875, 2013. doi: 10.1109/LAWP.2013.2271532.

[62] Hideaki Okamoto, Koshiro Kitao, and Shinichi Ichitsubo. Outdoor-to-indoor propagation loss prediction in 800-MHz to 8-GHz band for an urban area. *IEEE Transactions on Vehicular Technology*, 58(3):1059–1067, 2009. doi: 10.1109/TVT.2008.927996.

[63] Nima Afraz, Frank Slyne, Harleen Gill, and Marco Ruffini. Evolution of access network sharing and its role in 5G networks. *Applied Sciences*, 9(21), 2019. ISSN 2076-3417. doi: 10.3390/app9214566. URL https://www.mdpi.com/2076-3417/9/21/4566.

[64] Quan Yu, Jing Ren, Yinjin Fu, Ying Li, and Wei Zhang. Cybertwin: An origin of next generation network architecture. *IEEE Wireless Communications*, 26(6): 111–117, 2019. doi: 10.1109/MWC.001.1900184.