LTE Network Planning and Traffic Generation

by

Dima Dababneh, B.Sc

A thesis submitted to
the Faculty of Graduate Studies and Research
in partial fulfillment of the requirements for the degree of
Master of Applied Science in Electrical and Computer Engineering

Ottawa-Carleton Institute for Electrical and Computer Engineering (OCIECE)
Department of Systems and Computer Engineering
Carleton University
Ottawa, Ontario, Canada
March 2013

The undersigned recommend to
the Faculty of Graduate Studies and Research
acceptance of the Dissertation

LTE Network Planning and Traffic Generation

Submitted by
Dima Dababneh

in partial fulfillment of the requirements for the degree of
Master of Applied Science in Electrical and Computer Engineering

---

Thesis Supervisor
Prof. Marc St-Hilaire

Chair, Department of Systems and Computer Engineering
Dr. Howard Schwartz

Carleton University
2013

Abstract

The main objective of this thesis is to propose automatic planning tools for the planning problem of Long Term Evolution (LTE) networks based on realistic traffic. More precisely, a set of equations is first proposed to generate realistic traffic profile while considering as many practical aspects as possible. Then, a mathematical model is created to solve the planning problem of the evolved packet core of LTE networks while minimizing the cost. Since the planning problem is NP- Hard, an approximate method based on the local search is also proposed. The algorithm provides a trade-off between the solution quality and the execution time. Numerical results show that the approximate approach is capable of providing good solutions (on average within 4.03% of the optimal solution) in shorter time duration than the exact approach.

Acknowledgements

First, I would like to thank God for giving me strength and patience to work on this thesis. It would not have been possible without His guidance and support.

I would like to express my sincere appreciation to the help and support of my thesis supervisor, Professor Marc St-Hilaire. His continuous supervision and support were valuable assets for the completion of this thesis. He was always there whenever I needed help, and I am glad that I got the chance to work with him.

I would also like to thank Dr. Christian Makaya for his wise suggestions and valuable comments throughout the whole period, and I am grateful to Eng. Mohammd Reza Pasandideh for his valuable inputs.

I can’t find words to express my gratitude to my parents; Dr. Sana Naffa and Dr. Faris Dababneh. They have always been a great source of encouragement and support. I am also blessed to have the support and the encouragement of my brothers Laith and Saif, and my husband to be Firas El Farr.

Last but not least, I want to thank all my friends and colleagues who were a great source of encouragement and motivation.

Table of Content

List of Figures

List of Tables

List of Acronyms

1G	First Generation
2.5G	Second and a half Generation
2G	Second Generation
3G	Third Generation
4G	Fourth Generation
AAS	Adaptive Antenna Switching
ABT	Asynchronous Backtracking
ACI	Adjacent Channel Interference
ACIR	Adjacent Channel Interference Ratio
ACLR	Adjacent Channel Leakage Power Ratio
ACR	Access Control Router
AcS	Active Sessions
ADP	Asynchronous Distributed Pricing
AMPS	Advanced Mobile Phone System
AMR	Adaptive Multiple Rate
AS	Access Stratum
AuC	Authentication Center
AWC	Asynchronous Weak Commitment
BH	Busy Hour
BHDSA	Busy Hour Data Session Attempt
BHSA	Busy Hour Session Attempt
BHVSA	Busy Hour Voice Session Attempt
BS	Base Station
CCO	Coverage and Capacity Optimization
CDMA	Code Division Multiple Access
CINR	Carrier to Interference and Noise Ratio
CN	Core Network
CPU	Central Processing Unit

CS	Circuit Switched
CSFB	Circuit Switched Fallback
DL	Downlink
DTNS	Delay Threshold Normalized Scheduler
eNB or eNodeB	Enhanced NodeB
EPC	Evolved Packet Core
EPS	Evolved Packet System
EPSB	Evolved Packet System Bearer
E-UTRAN	Evolved UTRAN
FDD	Frequency Division Duplexing
FDMA	Frequency Division Multiple Access
FFR	Fractional Frequency Reuse
FUP	Fair Usage Policy
GPRS	General Packet Radio Service
GSM	Global System for Mobile Communications
HeNB	Home eNB
HN	Home Network
HO	Handover Ratio
HRPD	High Rate Packet Data
HSS	Home Subscriber Server
IA	Intelligent Agent
IMS	IP Multimedia Subsystem
Inter-RAT	Inter-Radio Access Technology
KPI	Key Performance Indicators
LP	Linear Programming
LS	Local Search
LTE	Long Term Evolution
M/G/R-PS	M/G/R-Processor Sharing
M2M	Machine to Machine
MAB	Multi Armed Bandit
MAP	Mobile Application Part

MCS	Modulation and Coding Schemes
MIP	Mixed Integer Programming
MLB	Mobility Load Balancing
MME	Mobility Management Entity
MMF	Max-Min Fair
MRO	Mobility Robust Optimization
MTP	Max Throughput
MWM	Maximum Weighted Matching
NAS	Non Access Stratum
NPO	Network Performance Optimizer
OME	Operation Management Equipment
P2P	Peer to Peer
PCC	Policy and Charging Control
PCEF	Policy Control and Charging Enforcement Function
PCRF	Policy Control and Charging Rules Function
PF	Proportional Fair
PFR	Partial Frequency Reuse
P-GW	Packet Data Network (PDN)Gateway
QCI	QoS Class Identifiers
QPNS	Queue Packet Normalized Scheduler
RAS	Radio Access Station
RB	Resource Blocks
RRM	Radio Resource Management
RTF	Retransmission Factor
SA	Simulated Annealing
SAE	System Architecture Evolution
SB	Scheduling Block
SFR	Soft Frequency Reuse
S-GW	Serving Gateway
SINR	Signal-to-interference and noise ratio
SON	Self-Organizing Networks

SRVCC	Single Radio Voice Call Continuity
TDD	Time Division Duplexing
TFT	Traffic Flow Templates
TMCS	Tactical Mobile Communication Systems
TTI	Time Transmission Interval
TU	Top Up
UE	User Equipment
UL	Uplink
UMTS	Universal Mobile Telecommunications System
US-MRO	User Speed Mobile Robust Optimization
VAF	Voice Activity Factor
VN	Visitor Network
VoLTE	Voice over LTE
VSA	Virtual Sub-band Algorithm

Chapter 1 Introduction

Cellular communications have been experiencing a massive continuous progress for the past decades, and millions of people are using various services provided by cellular networks every day. The increasing demand for data services puts pressure on the network operators who have to invest heavily in their network infrastructure and in planning tools that provide a cost-efficient network design able to meet all the users' demands. The competitive market also necessitates network operators to provide the services at reasonable cost, which also makes the cost an important factor in the design process. Finding a planning tool that can support realistic traffic and can take both the cost and the quality of service into consideration is not an easy task.

Most of the work that has been done was mostly focused on the radio network of LTE, and it was based on estimated traffic that was not realistic or measured traffic that is time consuming. In this thesis, our goal is to develop a tool to generate realistic traffic that takes different aspects into consideration, and to propose two different algorithms to approach the planning problem of LTE core network.

In this chapter, we first provide a general overview of the evolution of cellular networks. Then, Section 1.1.2 presents the typical architecture of LTE network describing its elements, interfaces and links. The problem statement is formulated in Section 1.2 followed by the research objectives in Section 1.3. The methodology used to achieve the research objectives is discussed in Section 1.4 and Section 1.5 presents the contributions. Finally, Section 1.6 describes the thesis overview.

1.1 Background

Cellular networks play a major role in our life today. The dependency on the services provided by cellular networks is increasing. In addition to that, this type of networks aims to provide the users with flexibility, increased capacity, reduced power consumption, larger coverage area and

reduced interference. In Section 1.1.1, we will briefly describe the evolution of cellular networks starting from the first generation getting to the Long Term Evolution (LTE). Then, we will describe in details the LTE architecture for both the access network and the core network in Section 1.1.2.

1.1.1 Cellular Network Evolution

In early 1980s [37], the first generation (1G) of the analog cellular networks was introduced as Advanced Mobile Phone System (AMPS) and it was based on Frequency Division Multiple Access (FDMA). In 1990s, the second generation (2G) of mobile phone system emerged, and it was based on Global System for Mobile Communications (GSM) in Europe and on Code Division Multiple Access (CDMA) in the US. The 2G allowed limited data support compared to the second and a half generation (2.5G) networks that were extended from it. The 2.5G includes the General Packet Radio Service (GPRS), and it uses both circuit and packet switching; the former is used for voice, and the latter is used for data transmission. In fact, the high demand for data services and higher speeds led to the evolution of the third generation (3G) which also uses packet switching instead of circuit switching for data transmission, but can provide services with higher efficiency, better streaming services, and faster web browsing than 2.5G. During the recent years, 3G networks were overwhelmed by the amount of growth in services and applications such as media streaming. Hence, the fourth generation (4G) was introduced and two competing standards were proposed: WiMAX and Long Term Evolution (LTE). It eliminated circuit switching and utilized packet switching efficiently over the Internet to provide users with better performance. Despite the fact that WiMAX access technology had a short lifespan, LTE is getting more popular and it also provides an increased bandwidth allowing more applications to be used such as wireless online gaming, and high quality mobile video streaming.

1.1.2 LTE Architecture

LTE is the evolution of the radio access Universal Mobile Telecommunications System (UMTS) known as Evolved UTRAN (E-UTRAN), whereas the evolution of the non-radio part including the Evolved Packet Core (EPC) network is referred to as the System Architecture Evolution (SAE). Both the LTE and SAE form what we call the Evolved Packet System (EPS). Figure 1.1

shows the overall network architecture [1] with the different types of interfaces and the type of traffic carried on each link.

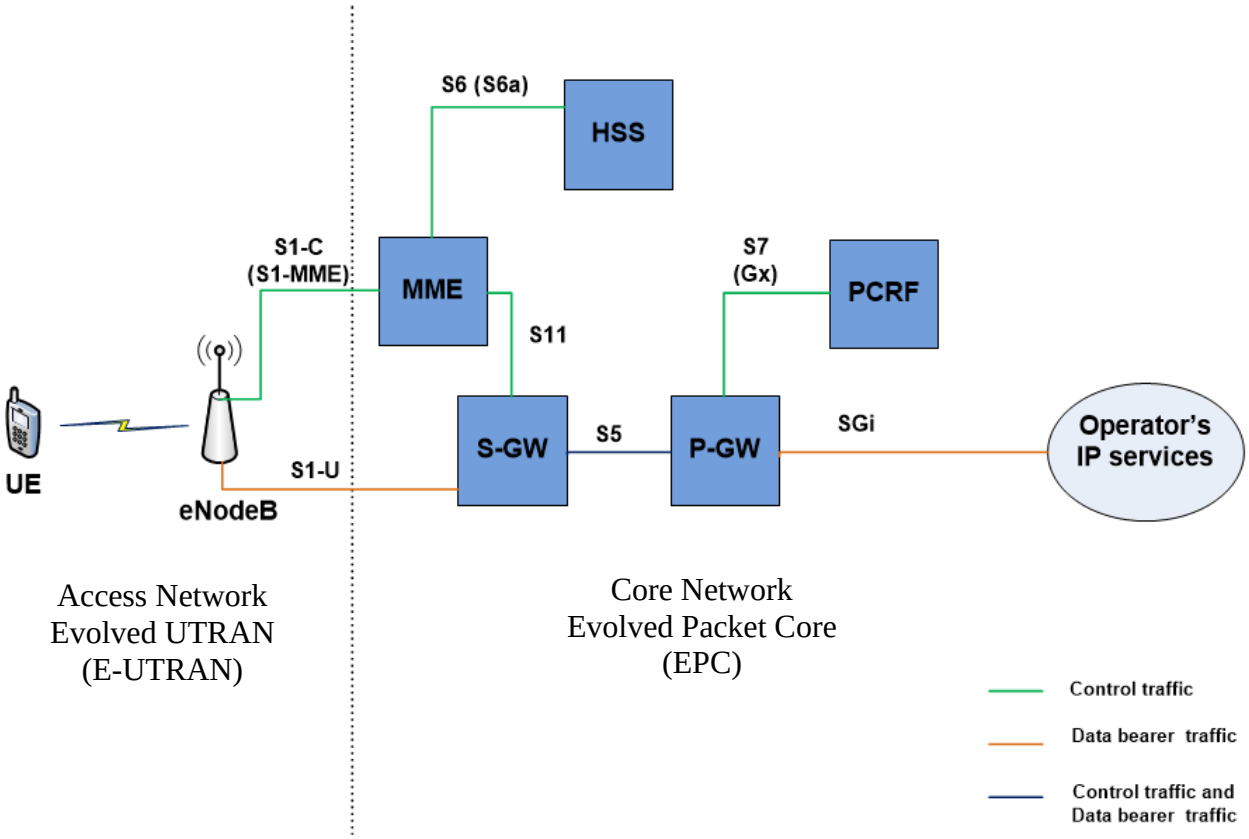

Figure 1.1: Typical architecture for LTE networks

The architecture of the access network is pretty simple as it is composed of a single type of node called enhanced NodeB (eNodeB or eNB). These advanced base stations are used to connect the User Equipment (UE) to the network. The E-UTRAN is considered a flat architecture because it has no centralized controllers. As shown in Figure 1.2, the E-UTRAN architecture consists of a network of eNodeBs that are interconnected with each other by X2 interfaces that allow communication between the different eNBs and connected to the EPC by S1 interfaces; S1-U user plane interface used between the eNB and the S-GW and S1-C (S1-MME) interface which allows communication between the eNB and the MME.

All the following active UE radio functions are run by the E-UTRAN [1][2]: (1) IP packet header compression; (2) Security and data encryption; (3) Connecting to the EPC which includes selecting and sending signals to MME as well as creating bearer path and routing data to S-GW;

and (4) Radio Resource Management (RRM) which covers all radio bearers functions such as admission control, mobility control, dynamic allocation of resources to UE, scheduling, and radio bearer control. On the other hand, the main task of the core network is to control the UE and establish the bearers [1]. It also provides the idle and active terminals with QoS, security, mobility and management and finally allows connectivity with external IP packet networks [2].

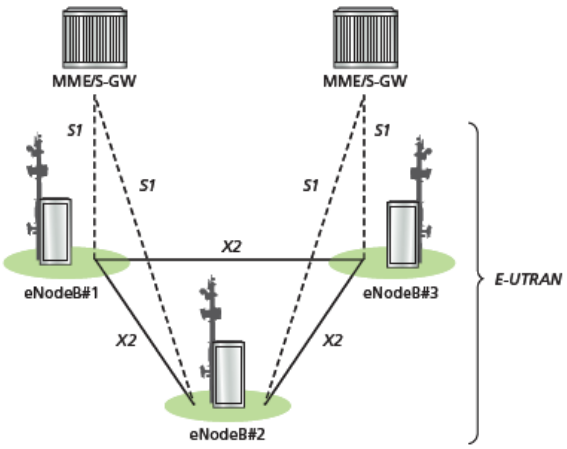

Figure 1.2: Overall E-UTRAN Architecture [1]

The architecture of the core network is a little bit more complex as is it composed of different pieces of equipment. As shown in Figure 1.1, the EPC is composed of five main nodes which are as follows [1].

- **Serving Gateway (S-GW):** It is the local mobility anchor that holds data when the UE are moving between eNodeBs during handover, and it deals with the user plane. S-GW is the connection between the radio part and the EPC. It is the hub on which all IP packets are transferred through; it transports the IP data traffic between UE and the external networks. Moreover, it keeps information about the bearers when the UE is idle and it works as a buffer for downlink data when the MME is initiating paging of the UE for bearers' reestablishment. S-GW has different administrative tasks in the network; it gathers data for charging such as the traffic on the link whether sent or received by a user. In addition, it works as mobility anchor for internetworking with other 3GPP

technologies like UMTS and GPRS. Logically, this gateway is connected to the PDN gateway.

- **Packet Data Network (PDN) Gateway (P-GW):** It is the interconnection point between the EPC and the external IP networks and it is mainly in charge of assigning and distributing the IP addresses for the UE, besides enforcing the QoS and flow based charging that is based on PCRF rules. The PDN gateway has the responsibility to distribute and sort out the IP packets in the downstream into different QoS based channels and bearers based on the Traffic Flow Templates (TFTs). It also has the ability to work as a mobility anchor for internetworking with non 3GPP technologies like High Rate Packet Data (HRPD) (aka 1xEV-DO) and WiFi. P-GW is considered the default gateway as well; it performs packet filtering and lawful interception which includes analyzing the signalling data in addition to the network management information.

- **Mobility Management Entity (MME):** It is the control node that is responsible for the signalling between the UE and the core network. It deals with the control plane, and it is considered the termination point of the Non Access Stratum (NAS) which plays a key role in initiating and maintaining the EPS bearers. It has a major role in registering UE in a network, handling mobility functions between UE and core network, and creating and keeping IP connectivity. NAS is carried over LTE-Uu which is the air interface between UE and eNB and the S1-MME interfaces between the eNB and the MME. On the other hand, Access Stratum (AS) protocols are those that run between eNodeBs and UE. In conclusion, NAS functions between a core network and the user equipment; whereas, AS functions between the radio network and the user equipment. There are two classifications for the main functions supported by the MME: (1) bearer management, and (2) connection management. The former is handled by the session management layer in the NAS protocol and is related to the establishing, maintaining and releasing bearers; but the latter is handled by the connection or mobility management layer in the NAS protocol and is related to establishing connection between the network and the UE along with providing security. In fact, the MME is responsible for [2]: (1) distributing the

paging messages to the eNBs, (2) security, (3) mobility control for users in idle state, (4) control of bearers, and (5) protecting NAS signalling integrity and ciphering.

- **Home Subscriber Server (HSS):** It holds dynamic information to keep track of the MME identities to which users are connected. HSS also includes data for the user's System Architecture Evolution (SAE) subscription such as the QoS profile and any roaming access restrictions. Moreover, it has the Packet Data Network (PDNs) information that allows users to connect to the PDN (e.g., Internet, IMS). It also plays a role in authentication and security due to its ability to integrate the Authentication Center (AuC) which formulates security keys and authentication vectors.

- **Policy Control and Charging Rules Function (PCRF):** One of its tasks is to control the Policy Control and Charging Enforcement Function (PCEF) functionalities that are flow-based and residing in the P-GW. The QoS authorization, which is consisting of the QoS Class Identifiers (QCI) and bit rates, defines the PCEF treatment for certain data flows in harmony and agreement with the user's subscription profile.

As shown in Figure 1.1, the names of the interfaces are standardized[71]. The S1-U interface connects the eNB and the S-GW for user plane traffic (i.e. bearers' tunneling, inter-eNB handover), also the S1-C (currently referred to as S1-MME) connects the eNB with the MME for control processes and signaling messages between the eNB and the MME. S11 connects the MME to the S-GW, S6 (currently referred to as S6a) connects the MME and the HSS, and S5 connects the S-GW to the P-GW. In addition to that, the S7 (currently referred to as Gx) connects the P-GW (PCEF) to the PCRF, and the SGi is the interface between the P-GW and the packet data network such as Internet or IMS (IP Multimedia Subsystem).

1.2 Problem Statement

The advancement in network complexity, the rapid growth in the number of users, and the increasing demand for data services urged service providers to heavily invest in their network infrastructure. Moreover, operators are fiercely competing to provide subscribers with good

quality services at reasonable prices. In order for the operators to be able to manage providing good quality services at reasonable prices, they need to have a good planning tool that provides optimal network design with minimum cost, taking into consideration the traffic handled by the network. Selecting a good planning tool, and providing a properly planned network saves money and time on operators. Several algorithms tackled the planning aspect of the radio network, but few approached the core network. Different tools were proposed to plan specific interfaces according to traffic dimensioning, but very few proposed tools to plan the whole network based on traffic flow and characteristics.

In general, the main goal of a network planning tool is to provide an optimal network with minimum cost based on realistic traffic. A realistic traffic in LTE networks must take into consideration different parameters such as bandwidth, number of simultaneous bearers, Busy Hour Session Attempt (BHSA), and signaling traffic. The more realistic the traffic, the better the planning tool is. The planning tool results in a topology that defines the number of core elements needed to support the traffic in addition to the type and the location of each element.

Due to the complexity of the networks, developing an exact method is considered NP- hard [84]. NP-hard problems do not find optimal solution in polynomial time. In other words, the exact method may not be able to find the optimal solution in finite reasonable amount of time. As a result, approximate algorithms are needed to solve the planning problem in less time and CPU usage.

The lack of previous work on the planning of the LTE core network (EPC) taking into consideration realistic traffic brings on the importance of the network planning tool developed in the thesis. A solution that achieves good quality in addition to optimized cost is found to be more appropriate and applicable to be used by operators and service providers.

1.3 Research Objectives

Based on the problem statement defined in the previous section, the main goal is to develop tools for the network planning of LTE EPC. More precisely, we want to achieve the following sub-objectives:

- Generate a realistic traffic profile that includes a variety of aspects such as bandwidth, signaling traffic, BHSA, and EPSB.
- Propose a mathematical model for the LTE core network planning that aims to find the minimum cost of the network while meeting the user demands. This method will provide optimal solutions at the cost of high complexity.
- Develop a heuristic algorithm to provide good solutions in an amount of time that is less than the exact methods.
- Compare the results from the two different methods in terms of cost and CPU time.

To achieve the previous objectives, the following methodology will be used.

1.4 Methodology

In any network planning problem, there are several parameters and constraints that need to be taken into consideration, which in turn increases the complexity of the planning problem. The core network planning was approached using the following methodology.

1. *Study different concepts of LTE*: This step is very important because it involves understanding the architecture of LTE networks as well as the functions of each core elements (i.e. MME, S-GW, HSS, P-GW, and PCRF) as well as the traffic flow starting from the eNB reaching to the Internet.
2. *Literature Review:* The importance of this step goes back to the need of understanding the previous work related to the planning of LTE networks. This step led to defining the network planning problem, and helped in understanding what has been done and what is missing.
3. *Generating realistic traffic*: In order to plan LTE networks, we need to know the requested amount of traffic as input. To that end, a set of equations were suggested to generate realistic traffic based on the maximum amount of traffic during the day taking into consideration different planning parameters and variables. A C/C++ code was developed to generate the traffic, and this traffic takes into consideration different aspects (i.e. signaling traffic, bandwidth, EPSB, BHSA).
4. *Developing a mathematical model:* The model involves a set of decision variables for the links and the nodes as well as traffic variables. The main objective function of the model is to minimize the network cost which includes the cost of the links, interfaces and nodes

taking into consideration different assignment, uniqueness and capacity constraints. This model will be used as a reference in order to evaluate the performance of the approximate algorithm.

5. *Implementation of the exact algorithm:* The objective function and the constraints were translated into a file of LP format, which employs the linear programming method and uses CPLEX for solving the problem. Due to the massive number of variables and constraints, there was a need to develop a C/C++ code to translate the objective function and the constraints into the LP format.
6. *Study of heuristic algorithms:* NP-hard problems can't find optimal solution in polynomial time, and such type of problems requires different types of algorithm which are referred to as heuristic algorithms. Heuristic algorithms tend to give a good solution in less amount of time; examples on heuristic algorithms that give approximate solutions are the Local Search, Tabu Search, Simulated Annealing, and Genetic algorithms.
7. *Implementation of the approximate algorithm*: To approach the NP-hard problem, the Local Search algorithm was adopted. A C/C++ code was developed to perform the local search, and achieve a total network cost in a reduced amount of CPU time.
8. *Comparison between the exact and the approximate solutions*: Comparison between the two different approaches is performed in terms of network cost and CPU time. The exact solutions were first generated and then compared with approximate solutions who generally give higher cost in less CPU time.

1.5 Contributions

Most of the work that has been done in LTE network planning focused only on the radio part of the network (i.e. transmission power, antenna tilt, interference reduction, energy saving, etc.). Moreover, the traffic that has been taken into consideration is either estimated or measured, and both methods are considered inefficient because they are either time consuming or not specific. This research added the following contributions to the area of LTE network planning:

- Generating a realistic traffic profile that includes a variety of traffic aspects such as bandwidth, signaling traffic, BHSA, and EPSB).

- Developing a mathematical model to plan the evolved packet core of networks and find the optimal network topology in terms of cost while maintaining good service.
- Proposing a Local Search (LS) algorithm to solve the network planning problem, and handle problems with higher complexity. The LS provides relatively good solutions in less amount of time than the exact methods, and a comparison between the two approaches is presented in the text.

1.6 Thesis Overview

A literature review related to the different aspects of LTE network planning is conducted in Chapter 2. Chapter 3 first describes the traffic generation tool followed by the mathematical model for the planning of the evolved packet core of LTE network. Then, an approximate algorithm based on the local search is also proposed. Chapter 4 presents the simulation results for both methods and a comparison in terms of network cost and CPU time. Finally, Chapter 5 summarizes the work and proposes future directions.

Chapter 2 Related work on LTE Planning

The fact that LTE is expected to accommodate different types of services such as real time applications, video streaming, VoIP, web browsing, online gaming and transactions and file transfer requires it to be designed carefully taking into consideration different factors [6]: (1) data rate; (2) latency; (3) capacity; (4) mobility support; (5) coverage; (6) spectrum flexibility. LTE demands high data rate, low latency, minimized cost, and low power consumption as well as backward compatibility with seamless handover, efficient migration from 3G systems, good quality of service, and enhanced handover. LTE offers flexibility to the service providers in assigning bandwidth depending on the amount of available spectrum. The handover procedure in LTE is better than the procedure in 2G networks since it minimizes interruption time, also the spectrum efficiency in the LTE Downlink (DL) is 3 to 4 times of that of HSDPA Release 6, and the Uplink (UL) is 2 to 3 times of that Release. Some of the LTE requirements are described in Table 2.1.

In this chapter, we review previous work that is related to the planning of LTE networks. First, in Section 2.1, we describe different approaches that were used to dimension LTE network interfaces based on LTE traffic. In Section 2.2, we describe the general planning of LTE and discuss different optimization and heuristic algorithms focusing on different areas in LTE such as power allocation, antenna directing and positioning in addition to resource scheduling. The concept of Self-Organizing Networks (SON) plays an important role in LTE network planning. Hence, the different algorithms and approaches based on SON and used in planning and optimizing the LTE network are described in Section 2.3.

Table 2.1: LTE performance requirements [6]

Metric	Description	Requirement
Peak data rate	The maximum data rate that can be offered by the system not taking into consideration radio interface parameters such as antenna configuration or transmission bandwidth [8].	DL:100 Mbps, UL: 50 Mbps (for spectrum of 20MHz)
Mobility support	The ability to provide mobility across the whole network for both low and high mobile speeds. It also takes into consideration maintaining high performance even for voice and real time services [9].	Optimized for low speed between 0 and 15 km/h. Generally up to 500 km/h
Control plane latency	It is the transition time from idle state to active state [7].	From idle to active is less than 100 ms
User plane latency	Aka transport delay [7]; the time it takes the IP packet to get from the source to the destination (UE/eNodeB). It is affected control signalling and different protocols when the UE is in active state.	Less than 5 ms
Control plane capacity	It includes the number of active users that can be supported by the system [8].	More than 200 users per cell (spectrum of 5MHz)
Coverage	Using and reusing sites and carrier frequencies to support UE.	5-100 km with minor degradation after 30 km
Spectrum flexibility	The ability to support spectrum allocations of different sizes, and to support different spectrum arrangements such as supporting both similar and different content delivery on the sane aggregated resources [9].	1.25, 2.5, 5, 10, 15, and 20MHz

2.1 Traffic Capacity Measurements

There are two major types of traffic: elastic traffic, and real time traffic. Elastic traffic, such as web browsing and FTP, is generated by non-real time applications and carried over TCP transmission protocol. On the other hand, real time traffic, such as streaming, conferencing and VoIP, is very sensitive to delay and require specific requirements to be transmitted.

Considering the two different types of traffic, Li et al. in [73] propose two different models for dimensioning traffic bandwidth for the S1 interfaces in LTE networks given the amount of traffic and the number of users in the cell. The main goal of dimensioning the bandwidth for the S1 interface is to minimize the cost of the network while maintaining a certain level of QoS for different services. The model suggested for elastic traffic is based on the M/G/R-PS (M/G/R-Processor Sharing) model, and it used to measure the mean time or throughput for TCP flows; whereas the model suggested for real time traffic uses the M/D/1 queuing model that estimates the network delay and performance.

Also, in [74], the M/G/R-PS model is used to dimension the bandwidth of elastic traffic for the LTE. This model measures the bandwidth of the eNB required to be handled by the interfaces to carry elastic traffic. The model guarantees end-to-end QoS by following the theory of process sharing which characterizes the traffic at the flow level, and the two main QoS concerns to be guaranteed are throughput and delay. The model is capable of characterizing the TCP traffic assuming each user has an individual flow for Internet services. The basic M/G/R-PS model is discussed in [76]; it is applied for dimensioning mobile networks as well as ADSL. The elastic traffic acts like a processor sharing system because all elastic traffic flows sharing the same link share the same amount of bandwidth and other resources.

Checko et al. [75] developed a traffic model based on predicted traffic values for 2015 in order to dimension the LTE backhaul network using three capacity planning methods: a delay based approach, a dimensioning formula-based approach, and an overbooking factor-based approach. The total amount of mobile data traffic predicted for 2015 is equal to 6,253,920 TB resulted by different applications such as video, web-browsing, Peer to Peer (P2P), VoIP, Machine to Machine (M2M), and gaming. Based on the forecasted values, the average user will transmit and receive 852 MB of data per busy hour [75]. The delay based approach allows increasing the capacity as much as needed as long as the delay requirements are satisfied. For the formula based approach, it calculates the bandwidth needed to support a number of users based on the peak aggregated throughput for those users. Last but not least, the overbooking factor based approach takes into consideration the probability of having the connection in an active state, and it states

that certain users are assigned capacities lower than the sum of their required capacities due to the fact that not all users are using all of their network resources.

The fact that the resource allocation is based on queue status (i.e. packet drop, packet delay, etc.) urged Lizos et al. [82] to develop two packet schedulers for LTE network taking into consideration the overall traffic flow evaluation. The two methods, Delay Threshold Normalized Scheduler (DTNS) and Queue Packet Normalized Scheduler (QPNS), are capable of accommodating high speed bursty traffic. These two methods don't associate the formulas with mathematical framework which makes the methods not practical, and there is no validation with real life. Moreover, the methods cannot be applied to each conventional eNB due to memory limitation and high complexity of the problem.

Jailani et al. [83] performed a research study in an area in Malaysia to collect data using the Network Performance Optimizer (NPO) tool, in particular they used traffic counters and indicators. The paper provides a dimensioning approach for LTE network based on the available LTE voice traffic taking the busy hour traffic as the best representation to evaluate the network performance and perform network dimensioning. The approach presented only deals with speech traffic and does not consider signaling, video or other applications.

Mainly, all the work that has been done in LTE network planning focused on the radio network, and not much work has been done on the core network. The main focus in network planning is on the eNB not on the core elements (i.e., MME, S-GW, HSS, P-GW, and PCRF). Dimensioning the air interface and the S1-U got the interest of planners but dimensioning other interfaces or elements were not approached. Bandwidth was the main factor taken into consideration in dimensioning, but signaling, BHSA, or EPSB were not discussed.

None of these models proposed methods to generate traffic; they basically dimensioned specific interfaces based on given traffic. In conclusion, the models described above cannot generate realistic traffic for the network; they also do not have the ability to plan a core network taking into consideration cost and quality of service. As a result, there was a need to develop a tool to

generate realistic traffic taking into account different aspects of traffic (i.e. bandwidth, EPSB, signaling and BHSA.

2.2 General Planning of LTE

There are many different approaches for LTE network planning. Some papers are concerned about the radio part and its parameters' optimization such as power allocation, radio resource scheduling, antenna down-tilt and BS positioning. Some papers are concerned about the radio part and its parameters' optimization such as radio resource scheduling, antenna down-tilt and BS positioning [15][16] or even traffic capacity planning approach for LTE radio networks [34]. Other papers tackle power issues such as applying intelligent agents to improve power management in LTE networks [31], and evaluating the performance of different coexistent LTE systems [33]. Cell coverage planning algorithm [32] tackles the issues of interference and throughput by optimizing the user uplink throughput. In general, LTE network planning involves a myriad of components such as antenna height, antenna inclination angle, Base Station (BS) transmit power, BS capacity, BS position and transmission bandwidth.

In [15], Li et al. tackled two important components namely BS positioning and BS power allocation in LTE networks. The method used for locating the BS position and allocating the initial power is the service search method that is based on the traffic in the planning region; the desired BS position is calculated and taken into consideration if the traffic achieved in the coverage area of the biggest BS radius is less than the maximum load and more than the minimum load. If the coverage rate of the covered traffic doesn't meet the requirement, then a smaller radius is chosen. The algorithm stops and the searching ends when the required coverage rate or smallest cell radius is achieved. The disadvantages of this method reside in the BS position; the coordinates of the BS position may be distributed in a straight line or BS may be distributed in some region while leaving another region. The other component is the BS power allocation which is approached by using the game theory with an enhancement known as Asynchronous Distributed Pricing (ADP) algorithm to increase the user quality of service and network performance. The ADP also assures maximum data transfer rate for edge users who

experience the strongest interference by taking into consideration the interference price, as well as initializing and updating the power and interference prices.

In [16], the Simulated Annealing (SA) meta-heuristic method is proposed to reduce the complexity of LTE network design by addressing the radio resource scheduling for multi-users on the downlink of the LTE systems. The proposed model involves allocating multiple radio resources of different users at the air interface and scheduling them simultaneously. The joint optimization model introduces a nonlinear optimization problem because of the need to jointly maximize the total bit rates for all users when regular optimization methods can be used without any guarantee to find the global optimality. In order to avoid local optimum, such a problem needs to be transformed to a linear one using extra auxiliary variables causing the solution space to increase and the cost to rise. The sub-optimal scheduler is used to jointly assign Modulation and Coding Schemes (MCSs), Scheduling Blocks (SBs), and users. This allocation is achieved in different stages; the first stage starts with allocating each SB to the highest bit rate user; and the second stage involves finding out the best MCS for each user. The main concept of this scheduler is used to reduce a problem of joint multiuser optimization into multiple equal single user optimization problems by assigning separate subset of SBs to each user. Different experiments were conducted in [16] to compare the performance of the Simulated Annealing (SA), global optimal and sub-optimal greedy algorithms. These experiments showed that the proposed SA method provides an optimal solution with reduced reasonable complexity.

The LTE radio network capacity planning approach proposed in [34] is an iterative process that aims to find the optimal capacity planning solution taking into consideration specific requirements and parameters. There are two different types of parameters: basic engineering and radio parameters, and optimization parameters. The former includes different parameters such as transmission power, and system bandwidth; whereas the latter considers issues such as antenna down-tilt, distance between sites, etc. The unified traffic process module converts the complex various traffic requirements into uniform information that takes into consideration QoS requirement and the number of users for every traffic type. After setting the optimization goal, the iterative process starts taking into account the dynamic simulation and the smart optimization. The dynamic simulation consists of four main processes: (1) carrier to interference

and noise ratio (CINR) modeling, which is the main method to determine precise signal quality distribution in networks, (2) user mapping, which is the process of assigning the CINR value for each terminal, (3) traffic adaption, which aims to fine-tune concurrent online user number and the behavior of generated traffic packet, and (4) traffic simulation. Similarly, smart optimization includes self-evaluation and automated search. The former checks if the capacity planning requirements are satisfied under certain conditions and the iteration stops and gives the results as soon as the optimal solution is found; whereas the latter sets the optimization parameters for the next iterations.

Due to the importance of energy saving and the need to improve its efficiency and consumption, Alcatel-Lucent invented a new energy efficient base stations known as light radio cubes or simply cubes. In [31], the suggested power management scheme adopts the concept of cubes. In the proposed LTE power management scheme, each cube is assumed to have an Intelligent Agent (IA) that makes decisions based on the traffic load information exchanged with its neighboring intelligent agents. The decisions involve adjusting transmission or radiation power, implementing beam-forming, and turning on or off the cubes, and they are taken based on the exchanged resource block utilization information and the reported locations of mobile users. The results show that implementing this scheme improves the power saving and reduces the power consumption by 50% to 70% based on the number of mobile users, while not affecting the network performance.

Due to the fact that different operators may need to deploy different Time Division Duplexing (TDD) and Frequency Division Duplexing (FDD) LTE systems in the same geographical area, the coexistence of these two systems is investigated and taken into consideration. In [33], single system, and inter-system interference are analyzed. In the single system scenario, there is no interference between users with different Resource Blocks (RB) because of the orthogonal uplink and downlink sub-carriers. However, users in different cells but in the same RB will be taken into account to calculate intra-system interference. On the other hand, inter-system interference occurs when two systems are functioning in the same neighboring spectrum. In fact, out-of band emission and spurious emission lead to Adjacent Channel Interference (ACI) which negatively affects the system capacity and performance. Different macro-cell propagation models based on

the vehicular test environment model and Hata model are used. The proposed fractional power control scheme compensates the path loss by controlling the transmit power of the mobile station, and the suggested link level performance model maps the SINR to throughput. In addition to that, the Adjacent Channel Leakage power Ratio (ACLR) model, discussed in [33], assumes that the mobile station uplink ACLR controls the Adjacent Channel Interference power Ratio (ACIR), which is the ratio of the source total transmitted power to the total interference power affecting receiver. Different parameters are taken into consideration, and many possible approaches, such as beam-forming and interference coordination, were used to reduce the interference and guarantee the coexistence of the different LTE systems.

Many approaches were adopted for dealing with cell coverage planning; nevertheless, they do not consider the scheduler complexity, the dynamic interference, or the cell edge user throughput planning. All these methods are based on basic mathematical models or general simulations. The LTE coverage planning approach that is based on the optimization of the uplink cell edge-user throughput, proposed in [32], allows cell planners to start with QoS requirements, initial restrictions, and required throughput of cell edge users. Another mathematical model is proposed to calculate the in-band interference parameters based on queuing theory; the in-band interference functionality is the responsibility of the MAC scheduler in the eNodeB and it is formed by the number of users in the same Resource Block (RB) in the same Time Transmission Interval (TTI) in surrounding neighbor cells. The advantage of this approach is that the expected cell edge user throughput is connected based on the in-band predicted result; cell planners have the freedom to dynamically adjust their models to their assumptions in different planning cases, and calculate the uplink cell edge guaranteed throughput rate.

Due to the fact that not much work was related to the LTE core network, there was a need to propose two methods to approach the core network planning problem, and those two methods are based on realistic traffic and can achieve minimal cost and good quality of service.

2.3 Planning of Self-Organized Networks

As we can see from the previous section, a lot of work has been done on the general planning of LTE networks. In addition to that, the concept of Self Organizing Networks (SON) takes part in LTE network planning [5], and plays a role in defining the cost of the network in operational and pre-operational stages [4]. The areas where most of the network planning work is focused are [10][11][12]: Coverage and Capacity Optimization (CCO), energy saving, interference reduction, physical cell ID assignment, Mobility Robust Optimization (MRO), and Mobility Load Balancing (MLB) optimization. In Section 2.3.1, SON related LTE access network planning is presented followed by a description of the core network planning in Section 2.3.2.

2.3.1 Radio Planning

In this section, we will describe the work related to the planning of the radio network based on the SON use cases. In addition, we will explain different methods and algorithms that were proposed to solve LTE radio planning issues and challenges.

2.3.1.1 Coverage and Capacity Optimization

The main objective of CCO algorithms is to optimize the network coverage while ensuring its continuity, and to maximize the capacity of the system, along with reducing interference and delay. In addition to that, the CCO algorithms enhance the cell edges' performance and increase the savings on drive tests.

Feng et al. suggested in [11] a process to deal with coverage and capacity that starts with collecting measurements from the eNBs, and then detect coverage and capacity problems. They proposed a planning tool that solves the problems and achieve optimization by adjusting the radio related parameters and passing them to the coverage and capacity optimization function. As a result, the coverage and capacity optimization function updates the parameters and makes them available to be used for operating the system. The process suggested is described in the proposed CCO Architecture depicted in Figure 2.1. The measurement collected from the eNBs and the UE

reports include the signal strength of the current cell and its neighbors, UE signaling and reporting, coverage triggered mobility counters, and traffic load measurements. On the other hand, the output of the optimization process is a set optimized radio configuration parameters, which includes downlink transmitting power and reference signal power offset in addition to the antenna tilt.

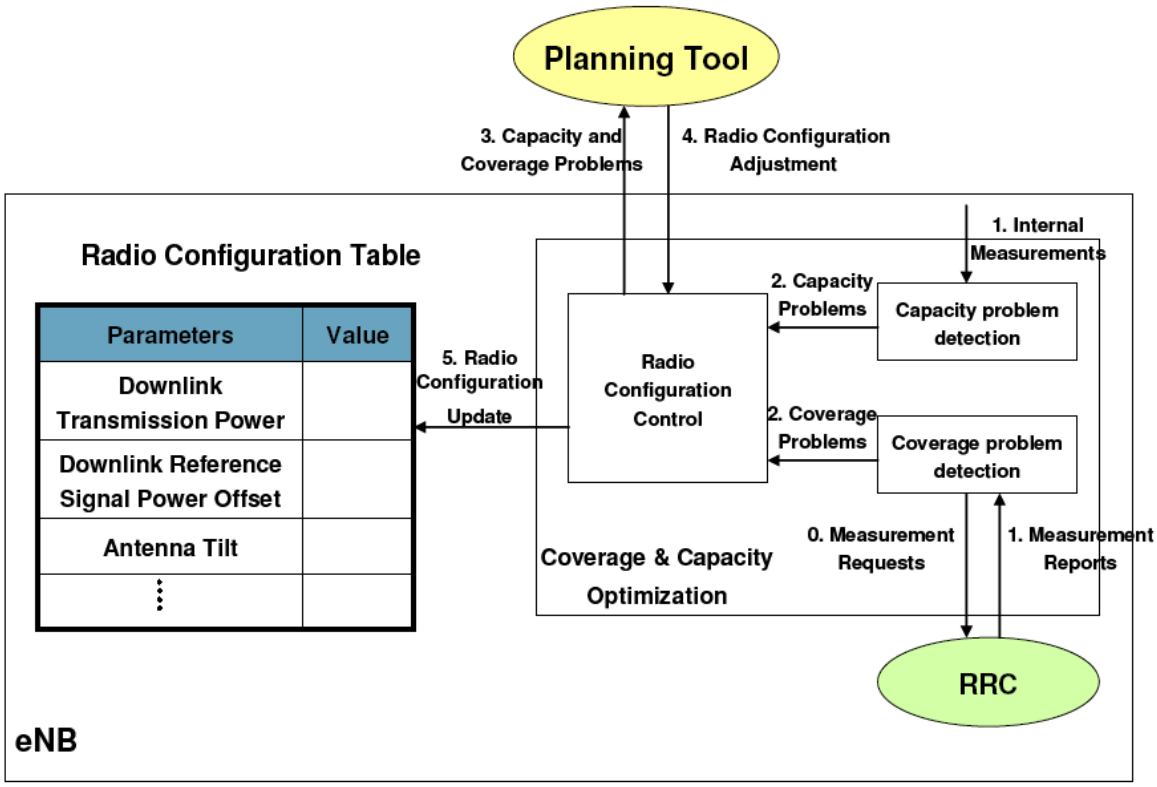

Figure 2.1: Possible Coverage and Capacity Optimization Architecture [11].

Another approach that is suggested is the capacity and coverage optimization model which introduces three high-level use cases for coverage and capacity optimization [12]; (1) E-UTRAN coverage holes with 2G/3G coverage; (2) E-UTRAN coverage holes without any other radio coverage; (3) E-UTRAN coverage holes with isolated island cell coverage which is approached in [12]. The isolated island cell area has its actual coverage smaller than the planned coverage; those uncovered areas are considered coverage holes that have to be identified and optimized by the optimization model of capacity and coverage. The case presented in this paper involves the automatic adjustment of antenna tilt based on the network traffic and the users' location. The function of coverage and capacity optimization is based on the performance measurements along

with the geographical binning; whereas, the behavior of the model is determined by the strength of the received signals that belong to specific users. An initial antenna tilt value is taken and signal strength is measured, compared with other antenna tilts values and their signal strength, and then the best coverage is chosen.

In [23], Combes et al. proposed a coverage and capacity self-optimization scheme that is based on α-fair schedulers including Proportional Fair (PF), Max Throughput (MTP) and Max-Min Fair (MMF) schedulers. This method considers adjusting the packet scheduling strategy to dynamically enhance coverage and capacity using different network Key Performance Indicators (KPIs) to choose the optimal α. The strategy used is stimulated by the problem of Multi Armed Bandit (MAB) which allows assigning the optimal α dynamically. As shown in the simulation results, the proposed coverage and capacity self-optimization scheme increases the coverage of users at cell edge while minimizing capacity loss in cell throughput as well as power consumption.

The approach presented in [24] uses antenna down-tilt adaptation to provide enhanced coverage and capacity optimization. The coverage and capacity optimization problem is approached by using a solution that is based on Fuzzy Q-Learning strategies that provide independent optimization process by presenting learning speed and convergence to optimal settings. Different cases are studied: stable and dynamic strategies. The former allows one cell to take action at a time; whereas, the latter enables many cells to take actions simultaneously at a time. Moreover, a hybrid strategy was introduced to merge between the advantages of the two strategies, and it proved to have better results since it is faster than the stable strategy, and converges faster and performs better than the dynamic strategy.

2.3.1.2 Energy Saving and Interference Reduction

Despite the fact that saving energy takes a massive role in cutting down the operational expenses, some approaches may affect the throughput and the performance of the network. The capacity offered by the network has to be as close as possible to the required traffic demand to ensure cutting the expenses while still ensuring the ability to provide good performance. Interference

reduction methods and energy saving approaches can be used to improve the performance of the network and enhance its capacity. Energy saving is based on many possible solutions such as transmission power adaptation, multi-antenna scheme adaptation, and switching on/off cells, and interference reduction can deploy the previous methods as well as frequency reuse techniques [11].

Xu et al. [13] suggested two approaches to deal with energy saving. In fact, they proposed two Home eNB (HeNB) adaptive transmission methods that aim to save energy and optimize coverage and capacity. The approaches deal with interference and power consumption problem caused by the fact that HeNB is supposed to send simultaneously even if there is no UE connected to it, and the UE is supposed to move to the HeNB macro coverage before moving to the HeNB. The first approach controls sending messages based on the connections with the HeNB, and the second approach suggests different transmission states for the HeNB based on the need of exchanging information and the type of information that has to be exchanged.

In [25], various calculations and simulations show that one of the methods that can be used for energy saving is adding femto cells to the macro cells deployment in the network. This method proposes a way that saves energy without highly affecting the throughput of the system. It improves the system throughput, performance and energy efficiency. The two techniques that were proposed in this paper are selective disconnection of cells and power reduction. The former adopts the concept of choosing particular cells to be switched off while maintaining a network free of coverage gaps and this approach achieves energy reduction, but it also affects the available throughput; whereas the latter reduces the transmitted power for all base stations. In fact, the power reduction approach performs better than the selective disconnection or switching off due to the fact that the power reduction approach can be applied to all overlapped cells. As stated in [25], a 4 dB power reduction decreases throughput of only 10% while achieving energy saving of 36%. On the other hand, switching off 15 cells out of 54 cells, reduces the throughput by 40% while achieving 26% of energy saving.

One of the problems in mobile stations of Tactical Mobile Communication Systems (TMCS) is that the single antenna that is used does not take into consideration the tactical operations.

Hence, it wastes power and capacity. In [26], this problem is addressed by taking the tactical situation into account and an Adaptive Antenna Switching (AAS) method is proposed to save energy, reduce power consumption, and to maximize operational efficiency and survivability. The proposed AAS structure includes Operation Management Equipment (OME), Access Control Router (ACR) and Radio Access Station (RAS) as well as the SON coordinator that manages the requests of the other self-organizing network functions. In addition to that, it has the AAS algorithm block which controls the arithmetic switching of the AAS.

Different scheduling and frequency allocation algorithms are used to minimize the inter-cell interference such as Soft Frequency Reuse (SFR), Partial Frequency Reuse (PFR), and Fractional Frequency Reuse (FFR) [19]. In addition to the previous techniques, the graph-based approach, which is considered a heuristic algorithm, is used in [18] to minimize inter-cell interference by allocating different resources for connected UE. In this technique, colors correspond to different set of frequencies, and each node is assigned a color that is not similar to any connected nodes.

One approach that leads to capacity and throughput improvements is using the inter-cell interference mechanisms based on managing the radio resources, and taking overload, resources priority and transmission power into consideration [20]. Another approach that maximizes the overall network utilization and reduces interference, proposed in [21], is based on soft fractional frequency reuse and adjusting transmit powers on a per-beam base. An algorithm that improves the system performance is proposed to reduce interference based on deploying a network that is predictable. The Virtual Sub-band Algorithm (VSA) adopts a new technique in which all beams are always switched-on, and this makes the transmit power and channels known and predictable.

2.3.1.3 Physical cell ID Assignment

Assigning a physical cell ID is necessary for the eNBs especially the ones that are newly installed, and the assignment problem is considered a complex problem since it needs certain requirements.

The approach presented in [17] maps the physical ID assignment problem to graph coloring method that colors nodes in a way that ensures not coloring any two nodes connected to the same edge with the same color and at the same time acquiring a minimum number of colors. The minimum number of colors is called the chromatic number and it is considered a NP-complete problem [85]. Physical ID assignment approach is presented in three different operations: (1) initial configuration which uses the extension of Welsh and Powell greedy algorithm; (2) incremental network expansion which uses graph coloring while restricting changes to small partitions of the graph; and (3) the confusion repair phase which uses the previous algorithm to change the physical ID of cells causing confusion.

In [18], a few asynchronous local search algorithms, as well as the complete algorithms Asynchronous Weak-Commitment (AWC) and Asynchronous Backtracking (ABT), are selected for evaluation. Four simple distributed local search algorithms will be used for graph coloring: (1) Bin; (2) Real; (3) Bin-Multi; and (4) Real-Multi. These algorithms can be classified according to: (1) the interference pricing between neighbors using the same resource; and (2) the number of alternatives tried by a node, when updating which resource to use. The best local search algorithms perform better than the complete ABT algorithm, because the latter relies on comprehensive search that is not possible with limited number of iterations and colors. AWC has a better performance than ABT especially when it comes to convergence; the reason is that AWC uses dynamic updating of priorities.

2.3.1.4 Mobility Robust Optimization

The HO procedure in LTE occurs between the serving eNB, and the target eNB. The former controls the serving cell as base station, and the latter controls a HO target cell. The measurement report sent from a UE to its serving eNB, that initiates the HO process, is triggered by different conditions such as power level.

The HO margin optimization algorithm, proposed in paper [22], observes the type of HO failure and tracks the cause of that failure before performing any changes in the UE mobility. Consequently, this algorithm can be used for spectacular changes in UE mobility by adjusting

parameters without the need to add UE mobility estimation functions. The algorithm adjusts the HO margin either by an increase or a decrease based on the selected foremost HO failure event; too early HO, too late HO, and HO to wrong cell, as well as the Ping-Pong HOs that is the process in which a UE goes back to the original serving cell in a specific amount of time after a HO. The HO margin optimization algorithm is robust against any changes in the mobility of UE and that was demonstrated by evaluating the algorithm and performing a parametric study taking into consideration the changes in the moving direction and the velocity of UE.

In [27], the MRO problem is approached by presenting the relationship between UE speed and Hyst parameter (handover parameter considered as a window frame). In fact, the proposed User Speed Mobile Robust Optimization (US-MRO) algorithm assigns different hysteresis parameters for UE with different speeds. The US-MRO algorithm enhances the performance, improves the HO success rate and boosts the user experience. The user's speed in LTE is divided into three levels: normal, medium and high; the information on each level directly impacts the HO optimization in MRO. In other words, the eNB has different speed levels to be selected by users. The different speed levels lead to different Hyst parameters that are reported to users. As a result, users choose the most appropriate HO parameters based on their speed.

The MRO algorithm defined in [28] identifies the inter-Radio Access Technology (inter-RAT) mobility configuration parameters (i.e. different thresholds, time to trigger (TTT), and filter coefficient) and investigates its problems (i.e. too late handover or unnecessary handover aka Ping-Pong). In addition to that, different KPIs are used to evaluate the performance of inter-RAT handover to same cell or different cell, too early or too late inter-RAT handover or even handover to wrong cell of new RAT. The intra-LTE handover occur when a UE is moving from the LTE source cell to the LTE destination cell, and the interference caused by a strong signal may lead to radio link failure due to too early or too late handover. However, the cell edge problem does not exist in inter-RAT handover due to the fact that source and target cells do not operate at the same frequency, also the area, where UE selects the good signal quality of either the source or the target cell, is large. The intra-LTE is caused by radio condition, whereas the inter-RAT is policy driven.

2.3.1.5 Mobility Load Balancing

In cellular networks, the cell load is unequal because traffic is random, time varying and usually unbalanced. Some cells may be overloaded with a great number of UE, while other cells' resources are not fully utilized. The current methods for network optimization improve the system capacity, and reduce the manual intervention in managing and optimizing the network; however, they fail to completely solve the LTE load balancing problem. Current methods allow load balancing by utilizing the less congested cells to serve UE located at the border of an adjacent overlapped cell that is more congested, but on the other side they have their drawbacks. In other words, traditional load balancing methods aims at enhancing throughput, delay and load balancing without taking into account the frequency of handovers.

According to [29], there are two different approaches for load re-distribution: (1) expanding the coverage area and increasing the pilot power to cover more UE, or (2) load migration from a heavily loaded cell to a less loaded neighboring cell. The first approach introduces different problems as well as increases the possibility of creating coverage holes; whereas the second approach enhances the resource utilization.

The main goal of MLB algorithm proposed in [11] is to optimize cell reselection and handover parameters in order to handle the imbalanced traffic load and reduce the number of handovers required to attain load balancing. The algorithm uses the eNBs to measure the load of their cells and then exchange the information. As a result, it distributes the load among the cells based on their needs and ensures that handover and cell reselection parameters are tuned in both cells.

Hu at al. [14] proposed a new MLB algorithm that takes into consideration the average delay of the system and the average number of handovers. The proposed MLB algorithm with penalized handovers aims to balance the unequal traffic load, enhance system performance and reduce the number of handovers. Hence, assigning a set of UE to a cell that provides better capacity, enhanced queue backlog, and higher data rate ensures the system stability and improved capacity. On the other hand, the handover process is costly and is not preferred to occur frequently. Hence, UE are preferred to be associated with their current serving cell. In addition to

that, to reduce the overhead caused by handovers, not more than one UE is accepted by a cell; and a UE cannot handover toward more than one cell. In this paper, assigning penalty on handover is suggested. The penalty is assigned to a pair of UE and cell in case there is a need for a handover that assigns UE to the cell. Also, the problem of choosing a set of UE for handovers is reduced to a Maximum Weighted Matching (MWM) problem that is solved by a greedy distributed algorithm providing low complexity and system overhead. The MLB suggested in [14] is a trade-off between the number of handovers, and the average queue backlog (average system delay); the larger the penalty factor is, the more the algorithm tends to reduce the number of handovers at the expense of having a larger average queue backlog and vice versa.

In [29], a capacity based MLB algorithm that aims to optimize the cell throughput, is proposed. The algorithm takes into consideration the load status of the source cell as well as the neighbor cells to choose the target cell for load balancing. The fact that MLB is executed in a cell cluster taking neighboring cells into consideration reduces the MLB confusion caused by the MLB collision of two overloaded neighboring cells requesting load balancing at the same time. The main goal of the proposed algorithm is to transfer the maximum number of users from overloaded cells to neighboring cells with the minimum number of rejections. In fact, the simulation results show that by adopting this algorithm the throughput is improved, user experience and network QoS are enhanced, and Ping-Pong effect is eliminated.

Similarly, the mobility load balancing algorithm introduced in [30] takes the load status of neighboring cells along with the source cell while dynamically adjusting the Radio Resource Management (RRM) parameters. However, this capacity based algorithm deals with the user signal-to-interference and noise ratio (SINR) rather than the cell throughput that is investigated in [29]. The algorithm proposed in [30] aims to minimize the number of displeased users, in addition to improving the overall user satisfaction by dynamically modifying the handover margin and handover parameters based on certain commands and conditions. As the simulation results show, the adjustment of handover margin improves the user satisfaction and reduces the number of unsatisfied users.

2.3.2 Core Network Planning

The amount of work that has been done on the core network is scarce compared to what has been done in the radio part. Nevertheless, few papers were found to be dealing with the core network. For example, the roaming and interoperability issues in the core layer have been investigated in [35] taking into consideration mobility management, routing, and real time charging. Also, Corici et al. [36] approached resource reservation issues and proposed mechanisms for machine type communication over the EPC.

In general, there are different cases for LTE deployment as well as diverse patterns for LTE evolution. Accordingly, a variety of technologies are proposed to perform roaming and attain interoperability. For example, in some cases Circuit Switched (CS) Fallback and Single Radio Voice Call Continuity (SRVCC) are used, yet there are many different cases where these two cannot work. Sanyal in [35] introduces many challenges in mobility management, message routing, policy control and real time charging. In mobility management, traditional networks use Mobile Application Part (MAP) protocol between the Visitor Network (VN) and the Home Network (HN) for managing mobility and location, as well as providing authentication. On the other hand, the LTE core network uses the Diameter protocol to perform the operations achieved by the MAP. In message routing, Diameter proxy is introduced because the Diameter protocol does not provide the network with the ability to route messages and map correctly to the destination IP address, so the Diameter proxy element enables routing and interoperability between different domains. In policy control and real time charging, it is concluded that there is no policy control and QoS enforcement in legacy networks, also there is a huge gap between LTE and 2G/3G real time charging models. The solutions proposed in [35] are based on conversion between Diameter and MAP, or even using different elements in the architecture (i.e. Diameter Proxy).

The development of a wide range of wireless devices (i.e. sensors and actuators) that are used in different sectors like health, transportation, education, security, among others; along with the increase in demand for smart phones and tablets boost the number of devices connected to the network. Consequently, the scalability needs to be improved, along with the mobility, charging

control and resource management as well as the ability to provide an increased throughput and a reduced delay. Each service, functionality and device need to reserve network resources despite the fact that they may be all handled jointly instead of being treated separately. Moreover, the increased number of attached devices increases the overhead in managing bearer and in the core network. Corici et al. [36] address the problem and propose an efficient resource reservation method represented in the 3GPP Evolved Packet Core (EPC), and based on grouping the devices according to their functionalities as well as predicting the amount of resources needed in core network. It highly depends on the event notification mechanisms as well as caching.

The idea behind the proposed method relates to the Policy and Charging Control (PCC) architecture that controls the resource reservation. In fact, the PCRF is in charge of making policy decisions according to the UE requirements. Basically, the suggested approach has two main procedures: service provisioning, and communication establishment and termination. In fact, it enhances the concept of bearers by introducing the function of time allowing the process to utilize core network resources more efficiently with less amount of signaling. Moreover, resource optimization is enhanced due to the fact that bearers are not limited for a single device, but a group of devices that has the same functionality. In addition to that, the ability to assign bearer information prior to communication reduces signaling.

2.4 Summary

In conclusion, the expansion of LTE network and the high demands on its services require operators to invest in network planning. It is important to implement mechanisms and adopt approaches that enhance network planning, and ensure planning a network that is capable of satisfying the needs of the users, and the requirements of the operators. Users need good quality services at reasonable cost; they also need to use the network without disruption or problems. As a result, the main goal for operators is to have a network that has optimal cost and still can deliver good quality services.

The effort spent on the radio portion of the network is measurable; however the core network, despite its importance, didn't get the same amount of attention. Many network planners who based their planning on traffic didn't focus on traffic generation, but rather used estimated or measured traffic values. In addition to that, the complexity of the network makes it more complex to be solved using simple methods. In our thesis, we developed a network planning tool that deals with the core network of LTE taking into consideration realistic traffic with its different aspects (i.e., EPSB, BHSA, signaling and bandwidth). We proposed two different algorithms to solve the core network planning problem (i.e., the exact and the approximate) in addition to that we developed a mathematical model that has a set of decision variables and constraints that will be explained in Chapter 3.

Chapter 3 LTE Network Model Formulation

In this chapter, we first present the type of traffic that is handled by the EPS, and then we propose a tool to generate realistic traffic in order to plan the evolved packet core of LTE networks. The next section describes the inputs, outputs, and the objective of the LTE planning problem. The mathematical model is explained thoroughly in this chapter, and both the exact method and the approximate method using the local search algorithm are described in details.

3.1 EPS Traffic

Due to the fact that traffic measurement plays an important role in planning a network and measuring its performance, there is a need to understand the traffic flow in the network and understand the different types of traffic carried on the links. It is also important to understand the types of links and interfaces used to carry the traffic. Traffic measurement is a major factor for network planning and design; it is as important as evaluating network capacity, number of nodes, network latency, and performance measurements. Traffic is simply defined as the amount of data carried over a link for a given period of time. In LTE, there is a classification based on the delay sensitivity that divides LTE traffic into 4 different classes [38]: conversational class, streaming class, interactive class, and background class.

The conversational class is considered the most delay sensitive since it carries real time traffic such as VoIP, and video conferencing. The user has the ability to control the length of the session [62], in other words the session ends whenever the user chooses to end the conversation. Compared to the conversational class, the streaming class is considered less delay sensitive, and it generally carries traffic for streaming purposes such as streaming audio or video. Some examples on the streaming class applications [62]: movies, news, education and training. Regarding the other two types of traffic, the interactive class is delay insensitive but not as much as the background class since it is considered the most delay insensitive class; the former deals

with interactive services such as e-mail and web browsing; whereas, the latter deals with background traffic such as background downloading of e-mails and databases.

Planning an efficient network that is based on realistic traffic is not an easy task, and to be able to develop a tool that considers realistic traffic there are certain traffic parameters that need to be introduced:

- **The number of subscribers:** This parameter represents the total number of subscribers that are currently covered by a given eNB.

- **The number of attached subscribers in Busy Hour (BH):** This parameter represents the number of LTE subscribers that were able to have a successful connection with the P-GW along with a successfully established default bearer and successfully allocated IP address. BH is known to be the busiest 60 minutes period of the day, in which the total traffic is the maximum throughout the day.

- **Busy Hour Data Session Attempt (BHDSA):** This parameter represents the number of data sessions attempted in a busy hour, and it is one of the main methods to measure the capacity of the network.

- **Busy Hour Voice Session Attempt (BHVSA):** This parameter represents the number of voice sessions attempted in a busy hour.

- **Bandwidth required for bearer sessions (BW):** This parameter characterizes the amount of throughput required for the users' services.

- **Simultaneous Evolved Packet System Bearer (EPSB):** This parameter shows the number of EPS bearer sessions occurring simultaneously in a busy hour. The EPSB is an established end-to-end connection between the UE and the P-GW to provide the users with the Internet services they need.

The eNBs traffic profile is calculated based on the subscriber traffic profile presented in Table 3.1 as well as other planning parameters presented in

Table **3.2** Usually, the information shown in Table 3.1 is provided by the operator or the service provider and it varies depending on data and voice plans provided by different operators; whereas, in

Table **3.2** some of the values were referenced and the others were assumed based on knowledge and logic.

Two main values that affect the dimensioning process assuming asymmetric services are presented in Table 3.1; the average downlink rate which was taken to represent a single direction of communication in dimensioning, as well as the monthly usage or data traffic per subscriber in a month. For example, Telus in Canada (in 2012) states that the current 4G LTE enable users to access the networks with speeds up to 75 Mbps with an expected average of 12-15 Mbps[39], while Bell LTE network in Canada (in 2012) is able to offer speeds up to 150 Mbps with an expected average speed of 18-40 Mbps [40]. Fido's data usage per subscriber starts from 100 MB up to 5 GB [41], and TELUS provides data usage that goes up to 5 GB [42].

Table 3.1: Dimensioning values - subscriber traffic profile

Average rate	25Mbps
Monthly usage	2GB/month/sub

Table 3.2: Planning parameters

Adaptive multiple rate [47] [49]	12.2 kbps
Mean session time [50]	180 sec
Handover ratio [49]	0.4
IP overhead percentage	50%
Dense area attached subscriber ratio	0.9
Active BH EPSB ratio	0.5
Average EPSB session duration (in seconds) [65]	900
Retransmission factor	0.25
S1U utilization factor [61]	0.8
Working days per month [45]	22
Working days traffic ratio [45]	0.9
Busy hour traffic ratio [50]	0.15
Voice Activity Factor (VAF) [48]	0.5
Burstiness [46]	0.25

A network must be capable of handling the highest amount of traffic during busy hour and it should be designed to support and provide a decent Quality of Service to all of its subscribers in a busy hour. In order to have a reasonable amount of traffic, different types of traffic were taken into consideration as previously mentioned; web browsing, file downloads, e-mail, messaging, conversation voice, conversation video, gaming, and streaming.

The average busy hour usage per subscriber is calculated using Equation 3.1 in which *k* is a constant that is equal to 1024*1024*1024*8 and is used to convert the monthly usage of GB provided in

Table **3.1** to bits. Working days traffic ratio represents a percentage of the amount of traffic that occurs during working days, and busy hour traffic ratio resembles a percentage of the amount of traffic that occurs during a busy hour. Average BH usage is measured in bits/subscriber/busy hour taking into consideration that a busy hour is an hour with the highest amount of traffic during the day.

$$\text{BH usage} = \frac{\text{monthly usage} * k * \text{working days traffic ratio} * \text{busy hour traffic ratio}}{\text{working days per month}} \quad \textbf{(3.1)}$$

Voice is supported in LTE using different techniques [55][63] such as Voice over LTE (VoLTE) and Circuit Switched Fallback (CSFB). IP Multimedia Subsystem (IMS) was supposed to be more available when LTE networks started; however, it was not as expected and it caused challenges in supporting Voice over LTE. For example, one of the issues faced by the LTE networks is a major software problem in one of the IMS elements which disrupted the LTE network services and affected the VoLTE service [77]. As a result, CSFB, which provides subscribers with voice services by using networks of previous generations such as GSM or UMTS was deployed. Due to the fact that the main goal is to support voice in LTE over IMS, VoLTE initiative was announced to develop a framework that supports voice over IMS in LTE [86].

Taking into account the VoLTE, there are several factors that control the voice bandwidth [52]: Codec (de/coder) and sample period, IP header, transmission medium and silence suppression. Adaptive Multiple Rate (AMR) codec increases the voice capacity and it uses multiple voice encoding rates ranging from 4.75 to 12.2 kbps [53][54]. For example, using the AMR 12.2 kbps codec rate causes a data rate of 30 kbps on top of IP [49].

Equation 3.2 is used to get the average throughput of the S1U interface per subscriber. Due to the fact that the S1U interface, which is the air interface between the eNB and the S-GW, carries different types of data, and voice sessions with different data rates; traffic is bursty, and as a result, the burstiness factor is included in the equation. In [43], burstiness is a representation of a group of packets with shorter gaps between other packets being handled before or after, and it has a value between 0 and 1. When the value gets closer to 0, the traffic gets more bursty.

Due to the fact that the S-GW is considered a mobility anchor for inter-eNB handovers, there is a need to include the Handover Ratio (HO) as indicated in the equation. Moreover, taking into consideration different types of applications and services, some may require packet retransmission in case of failure; therefore Retransmission Factor (RTF) is also included in this

equation; *k* is a constant that is equal to 3600 and is used to ensure dealing with S1UBW of rate bps since the BH usage acquired in Equation 3.1 is in terms of busy hour. Furthermore, the main two parts of the equation are controlled by the Voice Activity Factor (VAF) to ensure calculating the period in which voice is active, and other periods where other data applications are being handled. In addition, voice data constant represents the amount of data that needs the AMR codec for transmission, and since data is transmitted over IP, the IP overhead is taken into account.

$$\begin{aligned} S1UBW = {} & ((1-VAF)(BH\ usage/k) + (VAF * AMR * voice\ data\ constant)) \\ & * (1+HO\ ratio) * (1+RTF) * (1+IPoverhead) * (1+burstiness) \end{aligned} \quad \textbf{(3.2)}$$

As stated in [56], BHSA provides the number of session attempts during the busy hour. In [57], BHSA is calculated for each user by multiplying the busy hour traffic intensity, which represents amount of usage per subscriber in busy hour, by 3600 and dividing it by mean session duration. In Equation 3.3, BHSA represents the maximum number of busy hour session attempts for all users. In order to get the number of attempts in an hour for all the users, the number of attached subscribers is multiplied by the traffic intensity and 3600 seconds as defined in [51] and then divided by the mean session time. The unit for BHSA is in term of number of sessions per busy hour.

$$BHSA = \frac{attached\ subscribers * traffic\ intensity * 3600(sec)}{mean\ session\ time(sec)} \quad \textbf{(3.3)}$$

Due to the fact that there are two different types of BHSA (i.e. voice and data); we end up with two separate equations for BHVSA and BHDSA as shown in Equation 3.4 and Equation 3.5 respectively. The voice traffic intensity is equal to 20, and the data traffic intensity is averaged to 1.52 as provided in [59].

$$BHVSA = \frac{attached\ subscribers * voice\ traffic\ intensity * 3600(sec)}{mean\ session\ time(sec)} \quad \textbf{(3.4)}$$

$$\text{BHDSA} = \frac{\text{attached subscribers} * \text{data traffic intensity} * 3600(\text{sec})}{\text{mean session time}(\text{sec})} \quad \textbf{(3.5)}$$

In order to calculate the number of attached subscribers for the eNB, we need to multiply the total number of subscribers by the attached subscriber ratio for dense area (provided in

Table **3.2**) as shown in Equation 3.6. The attached subscriber ratio depends on the region, whether the area is crowded or the population in not considerable.

$$\text{Attached subscribers} = \text{subscribers} * \text{attached subscriber ratio} \quad \textbf{(3.6)}$$

A subscriber has the ability to have more than one simultaneous bearer session, and one of the advantages of this feature is that a user would be able to connect to a PDN for Internet service, and simultaneously connect for video or to another PDN (e.g., IMS) [60]. The number of simultaneous EPS dedicated bearer sessions handled by the S1U interface is calculated using Equation 3.7. It is the result of multiplying the number of attached subscribers by the average duration of each data bearer session and the active BH EPSB ratio which represents the percentage of active sessions and it is controlled by the operator and the capabilities of the network. In addition to that, the answer is divided by 3600 sec to make sure we get the value in a busy hour.

$$\text{eNBsimult. EPSB} = \frac{\text{attached subscribers} * \text{Average EPSB session duration (sec)} * \text{active BH EPSB ratio}}{3600(\text{sec})} \quad \textbf{(3.7)}$$

Prior to calculating the total throughput of the S1U interface, the Fair Usage Policy (FUP), which is a method to control accessing the offered services, needs to be described and added to the total throughput. The FUP is applied on users who deplete their quotas to limit the access speed as described in [44]. In fact, service provider makes sure the data provided to users is unlimited; however, speed is reduced until the end of the subscriber's billing cycle. Normally, applications that require high bandwidth such as video streaming will be affected; nevertheless, applications

such as emails and web browsing will not be highly affected due to the fact that they don't require high download rates. Assuming that 10% of attached subscribers depleted their quotas and had to use extra FUP data with size of 1GB, the FUP throughput can be achieved by using Equation 3.8. It is the result of multiplying the number of users using the FUP feature by the number of excess data per subscriber. *k* converts the amount of GB to bits, and its value is 1024*1024*1024*8, the rest of the values including 3600 are used to ascertain that FUPBW is in bps.

$$FUPBW = \frac{0.1 * \text{attached subscribers} * \text{extra data} * k * \text{working days traffic ratio} * \text{busy hour traffic ratio}}{\text{working days per month} * 3600} \quad \textbf{(3.8)}$$

Some users tend to use Top Up (TU) feature which is a significant factor that adds to the total throughput of the network. In fact, a number of users who depleted the quotas, and got their access speeds slowed down or their amount of data usage restricted may choose to purchase extra amount of data. Assuming a user pays $20 for 1 GB of data in a month, and the number of users who use the Top Up feature is 20% of the total number of attached subscribers, top up BW is shown in Equation 3.9. It is the result of multiplying the number of users using the TU feature by the excess data per subscriber. *k* converts the amount of GB to bits, and its value is 1024*1024*1024*8, the rest of the values including 3600 are used to ascertain that TopUpBW is in bps.

$$TopUpBW = \frac{0.2 * \text{attached subscribers} * \text{extra data} * k * \text{working days traffic ratio} * \text{busy hour traffic ratio}}{\text{working days per month} * 3600} \quad \textbf{(3.9)}$$

Equation 3.10 is used to measure the total amount of bandwidth being carried on the S1U interface assuming one bearer session for each attached subscriber. Taking into consideration that utilization is resulted by dividing the traffic load (bps) by the capacity (bps); where capacity is the maximum amount of load supported by the network. The S1U total BW is a result of adding the FUPBW and the TopUpBW to the amount of simultaneous EPS bearers multiplied by the average S1U throughput and the result of the summation is divided by the interface

utilization which ought to be less than 0.85. In order to achieve the total bandwidth in Mbps, the answer is divided by 1M.

$$\text{S1U total BW} = \frac{(\text{eNB simult. EPSB} * \text{S1U BW}) + \text{FUPBW} + \text{TopUpBW}}{\text{S1U interface utilization} * 1000000} \quad \textbf{(3.10)}$$

In terms of the control plane, the signaling and the control messages on the S1-C adds to the load of the network [58]. The control operations that are performed over the S1-C include signaling for attachment and detachment, as well as EPS bearer establishment and management, along with authentication requests and responses. Due to the fact that the eNB is directly connected to the MME, the MME has a massive load of transactions per second with both the HSS and the S-GW. Each element has a different amount of signaling because it requires a different amount of transactions per seconds for each of its operations. The elements have different functions and various processes so the amount of signaling is local and dependant on the equipment itself. Moreover, it varies depending on the vendor and the specs for the specific model of the element. The process in which HSS signaling is calculated is different than the process to calculate the signaling of the rest of the elements. Equation 3.11 and Equation 3.12 show the two ways to calculate the amount of signaling for the HSS, and the rest of the elements (MME, S-GW, P-GW and PCRF) respectively. The amount of transactions per second is not absolute; it simply depends on the element, the provider, and the model along with different parameters. Equation 3.11 presents the calculation of the total amount of signaling for the HSS, which is the result of the number of transactions per second per subscriber multiplied by the total number of subscribers; whereas in Equation 3.12, the total amount of signaling for the each element is calculated by multiplying the number of bearers during a busy hour by the number of transactions per second per bearer.

$$\text{Signaling HSS} = \text{subscribers} * \text{transactions per seconds per subscriber} \quad \textbf{(3.11)}$$

$$\text{Signaling core elements} = \text{EPSB} * \text{transactions per seconds per bearer} \quad \textbf{(3.12)}$$

For example, from [64] it can be concluded that the MME handles 9.3 transactions per second per bearer since it handles around 290,000 messages per second. For the P-GW, it handles 63,000 messages per second leading to 2 transactions per second per bearer. For the S-GW, the number of messages per second is 94,000 leading to 3 transactions per second per bearer. The fact that the S-GW is higher than the P-GW is that the latter doesn't have to deal with service request, release and paging messages. For the PCRF, as given in [65], the number of transactions per second per bearer is equal to 2. By using the Diameteriq smart signaling tool and taking into consideration the results shown in [66], the diameter messages per second in HSS are equal to 5,000 messages per second. Hence, it equals 6.2 transactions per second per subscriber multiplied by the total number of subscribers.

Taking into account the different types of traffic, the eNB traffic records presented in Table 3.3 shows different aspects of traffic measurements that were based on the previous equations and values.

Table 3.3: Typical example of the traffic record for 5 eNB

eNB Index	**Subscribers**	**Attached Subscribers**	**BHDSA (Data)**	**BHVSA (Voice)**	**BW (Mbps)**	**EPSB**
1	10,101	9,090	276,336	3,636,000	186	1136
2	18,577	16,719	508,257	6,687,600	343	2089
3	10,567	9,510	289,104	3,804,000	195	1188
4	15,336	13,802	419,580	5,520,800	283	1725
5	12,297	11,067	336,436	4,426,800	227	1383

The traffic record shown in Table 3.3 has different traffic profiles of 5 eNBs. The first column, eNB index, represents the index of the different eNBs to which the traffic belongs. The second column has the total number of subscribers covered by each eNB (i.e. subscribers whether they are active or idle), and it is randomly generated between 6,000 and 19,000. Attached subscribers, are represented in the third column, busy hour session attempt is presented for data and voice in columns four and five respectively. In column six, bandwidth measured in Megabits per second (Mbps) is presented; whereas in column seven the number of simultaneous sessions of dedicated EPS bearer is provided.

Once the traffic records are achieved and the amount of traffic being handled by the network is analyzed, a realistic traffic profile is ready and the planning process of the evolved packet core of LTE networks can start as described in Section 3.3. A set of mathematical notations is described along with a set of decision variables, and parameters to deal with the model and its features.

3.2 Planning Problem: Inputs, Outputs, and Objectives

For the model formulation and the planning problem, we assume that the following information is given:

- The different types of links and network elements (i.e., MME, S-GW, P-GW, HSS, and PCRF), as well as their capacities and costs.
- The eNBs' locations.
- The different potential locations of the core network elements (i.e., MME, S-GW, P-GW, HSS, and PCRF).
- The subscribers' traffic profiles provided by the operator.
- The number of subscribers covered by each eNB.
- The facilities installation cost (i.e., manual labour, electric installations, access floor spaces, cabling and grounding, cable runways, power and cooling equipment, equipment racks, etc.).
- The planning parameters.

The output of the planning problem focuses on the development of an automatic planning tool for the design of the evolved packet core of LTE networks. More precisely, it involves:

- Planning the topology of the network.
- Defining the number of core network elements (i.e., MME, S-GW, P-GW, HSS, and PCRF).
- Defining the number of links and interfaces needed to connect the core elements.
- Specifying the type of links and interfaces.
- Specifying the type of core network elements.
- Selecting the locations of the different core elements.

The main goal of the core network planning problem is to find the optimal network configuration with the minimum infrastructure cost which includes the cost of the nodes, links, and interfaces.

3.3 Model Formulation

After defining the objective of the problem, and the set of constraints, and collecting all needed information about the nodes, links, and interfaces, the model can be formulated. The features of the nodes, links and interfaces are translated into a mathematical representation. This translation process from analyzed information into mathematical modeling is known to be the model formulation; formulating a model for Linear Programming problems (LP) is expressing the objective, and each constraint algebraically in terms of the decision variables. The objective is to minimize the cost of the network while keeping good quality of service and reasonable network performance. In terms of traffic, each element of the core network has its own traffic types; hence, the eNB is considered the point which carries all the different types of traffic.

3.3.1 Notations

The Evolved Packet System (EPS) architecture depicted in Figure 3.1 presents the different elements of the LTE network for the flat access network and the core network; it also shows the links between those elements. In fact, Figure 3.1 represents the graphical representation of the notation which is composed of sets, decision variables, traffic variables, and cost parameters.

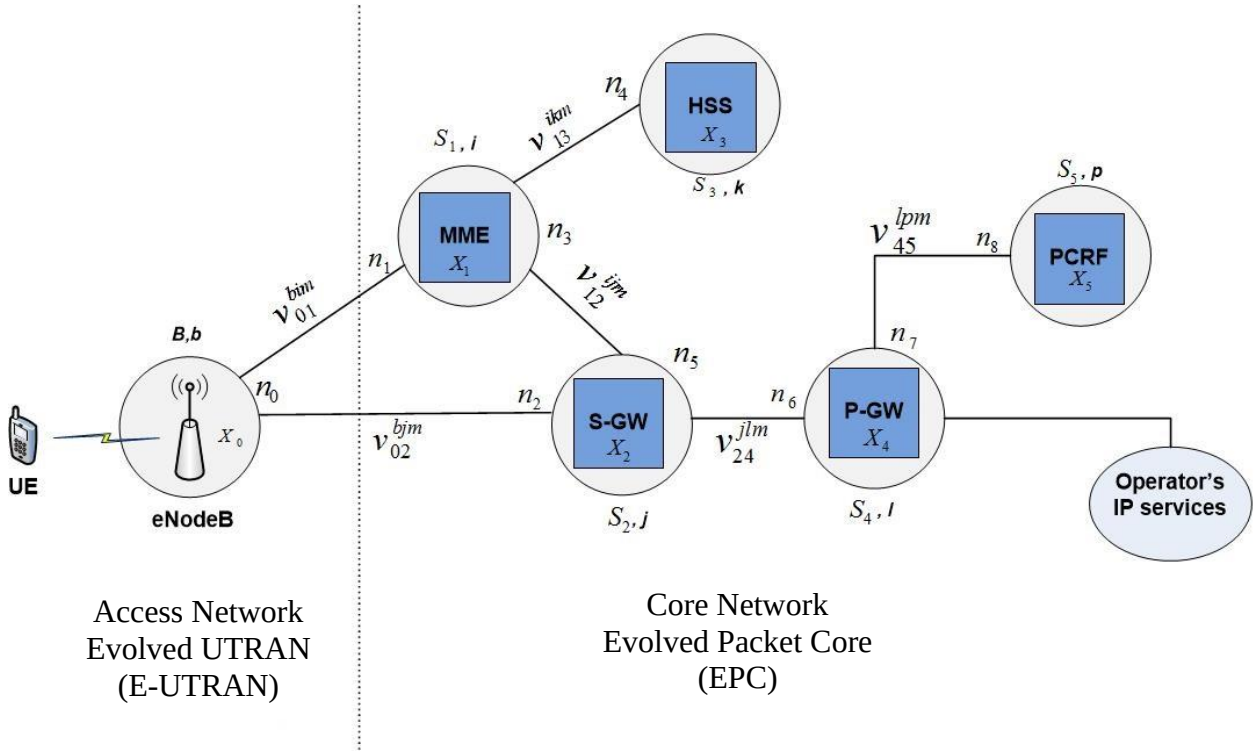

Figure 3.1: Graphical representation of the notation

3.3.1.1 Sets

- B, the set of eNodeBs that are already installed in different locations.
  - μ_b^c, the traffic of type $c \in C$ for an eNB $\in B$.
- C, the set of traffic characteristics such that $C = \{c_0, c_1, ..., c_4\}$ and where:
  - c_0: Number of subscribers.
  - c_1: Busy Hour Session Attempt (BHSA).
  - c_2: Traffic bandwidth (Mbps).
  - c_3: Number of attached subscribers.
  - c_4: Simultaneous Evolved Packet System Bearers (EPSB).

- $M_{01}, M_{02}, M_{12}, M_{13}, M_{24}$, and M_{45} are respectively the sets of links that can be used to connect eNBs with MME, eNBs with S-GW, MME with S-GW, MME with HSS, S-GW with P-GW, and P-GW with PCRF.

- w^m : Capacity of the link or the interface of type m. For example, link of type Fast Ethernet or Gigabit Ethernet.

- S_1, S_2, S_3, S_4, and S_5 are respectively the sets of potential sites to install the MME, S-GW, HSS, P-GW, and PCRF.
- T_1, T_2, T_3, T_4 and T_5 are respectively the set of types of network elements (i.e. MME, S-GW, HSS,P-GW, and PCRF) where:
  - SUB^t : Subscribers' capacity for a node of type *t*.
  - $BHSA^t$: BHSA capacity for a node of type *t*.
  - β^t : Switching fabric capacity in Mbps for a node of type *t*.
  - $EPSB^t$: Simultaneous Evolved Packet System Bearers (EPSB) capacity for a node of type *t*.
  - $ASUB^t$: Attached Subscribers' capacity for a node of type *t*.
  - SIG^t : Signaling capacity for a node of type *t*.
  - $EPSBW$: Throughput per bearer.
  - $SUBW$: Throughput per subscriber.
  - $TPSPB$: Transactions per second per bearer.
  - TPS : Transactions per second per subscriber.
  - $\eta_0^t, \eta_1^t, \eta_2^t, \eta_3^t, \eta_4^t, \eta_5^t, \eta_6^t, \eta_7^t$ and η_8^t are respectively the maximum number of interfaces that can be installed in a node of type *t* to connect the eNB to the MME (S-GW), the MME to the eNB, the S-GW to the eNB, the MME to the HSS (S-GW), the HSS to the MME, the S-GW to the MME (P-GW), the P-GW to the S-GW, the P-GW to the PCRF, and the PCRF to the P-GW.

3.3.1.2 Decision Variables

Links

- ν_{01}^{bi}, a binary variable such that $\nu_{01}^{bi} = 1$ if and only if the eNB ($b \in \mathrm{B}$) is connected to the MME installed at site $i \in S_1$.

- v_{01}^{bim}, the number of links of type $m \in M_{01}$ connecting the eNB ($b \in \mathrm{B}$) to the MME installed at site $i \in S_1$.
- v_{02}^{bj}, a binary variable such that $v_{02}^{bj} = 1$ if and only if the eNB ($b \in \mathrm{B}$) is connected to the S-GW installed at site $j \in S_2$.
- v_{02}^{bjm}, the number of links of type $m \in M_{02}$ connecting the eNB ($b \in \mathrm{B}$) to the S-GW installed at site $j \in S_2$.
- v_{12}^{ij}, a binary variable such that $v_{12}^{ij} = 1$ if and only if the MME installed at site $i \in S_1$ is connected to the S-GW installed at site $j \in S_2$.
- v_{12}^{ijm}, the number of links of type $m \in M_{12}$ connecting the MME installed at site $i \in S_1$ is connected to the S-GW installed at site $j \in S_2$.
- v_{13}^{ik}, a binary variable such that $v_{13}^{ik} = 1$ if and only if the MME installed at site $i \in S_1$ is connected to the HSS installed at site $k \in S_3$.
- v_{13}^{ikm}, the number of links of type $m \in M_{13}$ connecting the MME installed at site $i \in S_1$ to the HSS installed at site $k \in S_3$.
- v_{24}^{jl}, a binary variable such that $v_{24}^{jl} = 1$ if and only if the S-GW installed at site $j \in S_2$ is connected to the P-GW installed at site $l \in S_4$.
- v_{24}^{jlm}, the number of links of type $m \in M_{24}$ connecting the S-GW installed at site $j \in S_2$ to the P-GW installed at site $l \in S_4$.
- v_{45}^{lp}, a binary variable such that $v_{45}^{lp} = 1$ if and only if the P-GW installed at site $l \in S_4$ is connected to the PCRF installed at site $p \in S_5$.
- v_{45}^{lpm}, the number of links of type $m \in M_{45}$ connecting the P-GW installed at site $l \in S_4$ to the PCRF installed at site $p \in S_5$.

Nodes

- χ_1^{it}, a binary variable such that $\chi_1^{it} = 1$ if and only if a MME of type $t \in T_1$ is installed at site $i \in S_1$.
- χ_2^{jt}, a binary variable such that $\chi_2^{jt} = 1$ if and only if a S-GW of type $t \in T_2$ is installed at site $j \in S_2$.
- χ_3^{kt}, a binary variable such that $\chi_3^{kt} = 1$ if and only if a HSS of type $t \in T_3$ is installed at site $k \in S_3$.
- χ_4^{lt}, a binary variable such that $\chi_4^{lt} = 1$ if and only if a P-GW of type $t \in T_4$ is installed at site $l \in S_4$.
- χ_5^{pt}, a binary variable such that $\chi_5^{pt} = 1$ if and only if a PCRF of type $t \in T_5$ is installed at site $p \in S_5$.

3.3.1.3 Traffic Variables

The traffic flow is asymmetric in both directions on a link between two network elements. In other words, the uplink and the downlink don't have the same amount of traffic. The downlink is considered in this thesis since most of the traffic is sent on the downlink direction.

- f_{01}^{cbi}, the traffic of type $c \in C$ on the link from the eNB $b \in B$ to a MME installed at site $i \in S_1$.
- f_{02}^{cbj}, the traffic of type $c \in C$ on the link from the eNB $b \in B$ to a S-GW installed at site $j \in S_2$.
- f_{12}^{cij}, the traffic of type $c \in C$ on the link from the MME installed at site $i \in S_1$ to a S-GW installed at site $j \in S_2$.
- f_{13}^{cik}, the traffic of type $c \in C$ on the link from the MME installed at site $i \in S_1$ to a HSS installed at site $k \in S_3$.

- f_{24}^{cjl}, the traffic of type $c \in C$ on the link from the S-GW installed at site $j \in S_2$ to a P-GW installed at site $l \in S_4$.
- f_{45}^{clp}, the traffic of type $c \in C$ on the link from the P-GW installed at site $l \in S_4$ to a PCRF installed at site $p \in S_5$.

3.3.1.4 Cost Parameters

Links

- a_{01}^{bim}, the link and interface costs (including installation cost) for connecting an eNB installed at site $b \in B$ to a MME installed at site $i \in S_1$ through a link and interface of type $m \in M_{01}$.
- a_{02}^{bjm}, the link and interface costs (including installation cost) for connecting an eNB installed at site $b \in B$ to a S-GW installed at site $j \in S_2$ through a link and interface of type $m \in M_{02}$.
- a_{12}^{ijm}, the link and interface costs (including installation cost) for connecting a MME installed at site $i \in S_1$ to a S-GW installed at site $j \in S_2$ through a link and interface of type $m \in M_{12}$.
- a_{13}^{ikm}, the link and interface costs (including installation cost) for connecting a MME installed at site $i \in S_1$ to a HSS installed at site $k \in S_3$ through a link and interface of type $m \in M_{13}$.
- a_{24}^{jlm}, the link and interface costs (including installation cost) for connecting a S-GW installed at site $j \in S_2$ to a P-GW installed at site $l \in S_4$ through a link and interface of type $m \in M_{24}$.
- a_{45}^{lpm}, the link and interface costs (including installation cost) for connecting a P-GW installed at site $l \in S_4$ to a PCRF installed at site $p \in S_5$ through a link and interface of type $m \in M_{45}$.

Nodes

- $b_1^t, b_2^t, b_3^t, b_4^t$, and b_5^t are respectively the cost (including installation cost) of a MME of type $t \in T_1$, of a S-GW of type $t \in T_2$, of a HSS of type $t \in T_3$, of a P-GW of type $t \in T_4$ and of a PCRF of type $t \in T_5$.

3.4 Cost Function

The total cost of the network consists of two parts: the cost of the links and interfaces as well as the cost of the nodes. The cost of the links and interfaces, denoted by C_L is given by the following equation:

$$\begin{aligned} C_L(v) = & \sum_{b\in B}\sum_{i\in S_1}\sum_{m\in M_{01}} a_{01}^{bim} v_{01}^{bim} + \sum_{b\in B}\sum_{j\in S_2}\sum_{m\in M_{02}} a_{02}^{bjm} v_{02}^{bjm} \\ & + \sum_{i\in S_1}\sum_{j\in S_2}\sum_{m\in M_{12}} a_{12}^{ijm} v_{12}^{ijm} + \sum_{i\in S_1}\sum_{k\in S_3}\sum_{m\in M_{13}} a_{13}^{ikm} v_{13}^{ikm} \\ & + \sum_{j\in S_2}\sum_{l\in S_4}\sum_{m\in M_{24}} a_{24}^{jlm} v_{24}^{jlm} + \sum_{l\in S_4}\sum_{p\in S_5}\sum_{m\in M_{45}} a_{45}^{lpm} v_{45}^{lpm} \end{aligned} \quad \textbf{(3.13)}$$

The cost of the nodes, denoted by C_N, is given by the following equation:

$$C_N(x) = \sum_{t\in T_1} b_1^t \sum_{i\in S_1} \chi_1^{it} + \sum_{t\in T_2} b_2^t \sum_{j\in S_2} \chi_2^{jt} + \sum_{t\in T_3} b_3^t \sum_{k\in S_3} \chi_3^{kt} + \sum_{t\in T_4} b_4^t \sum_{l\in S_4} \chi_4^{lt} + \sum_{t\in T_5} b_5^t \sum_{p\in S_5} \chi_5^{pt} \quad \textbf{(3.14)}$$

3.5 The Model

The model for the LTE core network planning problem can now be given.

$$\text{Minimize } (C_L(v) + C_N(x)) \quad \textbf{(3.15)}$$

subject to the following constraints:

1. ***Uniqueness constraints*** impose that only one type of a specific node is installed at a certain location. Furthermore, nodes are dependent but not co-located. The uniqueness constraints of MME, S-GW, HSS, P-GW and PCRF are as follows:

MME-Type uniqueness constraints

$$\sum_{t\in T_1} \chi_1^{it} \le 1 \qquad (i \in S_1) \quad \textbf{(3.16)}$$

S-GW-Type uniqueness constraints

$$\sum_{t\in T_2} \chi_2^{jt} \le 1 \qquad (j \in S_2) \quad \textbf{(3.17)}$$

HSS-Type uniqueness constraints

$$\sum_{t\in T_3} \chi_3^{kt} \le 1 \qquad (k \in S_3) \quad \textbf{(3.18)}$$

P-GW-Type uniqueness constraints

$$\sum_{t \in T_4} \chi_4^{lt} \leq 1 \qquad (l \in S_4) \tag{3.19}$$

PCRF-Type uniqueness constraints

$$\sum_{t \in T_5} \chi_5^{pt} \leq 1 \qquad (p \in S_5) \tag{3.20}$$

2. ***Assignment constraints*** impose that a lower layer node is assigned to only one upper layer node. In other words, if a type of a node is installed, it will be connected to at most one upper layer node. The assignment constraints of the lower layer nodes (i.e., eNB, MME, S-GW, and P-GW) are as follows:

eNB assignment constraints

$$\sum_{i \in S_1} \nu_{01}^{bi} = 1 \qquad (b \in B) \tag{3.21}$$

$$\sum_{j \in S_2} \nu_{02}^{bj} = 1 \qquad (b \in B) \tag{3.22}$$

MME assignment constraints

$$\sum_{j \in S_2} \nu_{12}^{ij} = \sum_{t \in T_1} \chi_1^{it} \qquad (i \in S_1) \tag{3.23}$$

$$\sum_{k \in S_3} \nu_{13}^{ik} = \sum_{t \in T_1} \chi_1^{it} \qquad (i \in S_1) \tag{3.24}$$

S-GW assignment constraints

$$\sum_{l \in S_4} \nu_{24}^{jl} = \sum_{t \in T_2} \chi_2^{jt} \qquad (j \in S_2) \tag{3.25}$$

P-GW assignment constraints

$$\sum_{p \in S_5} \nu_{45}^{lp} = \sum_{t \in T_4} \chi_4^{lt} \qquad (l \in S_4) \qquad \textbf{(3.26)}$$

The importance of looking into the ***capacity constraints*** of LTE core elements goes back to the urge to identify the needed number of the different elements. We introduce capacity constraints at three separate levels: interface level, node level, and link level. In general, there are four different basic types of capacity constraints [72]: throughput, transactions, subscribers, and bearers. Throughput is the total amount of data load that a node can handle, and it is considered data plane limitation, whereas transactions are signaling messages related to control plane. Subscribers represent the number of subscribers that can be handled by a node, and there are different types of subscribers that can be handled by a node such as active vs. idle (i.e., without ongoing media session), or activated vs. configured subscribers who are not yet activated. Last but not least is the number of bearers; a bearer is a data connection and there are two types which are default and dedicated bearers. The default bearer is best effort and its obligatory for any attached user, whereas the dedicated bearer is established based on the need. The next paragraphs look at the capacity issues of each element of the core network.

MME is responsible for bearer setup, mobility, and authentication, and it is considered a control plane node. A big capacity issue is the number of transactions (or control messages) as well as the number of subscribers that are attached to the MME or in MME's temporary subscriber database. MME is in charge of S-GW and P-GW selection.

HSS is a database for subscribers' data and it is concerned about the control plane in particular. The connection between the MME and the HSS is control plane only. It generates security information for mutual authentication, ciphering, and integrity checking. HSS also provides information about user's location. Two legitimate concerns for the HSS are the number of transactions and the number of subscribers.

S-GW works as a mobility anchor for the traffic being carried on different eNBs, and it also forwards packets between the eNB and the P-GW. S-GW is focused on the user plane, and

throughput is the number one limitation. Regarding the number of transactions, it is a relative issue to consider as a main constraint, in this thesis we take the signaling traffic and the number of transactions into consideration. Generally, the S-GW is affected by 3 different operations: (1) setup and tear down of bearers, (2) mobility (inter-eNB handover), and (3) idle to connected transitions states of a mobile device.

P-GW has three major roles in QoS management, IP addresses allocation, and IP anchoring that guarantees not changing the IP address. P-GW is also responsible for lawful intercept and provides support for charging and policy enforcement as well as QoS management. THE P-GW implements the Policy and Charging Enforcement Function (PCEF). The number of subscribers is not considered a big constraint since the service providers are supposed to be able to accommodate a huge number, whereas throughput is the main constraint in this context. In terms of transactions, P-GW is affected to a limited extent by three control operations: (1) Setup/teardown of bearers, (2) QoS negotiation with the PCRF, and (3) inter S-GW mobility. The number of bearers is considered a limitation, and generally, it depends on the vendor.

The information provided in Table 3.4 is based on a study of different LTE core elements, and the main constraints affecting them [72]. Generally speaking, this information is not absolute, and it relies on the vendor. For some vendors' equipment, a certain limitation is more probable than another vendor.

Table 3.4: Capacity constraints affecting the LTE components

Core Element	Throughput (Data Plane)	Transactions (Control Plane)	Subscribers	Bearers
MME	✖	✓	✓	✖
HSS	✖	✓	✓	✖
S-GW	✓	✓	✖	✓
P-GW	✓	✖	✖	✓

3. ***Capacity constraints at the interface level*** impose that the number of links should not exceed the number of interfaces available in each element. For example, the number of links from a lower layer node to an upper layer node should not exceed the number of the interfaces on the upper layer node. Obviously, the link is only established if the upper layer node is already installed.

MME capacity constraints (at the interface level)

$$\sum_{m\in M_{01}}\sum_{b\in B} v_{01}^{bim} \le \sum_{t\in T_1} \eta_1^t \chi_1^{it} \qquad (i \in S_1) \tag{3.27}$$

$$\sum_{m\in M_{12}}\sum_{j\in S_2} v_{12}^{ijm} + \sum_{m\in M_{13}}\sum_{k\in S_3} v_{13}^{ikm} \le \sum_{t\in T_1} \eta_3^t \chi_1^{it} \quad (i \in S_1) \tag{3.28}$$

S-GW capacity constraints (at the interface level)

$$\sum_{m\in M_{02}}\sum_{b\in B} v_{02}^{bjm} \le \sum_{t\in T_2} \eta_2^t \chi_2^{jt} \qquad (j \in S_2) \tag{3.29}$$

$$\sum_{m\in M_{12}}\sum_{i\in S_1} v_{12}^{ijm} + \sum_{m\in M_{24}}\sum_{l\in S_4} v_{24}^{jlm} \le \sum_{t\in T_2} \eta_5^t \chi_2^{jt} \qquad (j \in S_2) \tag{3.30}$$

HSS capacity constraints (at the interface level)

$$\sum_{m\in M_{13}}\sum_{i\in S_1} v_{13}^{ikm} \le \sum_{t\in T_3} \eta_4^t \chi_3^{kt} \qquad (k \in S_3) \tag{3.31}$$

P-GW capacity constraints (at the interface level)

$$\sum_{m\in M_{24}}\sum_{j\in S_2} v_{24}^{jlm} \le \sum_{t\in T_4} \eta_6^t \chi_4^{lt} \qquad (l \in S_4) \tag{3.32}$$

$$\sum_{m\in M_{45}}\sum_{p\in S_5} v_{45}^{lpm} \le \sum_{t\in T_4} \eta_7^t \chi_4^{lt} \qquad (l \in S_4) \tag{3.33}$$

PCRF capacity constraints (at the interface level)

$$\sum_{m\in M_{45}}\sum_{l\in S_4} v_{45}^{lpm} \le \sum_{t\in T_5} \eta_8^t \chi_5^{pt} \qquad (p \in S_5) \tag{3.34}$$

4. ***Capacity constraints at the node level*** impose that the aggregated traffic of type $c \in C$ on a node should not exceed its capacity statement. At this level of constraints, we are not exactly following Table 3.4, but we are assuming more constraints as shown in the equations below.

MME capacity constraints (at the node level)

$$\sum_{b \in B} f_{01}^{cbi} \leq \sum_{t \in T_1} SUB^t \chi_1^{it} \qquad (c \in \{c_0\}, i \in S_1) \qquad \textbf{(3.35)}$$

$$\sum_{m \in M_{01}} w^m \sum_{b \in B} v_{01}^{bim} \leq \sum_{t \in T_1} \beta^t \chi_1^{it} \qquad (i \in S_1) \qquad \textbf{(3.36)}$$

$$\sum_{b \in B} f_{01}^{cbi} \leq \sum_{t \in T_1} BHSA^t \chi_1^{it} \qquad (c \in \{c_1\}, i \in S_1) \qquad \textbf{(3.37)}$$

$$\sum_{b \in B} f_{01}^{cbi} \leq \sum_{t \in T_1} EPSB^t \chi_1^{it} \qquad (c \in \{c_4\}, i \in S_1) \qquad \textbf{(3.38)}$$

$$\sum_{b \in B} f_{01}^{cbi} \leq \sum_{t \in T_1} ASUB^t \chi_1^{it} \qquad (c \in \{c_3\}, i \in S_1) \qquad \textbf{(3.39)}$$

$$\sum_{b \in B} TPSPB f_{01}^{cbi} \leq \sum_{t \in T_1} SIG^t \chi_1^{it} \qquad (c \in \{c_4\}, i \in S_1) \qquad \textbf{(3.40)}$$

S-GW capacity constraints (at the node level)

$$\sum_{b \in B} f_{12}^{cbj} \leq \sum_{t \in T_2} EPSB^t \chi_2^{jt} \qquad (c \in \{c_4\},\ j \in S_2) \qquad \textbf{(3.41)}$$

$$\sum_{m \in M_{02}} w^m \sum_{b \in B} v_{02}^{bjm} + \sum_{m \in M_{12}} w^m \sum_{i \in S_1} v_{12}^{ijm} \leq \sum_{t \in T_2} \beta^t \chi_2^{jt} \qquad (j \in S_2) \qquad \textbf{(3.42)}$$

$$\sum_{i \in S_1} f_{12}^{cij} \leq \sum_{t \in T_2} BHSA^t \chi_2^{jt} \qquad (c \in \{c_1\},\ j \in S_2) \qquad \textbf{(3.43)}$$

$$\sum_{i \in S_1} f_{12}^{cij} \leq \sum_{t \in T_2} ASUB^t \chi_2^{jt} \qquad (c \in \{c_3\},\ j \in S_2) \qquad \textbf{(3.44)}$$

$$\sum_{i \in S_1} TPSPB f_{12}^{cij} \leq \sum_{t \in T_2} SIG^t \chi_2^{jt} \qquad (c \in \{c_4\},\ j \in S_2) \qquad \textbf{(3.45)}$$

HSS capacity constraints (at the node level)

$$\sum_{i \in S_1} f_{13}^{cik} \le \sum_{t \in T_3} SUB^t \chi_3^{kt} \qquad (c \in \{c_0\},\ k \in S_3) \qquad \textbf{(3.46)}$$

$$\sum_{m \in M_{13}} w^m \sum_{i \in S_1} v_{13}^{ikm} \le \sum_{t \in T_3} \beta^t \chi_3^{kt} \qquad (k \in S_3) \qquad \textbf{(3.47)}$$

$$\sum_{i \in S_1} TPS f_{13}^{cik} \le \sum_{t \in T_3} SIG^t \chi_3^{kt} \qquad (c \in \{c_0\},\ k \in S_3) \qquad \textbf{(3.48)}$$

P-GW capacity constraints (at the node level)

$$\sum_{j \in S_2} f_{24}^{cjl} \le \sum_{t \in T_4} EPSB^t \chi_4^{lt} \qquad (c \in \{c_4\},\ l \in S_4) \qquad \textbf{(3.49)}$$

$$\sum_{m \in M_{24}} w^m \sum_{j \in S_2} v_{24}^{jlm} \le \sum_{t \in T_4} \beta^t \chi_4^{lt} \qquad (l \in S_4) \qquad \textbf{(3.50)}$$

$$\sum_{j \in S_2} f_{24}^{cjl} \le \sum_{t \in T_4} BHSA^t \chi_4^{lt} \qquad (c \in \{c_1\},\ l \in S_4) \qquad \textbf{(3.51)}$$

$$\sum_{j \in S_2} f_{24}^{cjl} \le \sum_{t \in T_4} ASUB^t \chi_4^{lt} \qquad (c \in \{c_3\},\ l \in S_4) \qquad \textbf{(3.52)}$$

$$\sum_{j \in S_2} TPSPB f_{24}^{cjl} \le \sum_{t \in T_4} SIG^t \chi_4^{lt} \qquad (c \in \{c_4\},\ l \in S_4) \qquad \textbf{(3.53)}$$

PCRF capacity constraints (at the node level)

$$\sum_{m \in M_{45}} w^m \sum_{l \in S_4} v_{45}^{lpm} \le \sum_{t \in T_5} \beta^t \chi_5^{pt} \qquad (p \in S_5) \qquad \textbf{(3.54)}$$

$$\sum_{l \in S_4} f_{45}^{clp} \le \sum_{t \in T_5} EPSB^t \chi_5^{pt} \qquad (c \in \{c_4\},\ p \in S_5) \qquad \textbf{(3.55)}$$

$$\sum_{l \in S_4} f_{45}^{clp} \le \sum_{t \in T_5} ASUB^t \chi_5^{pt} \qquad (c \in \{c_3\},\ p \in S_5) \qquad \textbf{(3.56)}$$

$$\sum_{l \in S_4} TPSPB f_{45}^{clp} \le \sum_{t \in T_5} SIG^t \chi_5^{pt} \qquad (c \in \{c_4\},\ p \in S_5) \qquad \textbf{(3.57)}$$

5. ***Capacity constraints at the link level*** impose that the traffic of type $c \in C$ of a node should not exceed the capacity of its links to upper layer node.

eNB – MME link capacity constraints

$$EPSBW f_{01}^{cbi} \le \sum_{m \in M_{01}} w^m v_{01}^{bim} \qquad (c \in \{c_4\}, b \in B, i \in S_1) \qquad \textbf{(3.58)}$$

eNB – S-GW link capacity constraints

$$f_{02}^{cbj} \le \sum_{m \in M_{02}} w^m v_{02}^{bjm} \qquad (c \in \{c_2\}, b \in B, j \in S_2) \qquad \textbf{(3.59)}$$

MME - S-GW link capacity constraints

$$EPSBW f_{12}^{cij} \le \sum_{m \in M_{12}} w^m v_{12}^{ijm} \qquad (c \in \{c_4\}, i \in S_1, j \in S_2) \qquad \textbf{(3.60)}$$

MME – HSS link capacity constraints

$$SUBW f_{13}^{cik} \le \sum_{m \in M_{13}} w^m v_{13}^{ikm} \qquad (c \in \{c_0\}, i \in S_1, k \in S_3) \qquad \textbf{(3.61)}$$

S-GW – P-GW link capacity constraints

$$f_{24}^{cjl} + EPSBWf_{24}^{c^* jl} \leq \sum_{m \in M_{24}} w^m v_{24}^{jlm} \quad (c \in \{c_2\}, c^* \in \{c_4\}, j \in S_2, l \in S_4) \qquad \textbf{(3.62)}$$

P-GW – PCRF link capacity constraints

$$EPSBWf_{45}^{clp} \leq \sum_{m \in M_{45}} w^m v_{45}^{lpm} \qquad (c \in \{c_4\}, l \in S_4, p \in S_5) \qquad \textbf{(3.63)}$$

6. ***Traffic flow conservation constraints*** state that sum of the input traffic type $c \in C$ of a node should be equal to the output traffic of the same type from the node.

$$\mu_b^c v_{01}^{bi} = f_{01}^{cbi} \qquad (c \in \{c_0, c_1, c_3, c_4\}, b \in B, i \in S_1) \qquad \textbf{(3.64)}$$

$$\mu_b^c v_{02}^{bj} = f_{02}^{cbj} \qquad (c \in \{c_2\}, b \in B, j \in S_2) \qquad \textbf{(3.65)}$$

The MME interacts with the HSS for user authentication, and profile download, whereas it interacts with the S-GW for EPS bearer establishment and management, tunnel control, paging, and handovers.

$$\sum_{b \in B} f_{01}^{cbi} = \sum_{j \in S_2} f_{12}^{cij} \qquad (c \in \{c_1, c_3, c_4\}, i \in S_1) \qquad \textbf{(3.66)}$$

$$\sum_{b \in B} f_{01}^{cbi} = \sum_{k \in S_3} f_{13}^{cik} \quad (c \in \{c_0\}, i \in S_1) \qquad \textbf{(3.67)}$$

$$\sum_{b \in B} f_{02}^{cbj} = \sum_{l \in S_4} f_{24}^{cjl} \qquad (c \in \{c_2\}, j \in S_2) \qquad \textbf{(3.68)}$$

$$\sum_{i \in S_1} f_{12}^{cij} = \sum_{l \in S_4} f_{24}^{cjl} \qquad (c \in \{c_1, c_3, c_4\}, j \in S_2) \qquad \textbf{(3.69)}$$

$$\sum_{j \in S_2} f_{24}^{cjl} = \sum_{p \in S_5} f_{45}^{clp} \qquad (c \in \{c_3, c_4\},\ l \in S_4) \qquad \textbf{(3.70)}$$

7. ***Additional constraints*** state that if a lower layer node is connected to upper layer node, there will be at least one link between them, but the number of links should not exceed the maximum number of interfaces on the lower layer node.

$$\nu_{01}^{bi} \le \sum_{m \in M_{01}} \nu_{01}^{bim} \qquad (b \in B,\ i \in S_1) \qquad \textbf{(3.71)}$$

$$\nu_{01}^{bi} \max\{\eta_0^t\} \ge \sum_{m \in M_{01}} \nu_{01}^{bim} \qquad (b \in B,\ i \in S_1,\ t \in T_0) \qquad \textbf{(3.72)}$$

$$\nu_{02}^{bj} \le \sum_{m \in M_{02}} \nu_{02}^{bjm} \qquad (b \in B,\ j \in S_2) \qquad \textbf{(3.73)}$$

$$\nu_{02}^{bj} \max\{\eta_0^t\} \ge \sum_{m \in M_{02}} \nu_{02}^{bjm} \qquad (b \in B,\ j \in S_2,\ t \in T_0) \qquad \textbf{(3.74)}$$

$$\nu_{12}^{ij} \le \sum_{m \in M_{12}} \nu_{12}^{ijm} \qquad (i \in S_1,\ j \in S_2) \qquad \textbf{(3.75)}$$

$$\nu_{12}^{ij} \max\{\eta_3^t\} \ge \sum_{m \in M_{12}} \nu_{12}^{ijm} \qquad (i \in S_1,\ j \in S_2,\ t \in T_1) \qquad \textbf{(3.76)}$$

$$\nu_{13}^{ik} \le \sum_{m \in M_{13}} \nu_{13}^{ikm} \qquad (i \in S_1,\ k \in S_3) \qquad \textbf{(3.77)}$$

$$\nu_{13}^{ik} \max\{\eta_3^t\} \ge \sum_{m \in M_{13}} \nu_{13}^{ikm} \qquad (i \in S_1,\ k \in S_3,\ t \in T_1) \qquad \textbf{(3.78)}$$

$$\nu_{24}^{jl} \le \sum_{m \in M_{24}} \nu_{24}^{jlm} \qquad (j \in S_2,\ l \in S_4) \qquad \textbf{(3.79)}$$

$$\nu_{24}^{jl} \max\{\eta_5^t\} \ge \sum_{m \in M_{24}} \nu_{24}^{jlm} \qquad (j \in S_2,\ l \in S_4,\ t \in T_2) \qquad \textbf{(3.80)}$$

$$\nu_{45}^{lp} \le \sum_{m \in M_{45}} \nu_{45}^{lpm} \qquad (l \in S_4,\ p \in S_5) \qquad \textbf{(3.81)}$$

$$\nu_{45}^{lp} \max\{\eta_7^t\} \ge \sum_{m \in M_{45}} \nu_{45}^{lpm} \qquad (l \in S_4,\ p \in S_5,\ t \in T_4) \qquad \textbf{(3.82)}$$

8. ***Non-negativity constraints*** are limiting the domain of the variables used in the model.

$$f \in \Re_{+} \tag{3.83}$$

9. ***Integrity constraints*** state that the variables used in this context are either 0 or 1. In short they are displayed in binary format. On the other hand, the rest of the values are integer values and they are defined as general variables (N).

$$v \in \mathrm{B}; \qquad v \in \nu_{01}^{bi}, \nu_{02}^{bj}, \nu_{12}^{ij}, \nu_{13}^{ik}, \nu_{24}^{jl}, \nu_{45}^{lp} \tag{3.84}$$

$$x \in \mathrm{B}; \qquad x \in \chi_{1}^{it}, \chi_{2}^{jt}, \chi_{3}^{kt}, \chi_{4}^{lt}, \chi_{5}^{pt} \tag{3.85}$$

The presented mathematical model has a linear objective function and linear constraints, and to be able to solve the model there is a need to translate the objective function and the constraints into Linear Programming (LP) file format that can be solved using CPLEX [79]. Unfortunately, the planning problem presented above is NP-hard and as a result, exact methods can find optimal solutions only when the network is simple and the problem size is small. As the network size increases, the Central Processing Unit (CPU) time increases and the computational complexity of the network design process increases to a point that it makes exact methods impractical to use. Heuristic approach presented in Section 3.6 provides approximate solutions that can be used instead.

3.6 Heuristic Approach

Since the exact formulation of the problem shown in the previous section is NP-hard [87][88], approximate algorithms are needed to solve the problem. Heuristic algorithms are preferred for solving many network design problems because they provide a solution close to optimal solution in a reasonable amount of time. Despite that fact that good solutions can be obtained with heuristic techniques, they do not always guarantee that the best solution will be found. In this

section, the Local Search (LS) algorithm was adopted to solve the planning problem of the LTE core network.

LS algorithm starts from an initial solution and iteratively tries to replace the current solution with a better solution within the neighborhood of the current solution. The iterative improvement only replaces the current solution with a better one. However, the algorithm stops and returns the current solution when no better solutions can be found in the neighbourhood anymore. The neighbourhood is defined as the set of all possible solutions that can be reached from the current solution by performing a move (i.e., changing the type of the element, or uninstalling the element in unneeded locations). The LS algorithm tends to get trapped in the first local minimum which may or may not be a global minimum.

The initial solution for this approach is set to have all the network elements installed in all their potential locations with the types of highest capacities. The initial solution is taken as the current solution, and its neighbourhood is explored to find a better solution and then result in a topology that provides the number of elements as well as the types, and the locations. The resultant topology provides good solution in terms of cost in relatively short CPU time duration.

The flow chart represented in Figure 3.2 explains the main steps of how the LS algorithm works. It describes how the LS algorithm starts with an initial solution and then, iteratively explores the neighborhood until it is not possible to find a better solution. The neighborhood of a given solution is obtained by performing a move. As explained previously, a move in this context represents either uninstalling an element or changing its type. For example, the neighbourhood of the initial solution (where all the elements are installed in their potential locations) is formed by all the solutions that can be obtained by removing or changing the type of an element. The cost of each neighbour is calculated and the neighbour with the lowest cost is compared to the cost of the initial solution. If the cost of the selected neighbour is lower than the initial cost, then the move is accepted and the solution is taken as the best solution found so far and a new iteration can start. However, if the cost is greater, the LS algorithm stops and returns the current solution. This means a local minimum is found and no further improvements can be made.

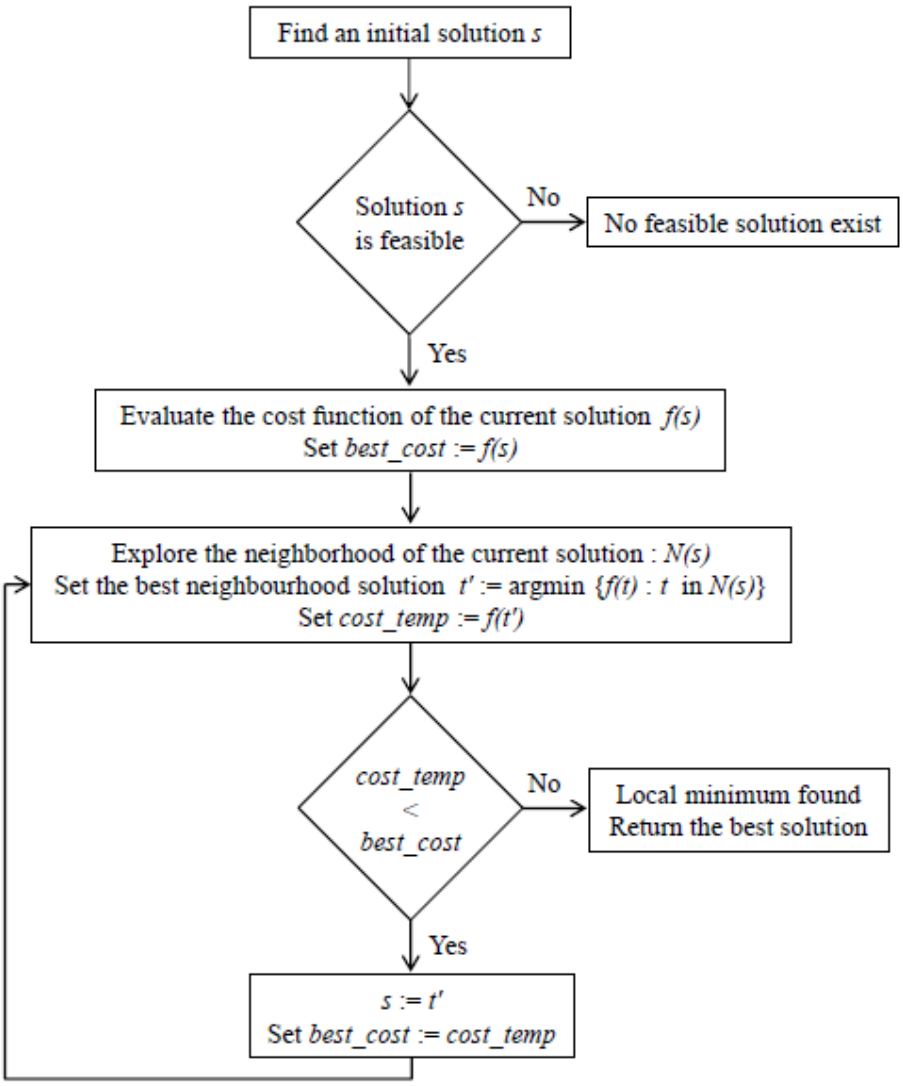

Figure 3.2: Local Search (LS) algorithm[80]

In order to solve the problem of assigning links between the elements, we divided the problem into 6 assignment sub-problems: the eNB to the MME, the eNB to the S-GW, the MME to the S-GW, the MME to the HSS, the S-GW to the P-GW, and the P-GW to the PCRF.

Starting with the first two sub-problems; the number of links, and elements as well as their capacities totally depend on the amount of the eNBs traffic. The eNBs traffic is afterwards

aggregated at a higher level (i.e., MME and S-GW). As soon as the number and types of links connecting the eNBs with the MME, and the S-GW are known, two cost matrices of interconnection links between the eNB and the MME, as well as the eNB and the S-GW are constructed for each taking the constraints into consideration.

Each cost matrix of the two sub-problems has different combinations for potential link assignments between the eNB and the MME (S-GW). The cost matrix is the input for the LAPJV algorithm, discussed in [81], which helps finding the optimal interconnections between all the eNBs and the MMEs (S-GWs) for each sub-problem. The same way those two sub-problems were approached, the other sub-problems are approached.

Chapter 4 Computation Results and Analysis

Proper network planning plays an important role in LTE network processes. It is an essential process that involves a variety of parameters (e.g. planning parameters) with different requirements such as capacity, and cost. A well-planned network is achieved by generating a variety of possible topologies with different types of elements that are distributed in different locations, and then selecting the topology that satisfies the goal of the network planner. Network optimization is of the same importance, and it also requires a clear objective as well as logical constraints, in order to achieve an optimal solution for the network.

4.1 Network Capacity Specifications

In this section, we present the different features for each of the core elements as well as the links and the interfaces connecting them. These specifications will be used to solve the planning problem in the next sections. For all equipment, there is a maximum capacity that cannot be crossed; if the traffic handled by the network is more than its capacity, then there is a need to maximize the capacity by using more elements or by choosing elements with higher capacities. There are many LTE/EPC technology suppliers: Alcatel-Lucent, Cisco Systems, Ericsson, Huawei, Motorola, Nokia-Siemens, among others. Hence, the features presented in this section are not absolute due to the fact that they vary among different suppliers, and different models. Table 4.1 presents the different features of the MME types. The two types are compared according to the features that matter most to the MME. Type B can support more subscribers and attached subscribers than type A; it also provides more throughput as well as higher BHSA and EPSB capacities. The cost is directly proportional with the capacity of the element model; hence, type B is more expensive than type A.

Table 4.1: Features of the MME types

	Type A	Type B
Subscribers	15,000,000	20,000,000 [67]
Attached subscribers	13,500,000	18,000,000
Throughput (Gbps)	55	75
BHSA	405,000,000	540,000,000 [69]
EPSB	15,000,000	20,000,000
Cost ($)	375,000	500,000

The two types of S-GW and their features are presented in Table 4.2. Type A and B are compared in terms of EPSB, attached subscribers, throughput, BHSA and cost. The cost of type B is higher than type A due to the fact that type B can support more attached subscribers, provide higher throughput, and handle a higher number of EPSB and BHSA.

Table 4.2: Features of the S-GW types

	Type A	Type B
EPSB	3,150,000	4,200,000 [68]
Attached subscribers	13,500,000	18,000,000
Throughput (Gbps)	55	75
BHSA	405,000,000	540,000,000
Cost ($)	450,000	600,000

Table 4.3 presents the two different types of P-GW along with their features. Taking into consideration the features that matter the most to the P-GW, and comparing them, we can conclude that the type B can provide higher throughput than type A and also support a larger number of attached subscribers. Moreover, it can handle a higher number of EPSB and BHSA. As a result of the higher capacity of type B, its cost is higher than the cost of type A.

Table 4.3: Features of the P-GW types

	Type A	Type B
EPSB	3,150,000	4,200,000 [68]
Attached subscribers	13,500,000	18,000,000
Throughput (Gbps)	55	75
BHSA	405,000,000	540,000,000
Cost ($)	450,000	600,000

Two different types of HSS and their features are presented in Table 4.4. The cost of the HSS is detailed to include the hardware cost as well as the engineering and the software costs. The

software cost totally depends on the number of CPUs on the server; the cost of each CPU is equal to $20,000. On the other hand, the engineering cost includes the engineering, training and integration costs and it is usually counted per day. Compared to type A, type B has a higher number of CPUs, and it can handle a higher number of subscribers. In addition to that, the throughput it can provide is higher than the throughput provided by type A.

Table 4.4: Features of the HSS types

	Type A	**Type B**
Throughput (Gbps)	25	55
Subscribers	15,000,000	20,000,000 [70]
Software cost ($)	160,000 (8 CPUs)	240,000 (12 CPUs)
Hardware cost ($)	150,000	200,000
Engineering cost ($)	50,000	60,000
Total cost ($)	360,000	500,000

The PCRF is mainly a piece of software that needs to be installed on powerful servers such as Linux and Solaris servers. The total cost includes the hardware cost, the software cost, and the operations cost. The software cost depends on the number of Active Sessions (AcS) that can be supported by the PCRF; there is a base price that is added to the number of active sessions multiplied by the cost for each active session. In addition to that, there is a need for a database where the cost is proportional to the number of CPUs, and this is also added to the software cost. The operations' cost is measured per day and it includes the trainings and the engineering cost, whereas the hardware cost is provided by the vendor based on the model and its capacity. As can be seen in Table 4.5, type B has a higher capacity than type A in terms of throughput, number of attached subscribers, and the number of EPSB. Likewise, the cost of a PCRF of type B is higher than the one of type A.

Table 4.5: Features of the PCRF types

	Type A	Type B
Throughput (Gbps)	25	55
Attached Subscribers	13,500,000	18,000,000
EPSB	3,150,000	4,200,000
Software cost ($)	800,000 (250,000 AcS)	1,450,000 (500,000 AcS)
Hardware cost ($)	150,000	200,000
Engineering cost ($)	50,000	60,000
Total cost ($)	1,000,000	1,710,000

Table 4.6 represents the number of access interfaces that are installed to connect the eNB with the higher level elements (i.e., MME and S-GW); type A is able to support up to 24 interfaces, and type B can handle up to 48 interfaces. According to that, the cost of type A is higher than the cost of type B. Table 4.7 represents the core interfaces which connect the core elements with each other, such as connecting the MME with the HSS, or connecting the S-GW to the P-GW; type A is able to support up to 20 interfaces, and type B can handle up to 30 interfaces.

Table 4.6: Features of access interfaces' types

	Type A	Type B
Number of interfaces	24	48
Cost ($)	375	750

Table 4.7: Features of core interfaces' types

	Type A	Type B
Number of interfaces	20	30
Cost ($)	200	500

There are many types of links that connect elements together; different rates and different materials of cables such as fiber, and copper. Regardless of the cable material, Table 4.8 provides two different types of links with two different capacities. Link capacity is measured in terms of bits per second; for example, Ethernet speed is equal to 10 Mbps, and Fast Ethernet is equal to 100 Mbps. Type A represents a link that is able to send frames at a rate of 1 Gbps. On the other hand, type B represents 10 Gigabit Ethernet which means the link is capable of transmitting frames at a rate of 10 Gbps.

Table 4.8: Features of links types

	Type A	Type B
Link capacity (Gbps)	1	10
Link cost ($/km)	750	1500

Signaling in LTE is a complicated task; it cannot be measured directly since it highly depends on the element itself along with the traffic affecting it. Table 4.9 presents the signaling capacity for the MME, S-GW, P-GW, and PCRF. The signaling capacity is measured in terms of transactions per second, and this value is calculated by multiplying the number of bearers in a busy hour by the number of transactions per second per bearer.

Table 4.9: Signaling capacity of core elements

	Transaction per sec per bearer	Signaling capacity (transaction per sec)	
		Type A	Type B
MME	9.3	139,500,000	186,000,000
S-GW	3	9,450,000	12,600,000
P-GW	2	6,300,000	8,400,000
PCRF	2	6,300,000	8,400,000

The process of signaling measurement for the HSS is different. The signaling capacity is measured in transactions per second as in other elements, but here the signaling capacity is a result of multiplying the number of subscribers by the number of transactions per second per subscriber. The number of transactions per subscriber is not absolute as explained previously, and the number of HSS transactions per second per subscriber is equal to 6.2. As shown in Table 4.10, the signaling capacity for the HSS of type B is higher than type A due to the fact the former supports less number of subscribers.

Table 4.10: Signaling capacity of HSS

	Type A	Type B
Subscribers	15,000,000	20,000,000
Signaling (Transactions per sec)	93,000,000	124,000,000

4.2 Detailed Example

In this section, we present a detailed example in order to show how the planning tool works and the difference between the exact method and the local search algorithm.

The network to be planned is located on a 10km by 10km area as shown in Figure 4.1. It includes 15 actual eNBs that need to be connected to the EPC network (i.e. MME and S-GW). The number of potential locations for the MME is equal to 5, and the number of potential locations for the S-GW is 5. The MME has to be connected to the S-GW as well as the HSS which has 5 potential locations. In addition to that, the S-GW needs to be connected to the P-GW that has 5 potential locations, and finally, the P-GW needs to be connected to the PCRF which also has 5 potential locations as shown in the graph. All these locations were generated randomly by a small C/C++ program.

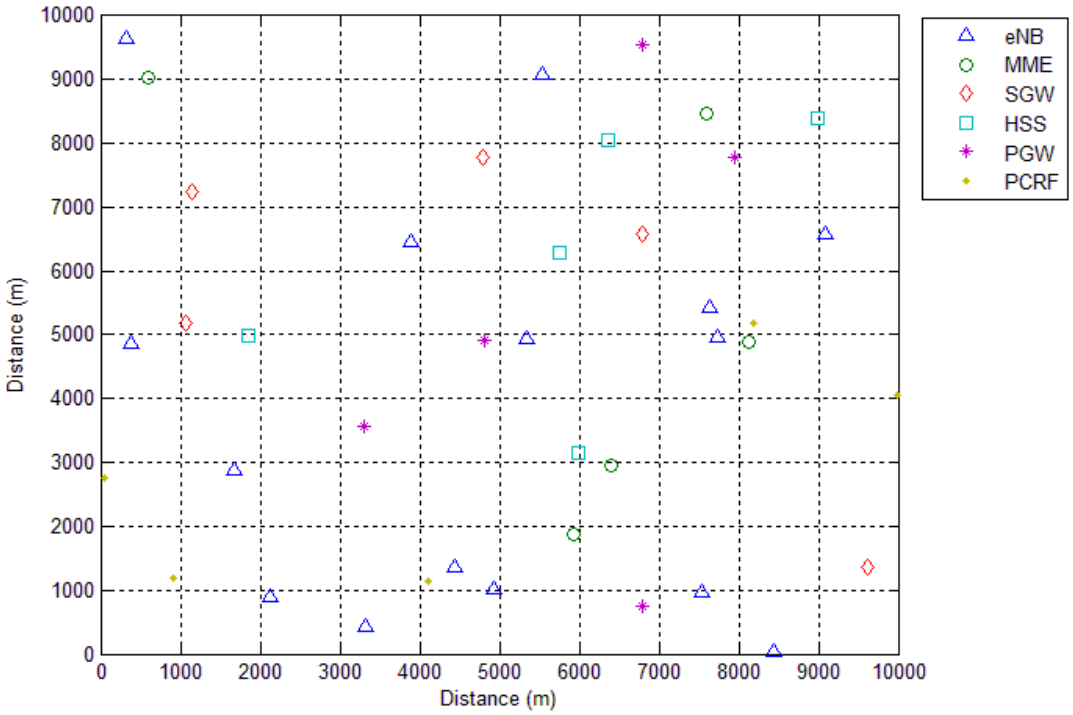

Figure 4.1: eNB locations and potential locations for the core elements

A traffic profile for each of the 15 eNB is generated, as shown in Table 4.11, by using the equations presented in Chapter 3 to generate different types of traffic. The traffic profile involves the total number of subscribers covered by the eNB, and the number of Busy Hour Session Attempts (BHSA) that is used for measuring the performance of the elements in terms of voice and data traffic. In addition, it also contains the traffic bandwidth in Mbps, the number of

attached subscribers who are subscribers connected to the Internet, and the number of simultaneous Evolved Packet System Bearers (EPSB). The location of each eNB is also presented in Table 4.11.

Table 4.11: Traffic profile for the eNBs

eNB Index	Location (m)		Subscribers	Attached Subscribers	BHSA	BW (Mbps)	EPSB
	X	Y					
1	5,534	9,078	9,850	8,865	3,815,496	181	1,108
2	3,895	6,467	12,941	11,646	5,012,825	238	1,455
3	8,436	42	6,756	6,080	2,617,004	124	760
4	5,349	4,938	15,893	14,303	6,156,312	293	1,787
5	9,082	6,593	7,914	7,122	3,065,567	146	890
6	1,676	2,875	14,285	12,856	5,533,437	263	1,607
7	2,132	905	7,702	6,931	2,983,446	142	866
8	4,936	1,026	17,659	15,893	6,840,390	325	1,986
9	7,727	4,961	10,194	9,174	3,948,747	188	1,146
10	7,646	5,440	8,075	7,267	3,127,932	149	908
11	7,538	974	10,504	9,453	4,068,829	193	1,181
12	388	4,868	9,970	8,973	3,861,979	183	1,121
13	330	9,655	11,363	10,226	4,401,571	209	1,278
14	4,438	1,356	12,300	11,070	4,764,528	226	1,383
15	3,332	437	12,244	11,019	4,742,835	225	1,377

4.2.1 Exact Method

In the mathematical model presented in Chapter 3, the decision variables are a mix of real numbers and integers; hence, mixed integer programming approach is adopted and CPLEX optimizer 12.5.0.0 is used to solve the model. CPLEX [79] is a Mixed Integer Programming (MIP) optimizing tool that solves linear optimization problems by satisfying the objective function (i.e., minimizing or maximizing) following different constraints and bounds. The branch and cut algorithm is used by CPLEX to solve the problems which may result in feasible or infeasible solution. The next paragraphs present a solution for the detailed example using the exact method and CPLEX tool.

The optimal solution achieved for this specific problem is presented in Figure 4.2; it contains 15 eNBs connected to one MME of type A as presented in Table 4.1, and one S-GW of type A as presented in Table 4.2. The MME is connected to one HSS of type A as presented in Table 4.4

and the S-GW is connected to one P-GW of type A as presented in Table 4.3. In addition to that, the P-GW is connected to one PCRF of type A as presented in Table 4.5. The MME is connected to the S-GW by a link of type B as presented in Table 4.8, whereas it is connected to the HSS by a link of type A as shown in Table 4.8. Finally, the S-GW is connected to the P-GW using a link of type B, and the P-GW is also connected to the PCRF using a link of type B, and both of the link' types are presented in Table 4.8.

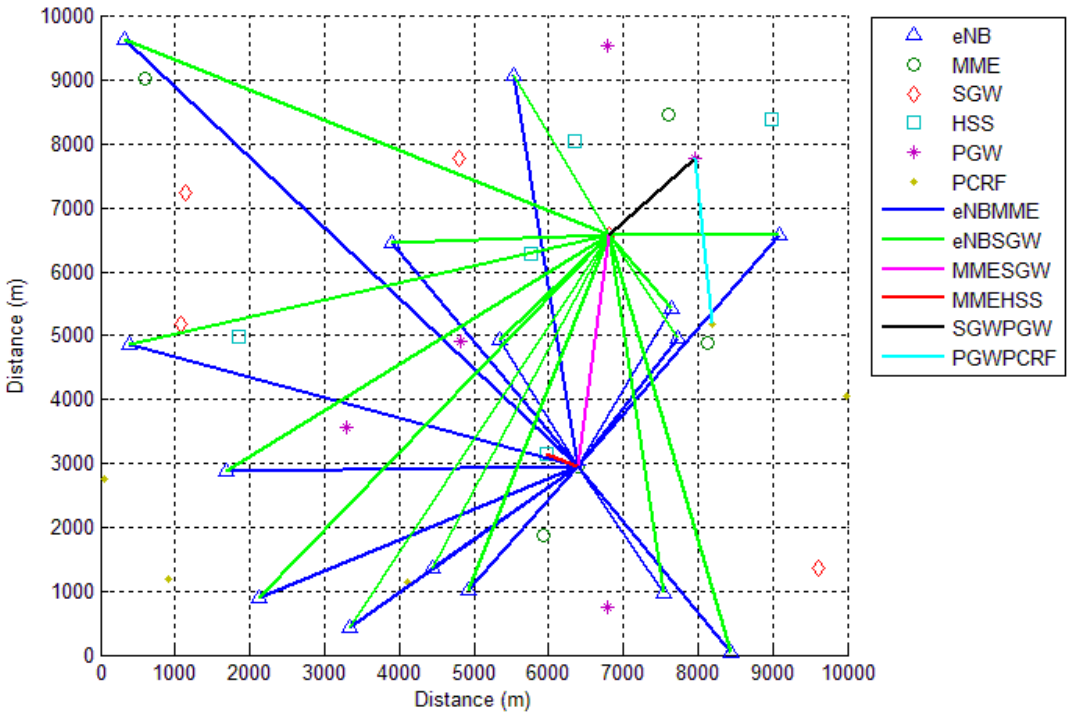

Figure 4.2: The optimal solution

The colored lines represent the links between the network elements based on the optimal solution giving the minimum cost (i.e. cost of the nodes and the links connecting them, taking into consideration the distances between them). As noticed, the 15 eNBs are connected to one MME (blue lines) for the purpose of control and management, and connected to the S-GW (green lines) for the purpose of accessing the PDN network to use its services. The pink line represents the connection between the MME and the S-GW, the red line shows the connection between the MME and the HSS, and the black line is the link connecting the S-GW with the P-GW. Furthermore, the turquoise line represents the connection between the P-GW and the PCRF which is used for policing and quality of service in addition to many other useful functions. The

elements that are not connected in Figure 4.2 are the potential locations that were not selected in the design process since they were not part of the optimal solution.

The optimal solution represented in Figure 4.2 has a total cost of $2,773,545 and was obtained after exploring 2,760 nodes in 221,168 iterations in a time of 52 seconds. In terms of traffic flow, each link carries specific types of traffic flows, and each element is concerned about certain features. Nevertheless, the traffic flow must be conserved among the links. As an example, the traffic generated from the 15 eNBs and entering the MME and the S-GW is described and explained. For the user data plane carried over the S1-U interface between the eNB and the S-GW, the bandwidth traffic flow conservation achieved from the simulation is represented in Table 4.12.

Table 4.12: Bandwidth flow conservation

eNB index	Traffic entering the S-GW (Mbps)	Traffic exiting the S-GW (Mbps)
1	181	3,085
2	238	
3	124	
4	293	
5	146	
6	263	
7	142	
8	325	
9	188	
10	149	
11	193	
12	183	
13	209	
14	226	
15	225	

The flow conservation for the number of subscribers is presented in Table 4.13, the flow conservation of busy hour session attempts is shows in Table 4.14, the flow conservation for the number of attached subscribers is presented in Table 4.15, and the flow conservation for the number of simultaneous EPS bearers is shown in Table 4.16.

The results shown in Table 4.12, Table 4.13, Table 4.14, Table 4.15, and Table 4.16 make sure that a specific type of traffic entering an element is equal to the traffic exiting that specific core

element, and this applies to all the core network components and not only the elements presented in this section. In general, the traffic flow depends on the topology, the number of elements and the capacity of each element.

Table 4.13: Conservation of subscribers' number

eNB index	Traffic entering the MME	Traffic exiting the MME
1	9,850	167,650
2	12,941	
3	6,756	
4	15,893	
5	7,914	
6	14,285	
7	7,702	
8	17,659	
9	10,194	
10	8,075	
11	10,504	
12	9,970	
13	11,363	
14	12,300	
15	12,244	

Table 4.14: Busy hour session attempts flow conservation

eNB index	Traffic entering the MME	Traffic exiting the MME
1	3,815,496	64,940,898
2	5,012,825	
3	2,617,004	
4	6,156,312	
5	3,065,567	
6	5,533,437	
7	2,983,446	
8	6,840,390	
9	3,948,747	
10	3,127,932	
11	4,068,829	
12	3,861,979	
13	4,401,571	
14	4,764,528	
15	4,742,835	

Table 4.15: Attached subscribers flow conservation

eNB index	Traffic entering the MME	Traffic exiting the MME
1	8,865	150,878
2	11,646	
3	6,080	
4	14,303	
5	7,122	
6	12,856	
7	6,931	
8	15,893	
9	9,174	
10	7,267	
11	9,453	
12	8,973	
13	10,226	
14	11,070	
15	11,019	

Table 4.16: Flow conservation of EPS bearers' number

eNB index	Traffic entering the MME	Traffic exiting the MME
1	1,108	18,853
2	1,455	
3	760	
4	1,787	
5	890	
6	1,607	
7	866	
8	1,986	
9	1,146	
10	908	
11	1,181	
12	1,121	
13	1,278	
14	1,383	
15	1,377	

The signaling does not have to be conserved among the links due to the fact that it depends on the element itself, and it varies from an element to another. The amount of signaling for the HSS depends on the number of subscribers and the number of transactions per seconds per subscriber, whereas the other elements (i.e. MME, S-GW, P-GW, and PCRF) depends on the number of simultaneous bearers and the number of transactions per seconds per bearer varying based on the element specifications. The signaling traffic carried and generated by each core element is

presented in Table 4.17; in general, the values attained are dependent on the traffic profile achieved from the simulation along with the equipment signaling specs that are provided by different vendors.

Table 4.17: Signaling traffic

Core element	Signaling (transactions per sec)
MME	175,332
S-GW	56,559
HSS	1,039,430
P-GW	37,706
PCRF	37,706

4.2.2 Approximate Method

The same problem as described previously is now approached using the LS algorithm, and the result of solving the problem using this approach is presented in Figure 4.3. The cost of the solution found is equal to $2,776,771 achieved in a CPU time equal to 35 seconds. This represents a difference of $3,226 (i.e. 0.12%) and 17 seconds with respect to the optimal solution found with CPLEX. In general, the difference in cost may be caused by various factors such as installing different numbers or types of elements/links, or selecting different locations for the core elements, which in turn affects the cost of the links which is based on distance. In this example, the difference in cost is due to selecting different locations for the P-GW and the PCRF, which in turn affects the distances between these elements and any other elements connected to them.

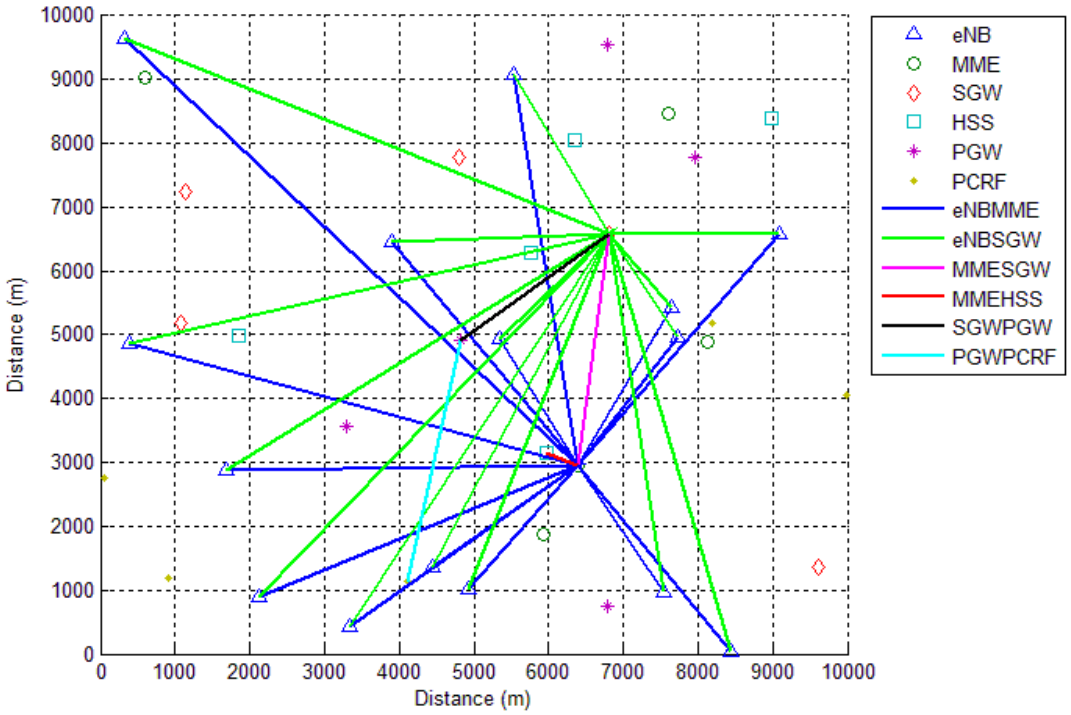

Figure 4.3: Solution found with the Local Search

The topology attained consists of 15 eNBs connected to one MME of type A as described in Table 4.1 and one S-GW of type A as shown in Table 4.2. The MME is connected to the S-GW, and it is connected to HSS of type A as listed in Table 4.4. Moreover, the S-GW is connected to P-GW of type A as represented in Table 4.3, and the P-GW is connected to PCRF of type A as described in Table 4.5. The links connecting the MME to the S-GW are of type B as presented in Table 4.8, whereas the links connecting the MME to the HSS are of type A as shown in Table 4.8. In addition to that, the S-GW is connected to the P-GW using a link of type B, and the P-GW is also connected to the PCRF using a link of type B; Table 4.8 describes the different types of links installed between the elements.

In terms of traffic flow, there is only one piece of each core element in the topology. As a result, the traffic flow is the same as the traffic flow in the exact method. In other cases, the traffic flow may not be the same because the number of elements is not limited to one.

4.3 Result Analysis

Different problems of different sizes were simulated using the exact algorithm, and then the cost and the CPU time were compared with the ones resulted from applying the approximate algorithm. The optimal cost and the time are definitely dependent on the number of elements, connections between them and on the complexity of the problem. On top of all, it depends on the approach used to solve the problem (i.e., the exact algorithm using CPLEX or the approximate algorithm using LS).

In this section, we present two different sets of problems: small scale problems and larger scale problems. For each set, several different sizes were generated and for each size, three instances of the problem were generated so that averages can be taken. The two sets are approached using the exact method and the approximate method and results are compared in terms of cost and CPU time. First, we present the small scale problems and then we describe the larger scale problems in order to show how the two algorithms adapt to changes.

4.3.1 Solving Small Scale Problems

The first set of problems has a range of eNBs between 10 and 200, and 5 potential locations for each core element distributed in an area of 400 km². We stopped at 200 eNB because after this point, the exact algorithm started giving infeasible solution since 5 potential locations were not enough to accommodate more eNBs.

The various problem sizes are presented in Table 4.18; there are 21 problems that have the same number of core elements potential locations. The first, second, and third instances are represented in Table 4.19, Table 4.20 and Table 4.21 respectively. The problems are presented including the cost and the CPU time of the exact solution compared to the cost and the CPU time of the approximate solution.

Table 4.18: Problem sizes for the small scale problems

Problem	Length(km)	eNB	MME	S-GW	HSS	P-GW	PCRF	Constraints
1	20	10	5	5	5	5	5	915
2	20	15	5	5	5	5	5	1,095
3	20	20	5	5	5	5	5	1,275
4	20	25	5	5	5	5	5	1,455
5	20	30	5	5	5	5	5	1,635
6	20	35	5	5	5	5	5	1,815
7	20	40	5	5	5	5	5	1,995
8	20	45	5	5	5	5	5	2,175
9	20	50	5	5	5	5	5	2,355
10	20	55	5	5	5	5	5	2,535
11	20	60	5	5	5	5	5	2,715
12	20	65	5	5	5	5	5	2,895
13	20	70	5	5	5	5	5	3,075
14	20	75	5	5	5	5	5	3,255
15	20	80	5	5	5	5	5	3,435
16	20	85	5	5	5	5	5	3,615
17	20	90	5	5	5	5	5	3,795
18	20	95	5	5	5	5	5	3,975
19	20	100	5	5	5	5	5	4,155
20	20	150	5	5	5	5	5	5,955
21	20	200	5	5	5	5	5	7,755

In Table 4.19, Table 4.20, and Table 4.21, the first column represents the problem number which correlates to the problem size described in Table 4.18. The second and third columns represent respectively the cost and CPU time obtained with CPLEX; whereas columns 4 and 5 represent respectively the cost and CPU time obtained with the local search. The cost gap between CPLEX and LS is represented as a percentage of the cost difference with respect to the optimal solution in column 6. The CPU time gap, shown in column 7, represents the percentage of CPU time difference between CPLEX and LS with respect to the time taken by CPLEX to find the optimal solution. From the results shown in the tables, we can see that the cost returned by CPLEX is always lower than the cost returned by LS. In fact, out of the 63 problems, the local search was not able to find the optimal solution for any of the problems although 24 were less than 1% away from the optimal solution. As far as the execution time is concerned, the time taken by the local search to find the solution is always less than the time taken by CPLEX and seems a lot more predictable.

Table 4.19: Small scale problems: Results for the first instance

	CPLEX		LS		Cost Gap	Time Gap
Problem	**Cost ($)**	**Time (sec)**	**Cost ($)**	**Time (sec)**	**(%)**	**(%)**
1	2,812,219	28	2,836,092	15	0.85	46.43
2	2,881,622	52	2,896,115	34	0.50	34.62
3	2,966,370	256	2,971,526	26	0.17	89.84
4	3,309,677	2,333	3,334,525	35	0.75	98.50
5	3,389,408	1,003	3,395,071	37	0.17	96.31
6	3,425,122	748	3,432,892	44	0.23	94.12
7	3,538,851	832	3,551,926	50	0.37	93.99
8	3,544,229	479	3,559,578	52	0.43	89.14
9	4,523,141	2,803	4,651,346	58	2.83	97.93
10	4,378,228	9,170	4,527,332	62	3.41	99.32
11	4,431,043	10,900	4,684,541	66	5.72	99.39
12	4,515,204	5,239	4,758,472	69	5.39	98.68
13	4,502,103	7,954	4,724,773	74	4.95	99.07
14	4,872,140	408	5,180,201	75	6.32	81.62
15	4,953,728	2,664	5,241,232	78	5.80	97.07
16	5,001,638	4,357	5,386,149	82	7.69	98.12
17	5,238,095	206	6,329,327	75	20.83	63.59
18	5,332,394	1,075	6,461,308	90	21.17	91.63
19	6,108,294	626	7,377,586	94	20.78	84.98
20	7,935,934	5,080	9,302,337	136	17.22	97.32
21	9,921,530	848	12,790,040	152	28.91	82.08

Table 4.20: Small scale problems: Results for the second instance

	CPLEX		LS		Cost Gap	Time Gap
Problem	**Cost ($)**	**Time (sec)**	**Cost ($)**	**Time (sec)**	**(%)**	**(%)**
1	2,788,430	37	2,801,838	14	0.48	62.16
2	2,865,937	37	2,874,759	18	0.31	51.35
3	2,937,003	55	2,937,245	26	0.01	52.73
4	3,310,522	317	3,319,783	32	0.28	89.91
5	3,328,432	2,203	3,337,290	39	0.27	98.23
6	3,432,039	1,402	3,442,963	41	0.32	97.08
7	3,495,506	1,143	3,512,397	49	0.48	95.71
8	3,656,193	2,223	3,676,428	49	0.55	97.80
9	4,416,708	555	4,588,202	60	3.88	89.19
10	4,430,559	17,726	4,661,310	61	5.21	99.66
11	4,423,343	1,284	4,731,476	66	6.97	94.86
12	4,480,562	1,159	4,722,968	71	5.41	93.87
13	4,547,493	2,136	4,805,608	75	5.68	96.49
14	4,911,057	8,544	5,216,642	75	6.22	99.12
15	4,928,707	640	5,447,419	80	10.52	87.50
16	4,996,732	1,742	5,259,686	102	5.26	94.14
17	5,257,271	1,339	6,236,378	77	18.62	94.25
18	5,248,573	3,463	6,498,165	81	23.81	97.66
19	6,011,952	2,990	7,273,825	94	20.99	96.86
20	8,043,552	10,308	9,117,385	147	13.35	98.57
21	9,713,921	5,592	13,056,299	146	34.41	97.39

Table 4.21: Small scale problems: Results for the third instance

	CPLEX		LS		Cost Gap	Time Gap
Problem	**Cost ($)**	**Time (sec)**	**Cost ($)**	**Time (sec)**	**(%)**	**(%)**
1	2,783,039	110	2,788,431	14	0.19	87.27
2	2,891,017	85	2,903,713	19	0.44	77.65
3	2,878,805	218	2,881,690	26	0.10	88.07
4	3,242,382	691	3,258,847	34	0.51	95.08
5	3,310,024	1,362	3,319,201	40	0.28	97.06
6	3,375,127	641	3,377,286	46	0.06	92.82
7	3,465,354	1,937	3,478,035	48	0.37	97.52
8	3,508,533	496	3,533,296	53	0.71	89.31
9	4,285,722	4,510	4,591,037	60	7.12	98.67
10	4,358,690	10,851	4,572,642	63	4.91	99.42
11	4,558,446	10,161	4,892,046	66	7.32	99.35
12	4,546,185	1,829	4,870,370	65	7.13	96.45
13	4,598,871	5,329	4,812,337	75	4.64	98.59
14	4,948,463	1,216	5,317,276	75	7.45	93.83
15	5,026,152	5,448	5,511,027	80	9.65	98.53
16	5,261,779	4,144	5,628,621	81	6.97	98.05
17	5,129,513	358	6,272,514	86	22.28	75.98
18	5,273,776	416	6,285,171	85	19.18	79.57
19	6,113,950	56,225	7,629,748	95	24.79	99.83
20	7,932,045	67,544	9,127,541	200	15.07	99.70
21	9,812,103	2,150	12,836,165	156	30.82	92.74

Among the three different instances, the lowest cost acquired by CPLEX is $2,783,039 and the highest is $9,921,530; whereas the lowest cost attained by LS is $2,788,431 and the highest is $13,056,299. Regarding time duration, among the three different instances the shortest time duration that took CPLEX to find optimal solution is 28 seconds, and the longest time duration is 67,544 seconds. For the LS, the shortest time duration taken by LS to provide a good solution is 14 seconds and the longest time duration is 200 seconds.

The average costs for the three different instances are graphically represented in Figure 4.4. As can be concluded from the figure, the cost of the network achieved by using the exact method (i.e., CPLEX) is always lower than the cost of the network achieved by using the approximate method (i.e., LS). The problem size affects the cost gap, it can be noted that the cost gap for the first 8 problems (first 24 instances) is less than 1%, whereas the cost gap for the problems 9 to 16 (second 24 instances) do not exceed 11% and the cost gap for problems 17 to 21 (last 15 instances) is less than 35%.

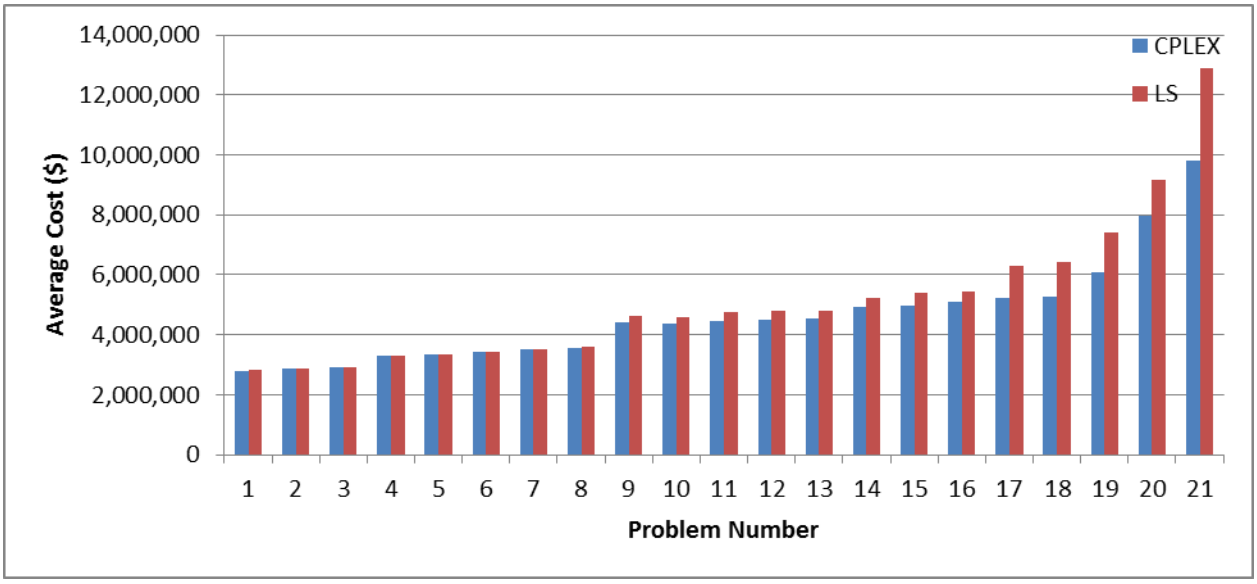

Figure 4.4: Average cost comparison

The average time durations for the three different instances are graphically represented in Figure 4.5 for the CPLEX method as well as the LS approach. The figure shows that the CPU time taken by the approximate method (i.e., LS) is much less than the time taken by the exact method (i.e., CPLEX).

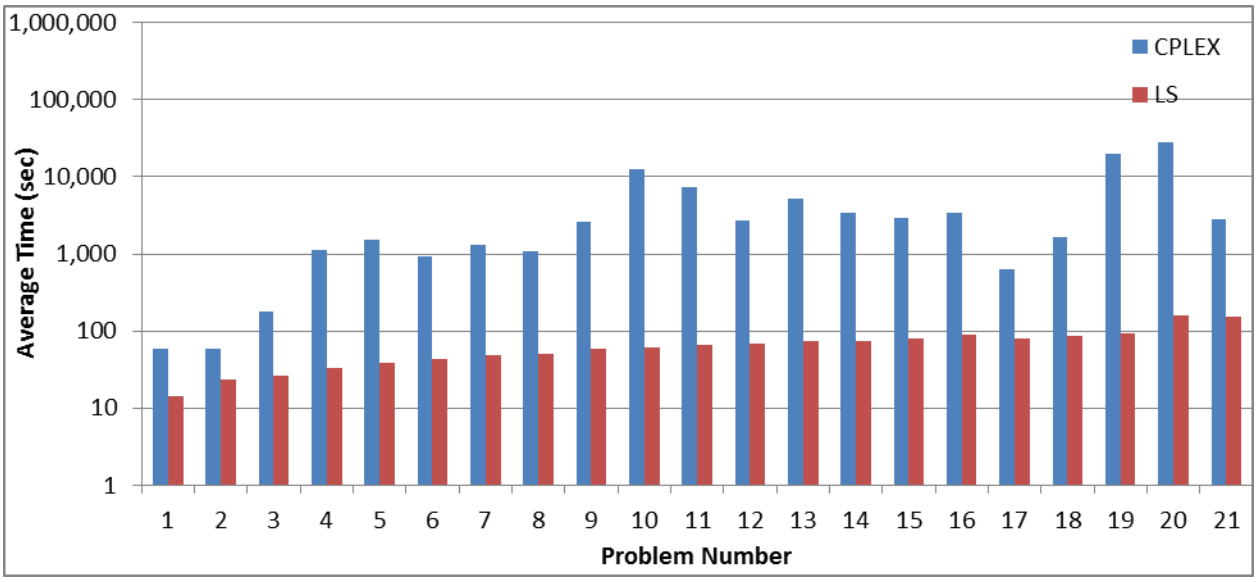

Figure 4.5: Average CPU time comparison

Table 4.22 illustrates the cost gaps of the three different instances. The mean for the three cost gaps is calculated and presented in the fifth column along with a confidence interval of 90% in the sixth column. The mean of the mean is 7.74, and taking a confidence interval of 90% the mean would fall into an interval of ±2.62. The cost gap highly depends on the size of the problem. In general, larger problem sizes lead to higher cost gaps. However, that may not be the case for each single problem as much as it is the general behaviour for all the problems. As can be concluded from Table 4.22, 24 problems out of 63 problems (i.e., 21 problems each with 3 instances) have a cost gap less than 1%, and the cost gap increases as the size of the problem increases.

Table 4.22: Cost gaps results comparison

Problem	First Cost Gap (%)	Second Cost Gap (%)	Third Cost Gap (%)	Mean	Confidence Interval
1	0.85	0.48	0.19	0.51	±0.36
2	0.50	0.31	0.44	0.42	±0.11
3	0.17	0.01	0.10	0.09	±0.09
4	0.75	0.28	0.51	0.51	±0.26
5	0.17	0.27	0.28	0.24	±0.07
6	0.23	0.32	0.06	0.20	±0.14
7	0.37	0.48	0.37	0.41	±0.07
8	0.43	0.55	0.71	0.56	±0.15
9	2.83	3.88	7.12	4.61	±2.44
10	3.41	5.21	4.91	4.51	±1.05
11	5.72	6.97	7.32	6.67	±0.92
12	5.39	5.41	7.13	5.98	±1.09
13	4.95	5.68	4.64	5.09	±0.58
14	6.32	6.22	7.45	6.66	±0.74
15	5.80	10.52	9.65	8.66	±2.74
16	7.69	5.26	6.97	6.64	±1.36
17	20.83	18.62	22.28	20.58	±2.01
18	21.17	23.81	19.18	21.39	±2.53
19	20.78	20.99	24.79	22.19	±2.46
20	17.22	13.35	15.07	15.21	±4.58
21	28.91	34.41	30.82	31.38	±3.04
Mean				7.74	±2.62

The CPU time gaps for the three different instances are represented in Table 4.23. The mean for the three time gaps is calculated and presented in column 5. In addition to that, a confidence

interval of 90% is shown in column 6. The mean of the mean is 90.06, and taking a confidence interval of 90% means that the mean would fall into an interval of ±3.40.

Table 4.23: Time Gaps results comparison

Problem	First Time Gap (%)	Second Time Gap (%)	Third Time Gap (%)	Mean	Confidence Interval
1	46.43	62.16	87.27	65.29	±22.46
2	34.62	51.35	77.65	54.54	±23.65
3	89.84	52.73	88.07	76.88	±22.82
4	98.50	89.91	95.08	94.50	±4.71
5	96.31	98.23	97.06	97.20	±1.05
6	94.12	97.08	92.82	94.67	±2.38
7	93.99	95.71	97.52	95.74	±1.92
8	89.14	97.80	89.31	92.08	±5.40
9	97.93	89.19	98.67	95.26	±5.75
10	99.32	99.66	99.42	99.47	±0.19
11	99.39	94.86	99.35	97.87	±2.84
12	98.68	93.87	96.45	96.33	±2.62
13	99.07	96.49	98.59	98.05	±1.50
14	81.62	99.12	93.83	91.52	±9.78
15	97.07	87.50	98.53	94.37	±6.53
16	98.12	94.14	98.05	96.77	±2.48
17	63.59	94.25	75.98	77.94	±16.81
18	91.63	97.66	79.57	89.62	±10.04
19	84.98	96.86	99.83	93.89	±8.57
20	97.32	98.57	99.70	98.53	±1.30
21	82.08	97.39	92.74	90.74	±8.56
Mean				90.06	±3.40

As can be concluded, LS can provide good solutions with an amount of time that is much less than CPLEX. The LS method proves to be a powerful heuristic approach that solves planning and optimization problems in reasonable amount of time. As the complexity of the network increases, CPLEX may tend to find it harder to acquire optimal solutions within the time limit (set to 108,000 seconds). The second example presented in Section 4.3.2 provides larger scale problems that show how CPLEX may not be able to find optimal solutions within the time limit (set to 108,000 seconds).

4.3.2 Larger Scale problems

In the second set of problems, we want to see how the two methods will behave when the number of potential locations is increased. As a result, in this 2nd example, the number of potential locations is increased from 5 to 10 for each core element. The range of eNBs varies between 10 and 70 in area of 900 km². The number of eNB is stopped at 70 because once the number of eNBs starts exceeding 25, finding the optimal solution was not a guarantee since most of the time, the time limit (set to 108,000 seconds) was reached.

Table 4.24 shows the 13 different problem sizes. The first column has the number of the problem, and the second column represents the length of the area. The third column represents the number of eNBs which is changing for each problem in this example. On the other hand, the number of potential locations for the core elements is fixed for all the problems as shown in columns 4 to 8.

Table 4.24: Problem sizes for the larger scale problems

Problem	Length (km)	eNB	MME	S-GW	HSS	P-GW	PCRF	Constraints
1	30	10	10	10	10	10	10	2,370
2	30	15	10	10	10	10	10	2,700
3	30	20	10	10	10	10	10	3,030
4	30	25	10	10	10	10	10	3,360
5	30	30	10	10	10	10	10	3,690
6	30	35	10	10	10	10	10	4,020
7	30	40	10	10	10	10	10	4,350
8	30	45	10	10	10	10	10	4,680
9	30	50	10	10	10	10	10	5,010
10	30	55	10	10	10	10	10	5,340
11	30	60	10	10	10	10	10	5,670
12	30	65	10	10	10	10	10	6,000
13	30	70	10	10	10	10	10	6,330

Three different instances of each problem size are represented in Table 4.25, Table 4.26, and Table 4.27. In each of these tables, the results of the exact method and the approximate algorithm are compared. The first column represents the problem number. The second and third columns represent the CPLEX cost and the CPU time respectively; whereas column 4 and 5 represent the LS cost and the CPU time respectively. The cost difference between CPLEX and LS is

represented as percentage with respect to the optimal solution in column 6. Finally, the CPU time gap, shown in column 7, represents the percentage of CPU time difference between CPLEX and LS with the respect to the time taken by CPLEX to find the optimal solution. From the results shown in the tables, we can see that the CPLEX cost is less than LS cost; whereas the time of LS is less than the time it takes the CPLEX to find a solution. Due to the fact that the gap is calculated between the optimal solution and the solution found by LS, it is not calculated for cases were the time limit (set to 108,000 seconds) was reached since there is no guarantee that the solution is optimal.

Table 4.25: First instance results comparison

	CPLEX		**LS**		**Cost Gap**	**Time Gap**
Problem	**Cost ($)**	**Time (sec)**	**Cost ($)**	**Time (sec)**	**(%)**	**(%)**
1	2,848,852	9,081	2,853,371	250	0.16	97.25
2	2,982,705	5,923	2,995,904	621	0.44	89.52
3	3,009,759	41,553	3,020,065	467	0.34	98.88
4	3,434,174	108,000 (TL)	3,444,163	584	-	-
5	3,431,961	108,000 (TL)	3,443,329	690	-	-
6	3,690,556	108,000 (TL)	3,739,717	870	-	-
7	3,742,632	108,000 (TL)	3,766,911	921	-	-
8	3,851,936	108,000 (TL)	3,861,777	982	-	-
9	4,646,915	108,000 (TL)	4,866,018	1,133	-	-
10	4,741,353	108,000 (TL)	4,944,191	1,209	-	-
11	4,832,876	108,000 (TL)	5,068,313	1,250	-	-
12	4,908,410	108,000 (TL)	5,221,154	1,393	-	-
13	5,191,203	108,000 (TL)	5,238,956	1,693	-	-

Table 4.26: Second instance results comparison

	CPLEX		LS		Cost Gap	Time Gap
Problem	**Cost ($)**	**Time (sec)**	**Cost ($)**	**Time (sec)**	**(%)**	**(%)**
1	2,842,104	22,233	2,851,228	242	0.32	98.91
2	2,929,215	3,760	2,933,316	593	0.14	84.23
3	3,017,277	108,000 (TL)	3,019,912	484	-	-
4	3,479,781	108,000 (TL)	3,484,548	584	-	-
5	3,552,273	108,000 (TL)	3,611,198	754	-	-
6	3,611,138	108,000 (TL)	3,844,966	886	-	-
7	3,706,368	108,000 (TL)	3,717,004	893	-	-
8	3,765,667	108,000 (TL)	3,810,961	1,026	-	-
9	4,566,764	108,000 (TL)	4,740,886	1,170	-	-
10	4,728,207	108,000 (TL)	4,875,961	1,164	-	-
11	4,752,682	108,000 (TL)	5,094,028	1,297	-	-
12	4,763,086	108,000 (TL)	5,139,424	1,323	-	-
13	4,949,414	108,000 (TL)	5,267,110	1,416	-	-

Table 4.27: Third instance results comparison

	CPLEX		LS		Cost Gap	Time Gap
Problem	**Cost ($)**	**Time (sec)**	**Cost ($)**	**Time (sec)**	**(%)**	**(%)**
1	2,898,656	1,608	2,906,994	242	0.29	84.95
2	3,005,487	11,916	3,027,680	631	0.74	94.70
3	3,038,894	27,149	3,044,948	487	0.20	98.21
4	3,411,022	108,000 (TL)	3,426,048	566	-	-
5	3,537,530	108,000 (TL)	3,549,686	777	-	-
6	3,595,439	108,000 (TL)	3,613,158	846	-	-
7	3,742,391	108,000 (TL)	3,745,125	915	-	-
8	3,934,613	108,000 (TL)	3,961,692	1,024	-	-
9	4,603,998	108,000 (TL)	4,944,902	1,116	-	-
10	4,633,610	108,000 (TL)	4,887,257	1,165	-	-
11	4,786,472	108,000 (TL)	5,024,519	1,242	-	-
12	4,826,025	108,000 (TL)	5,243,268	1,398	-	-
13	5,088,550	108,000 (TL)	5,594,177	1,529	-	-

Among the three different instances, the lowest cost obtained by CPLEX is $2,842,104 and the highest is $5,191,203; whereas the lowest cost attained by LS is $2,851,228 and the highest is $5,594,177.

The average costs for the three different instances are graphically represented in Figure 4.6. As can be concluded from the figure, the cost of the network achieved by using the exact method

(i.e., CPLEX) is lower than the cost of the network achieved by using the approximate method (i.e., LS). As can be noticed, the problem size is increasing and the cost gap seems to get a little bit larger. This is due to the fact that the search space gets bigger and more combinations become available when the problem size gets bigger and the number of elements increases. Therefore, it is difficult to fully explore the search space efficiently in a short computation time.

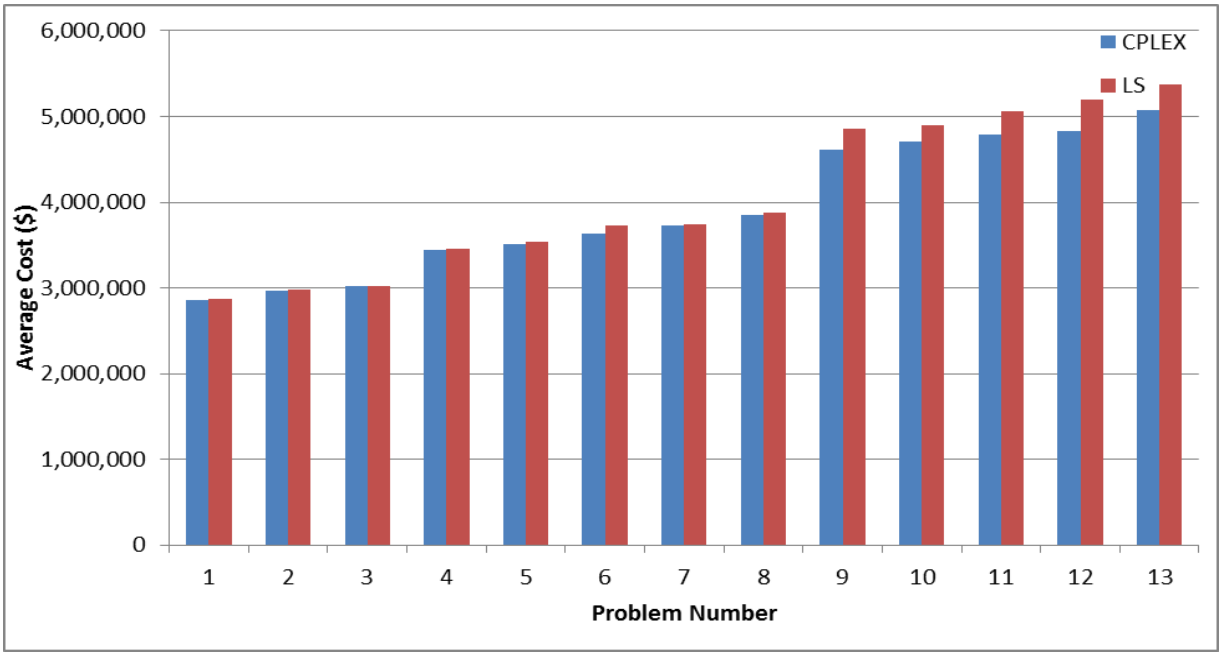

Figure 4.6: Average cost comparison

The average time durations for the three different instances of the second problem are graphically represented in Figure 4.7for the CPLEX method and the LS approach. In the results we achieved, the shortest time duration that took CPLEX to find optimal solution is 1,608 seconds; however, in many cases it reached the time limit (108,000 seconds) without the certainty of finding the optimal solution. It is important to note that without this time limit, CPLEX could take a lot longer before finding the optimal solution. From experience, if CPLEX cannot find the solution within the time limit, most likely it will take a lot of time to find it and as a result, would increase significantly the CPU time. For the LS, the shortest time duration taken by LS to provide a good solution is 242 seconds and the longest time duration is 1,693 seconds. The figure shows that the CPU time taken by the approximate method (i.e., LS) is much less than the time taken by the exact method (i.e., CPLEX).

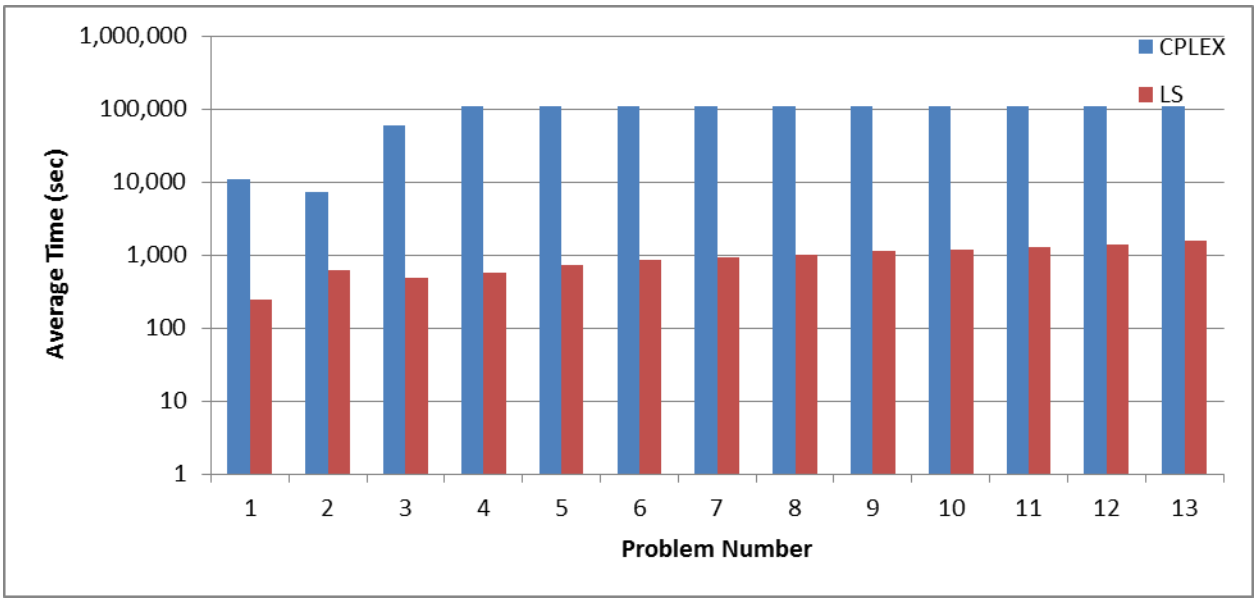

Figure 4.7: Average CPU time comparison

Table 4.28 shows a comparison between the three different instances in terms of cost gap. The first column represents the number of the problem. Column 2 represents the cost gap for the first instance, column 3 represents the cost gap for the second instance, and column 4 contains the third instance cost gap. The mean for the three cost gaps is calculated and presented in column 5. In addition to that, a confidence interval of 90% is shown in column 6. The mean of the mean is 0.32, and taking a confidence interval of 90% the mean would fall into an interval of ±0.11. This is considered very good but these problems are considered relatively small problems.

Table 4.28: Cost Gaps results comparison

Problem	First Cost Gap (%)	Second Cost Gap (%)	Third Cost Gap (%)	Mean	Confidence Interval
1	0.16	0.32	0.29	0.26	±0.09
2	0.44	0.14	0.74	0.44	±0.33
3	0.34	-	0.20	0.27	±0.22
4	-	-	-	-	-
5	-	-	-	-	-
6	-	-	-	-	-
7	-	-	-	-	-
8	-	-	-	-	-
9	-	-	-	-	-
10	-	-	-	-	-
11	-	-	-	-	-
12	-	-	-	-	-
13	-	-	-	-	-
Mean				0.32	±0.11

CPLEX may spend long time looking for optimal solutions, and in many cases the time limit can be reached without finding the optimal solution. However, the local search tends to find a good solution in much shorter time duration. The three different instances of the second problem are compared in Table 4.29 in terms of the time gap. The mean for the three time gaps is calculated and presented in column 5. In addition to that, a confidence interval of 90% is shown in column 6. The mean of the mean is 93.91, and taking a confidence interval of 90% the mean would fall into an interval of ±4.95.

Table 4.29: Time Gaps results comparison

Problem	First Time Gap (%)	Second Time Gap (%)	Third Time Gap (%)	Mean	Confidence Interval
1	97.25	98.91	84.95	93.70	±8.31
2	89.52	84.23	94.70	89.48	±5.71
3	98.88	-	98.21	98.55	±1.03
4	-	-	-	-	-
5	-	-	-	-	-
6	-	-	-	-	-
7	-	-	-	-	-
8	-	-	-	-	-
9	-	-	-	-	-
10	-	-	-	-	-
11	-	-	-	-	-
12	-	-	-	-	-
13	-	-	-	-	-
Mean				93.91	±4.95

In conclusion, the exact algorithm is able to give the optimal solution within the time limit for small problems, but as the problem size increases the algorithm may find it harder to find optimal solutions within the time limit. On the other hand, the approximate algorithm is capable of giving good solutions in relatively shorter time duration. In the next chapter, the problem is concluded along with the two different solving methods. In addition to that future work and various improvements are described.

Chapter 5 Conclusions and Future Work

Providing an optimal network that is capable of delivering good services is an important feature of any network planning tool. The main goal of a good network planning tool is to plan an efficient network that takes into consideration realistic traffic as well as cost and CPU processing time. Two different algorithms (i.e., the exact and approximate algorithms) were presented to meet the needs of the market, and to esnure that operators are capable of providing services at competitive prices.

The exact algorithm, resembled in a mathematical model implemented with the help of CPLEX, is used to provide an optimal network design with minimum cost and acceptable quality of service. However, the network planning problem is considered NP-Hard and the complexity keeps increasing with respect to the problem size. In fact, the exact algorithm can give optimal solution, but as the complexity of the problem increases, CPLEX finds it harder to give feasible solutions within the time limit. An approximate algorithm based on the local search is proposed to solve this problem and provide good solutions in less processing time. In general, the local search approach will find relatively good solutions.

Before we started planning, we generated realistic traffic that takes into account different values such as bandwidth, signaling, BHSA, and EPSB, and the planning problem started by using the exact method. We developed a mathematical model that has a clear objective function and a set of decision variables, uniqueness, assignment and capacity constraints that were applied at the link, interface and node levels. The goal of the mathematical model is to form a minimum cost network topology that is capable of handling the traffic. In fact, the model selects the locations of the core elements (i.e., MME, S-GW, HSS, P-GW, and PCRF), in addition to the types and quantity of core elements, links and interfaces based on the traffic.

The LS algorithm has the ability to solve network planning problems starting from simple problems reaching to complex problems.On average, the LS approach showed results that prove its ability to find solutions that are between 5.12%, 10.36% from the optimal solution in terms of cost, taking a confidence interval of 90%. In addition to that, LS proved that it solves the problems in less time than CPLEX with an improvement varying between 86.66% and 93.46%.

This model can be applied to the LTE and the LTE Advanced since there is no main difference in the architecture of the EPC. More work can be added in the future to enhance the network performance. CPLEX is a good tool; however in many cases, it was not able to ensure giving feasible solutions within the time limit in the medium range of problem complexity. Enhancing CPLEX performance can improve the results comparison by calculating gaps and differences between the exact and the approximate methods. In addition to that, the LS algorithm is a good heuristic algorithm, yet as shown by the results, there is no guarantee to find the optimal solution since it will get stuck in the first local minimum it will encounter. As a result, more advanced heuristic algorithms can be used such as Tabu Search algorithm to avoid the deficiency of the LS. Future work can also include applying SON algorithms for LTE network planning which is basically automating all the functionalities of network planning, optimization and handling failures. Some studies showed that using SON can increase the savings [3].

References

[1] "The LTE Network Architecture", strategic white paper, Alcatel- Lucent, 2009.

[2] Gerhard Fritze, "SAE: The Core Network for LTE", Ericsson, May 2012.

[3] "LTE Operations and Maintenance Strategy Using Self-Organizing Networks to Reduce OPEX", white paper, Motorola, April 2009.

[4] Louise Gabriel, Michel Grech, Fani Kontothanasi, Amit Mukhopadhyay, Marios Nicoloau and Ashoke Sharma, "Economic Benefits of SON Features in LTE Networks", IEEE Sarnoff Symposium , 2011.

[5] "Self-Organizing Network (SON): Introducing the Nokia Siemens Networks SON Suite – an efficient, future-proof platform for SON", white paper, Nokia Siemens networks, 2009.

[6] "Long Term Evolution (LTE): A Technical Overview", technical white paper, Motorola, June 2007.

[7] Sassan Ahmadi, "Mobile WiMAX: A Systems Approach to Understanding IEEE 802.16m Radio Access Technology" book, section 2.9.5 User-Plane/Control-Plane Latency and Handover Interruption Time, 2011, pp. 686.

[8] LTE; Requirements for further advancements for Evolved Universal Terrestrial Radio Access (E-UTRA) (LTE-Advanced) (3GPP TR 36.913 version 9.0.0 Release 9), ETSI technical report v.9, Feb 2010.

[9] 3rd Generation Partnership Project; Technical Specification Group Radio Access Network; Requirements for Evolved UTRA (E-UTRA) and Evolved UTRAN (E-UTRAN) (Release 9), 3GPP technical report v.9, Dec 2009.

[10] KN Premnath, Srinivasan Rajavelu, "Challenges in Self Organizing Networks for Wireless Telecommunications", International Conference on Recent Trends in Information Technology, June 2011, pp. 1331 – 1334.

[11] Sujuan Feng, Eiko Seidel, "Self-Organizing Networks (SON) in 3GPP Long Term Evolution", May 2008.

[12] I. Luketic, D. Simunic, T. Blajic, “Optimization of Coverage and Capacity of Self-Organizing Network in LTE”, the 34th International Convention MIPRO, May 2011, pp. 612 -617.

[13] Lixiang Xu, Chengjun Sun, Xiaoqiang Li, Chaegwon Lim, Hong He, “The Methods to Implement Self Optimization in LTE system”, International Conference on Communications technology and Applications, Oct. 2009, pp. 381-385.

[14] Honglin Hu, Jian Zhang, Xiaoying Zheng, Yang Yang, Ping Wu, “Self Configuration and Self-Optimization for LTE Networks”, IEEE Communications Magazine, Feb 2010, pp. 94-100.

[15] Zuozhou Li, Shudong Li, “LTE Network Planning Based on Game Theory”, International Conference on Computer Science and Service System (CSSS), June 2011, pp. 3963-3966.

[16] Mehmet E. Aydin, Raymond Kwan, Joyce Wu, and Jie Zhang, “Multiuser Scheduling on the LTE Downlink with Simulated Annealing”, IEEE 73rd Vehicular Technology Conference, May 2011, pp.1-5.

[17] Tobias Bandh, Georg Carle, Henning Sanneck, “Graph Coloring Based Physical-Cell-ID Assignment for LTE Networks”, the 2009 International Conference on Wireless Communications and Mobile Computing: Connecting the World Wirelessly, June 2009, pp. 116-120.

[18] Furqan Ahmed, Olav Tirkkonen, Matti Peltomaki, Juha-Matti Koljonen, Chia-Hao Yu, and Mikko Alava, “Distributed Graph Coloring for Self-Organization in LTE Networks”, Journal of Electrical and Computer Engineering Vol.10, Article ID 402831, Aug. 2010, 10 pages.

[19] Junsik Kim, Byunghan Ryu, Kyongtak Cho, Namhoon Park, “Interference Control Technology for Heterogeneous Networks”, the 5th International Conference on Mobile Ubiquitous Computing, Systems, Services and Technologies, Nov. 2011, pp. 290-295.

[20] ETSI Technical Report “LTE; Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Self-configuring and self-optimizing network use cases and solutions”, Sept. 2010.

[21] G. Wunder, M. Kasparick, A. Stolyar and H. Viswanathan, “Self-Organizing Distributed Inter-Cell Beam Coordination in Cellular Networks with Best Effort Traffic”, the 8th

international Symposium on Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks, June 2010, pp. 295–302.

[22] Koichiro Kitagawa, Toshihiko Komine, Toshiaki Yamamoto, Satoshi Konishi, "A Handover Optimization Algorithm with Mobility Robustness for LTE systems", IEEE 22nd International Symposium on Personal Indoor and Mobile Radio Communications (PIMRC), Sept. 2011, pp. 1647–1651.

[23] Richard Combes, Zwi Altman, and Eitan Altman, "On the use of packet scheduling in self-optimization processes: application to coverage-capacity optimization", 8th International Symposium on Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks (WiOpt, June. 2010, pp.98–107.

[24] Muhammad Naseer ul Islam, Andreas Mitschele-Thiel, "Reinforcement Learning Strategies for Self- Organized Coverage and Capacity Optimization", IEEE Conference on Wireless Communications and Networking, April 2012, pp. 2818–2823.

[25] Luis M. del Apio, Emilio Mino, Luis Cucala, Oscar Moreno, Ignacio Berberana, Esther Torrecilla, "Energy Efficiency and Performance in mobile networks deployments with femtocells", IEEE 22nd International Symposium on Personal Indoor and Mobile Radio Communications, Sept. 2011, pp. 107–111.

[26] Jungho Cho, "An Adaptive Antenna Switching Method for Energy Saving Based on Self Organizing Network in Tactical Mobile Communication System", 4th International Conference on Ubiquitous and Future Networks, July 2012, pp. 151-155.

[27] Zhenzhen Wei, "Mobility Robustness Optimization based on UE mobility for LTE system", International Conference on Wireless Communications and Signal Processing (WCSP), Oct. 2010, pp. 1–5.

[28] Ahmad Awada, Bernhard Wegmann, Dirk Rose, Ingo Viering, and Anja Klein, "Towards Self-Organizing Mobility Robustness Optimization in Inter-RAT Scenario", IEEE 73rd Vehicular Technology Conference, May 2011, pp. 1-5.

[29] Yao Wei, Mugen Peng "A Mobility Load Balancing Optimization Method for Hybrid Architecture in Self Organizing Network", IET International Conference on Communication Technology and Application, Oct. 2011, pp. 828-832.

[30] Weihao Lv, Wenjing Li, Heng Zhang, Yanguang Liu, “Distributed Mobility Load Balancing with RRM in LTE”, 3rd IEEE International Conference on Broadband Network and Multimedia Technology, Oct. 2010, pp. 457–461.

[31] Kongluan Lin, John Debenham, “Power Management in LTE Networks by Applying Intelligent Agents”, 6th International Conference on Broadband and Biomedical Communications, Nov. 2011, pp. 163–166.

[32] S. Louvros, K. Aggelis and A. Baltagiannis, “LTE Cell Coverage Planning Algorithm Optimizing Uplink User Cell Throughput”, International Conference on Telecommunications, June. 2011, pp. 51-58.

[33] Zheng Ruiming, Zhang Xin, Li Xi, Pan Qun, Fang Yinglong, Yang Dacheng, “Performance Evaluation on the Coexistence Scenario of two 3GPP LTE Systems”, IEEE 70th Vehicular Technology Conference, Sept. 2009, pp. 1-6.

[34] Jun Gu, Yufeng Ruan, Xi Chen, Chaowei Wang, “A Novel Traffic Capacity Planning Methodology for LTE Radio Network Dimensioning”, IET International Conference on Communication Technology and Application (ICCTA 2011), Oct. 2011, pp. 462–466.

[35] Rajarshi Sanyal, “Challenges in Interoperability and Roaming between LTE - Legacy core for Mobility Management, Routing, Real Time Charging”, 2011 Technical Symposium at ITU Telecom World, Oct. 2011, pp. 116-122 .

[36] Marius Corici, Jens Fiedler, Thomas Magedanz, Dragos Vingarzan, “Evolution of the Resource Reservation Mechanisms for Machine Type Communication over Mobile Broadband Evolved Packet Core Architecture”, 2011 IEEE Globecom Workshops, Dec. 2011, pp. 718-722.

[37] Shakil Akhtar, “2G-5G Networks: Evolution of Technologies, Standards, and Deployment”, Encyclopedia of Multimedia Technology and Networking, Second Edition, 2009.

[38] Qin-long Qiu, Jian Chen, Ling-di Ping, Qi-fei Zhang, Xue-Zeng Pan, “LTE/SAE Model and its Implementation in NS 2”, 5th International Conference on Mobile Ad-hoc and Sensor Networks, Dec. 2009, pp.299-303.

[39] Telus, http://www.telusmobility.com/en/ON/canada_travel/index.shtml?eVar6=link, Nov. 2012.

[40] Bell Canada LTE website, http://network.bell.ca/en/lte/, Nov. 2012.

[41] Fido http://www.fido.ca/web/content/monthly/internetdevicesplans, Nov 2012.

[42] Telus, http://www.telusmobility.com/en/ON/plans/promotional.shtml, Nov 2012.

[43] R. Krzanowski, "Burst of packets and Burstiness", v1.0, 66th IETF meeting, Oct 2006.

[44] "Enabling mobile broadband growth – Evolved Packet Core", technical white paper, Nokia Siemens, 2009.

[45] Mohammad Reza Pasandideh, "Automatic Planning of 3G UMTS All-IP Release 4 Networks with Realistic Traffic", 2011.

[46] Clint Smith, Daniel Collins, "3G Wireless Networks", McGraw-Hill Telecom Professional, 2002.

[47] Harvey Lehpamer, "Transmission Systems Design Handbook for Wireless Networks", Artech House, 2002.

[48] "Long Term Evolution (LTE): Overview of LTE Air-Interface Technical White Paper", technical white paper, Motorola, 2007.

[49] Ajay R Mishra, "Advanced Cellular Network Planning and Optimisation 2G/2.5G/3G evolution to 4G", Wiley, 2007.

[50] Cisco, "Traffic Analysis for Voice over IP", 2007.

[51] Kumiko Ono and Henning Schulzrinne, "One Server per City: Using TCP for Very Large SIP Servers", Principles, Systems and Applications of IP Telecommunications. Services and Security for Next Generation Networks, Henning Schulzrinne, Radu State, and Saverio Niccolini, Vol. 5310,Germany,pp. 133-148, 2008.

[52] Newport Networks, "VoIP Bandwidth Calculation", 2005.

[53] Vijay Garg, "Wireless Communication and Networking", Morgan Kaufmann, 2007.

[54] Harri Holma, Antti Toskala, "LTE for UMTS: Evolution to LTE-Advanced", Wiley, 2nd edition, 2010.

[55] Miikka Poikselka , Harri Holma, Jukka Hongisto, Juha Kallio, Antti Toskala, "Voice over LTE (VoLTE)", 2012.

[56] Mooi Choo Chuah, Qinqing Zhang, "Design and Performance of 3G Wireless Networks and Wireless LANs", 2006, pp. 50.

[57] Cisco, "Sizing Call Center Resources", Cisco IP Contact Center Enterprise Edition Releases 5.0 and 6.0 Solution Reference Network Design, 2006.

[58] Gottfried Punz, “Evolution of 3G Networks: The Concept, Architecture and Realization of Mobile Networks beyond UMTS”, 2010.

[59] Jun Gu, Yufeng Ruan, Xi Chen, Chaowei Wang, “A Novel Traffic Capacity Planning Methodology for LTE Radio Network Dimensioning”, Proceedings of International Conference on Communication Technology and Application (ICCTA2011), Oct 2011, pp. 462-466.

[60] Xiaoming Zhao, Wei Wu, Zhijun Cai, “Dynamic Aggregated Maximum Bit Rate for Evolved Packet System Non-guaranteed Bit Rate Quality of Service Enforcement and Network Bandwidth Utilization”, Research in Motion, Oct 2010.

[61] Syed Abdul Basit, “Dimensioning of LTE Network Description of Models and Tool, Coverage and Capacity Estimation of 3GPP Long Term Evolution radio interface”, 2009.

[62] ATDI, “Mobile LTE Network Design with ICS Telecom”, White Paper, Dec 2008.

[63] C. Gessner, O. Gerlach,“ Voice and SMS in LTE”, Rohde&schwarz , White Paper ,May 2011.

[64] “Signaling is growing 50% faster than data traffic”, Nokia Siemens Networks, White paper, 2012.

[65] Gabriel Brown, “Control Plane Scalability & Performance Requirements for 3G & LTE Networks”, White paper, Sept 2010.

[66] Diametriq, “Measuring the Explosion of LTE Signaling Traffic - A Diameter Traffic Model”, White paper, Sept 2012.

[67] Alcatel-Lucent 9471 Wireless Mobility Manager Mobility Management Entity/Serving GPRS Support Node, 2012.

[68] Gabriel Brown, “Performance & Monetization of the Evolved Packet Core”, Juniper Networks, Heavy Reading, Feb 2012.

[69] Juniper Networks Mobile Next Broadband Gateway Performance and Scalability Validation, Nov. 2011.

[70] “Alcatel-Lucent 1430 Unified Home Subscriber Service”, http://www.alcatel-lucent.com, Dec 2012.

[71] Pierre Lescuyer, Thierry Lucidarme, “Evolved Packet System: The LTE and SAE Evolution of 3G UMTS”, Wiley, 2008.

[72] LTE University videos, http://lteuniversity.com/ , Nov. 2012.

[73] X. Li, U. Toseef, T. Weerawardane , W. Bigos , D. Dulas; C. Goerg , A. Timm-Giel, A. Klug, "Dimensioning of the LTE S1 interface," Wireless and Mobile Networking Conference (WMNC), 2010 Third Joint IFIP, Oct. 2010, pp.1-6.

[74] X. Li, U. Toseef, T. Weerawardane , W. Bigos , D. Dulas; C. Goerg , A. Timm-Giel, A. Klug, "Dimensioning of the LTE access transport network for elastic internet traffic," Wireless and Mobile Computing, Networking and Communications (WiMob), 2010 IEEE 6th International Conference on , Oct. 2010, pp.346-354.

[75] Aleksandra Checko, Lars Ellegaard, Michael Berger, "Capacity planning for Carrier Ethernet LTE backhaul networks", Wireless Communications and Networking Conference (WCNC), 2012 IEEE, April 2012, pp.2741-2745.

[76] Karl Lindberger, "Balancing Quality of Service, Pricing and Utilisation in Multiservice Networks with Stream and Elastic Traffic", in Proc. of the International Teletraffic Congress (ITC 16), Edinburgh, Scotland, 1999.

[77] Kevin Fitchard, "IMS software bug caused Verizon LTE outage", http://connectedplanetonline.com/3g4g/news/ims-software-bug-caused-verizon-lte-outage-0519/, Feb. 2012.

[78] Teresa C. Piliouras, "Network Design Management and Technical Perspectives", 2nd edition, Auerbach publications, 2005.

[79] IBM, IBM ILOG CPLEX V12.1, User's Manual for CPLEX, IBM, Inc., 2009.

[80] M. St-Hilaire, S. Chamberland, and S. Pierre, "A local search heuristic for the global planning of UMTS networks" in Proceedings of the 2006 international conference on Wireless communications and mobile computing, ser. IWCMC '06. New York, NY, USA: ACM, 2006, pp. 1405-1410.

[81] R. Jonker and A. Volgenant, "A shortest augmenting path algorithm for dense and sparse linear assignment problems", Nov 1987, pp. 325-340.

[82] Konstantinos Lizos, Saied M.Abd El-atty, "A novel packet scheduling for high speed bursty traffic in LTE based-3G concepts", 2012 8th International Wireless Communications and Mobile Computing Conference (IWCMC), Aug. 2012, pp.671-676.

[83] Elias Jailani, Muhamad Ibrahim, Ruhani Ab Rahman, "LTE speech traffic estimation for network dimensioning", 2012 IEEE Symposium on Wireless Technology and Applications (ISWTA), Sept. 2012, pp.315-320.

[84] Dorit S. Hochbaum, “Approximation algorithms for NP-hard problems”, PWS Publisher Co., 1997.

[85] Richard Karp, “Reducibility among combinatorial problems,” in Complexity of Computer Computations, R. E. Miller and J. W. Thatcher, Eds.,. Plenum Press, 1972, pp. 85–103.

[86] GSMA PRD IR.92, “IMS Profile for Voice and SMS”, GSM Association, Dec. 2010.

[87] Marc St-Hilaire, Steven Chamberland, Samuel Pierre, "Uplink UMTS Network Design - An Integrated Approach", Computer Networks, vol. 50, no. 15, Oct. 2006, pp. 2747-2761.

[88] Samuel Pierre, Fabien Houeto, “A tabu search approach for assigning cells to switches in cellular mobile networks”, computer communications, vol. 25, 2002, pp. 464-477.