User Mobility and Quality-of-Experience Aware Placement of Virtual Network Functions in 5G

By

Palash Roy

Exam Roll: Curzon Hall-553

Registration No: 2015-516-785

Session: 2015-16

Anika Tahsin

Exam Roll: Curzon Hall-567

Registration No: 2015-116-798

Session: 2015-16

A thesis/report submitted for the degree of Bachelors of Science at University of Dhaka

Department of Computer Science and Engineering

Faculty of Engineering

January 5, 2020

Declaration of Authorship

We, hereby, declare that the work presented in this project is the outcome of the investigation performed by us under the supervision of Dr. Mamun-or-Rashid, Professor, Department of Computer Science and Engineering, University of Dhaka. We also declare that no part of this project has been or is being submitted elsewhere for the award of any degree or diploma.

Countersigned

.....

(Dr. Mamun-or-Rashid)

Supervisor

Signature

.....

(Palash Roy)

.....

(Anika Tahsin)

Abstract

Virtual Network Functions (VNFs) in cloud servers of Fifth Generation (5G) network systems are responsible for executing offloaded codes from mobile users. Placement of VNFs in the cloud is very complicated to get on-time execution service due to many reasons including users' mobility and resource heterogeneity, which often cause VNF relocations from one data center to another. Minimizing service delay (i.e., maximizing user Quality-of-Experience) for the user applications and the number of VNF relocations are the two main design goals of VNF placement problem; however, they do oppose each other. In this paper, we have formulated the above problem as a Multi-objective Integer Linear Programming (MILP), which is proven to be an NP-hard one. The proposed optimization framework trades-off between the number of VNF relocations and user Quality-of-Experience. We then develop an Artificial Intelligence based meta-heuristic Ant Colony Optimization (ACO) algorithm to achieve sub-optimal placement of VNFs within polynomial time. The performance analysis results, carried out in Cloudsim, depict that the proposed system outperforms the state-of-the-art works significantly in terms of user satisfaction and VNF relocation overhead.

Acknowledgements

All praise is to the Almighty, who is the most gracious and most merciful. There is no power and no strength except with Him.

Our deep gratitude goes to our thesis supervisor, Dr. Mamun-or-Rashid, Professor, Department of Computer Science and Engineering, University of Dhaka, for his proper guidance in our research field. He has shared his expert knowledge gathered from working in the field over an extensive period, and has been an integral support in our thesis work by constantly keeping updates and urging us to do something significant.

We are deeply indebted to Tamal Adhikary, Lecturer, Computer Science and Engineering, University of Dhaka and Dr. Md. Abdur Razzaque, Professor, Department of Computer Science and Engineering, University of Dhaka for their support and help for our work. The discussions with them on various topics have helped us to enrich our knowledge and concept regarding this work.

We want to thank our families and friends for their unwavering love and support. The opportunities that our parents have made possible for us determines the personalities we have built and the work that we produce today.

Lastly, we want to thank the Department of Computer Science and Engineering, University of Dhaka, its faculty, staff, and all other individuals related to the department. The department has facilitated us throughout our undergraduate program and subsequent thesis, and has also formed the base for our future endeavours.

Palash Roy
Anika Tahsin
January, 2020

Contents

List of Figures

List of Tables

List of Algorithms

Abbreviations

DC	**D**ata **C**enter
SDN	**S**oftware **D**efined **N**etwork
NFV	**N**etwork **F**unction **V**irtualization
VNF	**V**irtual **N**etwork **F**unction
VM	**V**irtual **M**achine
MCC	**M**obile **C**loud **C**omputing
MILP	**M**ulti-objective **I**nteger **L**inear **P**rogramming
ACO	**A**nt **C**olony **O**ptimization
QoS	**Q**uality-**o**f-**S**ervice
QoE	**Q**uality-**o**f-**E**xperience
GAP	**G**eneralized **A**ssignemnt **P**roblem
RAN	**R**adio **A**ccess **N**etwork
UE	**U**ser **E**quipment
S-GW	**S**erving **G**ate**w**ay
PDN-GW	**P**acket **D**ata **N**etwork **G**ate**w**ay
eNB	**e**volved **N**ode-**B**
IoT	**I**nternet **o**f **T**hings

Chapter 1

Introduction

1.1 Introduction

The Internet has become an inextricable part of our day-to-day life in recent times. The number of devices connected to the Internet is getting increased rapidly [1]. Almost all types of instruments from communication devices to home appliances like TV, washing machine, toaster, etc. have started to be connected to the Internet [2] [3]. The role of Fifth Generation (5G) cellular network is expected to be very promising for accommodating increasing reliability requirements on Internet-centric mobile applications [4] [5] [6]. 5G heterogeneous network (HetNet) is anticipated to provide more lucrative features such as higher throughput, lower latency, higher mobility range, massive device connectivity, higher network capacity and energy efficiency. It provides up to 20 times accelerated downloading and uploading speeds than 4G, 10 times lesser round trip latency and bandwidth up to 1 Gbps compared to only 20 Mbps in 4G [7].

Software Defined Network (SDN) and Network Function Virtualization (NFV) are the two influential key technologies which contribute significantly for developing the architectural design of 5G heterogeneous network [8] [9]. Virtualization is a middle layer technology between hardware and software layers which creates virtual representation of something such as virtual machines, servers, memory, network functions, etc. The NFV offers the advantage of segregating the network functions from proprietary hardware appliances and executing these functions in

software on standardized hardware instead. These decoupled network functions are referred to as Virtual Network Functions (VNFs) [10].

The NFV offered by Cloud Computing has obtained much popularity as constantly growing number of enterprises and individuals are offloading their workloads to cloud service providers and getting served by them [11]. This technology is also being taken into account by 5G mobile operators to deal with the increasing number of data traffics and data intensive applications [12]. In Mobile Cloud Computing (MCC), because of mobile devices having low computational power and battery lifespan, most of the applications are executed on various VNFs operating in different high computational data centers (DCs) in the cloud [13] [14].

1.1.1 5th Generation Network (5G)

5th Generation (5G) wireless network is very emerging technology in the recent world. 5G supports high frequency wave ranging from 3 to 300 GHz. It provides higher data rate, higher throughput and almost zero latency. 5G network technology provides 1 ms round trip latency which is almost 10 times faster than 4G. In 5G, larger base stations are split up to several small cells like micro cell, pico cell, femto cell. These small cells are known as miniature base stations. Due to this heterogeneous small cells, 5G provides enormous number of connected device that helps us to handle wireless traffic explosion. 5G provides energy aware smart base stations that reduces almost 90% energy than the previous one. Integration of both lower and higher frequency band, it brings revolutionary changes in the modern world. Lower bands handle basic coverage and higher bands are responsible for higher data rates. Device to device communication, IoT, vehicular communications, healthcare applications and all other novel emerging applications form the major driving force behind 5G [7].

1.1.1.1 Significant Applications of 5G

We can't think of our life without Internet. However, human interaction devices failed to provide better performance due to higher response time and lower uploading and downloading speed. 5G network is the key enablers to improve better performance of those devices. Some of the most significant applications of 5G are described below.

- **Smart Factory/ Industry 4.0:** 5G brings revolutionary change in industry. It helps us to take real time decisions, improves operational visibility and efficiency due to greater spectrum availability. Due to the real time visibility, 5G optimizes production by connecting all equipment in real time and avoids costly disruption.

- **Transportation:** 5G Internet of Things (IoT) brings significant changes in transport management and transit system. Users can now get real time traffic data within a few milliseconds. By improving efficiency and accuracy for asset tracking and driving management, it helps us to build smart cities. By collecting real time data using smart sensors, 5G IoT solutions help us better traffic and parking management.

- **Smart City:** Forth industrial revolution has created a wide scope for the evolution of smart cities by forming advanced urban architecture that uses data for enhancement of the quality of the daily life of citizens. However, the adoption of smart cities faces various problems and 5G wireless network can drive the smart urban services to a new level. With reduced latency, wider bandwidth, enhanced device connectivity and more reliability, deployment of 5G will enhance the functionalities of smart cities such as increased security, enhanced emergency response, better traffic management, energy efficient homes, developed city infrastructure management etc.

- **Health-care:** Due to higher data rates and lower latency, patients can take real time service and treatment remotely. Health-care companies are expanding their in-home services for aging and disabled people by monitoring their conditions remotely. 5G IoT-enabled wearable devices also help us to monitor fitness, health and wellness factor. This is another revolutionary change occurred by 5G.

- **Agriculture:** 5G-IoT enables smart farming solutions by introducing smart sensors, smart gateways and monitoring system to collect and analyze real time information. Crop observation, irrigation management, storage management, precision farming are the examples of smart farming system. Using sensor data, we can now improve the monitoring system and manage irrigation system to keep up-to demand of the larger populations.

- **Robotics & Drones:** 5G network devices have lower response time than 4G. For that reason, 5G enabled robots can analyze more data in real time,

absorb new things in few amount of time and can better communicate with the real world. Due to higher bandwidth, 5G enabled drones can shoot 4k or 360 degree videos more efficiently and bring emerging changes in the real world.

- **Artificial Intelligence:** Due to advancement of AI technology, complex deep learning based AI algorithms have been applied in today's application. This has been possible due to the petabytes of data generated by networks within few milliseconds. Autonomous cars, robotics, automation, several intelligent applications of mobile devices are driven by AI.

- **Entertainment & Multimedia :** At present, a huge amount of mobile internet traffic is being occupied by video download and streaming. 5G is anticipated to add a new dimension to entertainment sector by offering a high definition virtual world on mobile devices, high speed 4K video streaming with crystal clear audio quality, high quality live streaming without interruption, providing access to HD TV channels from mobile devices, offering improved Virtual Reality (VR) based gaming experience etc.

1.1.2 evolved Node-B

The term base station is very common in mobile telephony and wireless communication. In wireless communication, it is known as transceiver by which a lot of devices or mobile devices are connected. In 5G network, this base stations are split up to several small cells i.e., pico cells, micro cells, femto cells or miniature base station. These small cells are very useful in densely populated areas such as sports venues, airports, shopping centers and train stations. These cells are known as evolved Node-B (eNB). The previous base stations are known as Node-B which have minimum functionality. Due to having no control functionality, these are controlled by Radio Network Controller (RNC). However, in 5G, base stations have controllers that maintain the control functionality which simplifies the network architecture and offers lower response time. For that reason, these base stations are known as eNB.

1.1.3 Virtual Network Function

In Network Function Virtualization (NFV) technology, Virtual Network Functions (VNFs) are responsible for handling network functions that run on Virtual Machine (VM) rather than carried out by dedicated hardwares [10]. In our thesis, we consider VNF can handle all types of network functions like firewalls, social networks, high computational applications and all types of mobile applications that run in the Data Centers (DCs) rather than individual mobile devices. The main benefit of using VNF is that new services or applications have not be installed or configured manually in all hardwares. For that reason, VNFs can help us to improve network scalability, reliability and better use of network resources. It also helps us to reduce power consumption of the mobile devices and reduces operational cost.

1.1.4 Mobile Cloud Computing

Mobile Cloud Computing (MCC) uses cloud computing that provides extensive computational resources to mobile devices, network operators and cloud service providers. It saves huge amount of battery life of the mobile devices and improves reliability and flexibility by backing up information in the cloud. It also enables users to securely collect and integrate data from various sources. The basic difference between MCC and cloud computing is that cloud computing is simply storing and accessing data from the cloud whereas, MCC allows transmission of voice, data or any other services by using any wireless device without connecting to a physical link. MCC is very effective for high performance computational applications like image processing, natural language processing, sharing GPS/ network data, social networking, multimedia search and sensor data applications [15].

1.2 Motivation

5G integrated with Mobile Cloud Computing offers us high speed computation of applications in comparatively less amount of time and ensures improved user experience. However, the user experience and service performance depend highly on the proper placement of VNFs in the data centers. Due to user mobility, static placement of these VNFs will not work and will degrade the user experience as well

as overall performance. Therefore, dynamic and optimal allocation of the VNFs in the data centers satisfying all the performance parameters is a troublesome task in 5G because of user mobility.

Resource management in cloud computing has been well studied in many research problems [16] [17]. However, because of user mobility, allocation of VNFs for running user applications in different data centers is immensely challenging and difficult task in 5G. For optimal placement of VNFs, various parameters are needed to be considered such as total number of VNF relocation, communication delay as well as response time to get service, its cost, etc. If performance of one parameter is attempted to be improved, performances of some other parameters degrade as a consequence. For example, an attempt to minimize the number of VNF relocation results in increased communication delay as well as response time and vice-versa. So, optimal allocation of resources in different data centers satisfying all the performance parameters is a major challenge.

The problem of optimal allocation of the resources from the mobile devices to the DC has been well studied in many papers. However, these existing approaches in the literature encounter several limitations. In [18], the authors have studied the resource allocation problem in the case of static users but the approach cannot be feasibly used when there exists user mobility. For minimizing the load of the Virtual Machines (VMs), the authors in [19] have placed the VNFs efficiently in the DCs but minimizing VNF relocations and service cost have not been considered. However, in A-SGWR [20], minimization of the number of VNF relocations have been taken into account but not the response time for getting the service. In S-PL [20], the authors focused on minimizing the total response time but the total number of relocations and its overhead have been ignored.

Therefore, in our project, we propose optimal and meta-heuristic VNF placement algorithm in the data centers considering both the conflicting objectives that are, minimizing the number of VNF relocation and minimizing total response time to get service.

1.3 Problem Formulation

Fig. 1.1 shows the basic concept of user services architecture in 5G. The user equipments (UEs) are connected to respective evolved Node-Bs (eNBs) of their

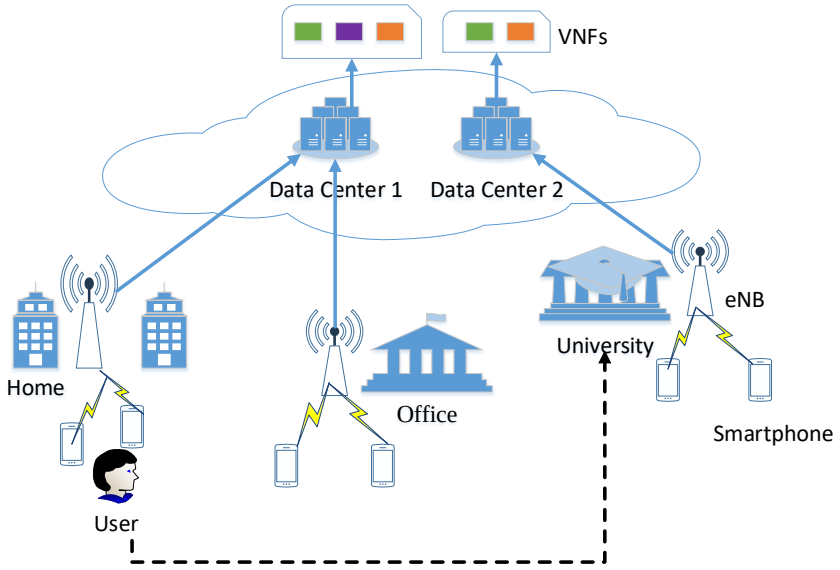

FIGURE 1.1: User Services Architecture in 5G

service area in the network. The cloud domain comprises a number of data centers that serve the eNBs in executing their user applications. Users of an eNB are served by VNFs of exactly one data center. In Fig. 1.1, for example, associated eNBs of the home and university is connected to data center 1 and data center 2, respectively. When any user moves from his/her home to university, the placement of the running VNFs of that user becomes a matter of concern.

The VNF relocation is only necessary in cases when a user moves between two eNBs connected to different data centers. In such case, VNF can be relocated from previous DC to service DC minimizing the communication path between the serving DC and new eNB, which in turn minimize the response time. Alternatively, user can get service from the DC on which VNF was running via the current serving DC. However, this second way causes higher response time. Therefore, minimizing VNF relocations and minimizing communication delay are two conflicting objectives. Artificial Intelligence (AI) is a branch of computer science which enables the development of computer programs which possess the ability to make decisions rationally and solve problem by learning from experiences and improving it gradually [21]. Due to the success of AI in solving complex control and decision-making problems, it is anticipated to contribute significantly for developing various aspects of 5G network and solving complicated problems. In this project, we bring trade-off between these two aforementioned conflicting objectives using Multi-objective Integer Linear Programming (MILP) and Ant Colony

Optimization (ACO) based VNF allocation algorithm.

1.4 Research Objectives

In this project, we have developed an optimal and meta-heuristic placement of VNFs in the DCs that trades-off between minimizing VNFs relocation and minimizing response time due to user mobility. The main objectives of this project are:

- Development of an optimal placement algorithm for the Virtual Network Functions (VNFs) in the DCs to increase QoE of the users.
- Maximizing QoE i.e., minimizing response time to get a service.
- Minimizing the number of VNF relocations i.e., migration cost to migrate a VNF from one DC to another.
- To investigate the boundary performances for maximizing QoE or decreasing the cost of operation.
- To study comprehensive performance of the proposed algorithm with the state-of-the-art works.

1.5 Contributions

In our work, our aim is to allocate the VNFs in the DC for the mobile users maintaining a trade-off between VNFs relocation and response time. The main contributions of this project are summarised as follows:

- We formulate the problem of allocating the VNFs associated with user mobility as a Multi-objective Integer Linear Programming (MILP) problem.
- We have brought trade off between minimizing the number of VNF relocations and minimizing the total response time (hereafter, we call our system TradeRC) to get the service ensuring user Quality of Service.

- Due to the NP-hardness of our proposed MILP system, we develop an Ant Colony Optimization (ACO) meta-heuristic based VNF placement algorithm. The operational principal of the proposed system is driven by learning from the previous experiences.
- We implement and simulate our proposed system TradeRC in CloudSim [22] and compare it with other state-of-the-art works. The results state that the user QoE in TradeRC has been increased by about 25% and relocation overhead decreased by about 15% compared to other state-of-the-art works.

1.6 Organization

The rest of the project report is organized as follows. Chapter 2 presents some related research works. Chapter 3 describes the system model, problem formulation and solution details of our proposed system. In this chapter, we also present ACO-based VNF allocation problem in the data center. Subsequently, Chapter 4 demonstrates the performance of the proposed optimization solution and comparison with other state-of-the-art works. Finally, Chapter 5 summarizes our work and outlines of the future work plans.

Chapter 2

Related Works

2.1 Introduction

VNF placement is a very emerging research problem in 5G heterogeneous network. The VNF placement problem can be modelled as a type of Virtual Machines (VMs) migration problem since VNFs are run on VM. A wide plethora of research works related to VM migration in distributed cloud or hybrid cloud have been addressed in [23–27]. As NFV integrated with SDN is becoming an emerging technology for development of the 5G network architecture, placement of VNF has gained much attention of many researchers. In this chapter, we discuss the existing works in the literature on 5G cellular network, Software Defined Network (SDN), mobility behaviour of the users and optimal VM/VNF migration in 5G cellular network. For our knowledge base, we have gone through several research articles and a few books on the topic. Here we discuss some thesis works and some recent articles dealing with VNF placement problem. Related works on this field are described as follows.

2.2 Centralized SDN Architecture for 5G

Addressed Problem: In order to satisfy the increasing demand of higher data rate and low latency cellular broadband applications, the need for 5G technology is growing rapidly. This paper proposes modification in the 5G architecture standardized by the Third Generation Partnership Project (3GPP) to apply the

principle of Software Defined Network (SDN). It removes the control functionality from Radio Access Network (RAN) to the network core leaving the base stations (known as eNB) as pure data plane node to reduce the signaling cost between the Radio Access Network (RAN) and the network core [28]. They justified effectiveness of their proposed architecture by evaluating Key Performance Indicator (KPI) of 5G networks such as network registration time and handover time.

Method Used: In this paper, they proposed a new architecture for 5G which applies idea of SDN to the RAN function. SDN separates data plane from the control plane. In their proposed architecture, they removed the RAN control functions such as Radio Resource Control (RRC) protocol layer and Radio Resource Management (RRM) functionalities from the gnBs to the network core.

In the proposed architecture, Access and Mobility Management Function (AMF) function from the core and the RRC layer with RRM function from the eNB which is the only control function of eNB are merged together and placed in the 5G core. This new evolved function in the core is known as enhanced AMF (eAMf).

Positive Aspects: The removal of RRC layer with RRM functionality from the eNB to the 5G core eventually reduces the signaling cost as less encoding decoding processes are performed for call registration, user verification etc. This is because of the direct delivery of RRC messages from UE (User Equipment) to the eAMF function embodied in 5G core without any processing at eNBs. The authors evaluated the registration and handover time by simulating both the architectures with the help of ns-3. The results show that proposed 5G architecture gives less registration and handover time the 3GPP's 5G architecture.

Challenges: This paper proposed SDN based 5G architecture but did not work with improvement of specific features of 5G network such as mobility management, resource management, load balancing etc. Improved Performance analysis result should have been presented comparing with other state-of-the-art works.

2.3 SDN Based Resource Management in 5G

Addressed Problem: In the case of heterogeneous network, resource management is one of the major issues. SDN based resource management algorithm is proposed for LTE and 802.11p taking into account velocity and mobility management in [29].

Method Used: Resource management is one of the major challenges for heterogeneous network. In this paper, an improved algorithm is proposed for data offloading in case of new call and handover.

The authors have assumed a network consisting of 802.11p Road Side Unit (RSU) which is for transmitting data between vehicles and roadside infrastructure and LTE nodeBs which are the base stations. Considering mobility issue, they have classified users in two types, pedestrians and vehicles. They have proposed different algorithms for traffic offloading between LTE and 802.11p. They have prioritized handover over new call.

- **Algorithm for resource management for NC (New Call)**: When a new call arrives, the type of user is determined by checking the velocity with a threshold value. Then the new call is either accepted or blocked checking the available channel and the guard channel .If number of available channels exceeds the guard channels it is blocked else accepted.

  In case of pedestrian, if the data to be processed exceeds the threshold amount then the data is offloaded between LTE and 802.11p else it is offloaded to LTE only. And in case of vehicles, if the data to be processed exceeds the threshold amount then the data is offloaded between LTE and 802.11p else it is offloaded to 802.11p only.

- **Algorithm for resource management for HO (Handover) calls**: The algorithm is similar as for new call except the blocking part. The check for available channel and comparing it with guard channel number for blocking or acceptance of that call is omitted here.

Positive Aspects: The offloading of resources between LTE and 802.11p helps to meet the requirements of delay and loss sensitive data flow. And the authors

have taken into consideration different types of users and velocities which improves the process of offloading.

Challenges: In this paper, the deployment process of SDN in 5G has not been discussed. The authors did not consider resource management for other applications besides new calls and handover calls.

2.4 Efficient Task Migration Policies for Cloud Computing Systems

Addressed Problem: Due to limited processing power and energy, mobile devices offload their tasks in the cloud. Efficient task management is very complicated due to dynamic nature of the tasks because the pattern of the task varies with time and location. The capacity of the VMs are is also time dependent that can be increased, decreased or constant with time. Besides these, user mobility provides additional challenges for task migration. In this paper [30], the authors have investigated the lightweight task migration techniques for the shared resources of the MCC. Due to user mobility, each time a user is attested to only certain number of servers those are attested to the current BS. Each time migration decisions are taken according to the load of the cloud and anticipated execution time.

Method Used: The execution of a task is dependent on number of different parameters like user mobility, task arrival time and effect of multitenancy. Three types of task migration techniques are developed in this paper.

- **Cloud-wide task migration,** where all migration decisions are taken by the central cloud that maximizes the objectives of the cloud provider. When migrating several conditions are maintained i) tasks of increasing data volume are given priority to migrate ii) tasks that have significant amount of residual processing burden are given priority to migrate iii) tasks that experience momentous multitenancy cost are preferred to migrate.
- **Server-centric task migration,** where migration control engine resides in each server and makes decision from that server where the task is currently

executing. It has lower complexity than cloud-wide task migration. Each server periodically check whether execution time can be improved by migrating a task to a new server. There needs to synchronization among the server.

- **Task-based migration,** where migration decision is taken by the task itself in order to minimizing its execution time.

Positive Aspects: In this paper, three types of solutions are proposed for task migration to minimize execution time in the cloud. These methods consider interaction of co-located tasks in the cloud and migration cost to migrate a task. By varying the number of tasks and link capacity, cloud wide task migration performs better than other two approaches in terms of total number of migrations and average task lifetime.

Challenges: The authors have not considered any application deadline to get the services. They also did not consider the DC's capacity to place a VNF in a DC to get on-time service. Users can take service from the current DC without migrating VNF from the previous one. They also didn't consider it. In task-centric migration method, an individual task is migrated to another cloud considering user mobility and load of the cloudlets. However, it failed to utilize cloud resources effectively and the performance of successful task execution within deadline is very poor.

2.5 User Mobility-Aware VNF placement in 5G

Addressed Problem: Due to user mobility, optimal placement of VNF (Virtual network Function) that forms a virtual network infrastructure (VNI) within the same data center is the main concern in [20]. On each data center, one or more VNFs of PDN-GWs (Public Data Network Gateways) and S-GWs (Serving Gateways) can be initiated that are created on demand and in a dynamic manner to form a VNI. Some points are considered to solve this problem:

- A UE(user equipment) can not have more than one S-GW at the same time.

- When an UE leaves its current service area, it needs to change its S-GW. This is called S-GW relocation and want to avoid this as much as possible.
- The path between UE and its corresponding PDN-GW has to be shortened.

This two conflicting objectives are deal with considering the mobility features and mobile users behavioural patterns.

Method Used: To solve this problem, two solutions are proposed. 1) In the first solution S-GW relocation cost is given more priority. 2) Second solution given priority on high QoE but shortening path between eNBs and PDN-GW.

- **A-SGWR: Avoiding S-GW Relocation** In this solution, min-max approach is used to minimize the S-GW relocation overhead. Initially maximum delay tolerated Delay_{MAX} by this network can be set infinity. In this case, optimal solution is obtained when all eNBs to the same data center reduce the S-GW relocation overhead.
- **S-PL: Shortening Path Length between eNBs and PDN-GW** The authors have wanted to optimize the path cost of the communication path between UEs and their respective PDN-GW VNF. Initially, maximum amount of S-GW relocation overhead SGWR_{MAX} by this network can be set infinity. In this case, the optimal solution would converge when the path cost of all eNBs associating to its nearest neighbour is minimal.

To evaluate this solutions, several simulation tool based on CPLEX, Matlab and CVX are used.

Positive Aspects: In this paper, a set of solutions to the problem of VNF placement on federated cloud to create efficient VNI are proposed. This proposed solutions tackle two conflicting objectives.

Challenges: In the first solution, communication delay can be infinite, which is not practical. In the second approach, number of relocations can be infinite. This is not feasible in real life. These are the major limitations in these two methods. The authors did not take into account resource capacity of each DC. They considered infinite amount of resources of each DC.

2.6 Dynamic Resource Allocation Exploiting Mobility Prediction in 5G

Addressed Problem: In this paper, to run high computing powerful application and extending battery life of the mobile devices, there is a possibility to offload computing task in the cloud. In Mobile Edge Computing (MEC), tiny sized data centers are placed in the eNBs [19]. Therefore, low latency applications or real time applications can easily get service from that DC. To compute a task in the cloud, a VM is created to allocated computed resources in the cloud. Due to users mobility, optimal allocation of the resources at the base station changes over time. The authors solved users mobility by VM migration and finding a new communication path between VM and UE by fixing the VM at its previous location.

Method Used: To find the user's next predicted position, they used SINR (Signal to Noise Inference Ratio) mapping. If the SINR value is increasing than the previous SINR value, the user is going far away. To offload a task when any user is moving, VM should be always ready to process that task. When offloading any task, we need uploading delay, processing delay, downloading delay, VM creation delay and VM starting delay. By using SINR value, they predicted the user's next possible position and placed the VM in that position.

Positive Aspects: The authors have shown that the average offloading delay is reduced in significant amount than other previously proposed algorithms. The authors have also shown that energy consumption is reduced by 9%.

Challenges: In this paper, tiny data centers are created in each base station. So, when any user moves form one eNB to another eNB, we need to relocate VM many number of times. Therefore, in this context, MEC is better for low latency service application. However, the number of relocations is increased. Creating these VMs and keeping tiny sized data center at each eNB is very costly. The authors have not considered any resources capacity of the DCs.

2.7 Research Gap

Many state-of-the-art works have proposed solutions to the VNFs placement problem. The authors in [18] have solved the problems for static users only. However, offloading the task in the data centers is very computationally expensive when the user is mobile. Subsequently, in order to minimize total resource cost, Zhang et al. [31] have proposed a framework for dynamic service placement in multiple data centers. They have only considered total response time between a DC and user location. However, they didn't consider any maximum number of resource allocation under a DC. Taleb et al. [20] have proposed algorithms dealt with two conflicting objectives, namely minimizing the path distance between PDN-GW and UE and minimizing the S-GW relocation by providing individual solutions for achieving these two objectives separately. In Shortening Path Length (S-PL) system, they have minimized the path delay between UE and PDN-GW. However, they did not impose any limit on the number of S-GW relocations tolerated in the network. On the other hand, In avoiding S-GW relocation (A-SGWR) system, they have tried to minimize the number of S-GW relocations. However, they did not impose any limitations on the amount of allowable communication delay in this case. Moreover, they have considered infinite capacity of the data centers for placing the VNFs. Considering all of the issues, we want to develop a solution bringing trade-off between the number of VNFs relocations and the user QoE due to user mobility.

2.8 Conclusion

From the above discussions, we can see that, a lot of works have been done for solving VNF allocation problem in 5G HetNet. However, none of the existing works focused on the efficient VNFs allocation system due to user mobility considering both minimizing number of relocations and response time. The number of relocations increases as a consequence of concentrating only on minimizing response times. Again, taking service from the previous DC reduces response time which in turn maximizes the number of relocations. These observations have driven us to make a framework that trades-off between minimizing the number of VNF relocations and increasing the user Quality-of-Experience.

Chapter 3

Proposed Approach

3.1 Introduction

In this chapter, we focus on the description of our proposed approach to solve the above mentioned VNFs placement problem. At first, the system architecture of VNF services in the 5G network and various assumptions are described. Then, an optimization framework to solve the VNFs placement problem, proof of NP-hardness of that proposed optimization model and a meta-heuristic algorithm for VNFs allocation have been presented in details.

3.2 System Model and Assumptions

Fig. 3.1 represents the system architecture of the network. The system architecture consists of two domains. One is the cloud domain and another one is the Radio Access Network (RAN) domain. The could domain comprises a set of data centers (DCs), D. These data centers have strong wired connection among them and the DCs can take services from one another through exploiting cloud confederation [32].

The RAN domain consists of a set of access points, i.e. base stations called evolved Node-B (eNB). Several users can be connected to an eNB through radio signal. A group of eNBs are connected to a base station controller. An eNB is connected to exactly one data center through the Serving Gateway (S-GW) and Packet Data

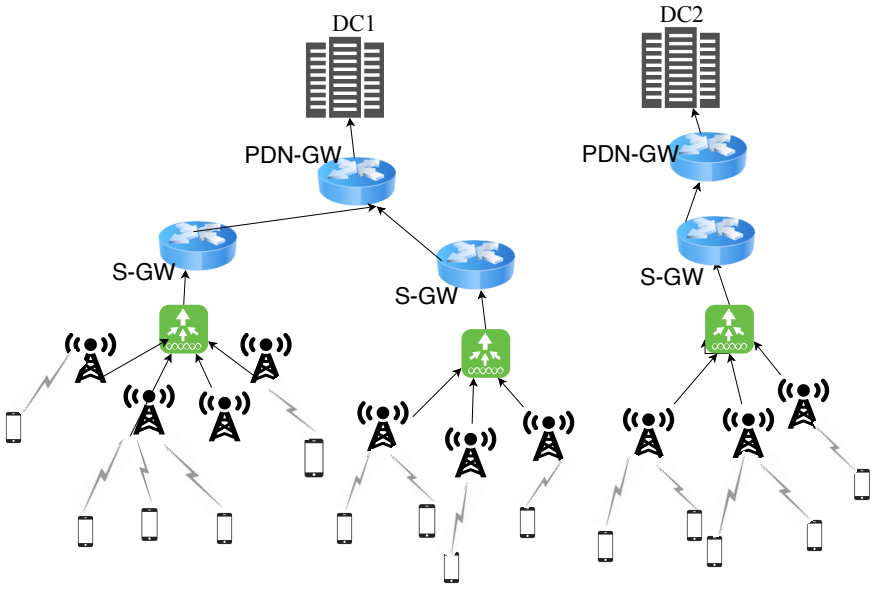

FIGURE 3.1: System Architecture of VNF services in 5G

Network Gateway (PDN-GW) of that DC [33]. Each DC has multiple number of eNBs connected to them. In some areas (busy areas), DCs serve huge number of eNBs and some areas (lightly loaded) DCs serve a small number of eNBs connected to them [34].

Different Virtual Network Functions (VNFs) are run on different data centers. Each DC has a fixed capacity of holding application/service oriented VNFs. An eNB gets service from the the data center it is connected to. N is the set of all eNBs that are connected to the DC where the system is running. The data center itself can run the respective VNF of a user under that eNB or it can manage the service from any neighbor data center. The second case incurs extra service cost. V_j is the set of all VNFs of eNB $j \in N$ that are needed to be considered for relocation and these VNF requests have been occurred for hand off between eNB $j \in N$ and any other eNB which is connected to a different DC. The major notations and their descriptions used to design TradeRC are listed in Table. 3.1.

3.3 Design of TradeRC

In this section, we present an optimization framework of our proposed TradeRC system and develop a meta-heuristic AI based Ant Colony Optimization (ACO) algorithm. The main focus of TradeRC system is to provide an optimal placement

TABLE 3.1: Notation Table

Notation	Description
D	The set of all data centers in the network
N	The set of all eNBs that are connected to the DC where the system is running.
V_j	The set of VNFs of eNB $j \in N$ that need to be considered for relocation
δ_{worst}	Maximum communication delay tolerated by the network
t_k	The communication delay between DC $k \in D$ and the DC where the system is running
t_j	The communication delay between eNB $j \in N$ and the DC where the system is running
S_f	Size of Virtual Network Function VNF $f \in V$
ϕ_k	Cost to relocate any VNF to DC $k \in D$
σ_k	Cost of taking service from DC $k \in D$
ζ_k	VNF holding capacity of DC $k \in D$
φ_f	Execution time of VNF $f \in V$
γ	Priority factor for VNF relocation
τ_f	Transfer time of VNF $f \in V$
X_k	Number of VNFs currently executing in DC $k \in D$
$Y_{k,j}^f$	Summation of relocation time, communication time and execution time because of allocating VNF $f \in V_j$ of eNB $j \in N$ to DC $k \in D$
$\mathcal{T}_o$	Initial pheromone value
$\mathcal{T}_{j,f}^k$	Pheromone value for placing VNF $f \in V_j$ of eNB $j \in N$ to DC $k \in D$
$\mathcal{H}_{j,f}^k$	Local heuristic value for placing VNF $f \in V_j$ of eNB $j \in N$ to DC $k \in D$
$p_{j,f,k}^z$	Probability of choosing DC $k \in D$ for placing VNF $f \in V_j$ of eNB $j \in N$ by ant z
ρ_l	Weight parameter for local pheromone update
ρ_g	Weight parameter for global pheromone update
$\Delta\mathcal{T}_{j,f}^k$	Global pheromone value for placing VNF $f \in V_j$ of eNB $j \in N$ to DC $k \in D$ in global solution.
α	Weight parameter for pheromone value
β	Weight parameter for local heuristic value

of the VNFs that are needed to be considered for relocation because of user mobility. The system will run in each DC to manage the VNF requests of the eNBs under it that have been come from other eNBs connected to a different DC because of hand off due to user mobility. We have two main objectives: 1) minimizing the number of VNF relocations and 2) minimizing total communication delay.

As these two are conflicting objectives, we have brought a trade off between them using priority factor for both. The priority factor can be changed according to the requirement of the network.

3.3.1 Optimal Placement of VNFs

For placing VNF $f \in V_j$ of eNB $j \in N$ to DC $k \in D$, the relocation time, denoted by $R^f_{k,j}$, is calculated as follows:

$$R^f_{k,j} = \{(1 - p^f_k) \times b^f_{k,j}\} \times H^f_k, \tag{3.1}$$

where, p^f_k is a binary variable. If VNF $f \in V_j$ is running on data center $k \in D$ before starting the solution, then p^f_k is 1 otherwise it is 0. Similarly, $b^f_{k,j}$ is also a binary variable. If we locate VNF $f \in V_j$ for eNB $j \in N$ in data center $k \in D$, then it is 1; otherwise, it is 0. So, if $(1-p^f_k)$ is 1 then we have an option to relocate VNF on another DC otherwise not. If the value of $\{(1 - p^f_k) \times b^f_{k,j}\}$ is 1, then we are relocating the VNF $f \in V_j$ to DC $k \in D$ and we need to calculate relocation time H^f_k for relocating VNF $f \in V_j$ to DC $k \in D$. The relocation cost H^f_k is calculated as,

$$H^f_k = (1 - n^f_k) \times \tau_f, \tag{3.2}$$

where, n^f_k is a binary variable. If the expected VNF $f \in V_j$ is running on DC $k \in D$, then it is 1 otherwise 0. We can get service for VNF $f \in V_j$ from that DC without creating any instance for that VNF. On the other hand, if n^f_k is 0, then there is no running VNF $f \in V_j$ on DC $k \in D$. So, we need to transfer that VNF from the previous DC. Therefore, if the value of $(1-n^f_k)$ is 1, then we transfer the VNF from the previous DC otherwise not. If this value is 1, then relocation cost is equal to the transfer time of that VNF. Transfer time of the VNF is calculated as,

$$\tau_f = \frac{S_f}{r}, \tag{3.3}$$

where, S_f is the size of the VNF and r is achievable data rate to transfer a VNF from one DC to another DC. Communication delay to get service for a VNF is calculated as,

$$C^f_{k,j} = b^f_{k,j} \times (t_j + t_k), \tag{3.4}$$

where, total communication delay is equal to summation of communication delay between eNB_j and the DC where the solution is running denoted by t_j and communication delay between DC $k \in D$ where we are locating the VNF $f \in V$ and the DC where the solution is running. If we take service from the own DC where the solution is running, then we need to calculate only t_j. But if we take service from another DC $k \in D$ then we also need to add t_k with it. Thus the objective function of our work is formulated as,
$Minimize$:

$$W = \sum_{j \in N} \sum_{f \in V_j} \sum_{k \in D} \{\gamma \times R^f_{k,j} \times \phi_k + (1 - \gamma) \times C^f_{k,j} \times \sigma_k\}. \tag{3.5}$$

Here, the objective function is formulated as a Multi-objective Integer Linear Programming (MILP) to be solved by Data center (DC). Minimizing total VNF relocation and minimizing total path cost in terms of communication delay are two conflicting objectives. The objective function is bringing trade off between relocation and communication (hereafter, we call our proposed system TradeRC) using weight factor γ. There is some extra cost to get services from the global DC than local DC. That is, if any eNB takes services from the distant DCs other than the DC to which they are directly connected, the sevice cost increases. There is also some cost to relocate VNF to any DC $k \in D$ denoted by ϕ_k . Minimizing the overall network cost is our main objective. There are some constraints that need to be satisfied. Subject to:

$$b^f_{k,j} = \{0, 1\}, \ \ \forall j \in N, \ \ \forall f \in V_j, \ \ \forall k \in D \tag{3.6}$$

$$\sum_{k \in D} b^f_{k,j} = 1, \ \ \forall j \in N, \ \ \forall f \in V_j \tag{3.7}$$

$$\sum_{f \in V_j} \sum_{k \in D} b^f_{k,j} = |V_j|, \ \ \forall j \in N \tag{3.8}$$

$$\sum_{k \in D} (R^f_{k,j} + C^f_{k,j} + \varphi_f) \leq \delta_{worst}, \ \ \forall j \in N, \ \ \forall f \in V_j \tag{3.9}$$

$$\sum_{j \in N} \sum_{f \in V_j} b^f_{k,j} \leq \zeta_k, \ \ \forall k \in D \tag{3.10}$$

Constraint in Eq. 3.6 is a binary constraint. If VNF $f \in V_j$ for eNB $j \in N$ is located in DC $k \in D$, then it is 1; otherwise 0. Constraint in Eq. 3.7 is an atomicity

constraint. It ensures that every VNF $f \in V_j$ for every eNB $j \in N$ exists exactly in only one data center $k \in D$. Constraint in Eq. 3.8 is the allocation constraint. It ensures that, all VNF $f \in V_j$ that comes from an eNB $j \in N$ should be allocated to some data centers. No VNF can be stayed unallocated to any DC. The QoS constraint in Eq. 3.9 ensures that, summation of communication delay, execution time and relocation time must be less than the maximum allowable application deadline. The allowable delay can vary from application to application. Execution time is calculated as;

$$\varphi_f = \frac{I_f}{MIPS_k}, \forall k \in D \tag{3.11}$$

The capacity constraint in Eq. 3.10 ensures that, total number of the VNFs of a DC can't be more than the maximum capacity of the DC.

Theorem 3.1. *VNF placement problem in TradeRC, formulated in Eq. 3.5, is NP-hard.*

Proof. The above optimization formulation is a Multi-objective Linear Optimization having two conflicting objectives, i.e, minimizing number of VNF relocations and communication time. Such problem is NP-hard and can not provide solutions in polynomial time for increasing number of VNFs and eNBs. The optimization problem can be reduced to a Generalized Assignment Problem (GAP) which is NP-Complete [35]. The Generalized Assignment Problem has the following components,

1. A set $T = \{1, ...n\}$ of tasks.
2. A set $M = \{1, ...m\}$ of agents.
3. A cost function, C, giving the cost of assigning a task to an agent.
4. A capacity function, A, giving the capacity used when a task is assigned to an agent.
5. A available capacity function, B, giving the available capacity of an agent.

The problem tries to minimize overall cost of performing the tasks which is given as,

Minimize :

$$\sum_{i \in T} \sum_{j \in M} C_{ij} X_{ij} \tag{3.12}$$

Subject to,

$$\sum_{j \in N} A_{ij} X_{ij} < B_i, \quad \forall i \in M \tag{3.13}$$

$$\sum_{i \in M} X_{ij} = 1, \quad \forall j \in T \tag{3.14}$$

$$X_{ij} \in \{0, 1\}, \quad \forall i \in M, \forall j \in T \tag{3.15}$$

The optimal VNF allocation problem can be reduced to Generalized Assignment Problem (GAP) by leveraging constraints and considering the problem for a single eNB. Let, $Z_{kf} = H_k^f(1 - p_k^f) \times \gamma \times \phi_k$. Considering only relocation part, the optimization problem takes the form,

Minimize :

$$\sum_{f \in V_j} \sum_{k \in D} Z_{kf} b_k^f \tag{3.16}$$

Subject to,

$$\sum_{f \in V_j} b_k^f \leq \zeta_k, \ \ \forall k \in D \tag{3.17}$$

$$\sum_{k \in D} b_k^f = 1, \quad \forall f \in V_j \tag{3.18}$$

$$b_k^f \in \{0, 1\}, \quad \forall f \in V_j, \forall k \in D \tag{3.19}$$

As we can reduce the problem into GAP, it can be safely ensured that the proposed formulation is at least as hard as GAP and no optimal solution is found in polynomial time for large networks. □

In support of evidence, we have carried out a simulation experiment. Fig. 3.2 shows the impact of the number of VNFs movement per eNB due to user mobility

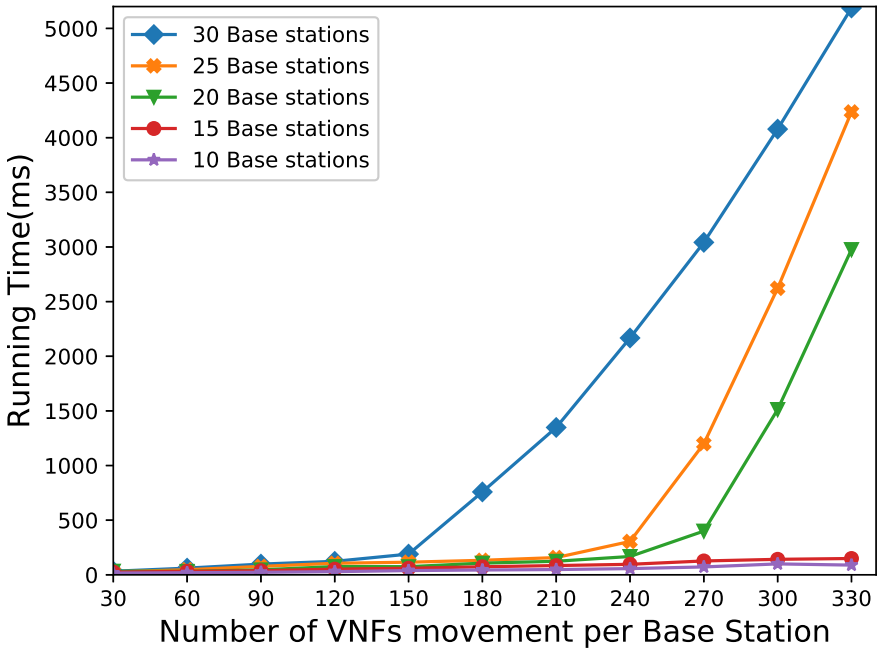

FIGURE 3.2: Impacts of number of VNFs movement per eNB on running time

on running time. To find the boundary values of the number of the eNBs per DC and number of VNFs movement per eNB in the environment, we simulate the objective function in Eq. 3.5 in the NEOS optimization server [36]. The result shows that, 10ms to 100ms are required for 10 to 15 eNBs with increased number of VNFs movement. As the number of eNBs and VNFs movement increases, algorithm running time increases exponentially. For 30 eNBs under a DC and 300 VNFs movement from an eNB, running time is on an average 4000 ms to 5000 ms which is not practical in real life to get a service.

3.3.2 Meta-Heuristic VNFs Placement

Due to the NP-hardness of the above optimization problem, we have provided meta-heuristic Ant Colony Optimization (ACO) based VNFs placement algorithm to solve this problem. We use meta-heuristic algorithm because it refers to master strategy that helps us to modify other heuristics to generate solutions. Heuristics are problem dependent and meta-heuristics are problem independent methods. For that reason, Meta-heuristics algorithms can be applied to a wide range of problems.

The ACO problem is an Artificial Intelligence (AI) based meta-heuristic algorithm that takes inspiration from the behaviour of real ant colonies. In this problem,

Algorithm 1 Ant Colony Based VNF Allocation at each data center $k \in D$

Input: eNB set N, VNF set V_j for each eNB $j \in N$ and data center set D.
Output: VNF-DC pair for each eNB

1: Initialize system parameters α, β, ρ_l, ρ_g
2: Initialize ants set A
3: Generate initial solution using Algorithm 2
4: Calculate initial pheromone value $\mathcal{T}_o$ using Eq. 3.20
5: Set maximum iteration MAX-IT
6: **while** ($iteration \leq$ MAX-IT) **do**
7: **for all** ant $a \in A$ **do**
8: **for all** eNB $j \in N$ **do**
9: **for all** VNF $f \in V_j$ **do**
10: Assign VNF $f \in V_j$ for eNB $j \in N$ to DC $k \in D$ using Eq. 3.23
11: **end for**
12: **end for**
13: **for all** VNF $f \in V_j$ **do**
14: Update the local pheromone value using Eq. 3.25
15: **end for**
16: **end for**
17: Update the global pheromone value using Eq. 3.26
18: $iteration = iteration + 1$
19: **end while**
20: **return** VNF-DC pairs for each eNB

a set of agents (virtual ants) are created. These ants have small memory. Each ant tries to build its solution using heuristic value. After that, these ants improve their solutions by exchanging information via pheromones. Each ant updates its local pheromone trail after building local optimal solution. Finally, all ants combine their local optimal solution to build a global optimal solution in distributed manner.

The proposed Ant Colony Optimization Algorithm for VNF placement has been presented in Algorithm 1. At first, various system parameters and ants set have been initialized in lines 1 and 2. Then we generate initial solution set using algorithm 2 in line 3 and calculate the initial pheromone value, $\mathcal{T}_o$, in line 4. Then we produce different solution sets of placing VNFs of all eNBs to suitable DCs for each ant based on pheromone values and local heuristic values in lines 7 to 12. Each time we get a local solution for an ant, we reduce the pheromone values for each element of that solution by local pheromone update in lines 13 to 15, so that variety of solution is generated by the ants. After this is done for all the ants of the ant set, we take the best solution set among the local ones as global solution and update the global pheromone value in line 17.

3.3.2.1 Initial Pheromone Calculation

When an ant moves from one place to another, it leaves a chemical, pheromone. This pheromone is used to mark these paths and helps the following ants to find their team members. In VNF allocation problem, pheromone represents the possibility of keeping a VNF in a DC. Each ant starts with an initial pheromone value for each VNF to DC pair. The initial solution is generated using First-Fit VNF (FF-VNF) allocation algorithm approach, which is listed in Algorithm 2.

In algorithm 2, we form an initial solution set, $\mathcal{S}_o$, by placing the VNFs of all the eNBs to the DCs based on capacity of the DCs and satisfying application deadline on a First-Fit basis. The inputs for this algorithm are set of eNBs under the data center where the algorithm is running denoted by N, the set of VNFs of eNB $j \in N$ that are needed to be considered for relocation due to user mobility denoted by V_j and the set of data centers D. The output of the algorithm is the initial solution set, $\mathcal{S}_0$, that represents in which data center the VNF of a certain eNB is placed initially. At first, the initial solution set $\mathcal{S}_0$ is initialized to an empty set in line 1. Then the first loop begins from line 2 that is iterated for all the eNBs under the data center where the algorithm is running and it finishes in line 12. Inside this first loop, there starts a second loop from line 3 that iterates for all the considerable VNFs of eNB $j \in N$ and it ends in line 11. Again, inside this second loop, there is a third loop beginning from line 4 that iterates for all the data centers in the network and it finishes in line 10. Inside this third loop, there

Algorithm 2 First-Fit VNF Allocation at each data center $k \in D$

Input: eNB set N, VNF set V_j for each eNB $j \in N$ and data center set D
Output: Set of VNF-DC pair for each eNB in the initial solution $\mathcal{S}_0$

1: $\mathcal{S}_0 \leftarrow \emptyset$
2: **for all** eNB $j \in N$ **do**
3: **for all** VNF $f \in V_j$ **do**
4: **for all** DC $k \in D$ **do**
5: **if** $(X_k < \zeta_k$ && $Y_{k,j}^f \leq \delta_{worst})$ **then**
6: $\mathcal{S}_0 \leftarrow \mathcal{S}_0 \bigcup \{(j, f, k)\}$
7: $X_k = X_k + 1$
8: Break
9: **end if**
10: **end for**
11: **end for**
12: **end for**
13: **return** $\mathcal{S}_0$

is a condition in line 5 that is, if the number of VNFs that are currently placed at data center $k \in D$ is less than the VNF holding capacity of of that data center and the response time in case of placing the respective VNF $f \in V_j$ of eNB $j \in N$ to that data center $k \in D$ meets the application deadline, that data center is selected for VNF assignment. In line 6, the assignment is added to initial solution set $\mathcal{S}_0$ and the number of existing VNFs of the selected data center is increased by one in line 6. And the third inner loop is stopped doing futrher iteration for other data centers in line 8. Finally, this algorithm returns the initial solution set, $\mathcal{S}_0$.

The initial pheromone value is calculated by summing the total relocation time and total communication delay of the system and taking the inverse of that value. Initial pheromone value is calculated as follows,

$$\mathcal{T}_o = \sum_{j \in N} \sum_{f \in V_j} \sum_{k \in D} \frac{1}{R^f_{k,j} + C^f_{k,j}} \times \mathcal{Y}^k_{j,f} \tag{3.20}$$

where $\mathcal{Y}^k_{j,f}$ is a binary variable. It is defined as,

$$\mathcal{Y}^k_{j,f} = \begin{cases} 1, & \text{if } (j, f, k) \in \mathcal{S}_0 \\ 0, & \text{otherwise} \end{cases} \tag{3.21}$$

where, $\mathcal{S}_0$ is the initial solution set. If VNF $f \in V_j$ of eNB $j \in N$ is placed to DC $k \in D$ in $\mathcal{S}_0$, then the value is 1, otherwise 0.

The rationale behind the choice is that, we want to provide more pheromone on the path where relocation and communication occur in the initial solution. The main reason to do inverse operation is that, the lower the value of the summation of relocation time and communication delay, we give more pheromone on that path and vice versa.

3.3.2.2 Determining Heuristic Value

For building solution in Ant Colony Optimization, each ant takes the decision of placing a VNF to a data center combinedly based on pheromone values and local heuristic value. This heuristic value is influential as it contributes to the selection of data center to place a VNF for constructing solution. As our goal is to minimize the total number of VNF relocation and minimize communication delay as well as response time, i.e, bringing trade off between the two conflicting objectives, we

have determined our heuristic value in such a way that these two objectives are satisfied. Our local heuristic is defined as,

$$\mathcal{H}_{j,f}^{k} = \frac{1}{\gamma \times R_{k,j}^{f} \times \phi_k + (1-\gamma) \times C_{k,j}^{f} \times \sigma_k} \tag{3.22}$$

The heuristic value for placing VNF $f \in V_j$ of eNB $j \in N$ to data center $k \in D$ is thus determined by Eq. 3.22. As mentioned earlier, γ is the priority factor given to VNF relocation and $(1-\gamma)$ to communication time. From Eq. 3.22, it can be ensured that the lesser the weighted sum of communication and relocation cost for placing a VNF to a data center, the higher will be the value of heuristic. And the favorability of choosing that data center to place a VNF by an ant will be increased.

3.3.2.3 Selection of Data Center

Each ant $a \in A$ selects the best suitable DC $k \in D_c$ to place VNF $f \in V_j$ of eNB $j \in N$ using the following rule, which is called pseudo-random-proportional action rule. Here, D_c is the set of candidate DCs which have capacity available to place that VNF and $D_c \subseteq D$. The pseudo-random-proportional action rule can be defined as,

$$s = \begin{cases} \underset{k \in D_c}{\operatorname{argmax}} \left(\left[\mathcal{T}_{j,f}^{k}\right]^{\alpha} \times \left[\mathcal{H}_{j,f}^{k}\right]^{\beta} \right), \text{if q} \leq q_0 \text{ (exploitation)} \\ S, \qquad\qquad\qquad \text{otherwise (exploration)} \end{cases} \tag{3.23}$$

where. q_0 is a system parameter on the range [0,1] and q is a random variable that is uniformly distributed in [0,1]. $\mathcal{T}_{j,f}^{k}$ is the pheromone value for allocating VNF $f \in V_j$ for eNB $j \in N$ to DC $k \in D_c$. According to the rule, an ant chooses the most suitable DC $k \in D_c$ to place a VNF $f \in V_j$ of eNB $j \in N$ in terms of heuristic value and pheromone trail value with q probability. Here, α and β denote the relative importance of pheromone value and heuristic value, respectively. Therefore, when $q \leq q_0$, an ant selects a DC $k \in D_c$ for VNF $f \in V_j$, where the quantity of $\left[\mathcal{T}_{j,f}^{k}\right]^{\alpha} \times \left[\mathcal{H}_{j,f}^{k}\right]^{\beta}$ gives the highest value among all possible data centers. On the other hand, in case of exploration of an ant z chooses DC $k \in D_c$ to place a VNF $f \in V_j$ of eNB $j \in N$ with the following probability [37].

$$p^z_{j,f,k} = \begin{cases} \dfrac{\left(\left[\mathcal{T}^k_{j,f}\right]^\alpha \times \left[\mathcal{H}^k_{j,f}\right]^\beta\right)}{\sum_{d \in D_c}\left(\left[\mathcal{T}^d_{j,f}\right]^\alpha \times \left[\mathcal{H}^d_{j,f}\right]^\beta\right)}, & \text{if k} \in D_c \\ 0, & \text{otherwise} \end{cases} \tag{3.24}$$

This probability indicates that the DC $k \in D_c$ that gives us highest probability among all candidate DC $d \in D_c$, we place the VNF $f \in V_j$ in that DC.

3.3.2.4 Pheromone Update

The process of updating pheromone values locally and globally are presented below:

Local Update When an ant places a VNF in DC $k \in D$, it instantly updates the pheromone trail. Local pheromone value is updated with respect to initial pheromone value. It is updated by each ant as follows,

$$\mathcal{T}^k_{j,f}(t+1) = \rho_l \times \mathcal{T}_0 + (1-\rho_l) \times \mathcal{T}^k_{j,f}(t), \tag{3.25}$$

where, ρ_l is a system parameter indicating the relative importance of historical pheromone value and current pheromone value. By local pheromone update, every time an ant allocates a VNF in a specific DC, its pheromone trail is decreased. As a result it becomes less desirable for other ants of the colony. It encourages exploration and helps to increase variety of solutions constructed by the colony.

Global Update Global pheromone value is updated when all ants construct their local optimal solutions and then they update the global one. The global optima is formed by selecting the best one among the local solutions produced by all the ants. Let, the global solution set is $\mathcal{G}$. Global pheromone value is updated as the following equation,

$$\mathcal{T}^k_{j,f}(t+1) = \rho_g \times \Delta\mathcal{T}^k_{j,f} + (1-\rho_g) \times \mathcal{T}^k_{j,f}(t), \tag{3.26}$$

where, ρ_g is a system parameter which indicates the relative importance of $\Delta\mathcal{T}^k_{j,f}$ and $\mathcal{T}^k_{j,f}(t)$ in Eq. 3.26. $\Delta\mathcal{T}^k_{j,f}$ is the global pheromone value for updated global solution $\mathcal{G}$. The value of $\Delta\mathcal{T}^k_{j,f}$ is determined by,

$$\Delta\mathcal{T}_{j,f}^{k} = \begin{cases} \mathcal{T}_{j,f}^{k}, & \text{if } (j,f,k) \in \mathcal{G} \\ 0, & \text{otherwise} \end{cases} \tag{3.27}$$

If VNF $f \in V_j$ of eNB $j \in N$ is placed to DC $k \in D$ in global solution $\mathcal{G}$, then the value of $\Delta\mathcal{T}_{j,f}^{k}$ will be $\mathcal{T}_{j,f}^{k}$, otherwise 0.

In brief, ACO meta-heuristic algorithm optimally selects a DC for a VNF from the previous learning experience of the ants using pheromone value. After completing one iteration, all ants update their pheromone value. In the next iteration, they use their historical information. After successful completion of the algorithm, all of the VNFs places in the data centers optimally.

3.4 Determination of Weight Parameter

In this subsection, we discuss on the chosen values of the weight parameters- γ, α, β, ρ_l, ρ_g and the number of ants. This research is a first step insight on trading-off between VNF relocation overhead and user QoE. Formulating a mathematical model for weight parameter γ in real-time might help us to increase the performance of the system. However, it requires extensive analysis on the depending parameters including network size, the service arrival rate, time-deadline of applications, etc., which demands a separate research work.

Similarly for determining the values of α, β, ρ_l, ρ_g and the number of ants, we depend on the numerous simulation experimental results and find that the following values give us better results for the given network. We set $\alpha = 5$, $\beta = 1$, $\rho_l = 0.3$, $\rho_g = 0.4$, $\gamma = 0.7$ and number of ants $= 20$ for all experiments.

3.5 Conclusion

In this chapter, we formulate our proposed TradeRC optimization model to bring trade-off between number of relocations cost and response time. For large networks, this model is proved to be an NP-hard one. For that reason, we have given a AI based meta-heuristic Ant Colony Optimization (ACO) algorithm to find near

optimal solution for the placement of VNFs in 5G data centers in polynomial time. And as we shall see in the next chapter, our proposed TradeRC outperforms the other state-of-the-art works in terms of user satisfaction and VNF relocation overhead.

Chapter 4

Performance Evaluation

In the previous chapter, we have discussed on the assumptions of the proposed system and formulate our TradeRC system for optimal placement of VNFs in the DCs. Due to the NP-hardness of that optimal placement, we have provided meta-heuristic ACO based VNFs placement algorithm. In this chapter, we present the detail performance evaluation result of our TradeRC system and analyze its effectiveness by comparing with other state-of-the-art works.

4.1 Introduction

In this chapter, we have compared the performance of proposed TradeRC system with the state-of-the-art work: A-SGWR [20] and S-PL [20] systems. We have also compared our system with the baseline greedy based FF-VNF method. To solve the VNF allocation optimization problem, we have used CPLEX solver at NEOS optimization server [36] (2x Intel Xeon E5-2698 @ 2.3-GHz 569 CPU and 92-GB RAM). To simulate our ACO base VNF allocation approach, we have used Cloudsim [22].

4.2 Simulation Environment

We consider a network consisting of 12 data centers. Each DC has primary memory between 2 to 16 GB and storage between 2 to 16 TB. Each DC has 200 - 400 VMs.

TABLE 4.1: Simulation Environment

Parameter	**Value**
Simulation area	2000 ×2000 m^2
Number of Data-Centers (DC)	12
Number of eNBs under a DC	4 - 25
Number of VNFs under an eNB	500 - 2000
Communication delay between DCs	10 - 200 msec
Communication delay between DC and eNB	2 - 5 msec
Date rate to transfer VNF between DCs	1 - 50 Mbps
Size of each Virtual Network Functions (VNFs)	100 - 300 KB
Weight factor (γ)	0.7
Importance of pheromone value (α)	5
Importance of heuristic value (β)	1
Local pheromone constant (ρ_l)	0.3
Global pheromone constant (ρ_g)	0.4
Number of ants	20
Maximum Iteration MAX-IT	200

These VMs are heterogeneous in size and capacity. There are several numbers of VNFs under a VM. Each VM has RAM between 512MB - 1024MB having clock speed 2.50GHz - 3GHz. Therefore, the processing power of a VM is 500 - 1000 instructions per second. Achievable data rate of a transmission link is randomly chosen with exponential distribution. We choose the range 1Mbps - 50Mbps for assigning data rates to individual links. The average of the results got from 50 simulation runs is used to plot the graph data points. For each of the simulation run, we have used different random seeds. Performance parameter values and ranges are summarized in Table 4.1.

4.3 Performance Metrics

We have compared the performance of the studied systems on the following metrics:

- Quality of Experience (QoE): QoE is defined as the inverse of average normalized response time to get service for a VNF from a particular DC. It helps us to quantify how fast an algorithm can extend services to the users.
- Number of relocations of the VNFs: VNF relocation is defined as the total number of migrations required to transfer a VNF from one DC to another.

This metric has direct impact on the overall performances of the network. Lower the value, higher is the user performance.

- User satisfaction is defined as the integrated metric in terms of the number of VNF relocations and user Quality-of-Experience. It is computed as the sum of the inverse of the normalized value of the number of relocations and the normalized value of QoE. This represents overall performance of the studied algorithms.

- VNF Relocation Overhead: Relocating a VNF from one DC to another requires extra cost related to time and resources. VNF relocation overhead is defined as the extra cost that is required to transfer a VNF from a DC to another and the corresponding service cost. It is measured as the extra cost divided by the total cost of the system.

4.4 Simulation Result

Here we have studied the performance of our proposed TradeRC system by varying the number of eNBs under a DC, number of VNFs movement under an eNB and capacity of different data centers.

4.4.1 Impacts of Number of VNFs Movement per eNB

In this experiment, we vary the number of VNF movements per eNB keeping the number of data centers under a network, number of eNBs under a DC and DC's VNFs holding capacity fixed at 12, 20 and 600, respectively.

Fig. 4.1 shows that, increasing the number of VNFs movement from an eNB QoE is decreased. In S-PL and in our TradeRC, Quality-of-Experience is very higher than A-SGWR and FF-VNF methods and perform close to each other. But in A-SGWR, response time is increased by huge amount after increasing the number of VNFs movement. This is because, A-SGWR considered only minimizing number of VNFs relocation. However, it didn't consider response time for a service. For that reason, QoE is very low in A-SGWR than S-PL and TradeRC methods. In A-SGWR, maximum amount of response time can be infinity which is not applicable

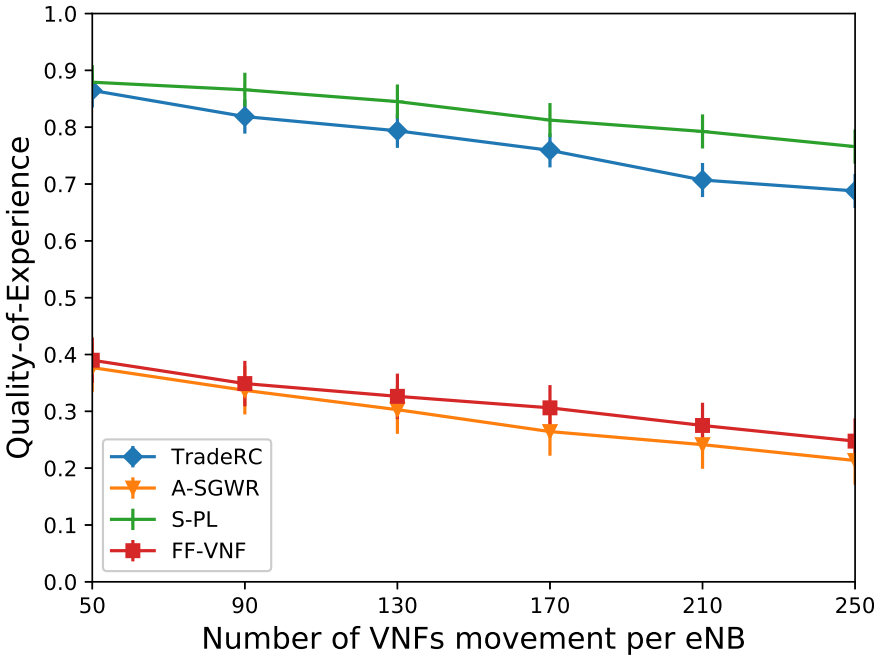

FIGURE 4.1: Impacts of QoE by varying the VNFs movement per eNB

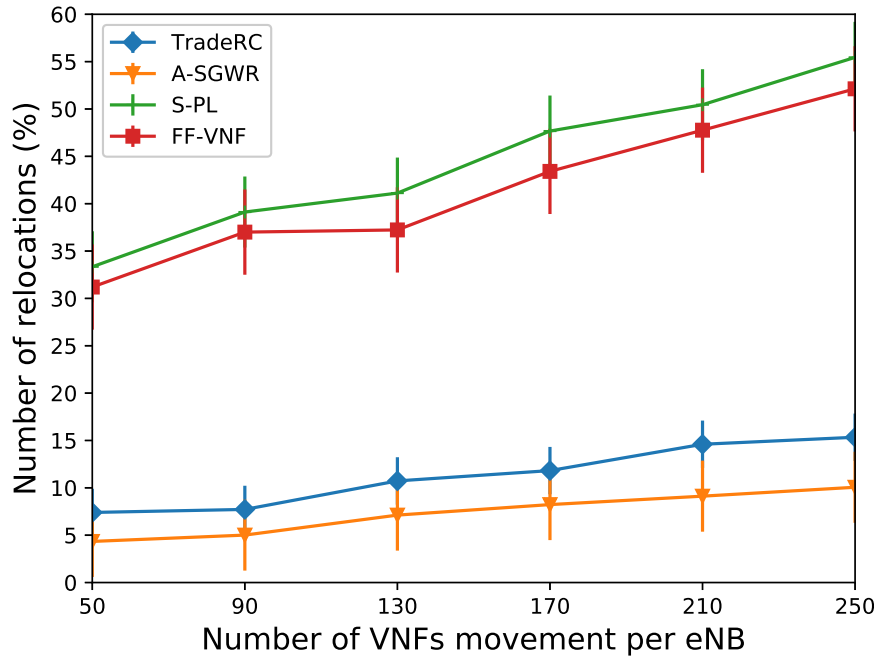

FIGURE 4.2: Impacts of number of relocations by varying the VNFs movement per eNB

in real-life scenario. Therefore, S-PL and TradeRC performs better in terms of QoE.

Fig. 4.2 shows that, as the number of VNFs movement from an eNB is increased, the number of relocation percentage is also enhanced. From the graph, we see that, for small networks number of relocation in S-PL and TradeRC is near to

each other. But for large networks, number of relocation in TradeRC is much less compared to S-PL solution. This is because, S-PL solution considered only minimizing response time. However, they didn't consider total number of relocation. In that system, maximum number of relocations can be infinity which is not practical. On the other hand, FF-VNF greedy method doesn't try to minimize the number of relocations and response time. For that reason, FF-VNF performs the worst than TradeRC in terms of number of relocations and QoE.

Fig. 4.3 indicates overall users' satisfaction is increased by increasing the number of VNFs movement from an eNB. For small types of networks, A-SGWR, S-PL and TradeRC perform near to each others. However, for increasing number of VNFs movement, user satisfaction in TradeRC is higher than other methods. In A-SGWR and S-PL, the authors have considered to improve only one parameter. However, for the other parameters they have not considered any requirement. In FF-VNF method, none of these two parameters are considered to be improved. Our system TradeRC works optimally for both of those parameters jointly. For that reason, user satisfaction is high in TradeRC.

Fig. 4.4 indicates that, with increasing number of VNFs movement, VNFs relocation overhead is also raised. This is because, we need to relocate higher number of VNFs from one DC to another DC with increasing number of VNFs movement.

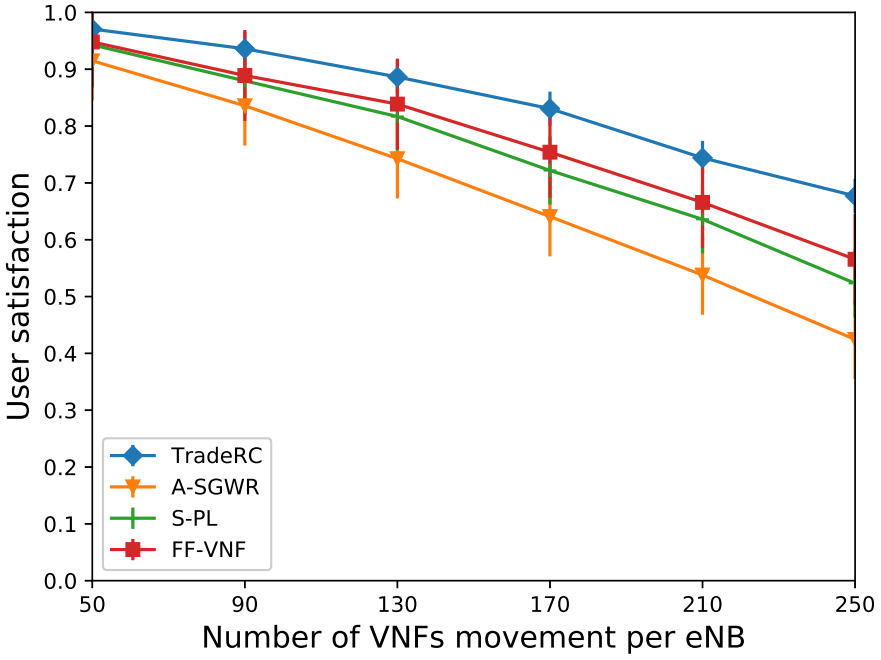

FIGURE 4.3: Impacts of user satisfaction by varying the VNFs movement per eNB

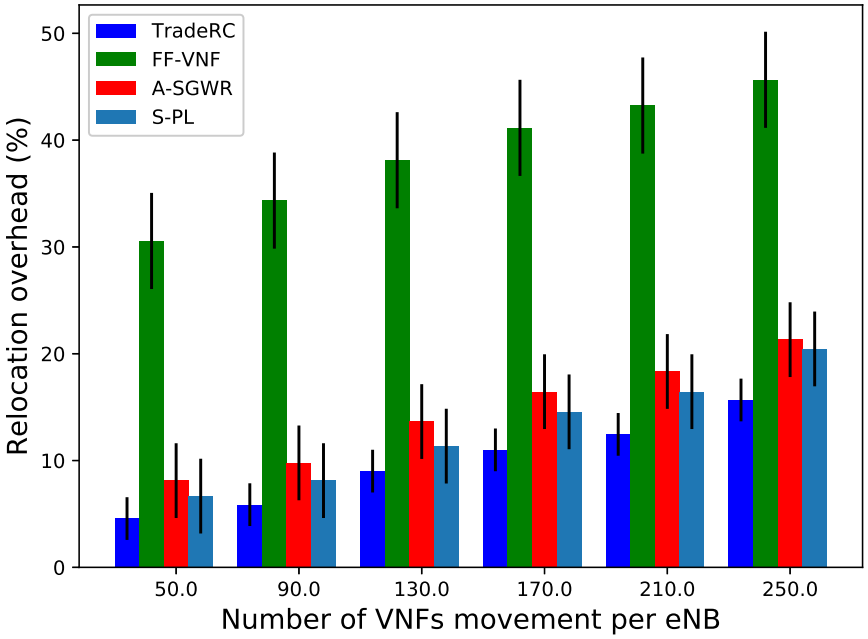

FIGURE 4.4: Impacts of VNF relocation overhead by varying the VNFs movement per eNB

Our proposed TradeRC works better than other methods. This is caused by the fact that, in A-SGWR the authors tried to place all VNFs in one DC where the algorithm is running. So we need extra cost to get service from that DC. On the other hand, in S-PL, the number of relocations is very much higher than TradeRC and A-SGWR. For that reason, relocation cost is very much higher in S-PL. In greedy based FF-VNF method, we place the VNF in that DC where the capacity is available. We don't consider minimization of any parameter. For that reason, our proposed TradeRC system works better than other methods in terms of VNFs migration overhead cost.

4.4.2 Impacts of Number of eNBs per DC

In this experiment, we vary the number of eNBs under a DC keeping the number of data centers under a network, number of VNFs movement per eNB and DC's VNFs holding capacity fixed at 12, 250 and 600, respectively.

Fig. 4.5 shows that, QoE is decreased with increased number of eNBs under a DC. In S-PL and our TradeRC systems, QoE is close to each other. However, response time is less in S-PL than TradeRC because it's main target is to minimize response time. For that reason QoE is very high in that approach. On the

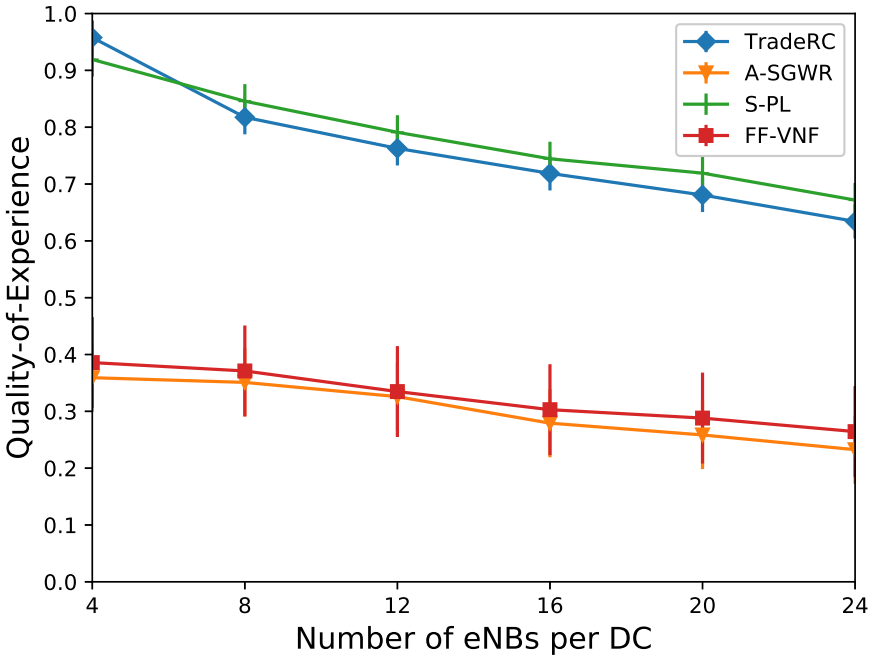

FIGURE 4.5: Impacts of QoE by varying the number of eNBs per DC

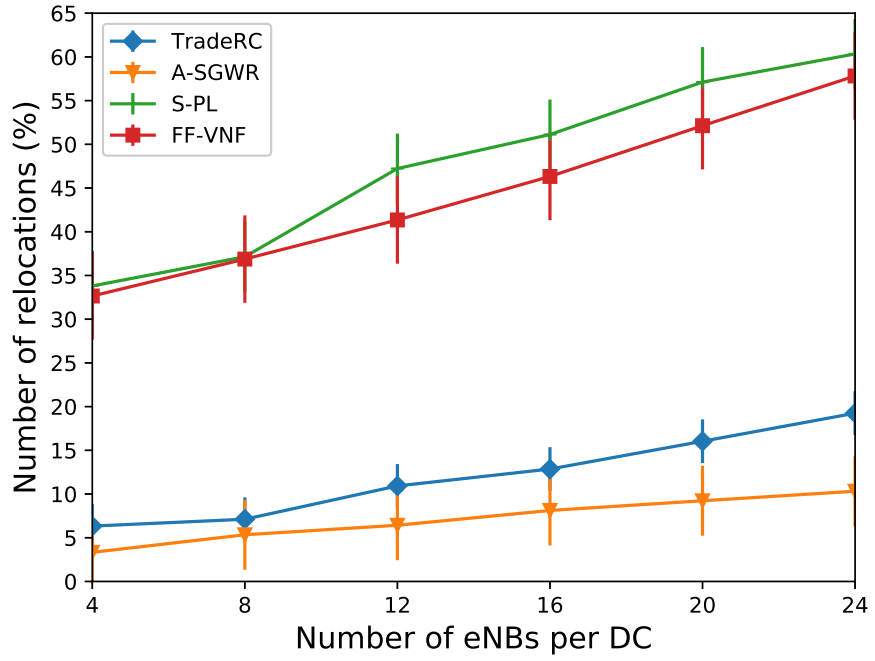

FIGURE 4.6: Impacts of number of relocations by varying the number of eNBs per DC

other hand, response time is very higher in A-SGWR and FF-VNF methods than other two systems. Because, A-SGWR considers only minimizing the number of VNF relocations. However, they didn't consider minimizing response time to get a service from a DC. In their system, response time can be infinity which is not

feasible in real life. In FF-VNF greedy solution, we have not considered to minimize the number of VNF relocations. If we have capacity in the DC, we place that VNF greedily in that DC. Therefore, QoE is very low in A-SGWR and FF-VNF methods.

Fig. 4.6 depicts that, number of VNF relocations is enhanced with the higher number of eNBs under a DC. As the number of eNBs is raised, the number of VNF relocations is also increased. In A-SGWR and TradeRC system, the number of VNF relocations is less than S-PL and FF-VNF approach and they perform close to each other. A-SGWR performs better than TradeRC because, it only considered to minimize the number of VNF relocations. In S-PL, number of relocations is very much higher. This is because, they have only considered to minimize the response time. In that system, number of relocations can be infinity which is not real life scenario. In FF-VNF greedy method, we have not consider to minimize the response time. For that reason, response time is higher in FF-VNF.

From Fig. 4.7, we observe that user satisfaction is decreased with the increased number of eNBs per DC. For small types of networks, all methods perform near to each other. However, for large networks user satisfaction is higher in TradeRC than other approaches. In A-SGWR and S-PL, the authors have considered to improve only one parameters. However, for the other parameters, they have not

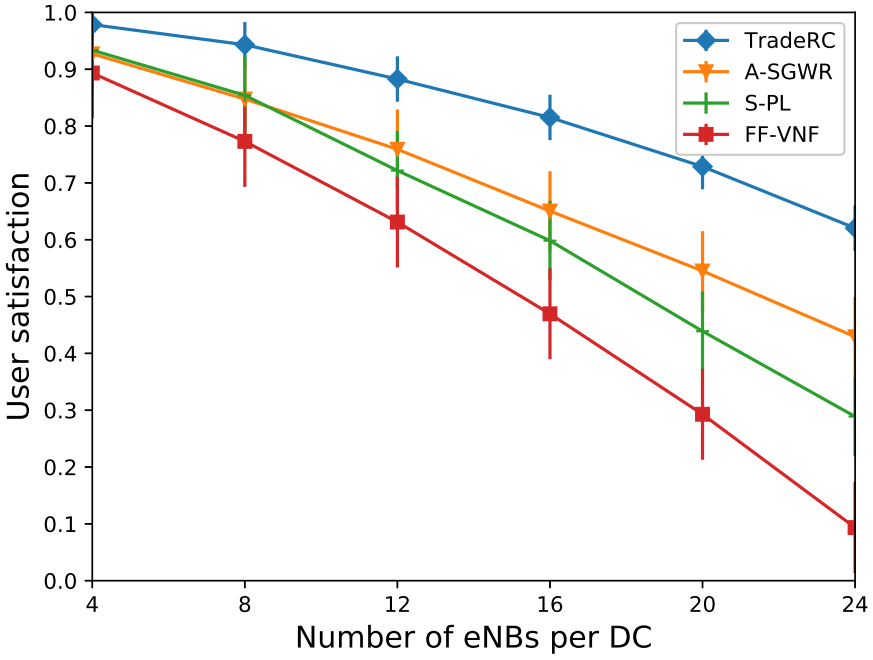

FIGURE 4.7: Impacts of user satisfaction by varying the number of eNBs per DC

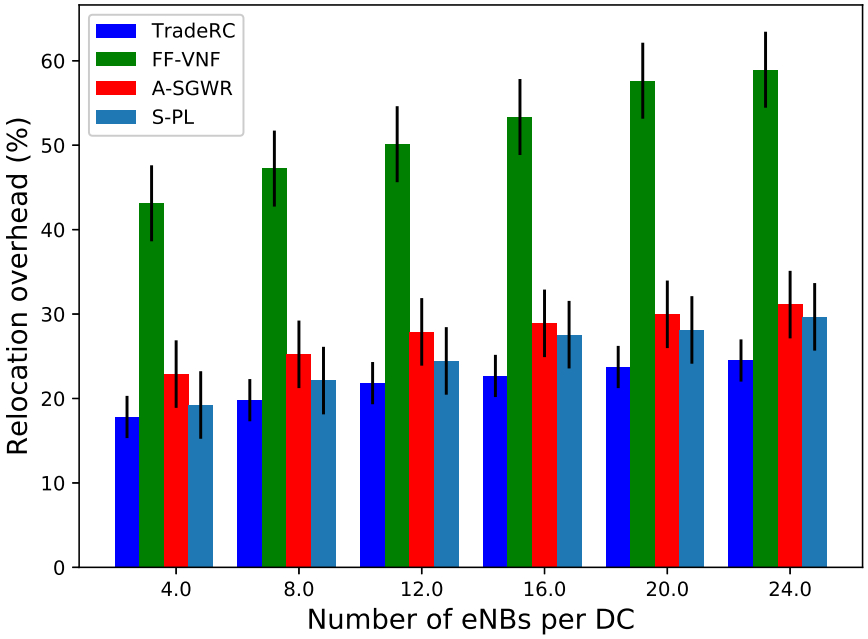

FIGURE 4.8: Impacts of VNF relocation overhead by varying the number of eNBs per DC

considered any requirement. The other parameter value can be infinity in that systems. For that reasons, user satisfaction is less in that systems. Our system TradeRC works optimally for both of those parameters jointly.

Fig. 4.8 indicates VNF relocation overhead by increasing number of eNBs under a DC. Although relocations overhead in TradeRC method increases with higher number of eNBs per DC, it performs better compared to other systems. This is because, in A-SGWR all VNFs are tried to migrate to that DC where the algorithm in running. In that system, relocation cost is very low but extra service cost is required to get service from that DC. In S-PL, the number of relocation is very much higher. For this reason, relocation cost is very higher in that system. Due to this reasons, VNF relocation overhead is higher than our TradeRC system. On the other hand, in FF-VNF greedy approach, we have not considered minimizing the number relocations and response time. For that reasons, migration overhead is very much higher in FF-VNF method than other approaches.

4.4.3 Impacts of VNF holding capacity of DC

In this experiment, we measure the impact of DC's VNFs holding capacity by varying the capacity of DC keeping the number of data centers under a network,

number of eNBs under a DC and number of VNFs movement per eNB fixed at 12, 20 and 250, respectively.

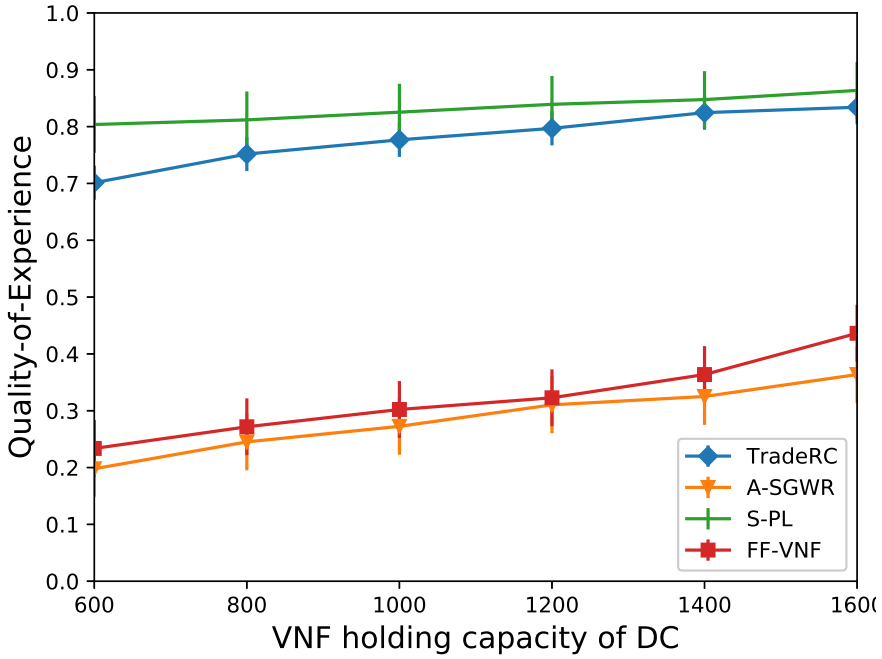

FIGURE 4.9: Impacts of QoE by varying VNF holding capacity of DC

Fig. 4.9 depicts that, by increasing the capacity of DC, QoE is increased. TradeRC and S-PL perform close to each other in terms of QoE. In S-PL, their main objective was to minimize response time. For that reason, S-PL performs better than other methods in terms of QoE. Again, TradeRC performs better than A-SGWR and FF-VNF method because in TradeRC, we try to bring trade off between response time and the number of relocations. On the other hand, response time can be infinite in A-SGWR system which is not practical. For that reason, QoE is very low in A-SGWR and FF-VNF methods. Therefore, S-PL and TradeRC perform better with respect to QoE than other two methods.

From Fig. 4.10 we see that, number of relocations is decreased with increasing the capacity of the DC. This is because, by increasing capacity, it means that DC has enough capacity and particular type of VNFs is available in that DC. So we don't need to relocate VNFs to get that particular type of service from another DC. A-SGWR and TradeRC perform better than other two approaches and A-SGWR perform better than TradeRC because in A-SGWR, it tried to minimize the number of relocations only. On the other hand, S-PL performs the worst because in S-PL the authors tried to minimize response time allowing infinite

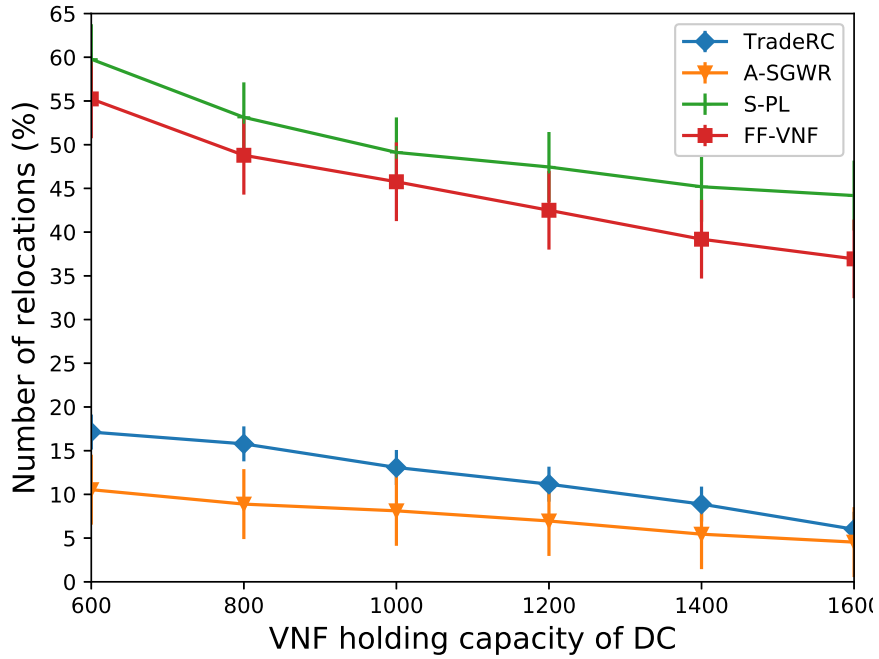

FIGURE 4.10: Impacts of number of relocations by varying VNF holding capacity of DC

number of relocations in their system which is not feasible in real life. In FF-VNF, we don't consider to minimize the number of relocation. For that reason, the number of relocations is very much higher in FF-VNF method.

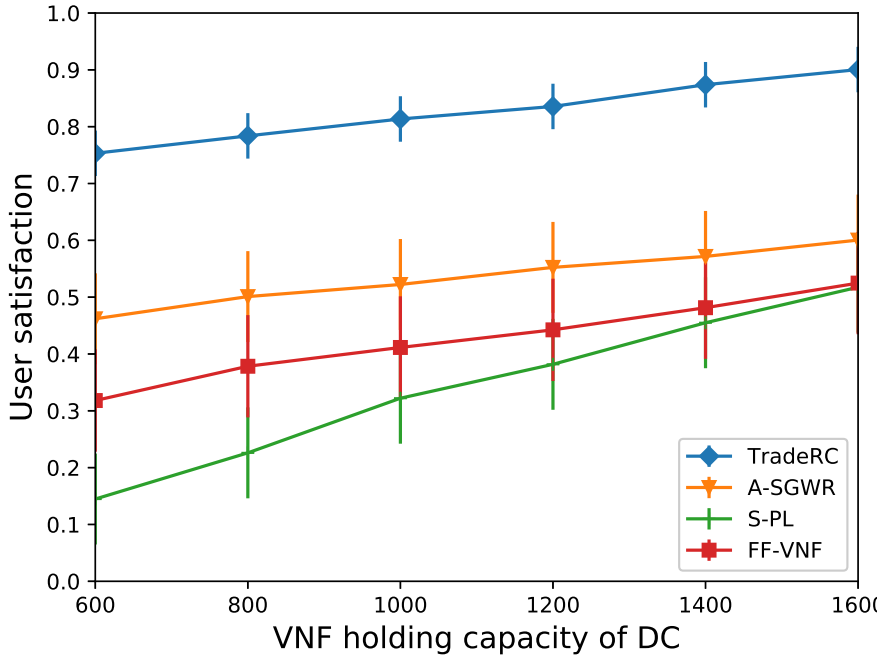

FIGURE 4.11: Impacts of user satisfaction by varying VNF holding capacity of DC

user satisfaction is improved with the increasing capacity of the DC which is depicted in Fig. 4.11. user satisfaction is also high in our proposed TradeRC system in that case. This is because, TradeRC brings trade-off between the two conflicting objectives and works optimally for both of those two parameters jointly. On the other hand, A-SGWR and S-PL tried to improve only one parameter value. In these systems, other parameter value can be infinity. For that reason, user satisfaction is less in A-SGWR and S-PL than TradeRC.

Fig. 4.12 shows that, VNF relocation overhead decreases with increased capacity of the DC. This is because, with increasing number of DC's capacity, number of reloactions decreases. In FF-VNF greedy method, extra service cost and relocations cost are huge. Because, in that method, the number of relocations is not minimized. In A-SGWR, all VNFs were tried to be placed in one DC so that the number of relocations is reduced. In that system, extra service costs are required to get service where the algorithm is running. On the other hand, In S-PL, the number of relocations can be infinite. For that reason, relocations cost are huge in that system. Therefore, migration overhead cost is higher in S-PL, A-SGWR and FF-VNF systems than our proposed TradeRC system.

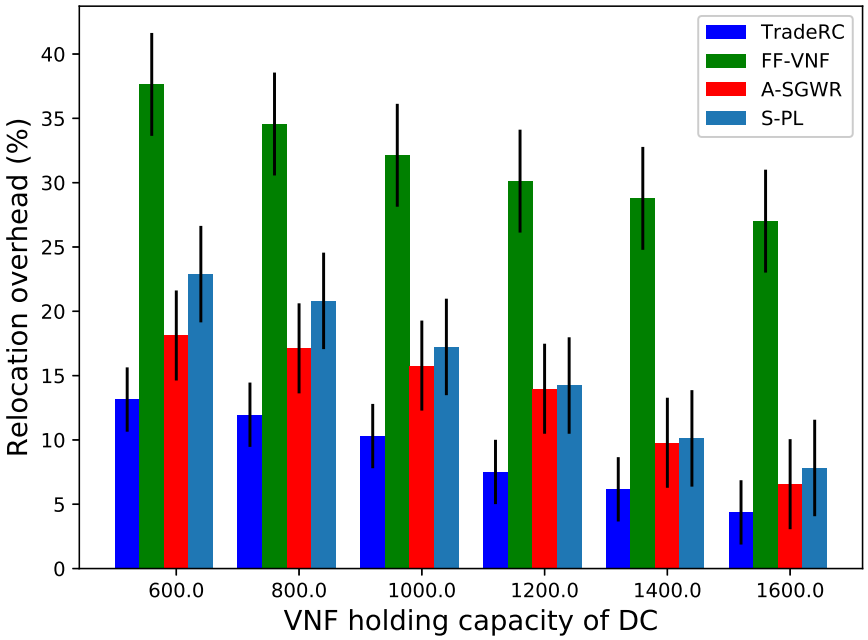

FIGURE 4.12: Impacts of VNF relocation overhead by varying VNF holding capacity of DC

4.5 Conclusion

From the above discussion and results, it can be implied that by varying the number of VNFs movement per eNB, our proposed TradeRC system outperforms the A-SGWR, S-PL and greedy based FF-VNF algorithm in terms of overall user satisfaction and VNF relocation overhead. We have also got improved results by varying the number of eNBs per DC and by varying the VNF holding capacity of a DC in terms of overall user satisfaction and VNF relocation overhead compared to other state-of-the-art works.

Chapter 5

Conclusion

In this chapter, we summarize our research work and present overall discussion of this thesis. We also state few guidelines for the researchers and our future plan.

5.1 Summary of the Research

With the advent of 5G network technology, users will experience higher throughput, lower latency, higher mobility range and faster uploading and downloading speed. Network Function Virtualization (NFV) and Software Defined Network (SDN) are the two major features of 5G technology. Due to the NFV technology, all functions of the mobile devices are run on cloud. This saves huge amount of battery life of the mobile devices. All users are connected to eNBs and these eNBs are connected to the DCs. Therefore, users' services are run in these DCs. When any user moves from one eNB to another eNB that are connected to different DC, the proper allocation of respected VNF is a major concern.

This project introduced a framework for optimal placement of VNFs in 5G data centers. Decreasing the response time for user code execution in VNFs of 5G data centers can be achieved by enabling VNF relocations; however, excessive migration causes communication and computation overhead. This work explored optimal trading-off approach in between two conflicting objectives- maximizing Quality-of-Experience and minimizing number of VNF relocations. For large networks, this placement problem was proven to be an NP-hard problem. Our proposed

AI based ACO solution maximized Quality-of-Experience and minimized relocation overhead, increasing user satisfaction. The experimental results have shown significant performance improvement in terms of user satisfaction and relocation overhead as high as 25% and 15%, respectively.

5.2 Discussion

5G network technology is very emerging technology in the next few years. It offers massive device connectivity, higher throughput and almost zero latency. Frequency spectrum of 5G is divided into millimeter waves. These millimeter waves provide fastest data rate than other previous technologies. We found that the current challenges for 5G technology are load balancing among the cloudlets, mobility issues, getting real time services and energy consumption of the small cells.

In this thesis, we have provided an optimal placement algorithms of VNFs in the DCs due to user mobility. Due to the movement of users from one eNB to another that is connected to different DC, how can users take services in that case is a matter of concern. If user takes service from the previous DC through DC to DC communication, this increases the response time significantly. However, this reduces migration cost of the whole network. On the other hand, if the VNF is migrated from the previous DC to current DC, this increases the migration cost but reduces reponse time to get a service. Therefore, if we improve the performance of one parameter, performance of some other parameters is decreased at the same time. So we need to bring trade-off between them. Previous research works [19, 20] focused on improving one parameter. In our research work, we have focused on optimal placement of VNFs in the DCs considering all of those parameters that minimizes total number of relocations and improves QoE at the same time. However, this optimal placement of VNFs is an NP-hard problem. For that reason, we have provided an AI based meta-heuristic Ant Colony Optimization (ACO) algorithm for VNF placement that provides a sub-optimal solution.

The *Computer Networking* and *Design and Analysis of Algorithm* courses have helped us a lot providing theoretical idea of the network architecture and to prove our formulation of optimal placement of VNFs as an NP-hard one. At first, problem formulation of the MILP was very difficult for us and we needed to spend lot of time in that phase. Besides this, getting introduced with NEOS optimization

server and to cope with new programming language CPLEX were much hectic for us. Spending plenty of time in the laboratories with our friends and researchers and reading a lot of papers of various sectors were the key to get success.

5.3 Future Work

In this thesis, we basically work on the optimal placement of VNFs in the DC by trading-off between minimizing the number of VNF relocations and maximizing QoE. While we get promising solutions compared to other state-of-the-art works, there are still several challenging issues that can further be investigated.

In the future, we want to develop mathematical modeling and analysis for determining the values of the system parameters dynamically so as to further enhance the performances. In our thesis, we have fixed the value of all wight parameter for all types of network. By varying the value of that weight parameter in respect to network size in real time, we can further enhance our performance.

In our thesis, we have not considered load balancing issue among the data-centers. So, one DC need to handle huge amount of VNF but another can't handle at all at the same time. This degrades the overall network performance. Load balancing among the the DCs can also be another research problem that will improve the service performances.

Bibliography

[1] D Evans. The internet of things: How the next evolution of the internet is changing everything. *Cisco Internet Business Solutions Group (IBSG)*, 1:1–11, 01 2011.

[2] Md Ahsan Habib, Sajeeb Saha, Md Abdur Razzaque, Md Mamun-or Rashid, Giancalro Fortino, and Mohammad Mehedi Hassan. Starfish routing for sensor networks with mobile sink. *Journal of Network and Computer Applications*, 2018.

[3] Sujan Sarker, Md Abdur Razzaque, Mohammad Mehedi Hassan, Ahmad Almogren, Giancarlo Fortino, and Mengchu Zhou. Optimal selection of crowdsourcing workers balancing their utilities and platform profit. *IEEE Internet of Things Journal*, 2019.

[4] M. Chen, Y. Hao, L. Hu, K. Huang, and V. K. N. Lau. Green and mobility-aware caching in 5g networks. *IEEE Transactions on Wireless Communications*, 16(12):8347–8361, Dec 2017.

[5] Min Chen, Yixue Hao, Hamid Gharavi, and Victor C.M. Leung. Cognitive information measurements: A new perspective. *Information Sciences*, 505:487 – 497, 2019.

[6] Min Chen, Yixue Hao, Meikang Qiu, Jeungeun Song, Di Wu, and Iztok Humar. Mobility-aware caching and computation offloading in 5g ultra-dense cellular networks. *Sensors*, 16(7):974, 2016.

[7] M. Agiwal, A. Roy, and N. Saxena. Next generation 5g wireless networks: A comprehensive survey. *IEEE Communications Surveys Tutorials*, 18(3):1617–1655, thirdquarter 2016.

[8] R. Khoder and R. Naja. Software-defined networking-based resource management in 5g hetnet. In *2018 IEEE Middle East and North Africa Communications Conference (MENACOMM)*, pages 1–6, April 2018.

[9] Md Rofiqul Islam, Md Muhaimin Shah Pahalovim, Tamal Adhikary, Md Abdur Razzaque, Mohammad Mehedi Hassan, and Ahmed Alsanad. Optimal execution of virtualized network functions for applications in cyber-physical-social-systems. *IEEE Access*, 6:8755–8767, 2018.

[10] NFVISG ETSI. Network functions virtualisation (nfv); terminology for main concepts in nfv. *Group Specification, Dec*, 2014.

[11] Sajeeb Saha, Md Ahsan Habib, Tamal Adhikary, Md Abdur Razzaque, and Md Mustafizur Rahman. Tradeoff between execution speedup and reliability for compute-intensive code offloading in mobile device cloud. *Multimedia Systems*, pages 1–13, 2017.

[12] Tarik Taleb. Toward carrier cloud: Potential, challenges, and solutions. *IEEE Wireless Communications*, 21(3):80–91, 2014.

[13] Shahryar Shafique Qureshi, Toufeeq Ahmad, Khalid Rafique, et al. Mobile cloud computing as future for mobile applications-implementation methods and challenging issues. In *2011 IEEE International Conference on Cloud Computing and Intelligence Systems*, pages 467–471. IEEE, 2011.

[14] M. Chen, Y. Hao, C. Lai, D. Wu, Y. Li, and K. Hwang. Opportunistic task scheduling over co-located clouds in mobile environment. *IEEE Transactions on Services Computing*, 11(3):549–561, May 2018.

[15] Niroshinie Fernando, Seng W. Loke, and Wenny Rahayu. Mobile cloud computing: A survey. *Future Generation Computer Systems*, 29(1):84 – 106, 2013. Including Special section: AIRCC-NetCoM 2009 and Special section: Clouds and Service-Oriented Architectures.

[16] N. Bobroff, A. Kochut, and K. Beaty. Dynamic placement of virtual machines for managing sla violations. In *2007 10th IFIP/IEEE International Symposium on Integrated Network Management*, pages 119–128, May 2007.

[17] H. N. Van, F. D. Tran, and J. Menaud. Sla-aware virtual resource management for cloud infrastructures. In *2009 Ninth IEEE International Conference on Computer and Information Technology*, volume 1, pages 357–362, Oct 2009.

[18] Marco Valerio Barbera, Sokol Kosta, Alessandro Mei, Vasile Claudiu Perta, and Julinda Stefa. Mobile offloading in the wild: Findings and lessons learned through a real-life experiment with a new cloud-aware system. *IEEE INFOCOM 2014 - IEEE Conference on Computer Communications*, pages 2355–2363, 2014.

[19] J. Plachy, Z. Becvar, and E. C. Strinati. Dynamic resource allocation exploiting mobility prediction in mobile edge computing. In *2016 IEEE 27th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, pages 1–6, Sep. 2016.

[20] T. Taleb, M. Bagaa, and A. Ksentini. User mobility-aware virtual network function placement for virtual 5g network infrastructure. In *2015 IEEE International Conference on Communications (ICC)*, pages 3879–3884, June 2015.

[21] Manuel Eugenio Morocho Cayamcela and Wansu Lim. Artificial intelligence in 5g technology: A survey. In *2018 International Conference on Information and Communication Technology Convergence (ICTC)*, pages 860–865. IEEE, 2018.

[22] Rodrigo N Calheiros, Rajiv Ranjan, Anton Beloglazov, César AF De Rose, and Rajkumar Buyya. Cloudsim: a toolkit for modeling and simulation of cloud computing environments and evaluation of resource provisioning algorithms. *Software: Practice and experience*, 41(1):23–50, 2011.

[23] Xiaoqiao Meng, Vasileios Pappas, and Li Zhang. Improving the scalability of data center networks with traffic-aware virtual machine placement. In *2010 Proceedings IEEE INFOCOM*, pages 1–9. IEEE, 2010.

[24] Rafael Moreno-Vozmediano, Rubén S Montero, and Ignacio M Llorente. Iaas cloud architecture: From virtualized datacenters to federated cloud infrastructures. *Computer*, 45(12):65–72, 2012.

[25] Deval Bhamare, Mohammed Samaka, Aiman Erbad, Raj Jain, Lav Gupta, and H Anthony Chan. Optimal virtual network function placement in multi-cloud service function chaining architecture. *Computer Communications*, 102:1–16, 2017.

[26] B. Addis, D. Belabed, M. Bouet, and S. Secci. Virtual network functions placement and routing optimization. In *2015 IEEE 4th International Conference on Cloud Networking (CloudNet)*, pages 171–177, Oct 2015.

[27] M. F. Bari, S. R. Chowdhury, R. Ahmed, and R. Boutaba. On orchestrating virtual network functions. In *2015 11th International Conference on Network and Service Management (CNSM)*, pages 50–56, Nov 2015.

[28] N. M. Akshatha, P. Jha, and A. Karandikar. A centralized sdn architecture for the 5g cellular network. In *2018 IEEE 5G World Forum (5GWF)*, pages 147–152, July 2018.

[29] R. Khoder and R. Naja. Software-defined networking-based resource management in 5g hetnet. In *2018 IEEE Middle East and North Africa Communications Conference (MENACOMM)*, pages 1–6, April 2018.

[30] L. Gkatzikis and I. Koutsopoulos. Mobiles on cloud nine: Efficient task migration policies for cloud computing systems. In *2014 IEEE 3rd International Conference on Cloud Networking (CloudNet)*, pages 204–210, Oct 2014.

[31] Q. Zhang, Q. Zhu, M. F. Zhani, and R. Boutaba. Dynamic service placement in geographically distributed clouds. In *2012 IEEE 32nd International Conference on Distributed Computing Systems*, pages 526–535, June 2012.

[32] Amit Kumar Das, Tamal Adhikary, Md Abdur Razzaque, Majed Alrubaian, Mohammad Mehedi Hassan, Md Zia Uddin, and Biao Song. Big media healthcare data processing in cloud: a collaborative resource management perspective. *Cluster Computing*, 20(2):1599–1614, 2017.

[33] Faqir Zarrar Yousaf, Johannes Lessmann, Paulo Loureiro, and Stefan Schmid. Softepc—dynamic instantiation of mobile core network entities for efficient resource utilization. In *2013 IEEE International Conference on Communications (ICC)*, pages 3602–3606. IEEE, 2013.

[34] Xiaofei Wang, Min Chen, Tarik Taleb, Adlen Ksentini, and Victor CM Leung. Cache in the air: Exploiting content caching and delivery techniques for 5g systems. *IEEE Communications Magazine*, 52(2):131–139, 2014.

[35] Michael R Garey and David S Johnson. *Computers and intractability*, volume 29. wh freeman New York, 2002.

[36] NEOS optimization server. http://www.neos-server.org/neos/. [Online; accessed 01-August-2019].

[37] Silvia Mazzeo and Irene Loiseau. An ant colony algorithm for the capacitated vehicle routing. *Electronic Notes in Discrete Mathematics*, 18:181–186, 2004.

List of Publications

International Journal Article

[1] Palash Roy, Anika Tahsin, Sujan Sarker, Tamal Adhikary, Md. Abdur Razzaque, Mohammad Mehedi Hassan, "User mobility and Quality-of-Experience aware placement of Virtual Network Functions in 5G", *Elsevier Computer Communications Journal 2019*, vol. 150, pages. 367 - 377, January 2020.
doi: https://doi.org/10.1016/j.comcom.2019.12.005.