A Source-Destination Based Dynamic Pricing Scheme to Optimize Resource Utilization in Heterogeneous Wireless Networks

Jeremiah Nzioka, MUTUNGI

Supervisor:

Ass/Prof. Olabisi Falowo

This thesis is submitted in fulfillment of the academic requirements

For the degree of

Master of Science in Electrical Engineering

in the Faculty of Engineering and The Built Environment

University of Cape Town

November 2016

As the candidate's supervisor, I have approved this dissertation for submission.

Name: A/Prof. Olabisi Falowo

Signed by candidate

Declaration

I hereby declare that: (1) the above thesis is my own unaided work, both in conception and execution, and that apart from the normal guidance of my supervisor, I have received no assistance apart from that stated below; (2) except as stated below, neither the substance or any part of the thesis has been submitted in the past, or is being, or is to be submitted for a degree in the University or any other University.

I am now presenting the thesis for examination for the Degree of MSc in Electrical Engineering. I also grant the University free license to reproduce the above thesis in whole or in part, for the purpose of research.

Name: Jeremiah N. Mutungi

Signature: Signed by candidate

Date:31/08/2016

EBE Faculty: Assessment of Ethics in Research Projects (Rev2)

Any person planning to undertake research in the Faculty of Engineering and the Built Environment at the University of Cape Town is required to complete this form before collecting or analysing data. When completed it should be submitted to the supervisor (where applicable) and from there to the Head of Department. If any of the questions below have been answered YES, and the applicant is NOT a fourth year student, the Head should forward this form for approval by the Faculty EIR committee: submit to Ms Zulpha Geyer (Zulpha.Geyer@uct.ac.za: Chem Eng Building, Ph 021 650 4791).
NB: A copy of this signed form must be included with the thesis/dissertation/report when it is submitted for examination

This form must only be completed once the most recent revision EBE EIR Handbook has been read.

Name of Principal Researcher/Student: JEREMIAH MUTUNGI Department: ELECTRICAL

Preferred email address of the applicant: jmziarah@gmail.com

If a Student: Degree: Msc Electrical Engineering Supervisor: Alfred Olabiyi Falowo

If a Research Contract indicate source of funding/sponsorship:

Research Project Title: A Source-Destination Based Dynamic Pricing Scheme to Optimize Resource Utilization in Heterogeneous Wireless Networks.

Overview of ethics issues in your research project:

Question 1: Is there a possibility that your research could cause harm to a third party (i.e. a person not involved in your project)?	YES	NO
Question 2: Is your research making use of human subjects as sources of data? If your answer is YES, please complete Addendum 2.	YES	NO
Question 3: Does your research involve the participation of or provision of services to communities? If your answer is YES, please complete Addendum 3.	YES	NO
Question 4: If your research is sponsored, is there any potential for conflicts of interest? If your answer is YES, please complete Addendum 4.	YES	NO

If you have answered YES to any of the above questions, please append a copy of your research proposal, as well as any interview schedules or questionnaires (Addendum 1) and please complete further addenda as appropriate. Ensure that you refer to the EIR Handbook to assist you in completing the documentation requirements for this form.

I hereby undertake to carry out my research in such a way that

- there is no apparent legal objection to the nature or the method of research; and
- the research will not compromise staff or students or the other responsibilities of the University;
- the stated objective will be achieved, and the findings will have a high degree of validity;
- limitations and alternative interpretations will be considered;
- the findings could be subject to peer review and publicly available; and
- I will comply with the conventions of copyright and avoid any practice that would constitute plagiarism.

Signed by:

	Full name and signature	Date
Principal Researcher/Student:	Signed by candidate	31/08/2016

This application is approved by:

Supervisor (if applicable):	Signed by candidate	31/08/2016
HOD (or delegated nominee): *Final authority for all assessments with NO to all questions and for all undergraduate research.*	Signed by candidate	31/8/16
Chair : Faculty EIR Committee For applicants other than undergraduate students who have answered YES to any of the above questions.		

To God for giving me the strength to complete this work,

My family for their continuous support and prayers,

My friends for their encouragements along the way.

Abstract

Mobile wireless resources demand is rapidly growing due to the proliferation of bandwidth-hungry mobile devices and applications. This has resulted in congestion in mobile wireless networks (MWN) especially during the peak hours when user traffic can be as high as tenfold the average traffic. Mobile network operators (MNOs) have been trying to solve this problem in various ways. First, MNOs have tried to expand the network capacity but have still been unable to meet the peak hour demand. Focus has then shifted to economic and behavioral mechanisms. The widely used of these economic mechanisms is dynamic pricing which varies the MWN resources' price according to the congestion level in the MWN. This encourages users to shift their non-critical traffic from the busy hour, when the MWN is congested, to off-peak hours when the network is under-utilized. As a result, congestion of the MWN during the peak hours is reduced. At the same time, the MWN utilization during the off-peak hours is also increased.

The current dynamic pricing schemes, however, only consider the congestion level in the call-originating cell and neglect the call-destination cell when computing the dynamic price. Due to this feature, we refer the current dynamic pricing schemes as source–based dynamic pricing (SDP) schemes in this work. The main problem with these schemes is that, when the majority of the users in a congested cell are callees, dynamic pricing is ineffective because callers and not callees pay for network services, and resources used by callers and callees are the same for symmetric services. For example, application of dynamic pricing does not deter a callee located in a congested cell from receiving a call, which originates from a caller located in an uncongested cell. Also, when the distribution of prospective callees is higher than that of callers in an under-utilized cell, SDP schemes are ineffective as callees do not pay for a call and therefore low discounts do not entice them to increase utilization. In this distribution, dynamic pricing entices prospective callers to make calls but since their distribution is low, the MWN resource utilization does not increase by any significant margin.

To address these problems, we have developed a source-destination based dynamic pricing (SDBDP) scheme, which considers congestion levels in both the call-originating and call-destination cells to compute the dynamic price to be paid by a caller. This SDBDP scheme is integrated with a load-based joint call admission control (JCAC) algorithm for admitting incoming

service requests in to the least utilized radio access technology (RAT). The load-based JCAC algorithm achieves uniform traffic distribution in the heterogeneous wireless network (HWN).

To test the SDBDP scheme, we have developed an analytical model based on M/M/m/m queuing model. New or handoff service requests, arriving when all the RATs in the HWN are fully utilized, lead to call blocking for new calls and call dropping for handoff calls. The call blocking probability, call dropping probability and percentage MWN utilization are used as the performance metrics in evaluating the SDBDP scheme. An exponential demand model is used to approximate the users' response to the presented dynamic price. The exponential demand model captures both the price elasticity of demand and the demand shift constant for different users.

The matrix laboratory (MATLAB) tool has been used to carry out the numerical simulations. An evaluation scenario consisting of four groups of co-located cells each with three RATs is used. Both SDP and the developed SDBDP schemes have been subjected under the evaluation scenario. Simulation results show that the developed SDBDP scheme reduces both the new call blocking and handoff call dropping probabilities during the peak hours, for all caller-callee distributions. On the other hand, the current SDP scheme only reduces new call blocking and handoff call dropping probabilities only under some caller –callee distributions (When the callers were the majority in the HWN). Also, the SDBDP scheme increases the percentage MWN utilization during the off-peak for all the caller-callee distributions in the HWN. On the other hand, the SDP scheme is found to increase the percentage MWN utilization only when the distribution of callers is higher than that of callees in the HWN.

From analyzing the simulations results, we conclude that the SDBDP scheme achieves better congestion control and MWN resource utilization than the existing SDP schemes, under arbitrary caller-callee distribution**.**

Acknowledgements

I would like to thank the almighty God for giving me the strength to carry out this research. I am grateful to my supervisor, A/Prof. Olabisi Falowo, for his tireless and priceless guidance throughout the entire research. I also convey my thanks to the entire Communications Research Group (CRG) for their constructive criticism.

I thank the MasterCard Foundation scholars program, for accepting and nurturing me in their global family of future African leaders. I am grateful to Daniel Muhoro, my line manager at Ericsson Kenya for his support during my final year of masters. A big thank you also to my colleagues and mentors at Ericsson Kenya, Sylvano Idaria, Antony Wandeto and John Ogutu.

Lastly, I would like to thank my parents Janet and John Mutungi, brothers; Josphat and Joseph Mutungi, Sisters; Josephine, Eunice, Rose, Carol and Irene for all the prayers, support and love they have shown me over the years.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

2G	Second Generation
3G	Third Generation
3GPP	Third Generation Partnership Project
4G	Fourth Generation
5G	Fifth Generation
AP	Access Point
BBU	Basic Bandwidth Unit
CAC	Call Admission Control
CAPEX	Capital Expenditure
CDMA	Code Division Multiple Access
CDR	Call Details Register
CSG	Closed Subscriber Group
DDM	Dynamic Discount Module
DPI	Deep Packet Inspection
EDGE	Enhanced Data Rates for GSM Evolution
eNB	Evolved Node B
FTP	File Transfer Protocol
GUI	Graphical User Interface
GSM	Global System for Mobile Communication
HCI	Human Computer Interaction
HSS	Home Subscriber Server
HeNB	Home Evolved Node B

HWN	Heterogeneous Wireless Network
JCAC	Joint Call Admission Control
JRRM	Joint Radio Resource Management
JSS	Joint Session Scheduling
LTE	Long Term Evolution
LTE-A	Long Term Evolution Advanced
M2M	Machine to Machine
MATLAB	Matrix Laboratory
MME	Mobility Management Entity
ML	Machine Learning
MNO	Mobile Network Operator
MWN	Mobile Wireless Network
NBS	Network Billing System
OPEX	Operational Expenditure
OSS	Operational Support Systems
PDP	Packet Data Protocol
PM	Performance Management
PSF	Price Sensitivity Factor
PSM	Price Sensitivity Measurement
QoS	Quality of Service
RAN	Radio Access Network
RAT	Radio Access Technology
RRM	Radio Resource Management

SDBDP	Source Destination Based Dynamic Pricing
SDP	Source Based Dynamic Pricing
THP	Traffic Handling Priority
TUBE	Time dependent Usage-based Broadband price Engineering
UE	User Equipment
UMTS	Universal Mobile Telecommunication System
USSD	Unstructured Supplementary Service Data
VoIP	Voice over Internet Protocol
WCDMA	Wideband Code Division Multiple Access
WiFi	Wireless Internet Fidelity
WLAN	Wireless Local Area Network
WTP	Willingness to Pay

Chapter 1 Introduction

1.1 Background

Mobile wireless network (MWN) traffic has been growing rapidly in the last few years and the trend is expected to continue as networks evolve to the latest technologies. A significant portion of this growth can be attributed to the increase in the number, variety, and capabilities of mobile handheld devices. The demand for mobile data and the availability of new content such as cloud applications which generate two-way traffic upload and download, is also contributing significantly to the increase in MWN traffic [1] [2]. The Cisco visual network index notes that internet traffic has increased fivefold in the years from 2009 to 2014. It further projects a threefold internet traffic increase from the year 2014 to 2019, with the peak period traffic growing more rapidly than the average internet traffic [1]. Bell labs also forecast that by 2017, end user traffic demand will have increased 3.7 times from 2012, with mobile traffic expected to have the highest growth [2]. This growth in traffic is leading to congestion in MWNs.

Mobile network operators (MNOs) are reacting to the growing MWN resources demand by acquiring additional spectrum, deploying small cells, adopting cell splitting techniques and new technologies such as Long Term Evolution (LTE), LTE-Advanced (LTE-A), fourth-generation (4G) and fifth-generation (5G) networks [5] [6]. All these strategies are resulting to an increase in the MNOs' capital expenses (CAPEX). However, revenues from the mobile traffic are not growing at a proportional rate [2].

The peak hour demand for MWN resources has not yet been met despite the adoption of the various new strategies. Also, the MWN resources remain under-utilized during the off-peak periods, when the demand is low. This has led to the use of behavioral and economic mechanisms to control congestion in MWNs during the peak period and increase MWN resource utilization during the off-peak periods [7]. Of these mechanisms, dynamic pricing has been the most widely adopted.

Dynamic pricing presents the user with varying prices to access MWN resources depending on conditions such as the current network utilization level or the amount of time spent in the network. Congestion control is achieved by giving the users an incentive to time shift their non-

critical traffic from the peak hours when the price is high to the off-peak hours when the price is low [8]. The traffic shifted from the peak-hours increases the utilization during the off-peak hours. As a result, the ratio of peak to off-peak MWN resources demand is reduced [9].

The existing dynamic pricing schemes compute the dynamic price based on the congestion level in the call-originating cell. The congestion level in the call-destination cell is neglected. In our work, we will refer to these schemes as source-based dynamic pricing (SDP) schemes. As a result, under caller-callee distributions where the callees are more than the callers, SDP schemes are ineffective since it is only callers who pay. Callees do not pay to receive calls and hence increase in price during the congested peak hours does not deter them from receiving calls.

1.2 Problem Statement

The demand for MWN resources is projected to increase as more powerful mobile handsets become available. Network capacity expansion has proved not to be cost effective in meeting the evergrowing demand for MWN resources since the capital expenditure (CAPEX) does not translate to increase in revenues. Therefore, altering the usage behaviour of consumers through offering dynamic prices is a more economical approach.

The current SDP schemes only consider the congestion level in the call-originating cell while neglecting the congestion level in the call-destination cell. The weaknesses of SDP schemes are:

1. In congested cells, they do not effectively control congestion in user distributions where the callee distribution is higher than that of the callers.
2. In under-utilized cells, they do not improve the MWN resource utilization in user distributions where the callee distribution is higher than that of the callers.

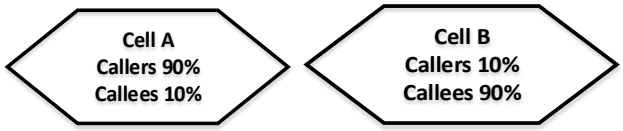
These two weaknesses are examined in detail next.

1.2.1 SDP Weakness during Congested Periods

In MWNs, a caller is the user making a call while a callee is the user receiving a call. Both the caller and the callee consume MWN's resources, and for symmetric services such as voice and video calls, the amount of MWN's resources consumed by a caller and a callee are equal. Usually,

a caller pays for a call while the callee does not. Typically, cells in a mobile network are composed of callers and callees in different proportions. For illustration purposes, we use a caller-callee ratio of 9:1 and 1:9, as shown in Figures 1-1a and 1-1b.

In Figure 1-1a, callers consume 90% of cell A's MWN resources while callees only consume 10%. Assuming peak-hour congestion level, and that users are price sensitive, application of dynamic pricing will increase the service price, and thereby deter more callers (90 % of the users) from calling. Thus, an increase in price in Figure 1-1a will control congestion by discouraging more callers from calling.



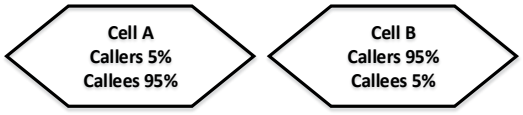
(a) (b)

Figure 1-1: Sample user distribution peak hour

However, in Figure 1-1b, 90% of cell B's MWN resources are consumed by callees while only 10% are consumed by callers. Application of dynamic pricing to increase service price will affect only 10% of users (callers). Callees, who cause 90% of the congestion, will not be affected by the high price because they do not pay. Thus, SDP schemes do not control congestion effectively in cases where the callee distribution is higher than that of the callers.

1.2.2 SDP Weakness during Under-utilized Periods

Let us consider two MWN cells, Cell A and B, both 0% utilized. Let us further assume that only 5% of Cell A's prospective users are callers while 95% are callees. In Cell B, 95% of the prospective users are callers while 5% are callees. These scenarios are illustrated in Figure 1-2a and b below.



(a) (b)

Figure 1-2: Sample user distribution off-peak hour

In the current SDP schemes, application of dynamic pricing in cell 1-2a will lower the price of calls. Only the 5% of the prospective users who are callers can respond to this low price. This increases the MWN resources utilization to a maximum of 5% if all the prospective callers respond to the low price. The remaining 95% do not receive any incentive since the caller is the payer and can only receive a discount in their originating cell. In cell 1-2b, application of dynamic pricing will entice 95% of the prospective users, who are callers, to make calls. This increases the MWN resource utilization to 95% if all the prospective callers respond. Therefore, SDP schemes can increase the MWN resource utilization in cell 1-2b but fail in cell 1-2a, during an under-utilized period.

To address these two problems, we propose a source-destination based dynamic pricing (SDBDP) scheme where the network load in both the call-originating and call-destination cells is considered in the computation of the dynamic price that callers pay. The proposed SDBDP scheme controls congestion during the peak period and improves MWN resource utilization during the off-peak periods, under any caller-callee distribution.

1.3 Research Objectives

The objectives of this work are:

1. Develop an SDBDP scheme which considers the network load in both the call-originating and the call-destination cells to compute the price for different classes of calls.
2. Formulate analytical models for both the developed SDBDP schemes and the existing SDP schemes.
3. Evaluate the performance of the SDBDP and SDP schemes. The performance metrics considered are call blocking probability, call dropping probability and MWN resource utilization.
4. Compare the performance of the SDBDP and SDP scheme using the determined performance metrics.

1.4 Research Methodology

In this work, a source-destination based dynamic pricing (SDBDP) scheme is integrated

with a load based joint call admission control (JCAC). The SDBDP scheme- JCAC integration aims to reduce congestion during the peak periods and improve MWN resource utilization during the off-peak periods. Numerical simulations are carried out in MATLAB tool for a 24-hour period, mirroring a typical day in an MNO's operation.

A linear discount function is formulated to compute the dynamic price based on the network load in the call originating and destination cells for the SDBDP scheme. A call destination cell congestion factor of zero in the formulated discount function yields the dynamic price for SDP scheme. A multi-dimensional Markov model based on *M/M/m/m* queue is used to model the load-based JCAC algorithm. An exponential demand model is used to simulate the user response to the computed dynamic prices, for both SDBDP and SDP schemes. The performance parameters used in the evaluation of both SDBDP and SDP schemes are the call blocking probability, call dropping probability, and MWN resource utilization.

1.5 Scope and Limitations

This work considers the dynamic pricing of voice services in an HWN. The developed SDBDP scheme is suitable in both homogeneous and heterogeneous networks. Two RATs and two classes of service are considered in the simulations. The RATs are assumed to be fully overlapped.

The developed SDBDP scheme is suitable in a single MNO environment. In a multi-operator environment, MNOs may not be willing to disclose their traffic data to their competitors.

It is assumed that core and transmission networks have sufficient capacity and the bottleneck mostly occurs at the radio access networks (RAN). Therefore, the SDBDP scheme addresses resource utilization problem at the RAN.

A load-based JCAC algorithm is used in the performance evaluation of both the existing SDP schemes and the SDBDP schemes. Only admission-level quality of service (QoS) is considered through evaluation of new-call blocking probability and handoff-call dropping probability.

Basic bandwidth units (BBUs) will be used to represent the different radio resources. These radio resources could be frequency channels, code sequence or timeslots depending on the multiple

access technology implemented in the air interface. A fixed capacity in the radio network cells is used in this work.

It is assumed that users are sensitive to price and will shift their demand from the peak hours when the network is congested to the under-utilized off-peak hours. This work will focus on using the SDBDP scheme to control congestion during the peak hours and increasing resource utilization during off-peak hours.

1.6 Contribution to knowledge

This work investigates the inclusion of call-destination cell network load in the dynamic price computation, which has been neglected in the current dynamic pricing schemes. The main contributions of this work are:

1. Congestion control during the congested peak periods for any caller-callee distributions.
2. Improvement of MWN resource utilization during the under-utilized periods, for any caller-callee distributions.

These contributions are contained in the author's publication stated below:

1. Jeremiah Mutungi and O.E Falowo, "A Source-destination Based Dynamic Pricing Scheme to Control Congestion in Heterogeneous Wireless Networks", proceedings of the 27th Annual IEEE international symposium on Personal, Indoor and Mobile Radio Communication (PIMRC), 4-7 September, Valencia, Spain
2. Jeremiah Mutungi and O.E Falowo, "A Source-destination Based Dynamic Pricing Scheme to increase Utilization in Heterogeneous Wireless Networks", proceedings of the Southern Africa Telecommunications Networks and Applications Conference (SATNAC), 4-6 September, Fancourt, South Africa.

1.7 Outline of Thesis

The rest of this work is structured as follows:

Chapter 2 presents a background and literature review on related works in dynamic pricing for circuit switched and packet switched cellular services. It also provides a review of technical concepts applied in this work. These are Heterogeneous wireless networks (HWN), joint call

admission control (JCAC), and radio resource management (RRM).

Chapter 3 presents the architecture and building blocks of the proposed scheme. Each of the building blocks is explained in detail. These are the Dynamic traffic monitoring (DTM) module, the Network Billing System (NBS), the user price notification module and the call details recording (CDR) module. The integration of the proposed module to an HWN is also discussed.

Chapter 4 presents the analytical system model and all assumptions made. A multi-dimensional Markov model based on M/M/m/m queue is presented. A load based JCAC algorithm adopted in this work is modeled. A demand model to examine the response of users to the dynamic is also presented. Finally, the system parameters used in the performance evaluation of both SDP and SDBDP schemes are presented.

Chapter 5 presents a numerical example and evaluation results for the SDP scheme and the proposed SDBDP scheme. Performance parameters namely; new call blocking probability, handoff call dropping probability and percentage system utilization for both SDP and SDBDP schemes under different caller-callee distributions are presented. Finally, a discussion of the results is presented.

Chapter 6 presents the conclusion. The SDBDP scheme is found to control congestion during the peak periods and improve MWN resource utilization during the off-peak periods for arbitrary caller-callee distribution. On the other hand, the current SDP schemes are found to control congestion during the peak period and improve resource utilization during the off-peak period only in cases where the caller distribution is higher than that of callees.

Chapter 7 presents recommendations for future work. Evaluation of the proposed SDBDP scheme can be done in a multi-operator environment. Also, the SDBDP scheme can be developed further for pricing of mobile data services.

Chapter 2 Literature Review

2.1 Heterogeneous Network

An HWN is composed of different co-existing radio access technologies (RATs). Users are able to utilize resources from any of the RATs, and roam from one RAT to another using multimode terminals [3][4].

Technological and economic changes are making it difficult for MNOs to maintain a steady customer base with a single type of access network [4]. MNOs are therefore adopting more than one access technology operating on both licensed and unlicensed spectrum. This is meant to serve the increasing number of subscribers with increasing demand for MWN resources. The different access technologies are deployed in an overlaid version to harness the wide variability in coverage, increased bandwidth, and reliability. These deployments are capable of providing different classes of service governed by their corresponding QoS capabilities. An example of an HWN composed of different access technologies is a cellular network such as 2G and 3G integrated with WiFi and co-existing in the same geographical area [3]. Figure 2-1 below shows such a network composed of 2G, 3G, WiFi and LTE technologies.

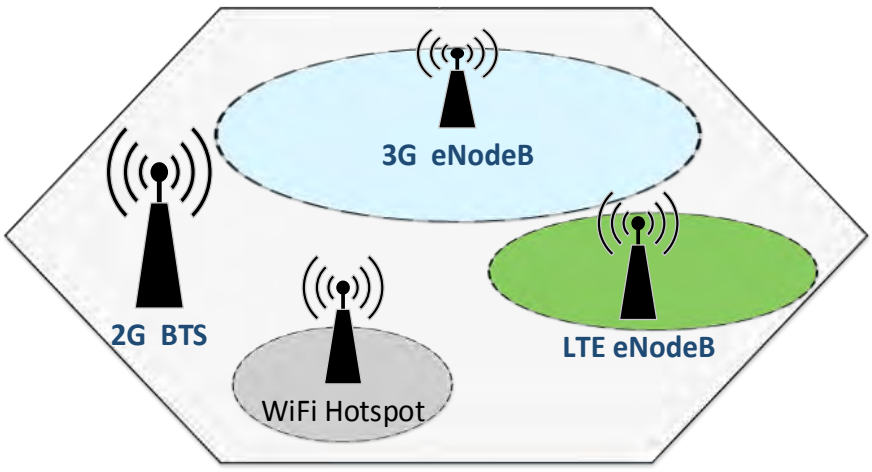


Figure 2-1: An HWN showing different access technologies

There has been an emerging trend whereby low power nodes are deployed within a macrocell, also forming an HWN [3]. According to the third generation project partnership (3GPP), this deployment results to a 3GPP HWN. A 3GPP HWN contains base stations and relay nodes points with different characteristics such as transmission power and radio frequency, of similar access technology, for instance, LTE [5]. The low-power nodes, also called small cells, increase the capacity in hotspots with high user demands as well as fill in areas not covered by the macrocell - both indoors and outdoors. An example of a 3GPP HWN composed of a macro eNB, pico eNB, femto eNB and relay eNB is shown in Figure 2-2. In this Figure, evolved node B (eNB) denotes the access point (AP) for the various wireless technologies. Femtocells are also called home eNB (HeNB).

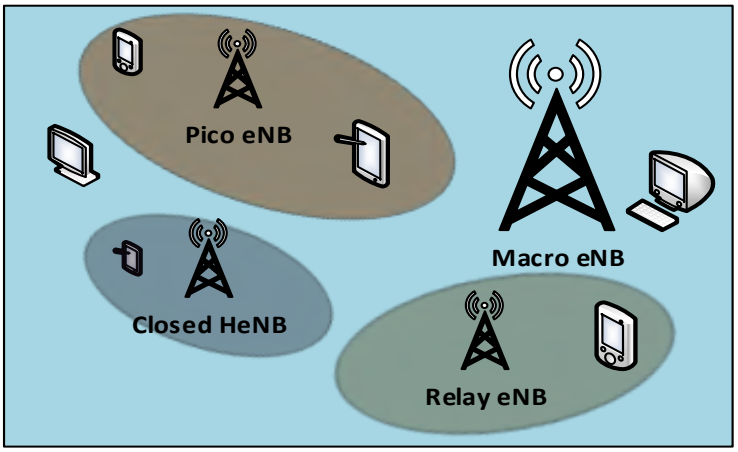


Figure 2-2: LTE heterogeneous network [3]

2.1.1 Motivation for Heterogeneous Wireless Networks (HWNs)

Adoption of HWNs has been necessitated by the following factors:

2.1.1.1 The Increasing Demand for MWN Resources

The increasing demand for MWN resources is forcing MNOs to look for innovative ways to expand their network capacity with a limited CAPEX investment [5]. Deployment of small cells has been found to be less costly than setting up macro-cells [6]. The small cells consume transmit power ranging between 100-2000 mW compared to a range of 5-40W required by the traditional macro-cell. Because of the low transmit power; the small cells' power amplifiers do not require

any cooling [4]. The reduction in transmit power also helps to avoid interference when uncoordinated deployment takes place. This enables users to deploy femtocells for home usage without interference from the traditional macrocell [5].

Picocells are ideal for improving conditions in coverage holes and cell edges while femtocells are used for the indoor environment. Unlike picocells, femtocells may be configured with a restricted association, allowing access only to its closed subscriber group (CSG) members [6]. In areas where wired backhaul is not available, relay nodes can be deployed where the air interface spectrum is used for backhaul connectivity on top of providing access to mobile terminals [5]. In such deployment, the relay node appears as user equipment (UE) to the macro base station and as a regular base station to the UE it serves.

2.1.1.2 Different RATs Optimized for Specific Services

Different RATs have different capabilities such as data rate, coverage area, QoS level and security levels [3]. Therefore, it is necessary to deploy two or more RATs in the same geographical area so as to reap the benefits of the complementary features provided by each. A good example of RATs with complementary features is an integrated UMTS-WLAN network. Users in this network can use the WLAN for data, benefitting from the high data rate offered by WLAN, and the UMTS network for voice services

2.1.1.3 Evolution of Wireless Technologies

The evolution of the wireless technology has led to different access technologies being deployed by MNOs. Most MNOs do not discard their existing technologies after the acquisition of new ones. This can happen due to the complementary nature of access technologies as well as some users having legacy mobile terminals which may not support the new technologies. For instance, many MNOs have installed 4G LTE networks while some of their subscribers do not have 4G capable mobile terminals. As a result, these MNOs are forced to retain their older technologies in order to satisfy these subscribers. The coexistence of these different technologies within the same geographical areas leads to the existence of HWNs.

2.1.2 Challenges of Heterogeneous Wireless Networks

The coexistence of different access base stations within the same geographical area poses challenges to users and MNOs. For efficient HWN operation, the following challenges have to be addressed:

2.1.2.1 Interworking of Different Access Nodes and Technologies

The different access networks in the HWNs should have overlapping coverage such that mobile users can connect to any of the access network based on some established criterion. This criterion could be determined by either the user or the MNO. For instance, admission into a RAT could be based on the type of service requested or the network utilization level. Because some access networks are better suited to provide certain services, users could be admitted to the access networks based on their service requests type. On the other hand, users could be distributed among the different access networks based on the RAT load level by use of algorithms such as load balancing JCAC algorithm. This will ensure uniform load sharing among all the RATs in the HWN.

2.1.2.2 Wireless Network Security

In an HWN, each access network is designed with its own security features. Some of these security features may not be compatible across all the RATs in an HWN. Some access networks have stronger security features than others and the support for any weaker feature is disabled. For instance, the security features of GSM have been enhanced in UMTS, correcting any perceived weaknesses [3]. It is important to standardize the security of each RAT so as to ensure that the entire HWN is fully secure.

2.1.2.3 Mobility Management

Since users are mobile within a geographical area, it is important to ensure that the handover mechanism between RATs is seamless. Therefore, both horizontal and vertical handoff should be supported while maintaining ongoing services' QoS.

2.1.2.4 Common Billing

A common billing platform is imperative for charging the different services in the HWN. When a handoff occurs from one RAT to another, a mobile user will consume MWN resources

from two RATs, for a single service request. The HWN billing must take into account the different amounts of MWN resources consumed from RAT.

2.1.2.5 Resource Management

In order to provide a good QoS to users for different service requests, a radio resource management (RRM) scheme has to be implemented in the HWN. The RRM scheme should ensure efficient utilization of scarce MWN resources, avoiding both congestion and underutilization of the MWN. Efficient RRM schemes enable MNOs to maximize the MWN resource utilization in order to get a good value for money from infrastructure investments.

2.1.2.6 Mobile Terminals

The presence of HWNs leads to the necessity of multi-mode terminals based on the number of RATs in the HWN [7]. Examples of multimode terminals are single mode, dual mode, triple mode and quad-mode. Multimode terminals support more than RAT. For instance a quad-mode mobile terminal supports four RATs. Multimode terminals, therefore, enable users to enjoy services from more than one RAT.

2.2 Radio Resource Management in Heterogeneous Networks

Radio resource management (RRM) involves efficient administration of the scarce MWN resources. It ensures the provision of adequate QoS for different user services. MWN technologies have been advancing over the years, from homogenous networks composed of a single RAT to HWNs composed of multiple RATs. RRM techniques have evolved with advances in MWN technologies from managing homogenous networks to managing the current heterogeneous networks.

In HWNs, MWN resources can be independently managed or jointly managed. When MWN resources are independently managed, each group of subscribers is confined to a single RAT. When the MWN resources are jointly managed, subscribers from any group can consume MWN resources from any of the available RAT, which supports its QOS requirements. However, a multimode mobile terminal is required to access MWN resources from the different RATs [3].

A joint RRM (JRRM) technique is responsible for the efficient management of a pool of radio resources from all the RATs in the HWN. These RATs are based on a specific multiple

access technique such as frequency, time, and code. The JRRM algorithm also facilitates the seamless operation of multiple RATs. As such users are admitted into the various RATs based on some established criterion such as QoS requirements or the network load level.

To achieve efficient utilization of the scarce MWN resources, JRRM uses various algorithms. The two key algorithms are JCAC and joint session scheduling (JSS). In a JCAC algorithm, service requests can be admitted into any of the available RATs in the heterogeneous wireless network. However, a single request cannot be split among two or more RATs. The JSS algorithm allows a single service request to be split among two or more RATs [3]. While the JSS algorithm is advantageous in that it offers more flexibility and sometimes higher data rates than the JCAC algorithm, it introduces a high level of signaling and complexity into the HWN.

The two basic functions of a JCAC algorithm are:

1. Admitting calls into the HWN while ensuring that the QoS of ongoing new and handoff calls is preserved.
2. Selecting the best RAT for admitting new and handoff service requests.

Figure 2-3 shows a JCAC algorithm performing the two functions stated above when it receives a service request from a mobile terminal. At call set-up, the mobile terminal sends a request to the JCAC module which implements the JCAC algorithm. The request contains the type of call (new or handoff) and the class of service. Based on the request details, the JCAC algorithm determines the most suitable RAT to admit the service request. The service request is admitted, if there is sufficient bandwidth in the selected RAT, or declined (blocked for new call and dropped for handoff call) in case of insufficient bandwidth. The admission decision is relayed back to the mobile terminal in the response [8].

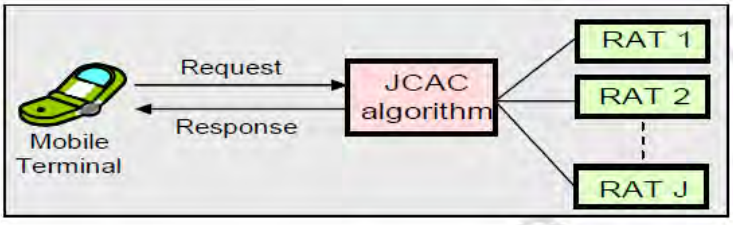


Figure 2-3: Call admission in heterogeneous wireless network [8]

For admitting service requests into the HWN, JCAC schemes are based on different algorithms. These JCAC algorithms determine how new and handoff calls are treated before admission into an HWN. Depending on the sensitivity of different service classes, prioritization of calls is always needed. To achieve prioritization, some resources are reserved for the high priority services. In the guard channel approach, channels are reserved for handoff calls which have a higher priority than new calls [9]. This is because it is more annoying to customers to have a call dropped than blocked. In bandwidth reservation policy, part of the total bandwidth is reserved for handoff calls [3][10]. For instance, if the total number of available bandwidth in a cell is *K* basic bandwidth units (bbu), and the bandwidth reserved for handoff calls is *T* bbu, a new call will only be admitted when *K-T* bbu bandwidth limit is not exceeded. Handoff calls on the other hand will be admitted to the cell when there is available bandwidth in the cell (handoff calls will be admitted even when *K-T* bandwidth limit is exceeded). This is illustrated in Figure 2-4.

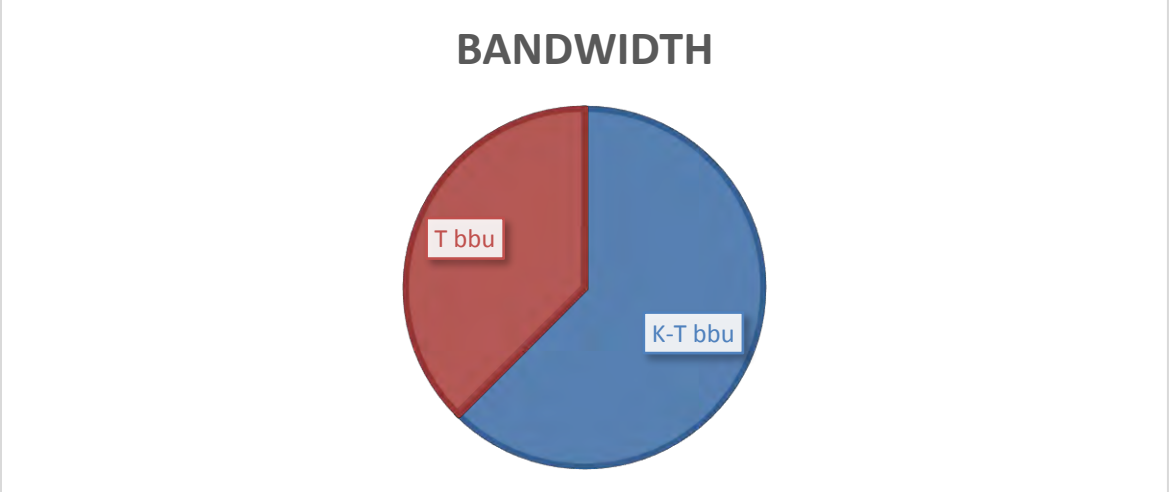


Figure 2-4: Bandwidth reservation technique for handoff calls

For RAT selection, JCAC algorithms also use different criteria for selecting the most suitable RAT into which to admit a call. This criterion can be based on either the users or the MNO's preference. In techniques that consider the user's preference, an indicator is submitted with the user's service request. The user preference could be based on factors that enhance the QoS such as maximum data rate, minimum delay, and least power consumption [3]. For JCAC selection techniques that rely on the MNO's preference, some of the factors considered are:

1. Revenue maximization,
2. MWN resource utilization maximization,
3. Traffic load-balancing among the various RATs.

We are going to examine various RAT selection techniques next.

2.2.1 Service Request Type Based RAT Selection Technique

This RAT selection technique chooses the most suitable RAT based on type of the incoming service request. This technique is based on the fact that some access technologies are optimized to support specific services. Therefore, there is a direct mapping between the RATs and the services [7]. For instance in an HWN consisting of two RATs (2G and WLAN) voice service requests will be admitted into the 2G cellular RAT while data service requests are admitted into the WLAN RAT. This is because WLAN is optimized for data services while the 2G cellular technology is designed to carry voice traffic.

Service request type based RAT selection techniques lead to a highly unbalanced network load. This is because some RATs can be free while others are congested depending on the types of user requests. Since user requests are handled by specific RATs, this technique also promotes unfair distribution of MWN resources among heterogeneous mobile terminals [7].

2.2.2 Service Cost Based RAT Selection Technique

Among the price-sensitive users, the cost of cellular service plays a significant role in making decisions whether to use cellular services. For such users, their preference is low-cost services. In service cost based RAT selection technique, a service request is admitted to the least expensive RAT [3]. The service cost differs from one RAT to another in the HWN. The service cost based RAT selection technique is advantageous to price sensitive users. However, this technique causes a highly unbalanced load in the HWN, especially in regions with price sensitive users. In such areas, the least expensive RAT will experience congestion while the more expensive RATs are under-utilized

2.2.3 Random RAT Selection Technique

Under this technique, incoming service requests are admitted randomly into the available

RATs. In an HWN with J RATs, the probability of selecting a particular RAT is $1/J$. This technique is simple to implement but suffers from unbalanced traffic distribution in the HWN. This could lead to high call blocking and call dropping even at times when some RATs are under-utilized [7].

2.2.4 Load Based RAT Selection Technique

Under this technique, service requests are admitted into the least loaded RAT in the HWN. This technique balances the load in the heterogeneous network by uniformly distributing the incoming traffic among all the RATs present in the HWN.

This scheme achieves a very high network stability [7]. However, in cases where some mobile terminals do not support the technologies presented by the various RATs in the HWN, this technique leads to an unfair distribution of MWN resources.

Falowo *et al* in [11] analyzed and compared service-based, random-based and load-based RAT selection techniques. The performance metric used for comparison was connection-level QoS derived from new call blocking and handoff call dropping probabilities. From the simulation results obtained, the load-based RAT selection technique was found to provide the best connection-level QoS while the random RAT selection technique provided the least connection-level QoS.

2.2.5 Statistical Based CAC Algorithm

Statistical based CAC techniques consider parameters such as QoS, cell population and channel efficiency to estimate the mobility of users between cells. This mobility estimation determine the amount of bandwidth to be allocated for new calls and handoff calls. Eipstein *et al* in [12] proposed a statistical CAC algorithm based on one step prediction scheme (OSPS). The OSPS technique predicts the necessary bandwidth reserved in the home cell and in the adjacent cells for the next time slot. The disadvantage of the OSPS technique is that it assumes that users can move to all adjacent cells with equal probability. This drawback results to wastage of bandwidth hence increasing the call-blocking and dropping probabilities.

2.2.6 Game Theory Based CAC Technique

Game theory provides a suite of analytical tools for analysing interactions of players with

conflicting interests [13]. Each player performs independent actions confined to their own strategy space which impacts both their payoffs and the other players' payoffs. Niyato *et al* in [14] propose a game theoretic CAC technique which considers various classes of service as players with Nash equilibrium being used to determine the amount of bandwidth to be offered to incoming connection. The payoffs are calculated as a function of perceived delay and output performance. Antoniou *et al* in [15] formulate a game using service providers as players, each with strategies in the form of amount of bandwidth offered. Nash equilibrium is then used to maximize the service providers' profits while satisfying the mobile users.

2.2.7 Machine-Learning (ML) CAC Technique

ML RAT selection techniques have the ability to learn the system behaviour from past data and estimate future behaviour based on the learned system model. ML CAC techniques achieve this by using automated and intelligent data analysis techniques to develop models from datasets. The developed models are used to predict the most optimal RAT to select. Bashar *et al* in [16] propose an ML based CAC mechanism which uses the concept of Multiple Entity Bayesian Network (MEBN). A call connection request consists of characteristics defined in terms of traffic descriptors and desired QoS. These characteristics are the inputs to ML algorithm. The output of the algorithm are values of QoS metrics like packet-loss, delay and jitter. An acceptance or rejection decision is made based of the output values. Based on the inputs and outputs, the ML based CAC is trained offline with a set of data for a desired period of time. This training data consists of cases where both inputs and outputs are known. When operating in online mode, the output of the ML based CAC, which are the predicted QoS metrics, are compared to the existing QoS from which an admission decision is made.

The overall ML based CAC performance is based on the prediction accuracy of the model used for training. This depends on the diversity of the cases for training it.

2.3 Congestion in Heterogeneous Wireless Networks

Despite the advancement in MWN technologies, congestion in HWN has been increasing over the years. Industry analysts project further increase in congestion for the next few years [1][2]. Some of the causes of congestion in HWNs are:

1. Greed among users of MWN resources.
2. Mobile devices with powerful computing capabilities.
3. Capacity hungry applications.
4. Cloud services.
5. Machine to machine (M2M) communication.

These causes are discussed in detail next.

2.3.1 Greed among Users of MWN Resources

MWN resources are shared among the subscribers of a particular MNO. A greedy subscriber can continuously transmit, limiting the amount of MWN resources available to the other subscribers. This behavior constitutes to tragedy of the commons in economics [17]. Also, if a user is allowed to transmit when the network is congested, the QoS of other users in the network such as packet delay and packet loss may become severely degraded. MNOs are therefore striving to maximize economic efficiency in a way that ensures the users who value the MWN resources the most get them. The mostly widely used economic mechanism is dynamic pricing which varies the cost of a call based on the network load or the duration of the call.

2.3.2 Mobile Devices with Powerful Computing Capabilities

In the recent years, there has been an increasing uptake of handheld devices equipped with powerful processors, high-resolution cameras, and large displays. This trend is supported by Moore's law which states that the number of components in an integrated circuit (IC) double every year [18]. Fitting more components in an IC enable electronic devices to be manufactured at a cheaper cost. The cost of powerful mobile devices is going down with the decreasing manufacturing cost enabling more people afford them.

The powerful mobile devices make it possible to stream high-quality videos, hence increasing the demand for MWN resources. Cisco has projected that the average monthly data usage per mobile device will rise from 150 MB in 2011 to 2.6 GB in 2016 [1]. This trend is confirmed by the introduction of new features like Siri, a voice recognition feature in Apple devices, which doubled the data consumption of Apple devices users, after introduction [19]. There is also increasing usage of laptops and notebooks fitted with wireless network dongles for

connection to the internet. This large number of mobile devices has increased demand for high-speed wireless network access. Since wireless spectrum is a limited resource, the increased demand for MWN resources has led to congestion in MWNs.

2.3.3 Capacity-hungry Applications

The growing popularity of powerful handheld mobile devices has led to rapid growth of bandwidth-hungry applications for social networking, music and personalized online news on top of file downloads and video streaming. These bandwidth-hungry applications increase the demand for MWN resources leading to congestion. Also, some of the applications in the market are not optimized for bandwidth consumption leading to more than ordinary bandwidth consumption [20].

2.3.4 Cloud Services and Machine to Machine (M2M) Communication.

Cloud-based services synchronize data across multiple devices leading to a surge in mobile traffic for MNOs. Examples of some of these services are iCloud, Dropbox and Amazon elastic cloud [21]. Similarly, M2M applications that generate data intermittently, for instance, sensors, actuators, and smart meters or continuously such as video surveillance often load MWN with large signaling overhead [22].

2.4 Controlling Congestion in HWNs

Several techniques for controlling congestion in HWNs have been suggested. In this section, we are going to discuss the key techniques used to control congestion.

2.4.1 Capacity Expansion through Spectrum Addition.

This involves expanding the MWN bandwidth of the MWN by acquiring additional radio spectrum. However, acquisition of new radio spectrum is limited by regulatory restrictions since spectrum is a limited resource. Moreover, there are explicit costs required when a new band cannot be accommodated by existing configuration. These costs include the purchase of new base station equipment, antennas, civil works and upgrades [23].

2.4.2 Traffic Offloading

Traffic offloading involves the transfer of MWN traffic from the traditional macrocell

network to small cell network such as femtocells. Small cells can offload both indoor and outdoor traffic and hence alleviate congestion in the macrocell.

Spectrum licensing is not required for traffic offloading. WiFi operates in licence-exempt spectrum while small cells operate in licensed spectrum in a way which does not significantly increase the overall spectrum requirement. The small cells achieves this by successfully coordinating with macro cells in the same channels or in dedicated channels with a high level of frequency reuse [23].

2.4.3 Imposing Penalties on Greedy Users.

In this method, the MNOs drop packets from greedy users, place a bandwidth cap or time limit their calls to alleviate congestion. However, imposing penalties on greedy users often fail to achieve economic efficiency. This is because such schemes do not provide guarantees that users who value the MWN resources the most are actually the ones who get them [24].

Despite the adoption of the various techniques discussed above, the problem of congestion in HWNs is still on the increase. Furthermore, the peak hour demand for MWN resources has not yet been met despite some huge infrastructure investments [22]. This has led to researchers considering the use of behavioral and economic mechanisms to control congestion in MWNs with dynamic pricing being widely adopted [25].

Dynamic pricing is based on decreasing or increasing the price of network services depending on some condition in the network, for instance, the network load or the length of the service request. Congestion control is achieved by giving the users an incentive to time shift their non-critical traffic from the peak hours when the price is high to the off-peak hours when the price is low [26]. As a result, the ratio of peak to off-peak demand of MWN resources is reduced [27]. Dynamic pricing in HWNs is discussed in details in the next section.

2.5 Dynamic Pricing in Heterogeneous Networks

Dynamic pricing involves the use of economic and behavioral strategies to control congestion as well as increase resource utilization in HWNs [28] . The price for utilizing MWN resources is determined dynamically according to the network load. The price is increased when

the network load is high to reduce the demand for MWN resources and decreased when the network load is low so as to increase the demand of MWN resources [17]. As such, dynamic pricing is used to promote a rational and efficient use of MWN resources by influencing the users' behavior.

Dynamic pricing also allows the MNO to modify a user's service priority according to the varying MWN resources demand. This generates more profit for the MNO. In some dynamic pricing schemes, users are able to choose their preferred class of service according to the price they are willing to pay [24].

The design of any dynamic pricing scheme depends primarily on two components. These components are:

1. User Demand Model: The demand model takes care of the demand behavior of the users. Different users will react differently to the same price, owing to their price sensitivity. This is termed as price elasticity of demand in economics, which is a measure of the responsiveness of a change in demand for a good or service to a change in price [29]. Different pricing schemes will use different demand models. For instance, some use exponential functions to represent the users' demand for MWN resources. Others use utility functions to represent the user's preference for a certain service [24].
2. Price function: This determines how the charge for any class of service is calculated for a given time, bandwidth or power [24]. Different price functions are derived for different dynamic pricing schemes. Some schemes use linear functions [10] while others use exponential functions [30].

The main objectives of dynamic pricing are:

1. Network Optimization: Dynamic pricing shifts traffic from the congested peak hours to the under-utilized off-peak hours. This reduces congestion during the peak hours and utilizes the free capacity during the off-peak hours.
2. Subscriber Acquisition: Dynamic pricing creates an option of making calls at low costs when the MWN is under-utilized. This encourages low-income subscribers to enjoy

services they would not have afforded if there were no discounts. Thus, dynamic pricing scheme lowers the entry barrier for low-income subscribers. In this way, dynamic pricing can be used to beat competition since an MNO is able to compete at lower price levels.

3. Maximization of MNO revenue: When an MNO attracts the low-income users by offering affordable services during the off-peak hours, their overall revenues increase. For existing subscribers, dynamic pricing increases their value for money since they are able to make longer calls when the discount is high.

2.6 Dynamic Pricing for Circuit Switched Services

For circuit switched services, the price per unit time or bandwidth is determined at call set-up and is fixed for the entire call duration. This price is determined according to some condition in the network such as network load. There are two types of calls namely; new call and handoff call. A new call happens when a new user requests connection to access MWN resources. A handoff call occurs when an active user moves from one cell to another. Since charging occurs at call set-up, handoff calls are not affected by the dynamic pricing as they are already charged in their cells of origin [3].

Dynamic pricing at the admission level controls the arrival rate of users into an HWN by offering monetary incentives. This serves to maintain the connection-level QoS at the desired threshold [17]. The parameters mainly used to evaluate performance of an HWN for circuit switched services are:

1. New call blocking probability- This is the probability that a new call is rejected due to unavailability of MWN resources to support it.
2. Handoff call dropping probability- This is the probability that a handoff call is dropped due to unavailability of MWN resources to support it.

2.6.1 Procedure for Circuit Switched Dynamic Pricing

The dynamic pricing procedure is shown in Figure 2-5. When a user makes a new call request, the base station calculates the dynamic price for a unit of time or bandwidth, depending on the current network load [17]. The computed dynamic price is then communicated to the user.

The user has the option of accepting or declining the presented dynamic price. If the user accepts the dynamic price, the call is established.

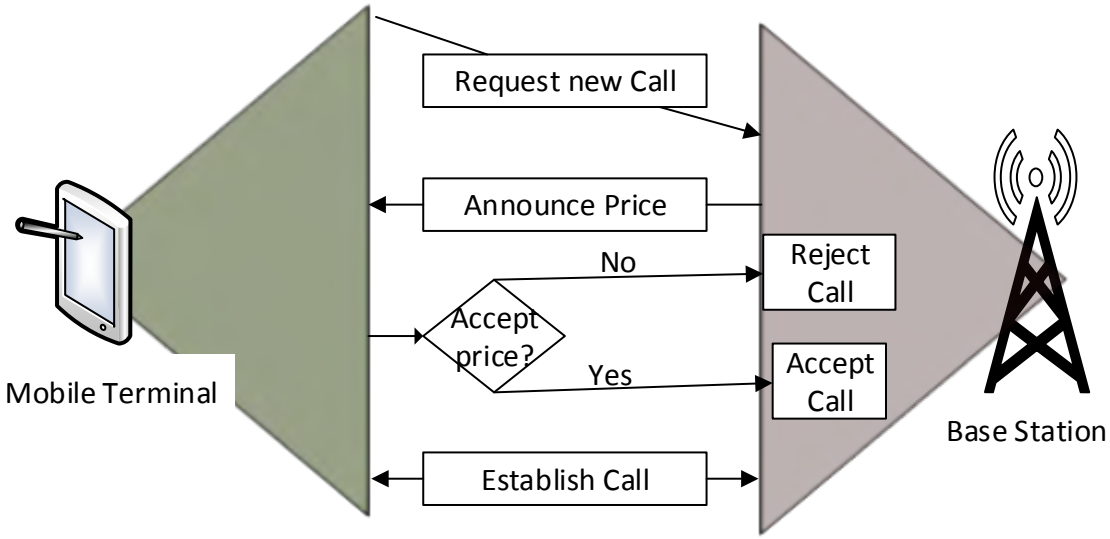


Figure 2-5: Admission level dynamic pricing procedure [17]

We are going to classify dynamic pricing in circuit switched networks into two; Dynamic pricing integrated with call admission control (CAC) and Dynamic pricing with implied CAC.

2.6.2 Dynamic Pricing Integrated with Admission Control

Admission control is a resource management strategy used to enforce connection-level QoS in HWNs. It limits new connections to a network based on the availability of idle MWN resources. As such, an admission control algorithm on its own does not ensure congestion control since it does not provide any incentive to promote rational and efficient resource usage. With admission control only, the new call blocking probability and handoff call dropping probabilities reach very high levels during the peak periods. To yield maximum benefits, admission control algorithms are integrated with dynamic pricing in order to offer monetary incentives to users to time-shift their non-critical traffic from peak periods to off-peak periods. Dynamic pricing integrated with admission control is able to:

1. Control peak period congestion by reducing the blocking and dropping probability for new and handoff calls respectively.

2. Enforce connection-level QoS by limiting the number of connections admitted into the HWN. New or handoff calls will only be admitted into the HWN when there are sufficient MWN resources to support them. This ensures that QoS of both new and ongoing calls is maintained.
3. Increase MWN resource utilization during the off-peak period when the HWN is under-utilized. The dynamic pricing scheme entices users to make calls by offering very high discounts.

A CAC algorithm is used in homogeneous wireless networks and a joint call admission control (JCAC) algorithm is used in HWNs. We proceed to review schemes which have integrated dynamic pricing with admission control.

In [10], Kabahuma *et al*. proposed a scheme that integrates dynamic pricing with JCAC to reduce peak hour congestion and increase MNO revenue. The proposed scheme is composed of three major components as shown in Figure 2-6. These components are JCAC module, the pricing entity and user reaction to price module. A next generation wireless network (NGWN), which is heterogeneous in nature, is used to evaluate the proposed scheme.

A load-based JCAC algorithm is used to distribute calls to the various RATs in the HWN based on their capacity. The pricing entity computes a dynamic price based on a linear discount function. A high discount is advanced when utilization of the HWN is low and vice-versa.

The demand for MWN resources by users is based on Van Westendorp price sensitivity measurement (PSM). The Van Westendorp PSM posits that there is a range of prices that a user is willing to spend, below which the service credibility is doubted [31]. The PSM model assumes that when the prices fall within a specific range, users will have a specific price sensitivity factor (PSF). Users' response to dynamic price is determined by the PSF. The PSF

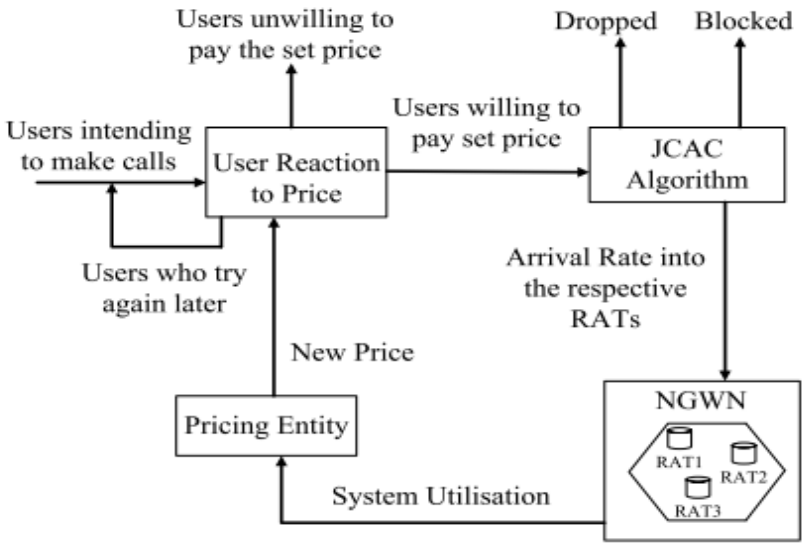


Figure 2-6: Dynamic pricing scheme integrated with JCAC schematic [10]

values are obtained from market research and vary by region. The PSM model adopted for this work is shown in Table 2-1. This model shows the price range, the user perception, and the corresponding PSF values.

Table 2-1:Price Sensitivity Measurement [10]

Price Range	User Perception	PSF
$0 \leq P_x \leq 0.3$	Very Cheap	4
$0.3 < P_x \leq 0.6$	Moderate	2
$0.6 < P_x \leq 0.9$	Expensive	1
$0.9 < P_x \leq 2$	Very Expensive	0.7

Assuming λ_{ij}^{o} to be the arrival rate of users before application of dynamic pricing and $\lambda_{ij}^{n}$to be the arrival rate of users after application of dynamic pricing of class *i* in RAT *j*, the PSM model is used to determine the user response to dynamic prices as shown in Equation 2.1.

$$\lambda_{ij}^{n} = PSF * \lambda_{ij}^{o} \quad (2.1)$$

This scheme, when compared to flat pricing, increases the MNO's revenue in addition to controlling congestion during the peak period. However, the efficiency of this scheme in optimizing MWN resource utilization under different caller-callee distribution has been ignored.

Papavassiliou *et al* in [30] have presented a congestion pricing scheme integrated with CAC. This scheme is meant to control congestion as well as maximize total user utility in an MWN. The authors have defined user utility as the function of call rejecting probability P_b. The call rejecting probability is further defined as the weighted sum of call blocking probability and call dropping probability.

The authors in [30] have shown that in a given wireless network, there exists some optimal new call arrival rate (λ_n^*) that maximizes the total utility of the user. The new call arrival rate (λ_n), therefore, ensures that MWN resources are utilized optimally. Congestion in the wireless network will occur when the new call arrival rate exceeds the optimal new call arrival rate ($\lambda_n > \lambda_n^*$). When the wireless network is congested, large numbers of new and handoff calls are blocked and dropped respectively. As a result, users receive lower QoS than expected. MNOs therefore strive to keep the new call arrival rate less than or equal to the optimal new call arrival rate in order to control congestion and maximize user utility.

A schematic of the scheme in [30] is shown in Figure 2-7.

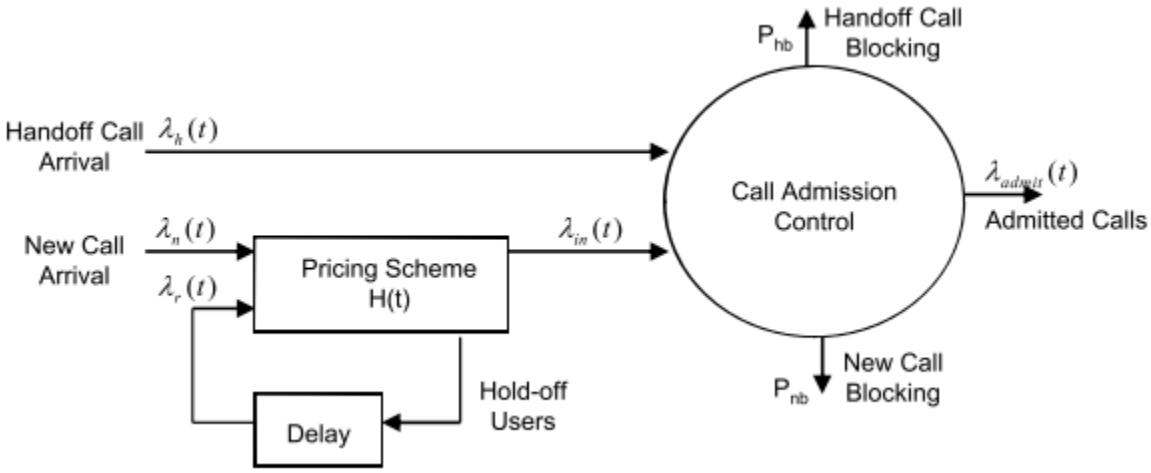


Figure 2-7: Congestion pricing integrated with CAC in [30]

As seen in Figure 2-7, the scheme is composed of two main entities namely; the Call Admission Control and the pricing scheme. The CAC accepts or rejects new and handoff calls based on the amount of available MWN resources. The pricing scheme has two modes of operations. These modes are:

1) When the new call arrival rate is less than the optimal new call arrival rate ($\lambda_n < \lambda_n^*$). In this mode, there are available MWN resources for both new and handoff calls and a fixed price is charged to every user.
2) When the new call arrival rate is higher than the optimal new call arrival rate ($\lambda_n > \lambda_n^*$). In this mode, the demand exceeds the available MWN and therefore the network is congested. A congestion charge depending on the MWN resource utilization is charged to every new caller. Users who accept the congestion price, which is higher than the fixed price, are served. On the other hand, users who find the congestion price too high defer their calls to a less congested period, where they pay the lower fixed price. These users, who defer their calls, are denoted by hold-off users in Figure 2-7 and generate retry traffic with arrival rate λ_r.

The scheme uses an exponential demand model suggested in [32] to describe the reaction of users to the change in price. This model is shown in Equation 2.2 where *D(cp(t))* is the percentage of users that will accept the congestion price, *fp* is the fixed price charged when $\lambda_n \leq \lambda_n^*$, and *cp(t)* is the price charged during congestion periods. The optimal new call arrival rate (λ_n^*), which ensures that the user utility is maximized while controlling congestion, is obtained from the demand function. The value of *cp(t)* is calculated such that the arrival rates $\lambda_n + \lambda_r$ (i.e., of new and retry users) are equal to the optimal arrival rate λ_n^*.

$$D\big(cp(t)\big) = exp\left[-\left(\frac{cp(t)}{fp} - 1\right)^2\right], cp(t) \geq fp \qquad (2.2)$$

This congestion pricing scheme integrated with CAC achieves better congestion control and obtains more revenue when compared to conventional flat-pricing schemes. However, this scheme only concentrates on preventing the network from getting congested by keeping $\lambda_n \leq \lambda_n^*$. It does not address the issue of the network being under-utilized during the off-peak hours since users are charged a fixed price with no incentive to increase their utilization. Moreover, the demand function adopted from [32] ignores the price elasticity of demand.

Hew *et al* in [33] suggested an improvement to the works of Hou et al. in [30]. They propose a congestion pricing scheme integrated with CAC to reduce congestion and increase revenue in a wireless network. Hew *et al* have considered multiple classes of service differentiated

by the QoS needs. They have used a Weibull distribution with mean φ_i and shape β_i, to model the users' willingness to pay (WTP) for different classes of service. The weibull distribution is versatile and can take up the characteristics of other types of distributions by altering the value of β_i. A new caller makes a connection request if his WTP is higher than or equal to the cost of the service as seen in Equation 2.3. In this equation, $cp_i(t)$ denotes the price of class *i* service at time *t*, b_i denotes the bandwidth units while u_i denotes the average call duration.

$$\varphi_i = \frac{cp_i(t) * b_i}{u_i} \tag{2.3}$$

Assuming *k* classes of service, new class *i* users who will be rejected by the CAC or have insufficient WTP will either retry later with probability α_{Rj} , opt-out to another service $j \neq i$ with probability α_{SOij} or abandon the system with probability α_{Aij}. New class *i* users who retry later and users of service class $j \neq i$ who opt-in are said to be in orbit, as shown in Figure 2-8.

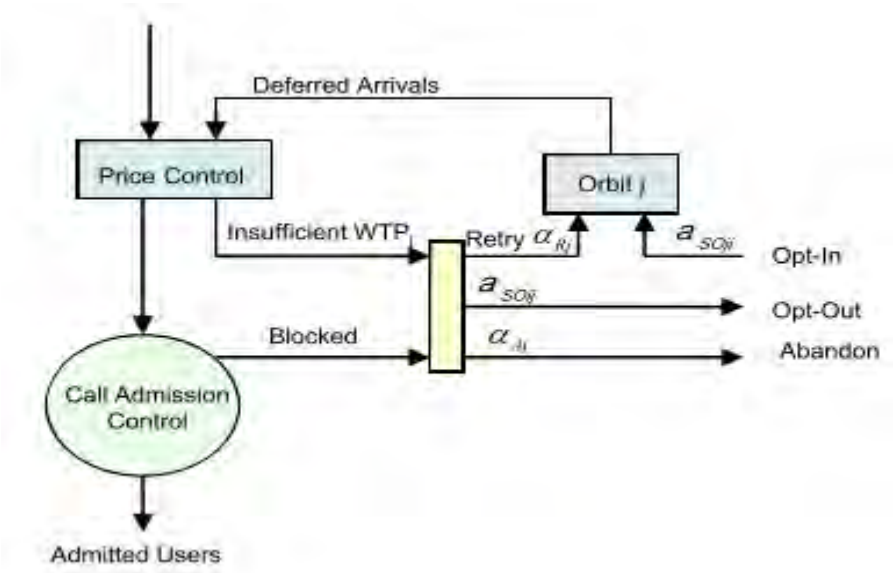


Figure 2-8: Schematic representation of CAC scheme in [33]

2.6.3 Dynamic Pricing with Implied Admission Control

Some dynamic pricing schemes limit the number of connections at admission level implicitly without application of any admission control algorithm. These dynamic pricing schemes regulate the usage of MWN resources without integration to a CAC for homogeneous wireless

networks or a JCAC for heterogeneous wireless networks. A review of two dynamic pricing schemes with implied admission control is presented next.

2.6.3.1 Differentiated Services Dynamic Pricing

Mandal *et al* in [34] have presented a dynamic pricing scheme for different classes of service in wireless networks. In their scheme, users are classified into *k* classes of service based on the QoS requirements. Each class *i* service *(1 ≤ i ≤ k)* is defined by the new call admission probability d_i where $d_1 > d_2 > d_3 \ldots \ldots > d_k$. The price of a call in the i^{th} class of service is varied according to the traffic load in the network in *m* discrete steps namely, $cp_{i1}, cp_{i2}, cp_{i3} \ldots\ldots cp_{im}$. In order to regulate demand, these discrete steps are within a price range of cp_i^{max} and cp_i^{min}. The upper and lower values of the range are chosen such that the revenues of the MNO are maximized.

The user demand under the dynamic price is modeled using an exponential model shown in Equation 2. 4 where $D_i(t)$ is the quantity of resources demanded by class *i* users, $b_i(t)$ is the price elasticity of demand, $a_i(t)$ is the demand shift constant while $cp_i(t)$ is the dynamic price of class *i* call at time *t* [17].

$$D_i(t) = a_i(t) * e^{-b_i(t)*cp_i(t)} \tag{2.4}$$

The price elasticity of demand and demand shift constant take different values for different times of the day, and can be determined by market studies on real demand for users of the different classes.

This dynamic scheme is advantageous in that the dynamic prices for the different classes of service are chosen from a predetermined set. As a result, users know the charge for a particular class of service. The main challenge comes when determining the upper and lower limits of the charge for the different classes of service. Choosing a high value for the upper limit may discourage usage during the congested periods, leading to revenue losses.

2.6.3.2 Auction Based Dynamic pricing

Yaiparoj *et al* in [35] have proposed a congestion pricing scheme for services in an enhanced data rates for GSM evolution (EDGE) wireless network. The authors argue that estimating the users demand, as Mandal *et al* in [34] discussed earlier, is too complex and time-consuming. This is because user demand can be a function of QoS, pricing structures as well as

applications and services that a user is running which differ from market to market. The authors in [27] have presented an auction based scheme to price different classes of service in an MWN. As a result, information about user demand is conveyed through bids submitted by users.

The proposed scheme is similar to smart market pricing, where users submit bids along with their service request [36]. However, this scheme differs from smart-market pricing scheme in that bids are submitted at the beginning of a call as opposed to attaching a bid with every packet. This scheme assumes the multi-unit Vickery bid, where users are allowed to bid for more than one unit of the same item. After bidding, the N highest bidders are admitted into the MWN where they pay the value of the highest losing bid [17]. All the submitted bids are required to be higher than or equal to an optimal price called the reserve price. The reserve price is computed such that the MNO maximizes the revenues from the utilization of the MWN resources. The reserve price is computed based on the network utilization; low during periods of under-utilization and high during periods of network congestion.

This scheme is advantageous in that user demand is conveyed through the submitted bids and not estimated like the scheme proposed by Mandal *et al* in [34]. Users are also less likely to shade bids in Vickery auction. Users shade bids by bidding below the true valuation of the MWN resources under auction so as to avoid subsequent loss of winning in auctions where users pay the highest bids [17]. However, this scheme has been designed for enhanced data rates for GSM evolution (EDGE) networks while the current wireless networks are heterogeneous in nature. Therefore, this scheme requires adjustments in order to be applicable in today's heterogeneous networks.

Mandal *et al* in [37] and [38] have proposed an auction based congestion based dynamic pricing scheme similar to [34]. The only difference with the auction scheme in [34] is the lack of a reserve price. Instead, the authors in [37] and [38] have proposed two multi-unit auction mechanisms namely, uniform pricing auction and discriminatory pricing auction. For uniform pricing auction [39], the M highest bidders pay the price of the lowest winning bidder. In discriminatory pricing auction [39], the M highest bidders are chosen and each gets charged their own bid. Different classes of service are considered based on their QoS requirement. The MNO maximizes their revenue by admitting the highest bidders according to the auction type to use

while meeting the QoS requirement of the different classes.

From simulation results, the auction based dynamic pricing schemes in [37] and [38] achieve more revenue than flat pricing. However, the discriminatory pricing auction suffers from bid shading. Uniform auction pricing scheme does not suffer from bid shading and achieves more revenue than discriminatory pricing auction. Another disadvantage of the schemes in [37] and [38] is the absence of reserve price. Reliance on bids to control congestion during peak hours and increase utilization during the off-peak hours is not efficient. This is because the users may under-bid the MWN resources resulting to revenue losses.

2.7 Power-Level Dynamic Pricing

In power-level dynamic pricing, users are dynamically charged according to their power consumption either in uplink or downlink. As a result, power-level dynamic pricing regulates the power usage in the network while at the same time controlling congestion during the peak hours [17]. Unlike admission level dynamic pricing, power-level dynamic pricing happens after call admission into the network and the price varies during a call duration based on the power consumption [40]. This is different from admission-level dynamic pricing where the price of a call is determined at call setup and remains constant for its entire duration [17]. The power-level dynamic pricing scheme is ideal for code division multiple access based (CDMA) wireless networks which suffer from interference caused by the power transmitted to the base stations from the user's handset and vice-versa [17][40]. The interference causes degradation in QoS of the different services offered by the MNO.

2.7.1 Advantages of Power-Level Dynamic Pricing

Power-level dynamic pricing is beneficial to both the MNOs and the users in the following ways:

1. It provides monetary incentives to the users for rational usage of power. This helps the MNO in mitigating interference and alleviating congestion in the MWN [17].

2. It increases the battery life of the mobile terminals by saving the transmission power from the users' mobile handsets. Therefore, users are able to charge their mobile handsets less regularly.

2.7.2 Modes of Power-Level Dynamic Pricing

In MWN, the downlink and uplink channels have different characteristics. Therefore, the power computations for the uplink and downlink usage are done differently [17]. For uplink traffic, the base station computes the price for every unit of power and sends this information to the mobile handsets. The user's mobile handset then adjusts the amount of power to use for uplink transmission based on the price received from the base station. This process is shown in Figure 2-9.

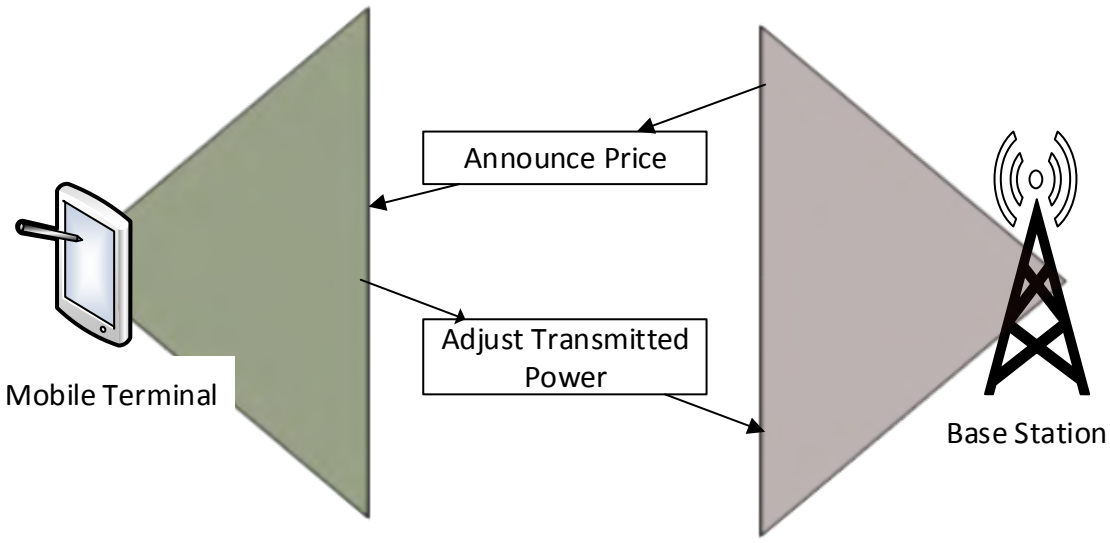


Figure 2-9: Uplink power-level dynamic pricing [17]

For downlink transmission, the base station announces the price per unit of power to the users' handsets. The handsets then decide the amount of power to use for downlink traffic from the base station. The handsets send the chosen power level to the base station, which transmits to the users' handsets according to the requested power level. This process is as seen in Figure 2-10.

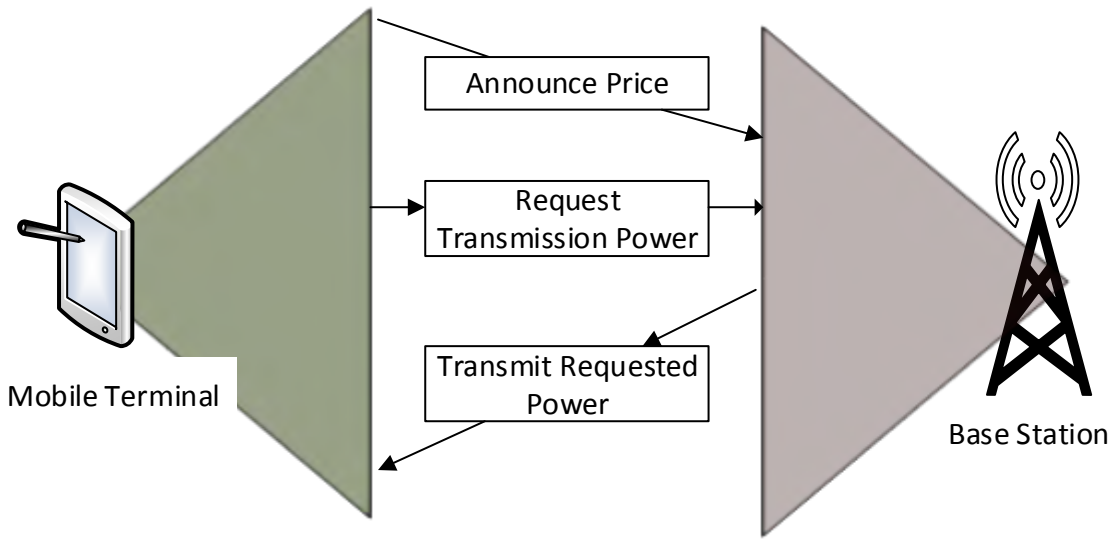


Figure 2-10: Downlink power-level dynamic pricing [17]

The power-level dynamic pricing scheme uses different variations of the uplink and downlink power control concepts. Based on this, power-level dynamic pricing schemes can be classified into two broad categories:

1. Uplink power-level control dynamic pricing.
2. Downlink power-level control dynamic pricing.

2.7.2.1 Uplink Power-Level Control Dynamic Pricing Scheme

In CDMA networks, the uplink power is limited by interference [40]. Therefore, users cannot increase their transmit power without bound to prevent interference to other users. Uplink power-level control dynamic pricing schemes aim to reduce interference in the uplink by incentivizing users to control their uplink transmission power [41].

Han *et al* in [42] have proposed an uplink power level dynamic pricing scheme in which the users behavior is modeled by a utility function. The utility function is, in turn, a function of the power transmitted by the user's handset. Each user decides on their uplink transmit power which maximizes their utility. According to the uplink transmit power, a dynamic price to be paid by the user is computed. The authors in [42] show that their proposed dynamic pricing scheme converges to an optimal power allocation that maximizes the social welfare.

Since the power computation happens at the user side, the scheme in [42] reduces signaling

from the user to the base station. However, the power computation at the user side drains the handset battery leading to low battery life. Furthermore, since users near the base station use less power to transmit than those further away, this scheme creates unfairness. For the same data rates, users at different locations of a cell are charged differently.

2.7.3 Downlink Power-Level Dynamic Pricing

In the downlink, the total power transmitted by the base station is limited to avoid causing interference to adjacent cells. Therefore, in downlink power level dynamic pricing schemes users are given monetary incentives to accept low downlink power from the base stations.

Siris *et al* in [41] have proposed a power-level dynamic pricing scheme which strives to control the uplink and downlink transmit power transmitted. They show that the transmission power from the base station to user characterizes resource usage. As a result, users decide on the downlink transmission power so as to maximize their utility. The utility is assumed to be a function of the user's average throughput. And users are then charged proportionally to their power usage.

Like the uplink power-level dynamic pricing scheme, users closer to the base station enjoy substantially lower prices than those further away. This is because users near the base station can afford to transmit at lower power which in turn attracts a lower price. Therefore, this scheme increases efficiency at the expense of fairness.

2.8 Pricing in Mobile Data Service

As mobile data gradually replaces voice as the core product in MNOs service offering, a wide range of mobile data pricing schemes to build its value proposition have been explored [43]. Traditionally, mobile data pricing has been dominated by flat rate and usage-based pricing. In flat rate mobile data pricing, the users pay a fixed charge to access mobile data services for a given time duration. The popularity of flat-rate mobile data pricing plans arises from the certainty they provide to users expected monthly bill and the ease of implementation.

In usage-based pricing, users are billed according to the volume of data they consume. Usage-based pricing is usually offered in two ways; pre-paid and post-paid data plans. In pre-paid data plans, users purchase data bundles valid for a given time period. The validity of the pre-paid

data bundles could be a week, month, 3 months etc. In post-paid data plan, users pay for the volume of data consumed, usually after the consumption.

Most MNOs in the world use usage-based pricing. AT&T, an MNO in the USA, offered a monthly data cap of 250 Gbytes with $10 overage for additional blocks of 50 Gbytes in 2011. In 2010, the same MNO was offering 3-tiered data plans: 200 Mbytes for $15, 2 Gbytes for $25 and 4 Gbytes for $45 [15].

Both flat-rate pricing and usage-based pricing ignore MWN congestion during the peak periods and under-utilization during the off-peak periods. Therefore, both schemes do not alter user behavior to optimize the MWN resource utilization. In order to reap maximum benefits from the growing mobile data demand, MNOs are adopting more innovative pricing schemes. These schemes are broadly referred to as smart data pricing, an umbrella term for a suite of pricing and policy practices, used to control user behavior based on the MWN conditions at the data consumption epoch [21]. These smart data pricing plans take into account the following key factors, most of which are new in the MWN landscape:

1. The growth in mobile traffic with high time elasticity of demand. This type of traffic can be scheduled to a less congested time, without the users' intervention, helping minimize congestion in the MWN [25]. Examples of traffic with high elasticity of demand is file downloads, cloud backups, and machine to machine communication that sends data intermittently.
2. The introduction of a congestion feedback control loops such that the current MWN utilization level is computed in real-time. The MWN utilization statistics are then communicated to the mobile devices. As such smart data pricing is able to factor in the current congestion level in the MWN when computing the mobile data prices.
3. The development of new system architectures that encompasses both economic theory of pricing models, systems engineering and human-computer interaction (HCI) aspects. This leads to the development of smart data pricing schemes that not only take into consideration the price sensitivity of the users, but also the presentation of the smart prices to the users in graphical user interfaces (GUI) or USSD messages.

2.8.1 Smart Data Pricing Review

The main motive of smart data pricing is to create the right incentives for mobile data users so that they modify their usage behavior. This is done with the objective of improving MWN resource allocation and utilization. However, creating the right incentive requires the MNOs to account for the users response to the prices offered [20]. Three of the most important factors to consider in smart data pricing are:

1. The granularity of the changes in price i.e. how often do the prices change?
2. Balancing the user's reluctance to real-time pricing and the inability of flat-rate or usage-based pricing to exploit the time elasticity of demand of the different applications in peak periods.
3. Balancing the trade-off between the users need for transparency and the control over usage and the need for automation in dynamic pricing scenarios [21].

The various smart data pricing plans are as seen in the taxonomy shown in Figure 2-11 and further explained in detail in this section.

2.8.1.1 Real-time Pricing

In this scheme, MNOs monitor the MWN resource utilization and increases the price of data services when congestion is observed. Also, when under-utilization is observed, the price for mobile data is reduced. This scheme establishes a feedback control loop between the MNO and the users.

The user device can be automated to cleverly respond to the dynamic price in real time pricing [21]. For instance, Sen *et al* in [25] have developed an application for the user device, which allows users to manually optimize their data usage or engage an "autopilot" mode in scheduling the data usage.

2.8.1.2 Peak Load Pricing

Parris *et al* in [44] have presented a time-dependent pricing model for peak load pricing. They have proposed a simulation-based model where specified periods of time are classified as peak and off-peak hours. The model considers a higher arrival rate and a higher price during the peak hours as compared to the off-peak hours. Every user accessing the network is associated with

a specific elasticity that defines if their request during the peak hour will be deferred to an off-peak hour. A request with elasticity 0 cannot be deferred while an elasticity of 1 can be deferred. When a user makes an elastic request during a peak period and the request can be completed based on the user's willingness to pay, the user is admitted into the MWN and charged at the peak period per-packet price. If the user has an insufficient budget, they are deferred to an off-peak period where they are charged a low per-packet price

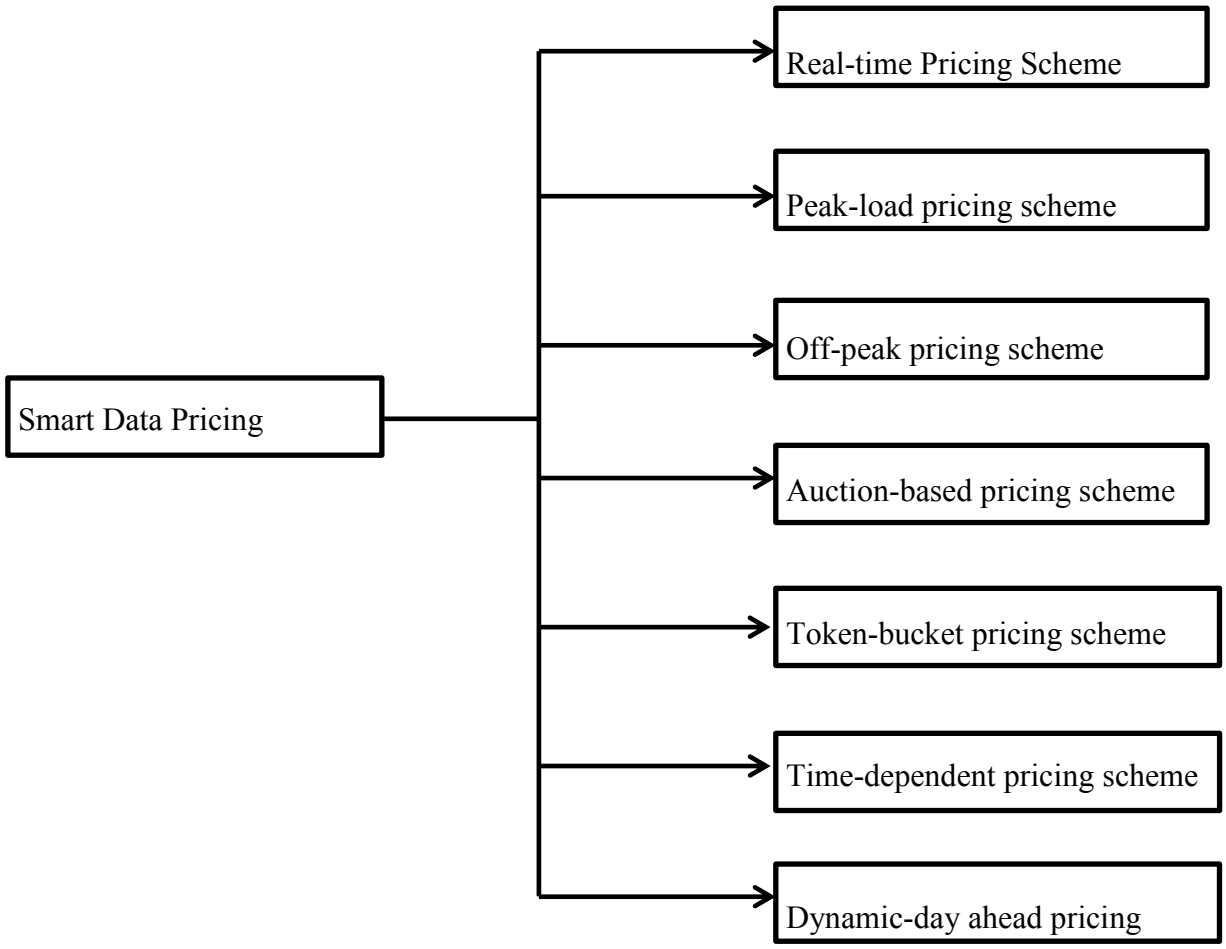


Figure 2-11: Smart data pricing schemes

This scheme is found to even out the demand for data services over time periods as well as generating higher revenues than non-peak load pricing. However, it segments users with low-income and low-elasticity, who are denied service. The segmentation hurts broadband adoption in

low-income groups.

2.8.1.3 Off-peak Price Discount

El-Sayed *et al* in [22] have studied off-peak price discount model for mobile data. In their model, discount incentives are advanced to users during the off-peak hours. This model is geared towards reducing the network peak load. A usage shift from the busy hour window to the off-peak window is noted. This usage shift implies savings in network capacity since it relieves cost pressure for network expansion with an adequate capacity build for the peak. Their model considers three parameters to quantify the impact of the off-peak discount on the network cost saving and revenue loss. These parameters are:

1. The duration of the off-peak window during which the off-peak incentives are offered.
2. Discount percentage per-megabyte during the off-peak hours.
3. Percentage of busy hour load that the users shift due to off-peak incentives, and its corresponding savings in network peak capacity demand.

2.8.1.4 Auction-based Pricing

Auction-based pricing enables users to explicitly specify their willingness to pay for handling packets from applications with different price elasticities [20]. Mackie-mason *et al* in [45] have presented a smart market pricing scheme for mobile data. The scheme is a closed control feedback loop where the price depends on the congestion level in the network. Each user will place a bid which is a reflection of their willingness to pay to send traffic in the network at the given time. A gateway will then admit the packets in a descending order of their bids while ensuring that the network performance is above the desired threshold. Users will be charged according to the minimum bid on the packet admitted to the network. Thus, a user admitted into the network will only pay for the cost of sending traffic through the network at the market clearing price.

2.8.1.5 Token Bucket Pricing

Token bucket pricing divides the day into peak and off-peak hours. Users pay a fixed monthly fee for internet access and receive a specific number of tokens which may be exchanged for service during peak hours. The service provider offers a high-quality service 1, which requires tokens and a normal quality service 2, which requires no tokens. Since tokens are limited, users

are incentivized to use service 1 when they desire a high utility. This increases the social welfare of the network since valuable resources get utilized for more valued services. This scheme requires client side automation to enable users to follow optimal policies for using their tokens.

2.8.1.6 Time-dependent Pricing

In time-dependent pricing, users are charged according to the time of day at which they consume mobile data. The prices change hourly. The users are incentivized to shift their usage to cheaper times of the day when the network is less congested which evens out the MWN resources demand throughout the day.

Time-dependent pricing does not explicitly price the MWN resources according to the network utilization. Instead, it relies on congestion estimates based on time of day and historical usage patterns. Also, the hours deemed as peak in time-dependent pricing are fixed. Therefore, traffic peaks arise in different parts of the network at different times, which can be hard to predict in advance. This ends up creating two peaks during the day; one during peak periods, for traffic that cannot wait for several hours for lower-price periods, and another peak during discounted off-peak periods for time-insensitive traffic. Such patterns have also been observed in dynamic pricing for voice calls in operational networks [46].

MTN Uganda and Uninor in India have implemented versions of time-dependent dynamic pricing. In their schemes, prices are updated hourly depending on the congestion conditions at the call's originating location [15].

2.8.1.7 Dynamic Day Ahead Pricing

Sangtae *et al* in [25] have presented a dynamic day ahead pricing scheme in the form of a system called time-dependent usage-based broadband price engineering (TUBE). Through the TUBE system, the authors in [25] have incorporated evolving user behavior, used user surveys to determine the system parameters and finally designed a supporting architecture. A nine-month trial consisting of 50 participants with iPads or iPhones has been done for AT&T subscribers in the USA. This trial showed that, when offered monetary discounts, users shifted their demand from peak to off-peak periods and even consumed more MWN resources during these off-peak periods

[27].

The architecture of the TUBE system uses a feedback control loop between the internet service provider (ISP) server which computes the prices offered to users, and the users who respond to the offered dynamic prices. The TUBE system feedback loop is shown in Figure 2-12.

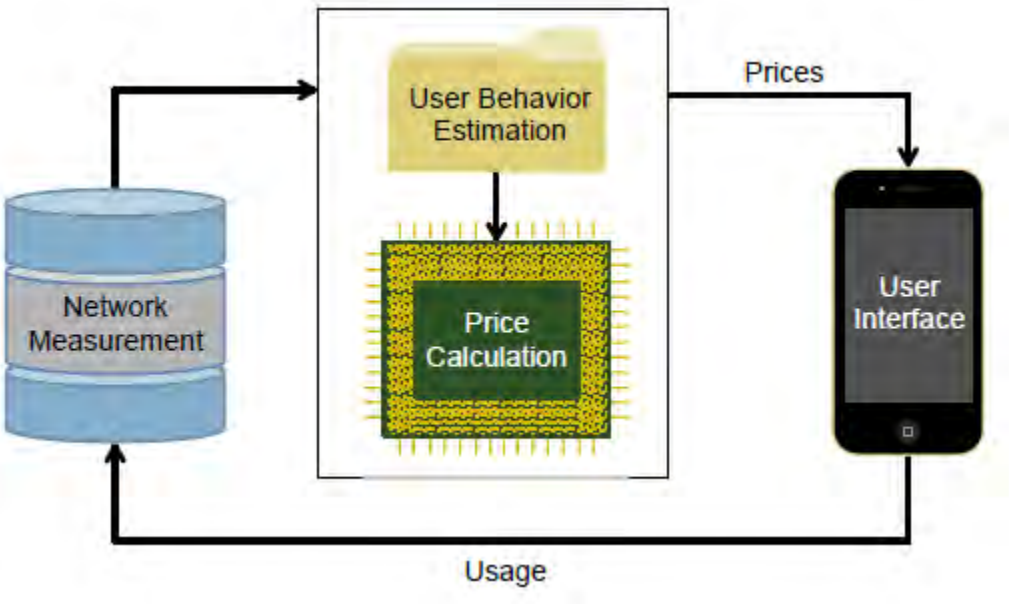


Figure 2-12: Dynamic day ahead pricing feedback loop [25]

The MNO performs network measurement and computes prices on a day ahead basis. The user device polls the prices from the MNO billing server at regular intervals and displays it in a GUI. Therefore, at any given time, users know the price for the next 24 hours. This system is beneficial in two ways; first, it enables users to plan ahead with certainty, secondly, it allows ISPs to adjust prices daily according to the updated user behavior estimates. Also, the advance notice of prices enables users to plan their bandwidth consumption over the next day if desired.

2.8.1.8 Application and content based pricing

This mode of pricing is analogous to certain types of tolls in road networks where drivers are charged different rates for different types of vehicles. In these road networks, trucks are charged a higher rate than small cars. In application-based pricing, charges depend on the application type. Thus, content and data plans are bundled together, for instance, Telus MNO in Canada offer bundled access to music and movie streaming [47]. Also, Orange UK offers users

the flexibility of choosing which media services to bundle [48].

Sponsored content is another form of application-based pricing. In sponsored content, MNOs subsidize certain types of data. Mobistar, a Belgium based MNO, offer free access to Facebook and Twitter but charge any other data service [49]. This is driven by technological advances such as deep packet inspection (DPI) which allow MNOs to differentiate traffic of some applications and offer sponsored content pricing [50].

2.9 Chapter Summary

This chapter has presented an overview of HWNs and the factors driving their adoption. The upsurge in MWN resource demand and evolution in wireless access technologies have been cited as the main factors motivating the uptake of HWNs by MNOs. However, deployment of HWN networks also comes with challenges. The main challenge identified is radio resource management across different RATs. Users should be guaranteed good QoS for all their service requests and access to different RATs within the HWN.

JCAC algorithms have been identified as key elements of radio resource management, controlling the admission of users into the network. These JCAC algorithms, on their own, have been found to be inefficient in controlling congestion during the peak hours and increasing utilization during the off-peak hours. This is because they only control the selection and subsequent admission of service requests into the various RATs within the HWN and do not incentivize users to change their usage behavior. As a result, JCAC algorithms have been integrated with dynamic pricing so as to control user behavior and maximize MWN resource utilization.

Mobile data service has been observed to be increasing over the years and gradually overtaking voice as the core service in MNOs offerings. This growth has further contributed to the increasing congestion in MWNs. Smart data pricing schemes have been proposed as a method of altering the user data consumption behavior. These smart data pricing schemes incentivize users to change their mobile data usage behaviors, especially shifting time insensitive traffic from the peak periods to off-peak periods and hence controlling congestion.

Chapter 3 Proposed Scheme

This chapter presents a source-destination based dynamic pricing (SDBDP) scheme that considers the congestion levels in both the call-originating and the call-destination cell to compute the dynamic price paid by the caller. The architecture of the scheme is presented. The SDBDP scheme integration to a heterogeneous wireless network is also discussed.

3.1 SDBDP Scheme Architecture

The SDBDP scheme is composed of the following five modules:

1. JCAC module.
2. Performance Management (PM) System
3. Dynamic Discount Module (DDM).
4. Network Billing System (NBS)
5. Dynamic price notification module.

The architecture of the SDBDP scheme is shown in Figure 3-1.

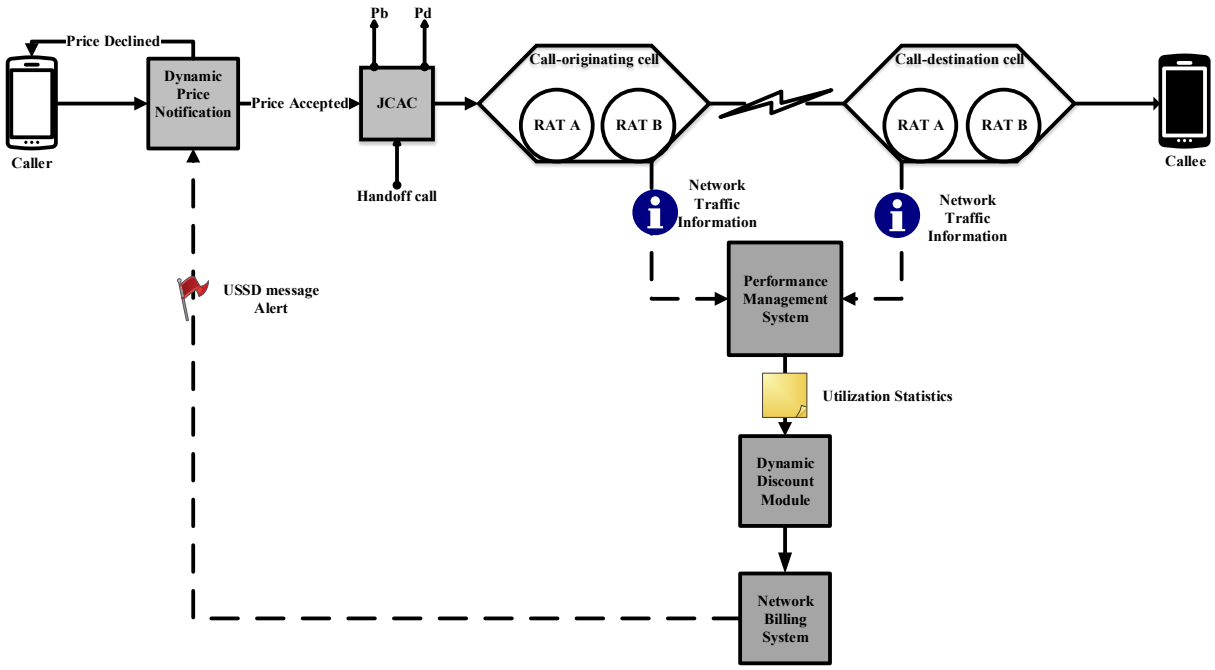


Figure 3-1: SDBDP scheme architecture

The functions of each of the five modules are described as follows:

3.1.1 JCAC Module

A load-based JCAC algorithm is used to admit both new and handoff calls to the least loaded RAT with sufficient bandwidth [10]. In the event of insufficient bandwidth in all the RATs, new calls will be blocked while handoff calls will be dropped. The JCAC algorithm also ensures that QoS of ongoing calls is not degraded by admitting calls based on RAT capacity [3] [10] [51] [52]. This is achieved by implementing bandwidth reservation and splitting of arrival rate functionalities.

3.1.1.1 Bandwidth Reservation

Handoff calls are given more priority than new calls since users are more sensitive to call blocking than call dropping. Also, depending on the QoS requirement, some classes of service require more bandwidth than others [3] [10]. The SDBDP scheme, therefore, reserves more bandwidth for handoff calls than for new calls, as well as high service classes with high bandwidth requirement. Table 3-1 shows the different traffic classes and a description of their bandwidth requirements [53].

Table 3-1: Traffic classes and their bandwidth requirement

Traffic Class	Description	Example Service
Conversational	Conversational pattern with very low delay and jitter. This is the most delay sensitive traffic class	VoIP and video conferencing
Streaming	Delay and jitter requirements not as strict as with conversational traffic class	Video on demand
Interactive	Enables prioritizing between packet data protocol (PDP) contexts, which allow end user service prioritizing. Interactive class is associated with a traffic handling priority (THP).	web browsing and telnet
Background	Best effort is acceptable for data delivery. This is the least sensitive traffic class	Email and file transfer protocol (ftp) services

Delay-sensitive traffic classes require more bandwidth than the other classes. Therefore, more bandwidth is reserved for conversational traffic class than the other classes. Assuming x classes of calls, bandwidth requirement for class i calls is denoted by B_i where $i=1, 2, 3....x$.

Considering that each cell has a specific capacity and both new and handoff calls share this capacity, we partition the incoming calls into *x* classes of service, indexed in a decreasing order of the bandwidth requirement as shown in Equation 3.1. Thus, class-*1* services require the largest bandwidth while class-*x* services require the least amount of bandwidth.

$$B_1 \geq B_2 \geq B_3 \ldots\ldots \geq B_x \tag{3.1}$$

In bandwidth reservation technique, the various classes of handoff and new calls are allocated different threshold bandwidth capacities. The threshold capacity for handoff calls is higher than that of new calls [3]. Similarly, the threshold capacity of class-*1* calls is higher than that of class-*2* calls. The threshold capacity of new and handoff class-*i* calls in RAT-*j* is denoted by T_{ij}^{n} and T_{ij}^{h} respectively. The threshold capacity comparison for different classes of service of both new and handoff calls is shown in Figure 3-2.

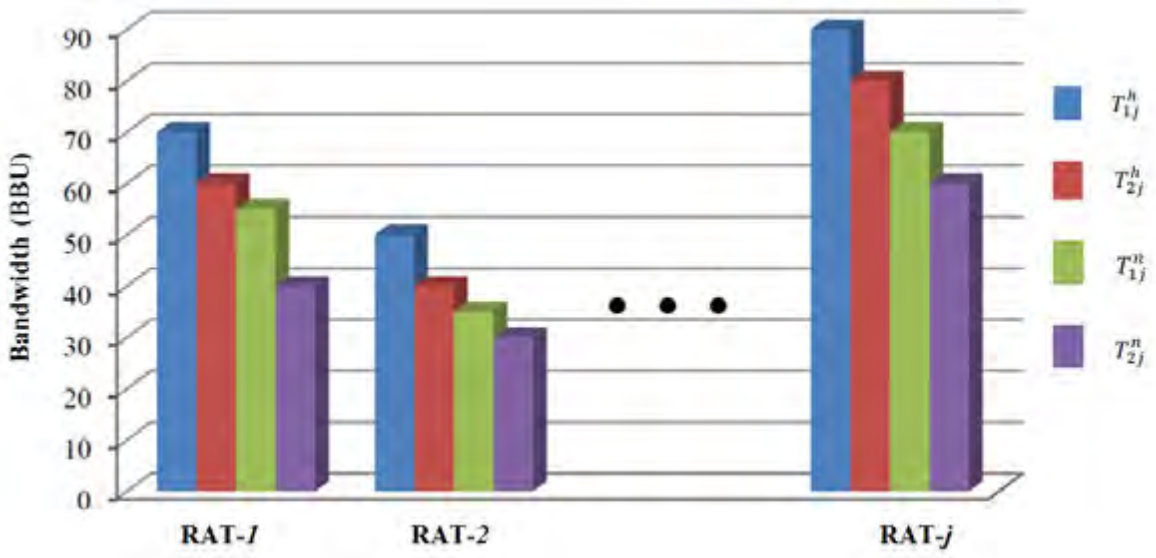


Figure 3-2 Threshold capacity comparison [4]

3.1.1.2 Splitting Arrival Rates

Let λ_i^n and λ_i^h denote the mean arrival rates for new and handoff class-*i* calls respectively into an HWN, with C_j being the capacity of RAT *j* and *y* the number of RATs. The load-based JCAC algorithm distributes new calls (λ_{ij}^n) as shown in Equation (3.2) and Figure 3-3

$$\lambda_{ij}^{n} = \frac{C_j}{\sum_{j=1}^{y} C_j} * \lambda_i^n \tag{3.2}$$

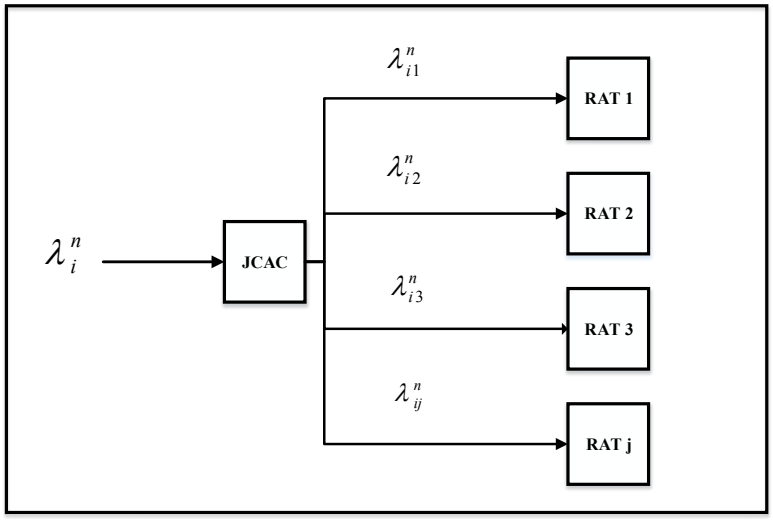


Figure 3-3: Splitting new calls into various RATs.

Handoff calls (λ_{ij}^{h}) are distributed as shown in Equation (3.3). This process of splitting the handoff calls arrival rates in to the different RATs is further shown in Figure 3-4.

$$\lambda_{ij}^{h} = \frac{C_j}{\sum_{j=1}^{y} C_j} * \lambda_i^h \tag{3.3}$$

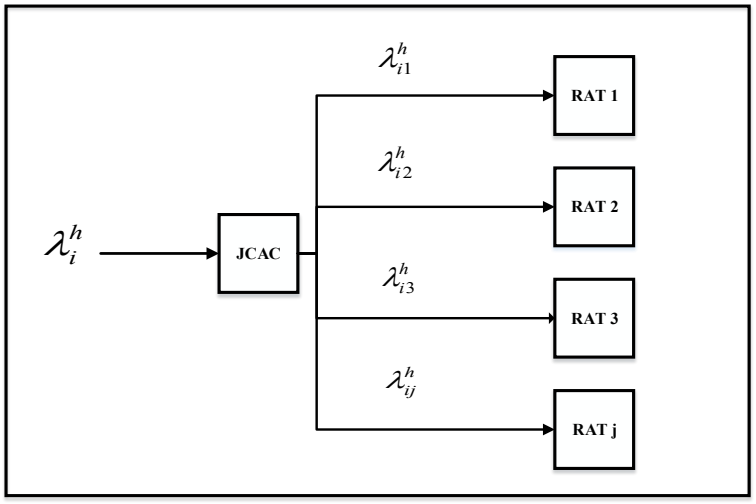


Figure 3-4: Splitting handoff-calls into various RATs.

3.1.2 Performance Management (PM) System

The PM system is usually part of operational support systems (OSS) and its key role is to keep track of the traffic passing through different network elements (NE) at any given instance. The traffic information helps the MNO to improve the performance and traffic management across the different domains in an MWN. The DDM module polls the cell utilization statistics from the PM system.

3.1.3 Dynamic Discount Monitoring (DDM) Module.

The DDM module receives MWN resources utilization data from the PM system. From the received network traffic in the call-originating and destination cells, the DDM module computes the congestion level for both cells. The computed congestion level is used to obtain the dynamic price to be paid by the caller.

Let *L* denote low congestion level, *M* denote medium congestion level and *H* denote high congestion level. The different congestion level scenarios for both the call-originating cell and the call-destination cell are illustrated in Table 3-2

Table 3-2: Illustration of congestion states in SDBDP scheme

		Destination cell		
		L	**M**	**H**
Originating cell	**L**	(1) LL	(2) LM	(3) LH
	M	(4) ML	(5) MM	(6) MH
	H	(7) HL	(8) HM	(9) HH

In scenario 1 (*LL*), a high discount will be advanced users to encourage usage of MWN resources in the call-originating and call-destination cells. For scenario 9 (*HH*), no discount will be advanced in order to reduce congestion by discouraging price-sensitive users from accessing the network. For scenarios 2 to 8, the discount advanced will be varying between the discount level

for scenarios 1 and 9. Users are offered a high discount when the network load in both the call-originating and the call-destination cell is low. When the network load in both the call-originating and the call-destination cell is high, little or no discount will be advanced to callers. When the network load in the call-originating cell is high and that of call-destination is low and vice-versa, a medium discount will be advanced to callers.

The SDBDP scheme encourages usage of MWN resources during the low network load periods while also discouraging usage under high load conditions. Therefore, utilization of networks resources will be increased during the off-peak period when the network is under-utilized and congestion will be reduced during the peak period when the network utilization is high.

Let W denote the average call service time, λ_i^n new class-i calls arrival rate and λ_i^h handoff class-i calls arrival rate. From the arrival rate of users into the network and the average service time, little's law is used to determine the number of users in the heterogeneous wireless network at any given time [54]. The average number of new class-i calls (n_i) in the system is given by Equation (3.4).

$$n_i = \lambda_i^n * W \tag{3.4}$$

Similarly, the average number of handoff class-i calls (h_i) is given by Equation (3.5).

$$h_i = \lambda_i^h * W \tag{3.5}$$

Let H_i denote the maximum price for class-i calls under SDBDP scheme, D_i the discount advanced to class-i calls, B_i the bandwidth required for a class-i call, P_i^n the dynamic price presented to callers initiating class-i calls, θ a scaling factor to set the discount level, Φ the destination cell congestion factor, n_{ij}^o and h_{ij}^o the ongoing new and handoff class-i calls respectively in RAT j for the call-originating cell. Let also n_{ij}^d and h_{ij}^d denote the ongoing new and handoff class-i calls respectively in RAT j for the call-destination cell, C_{total}^o and C_{total}^d the total capacity of the call-originating cell and call-destination cell respectively. The congestion level in both cells is then used to compute instantaneous dynamic discount as shown in Equation (3.6).

$$D_i = H_i * \theta_i * \left(\left(1 - \frac{\sum_{j=1}^{y} \left(n_{ij}^{o} + h_{ij}^{o} \right) * B_i}{C_{total}^{o}} \right) + \phi_i * \left(1 - \frac{\sum_{j=1}^{y} \left(n_{ij}^{d} + h_{ij}^{d} \right) * B_i}{C_{total}^{d}} \right) \right) \quad (3.6)$$

From the computed dynamic discount, the dynamic price is calculated as shown in Equation (3.7).

$$P_i^n = H_i - D_i \quad (3.7)$$

The operator defines the business rules to control the dynamic discount calculations in the DTM module. These rules include:

1. Setting the maximum and minimum discount for a particular class of service by adjusting the value of θ.
2. Setting the maximum price by adjusting the value of H_i .
3. Setting the call-destination cell congestion factor by adjusting the value of Φ.
4. Adjusting the cell capacities by configuring C_{total}^{o} and C_{total}^{d} .
5. Setting duration of display of the USSD notification.

Note: When $\Phi=0$, the call-destination cell congestion level is ignored in computing the dynamic price paid by the caller. This is equivalent to the SDP scheme, and its discount function is shown in Equation (3.8).

$$D_i = H_i * \theta_i * \left(1 - \frac{\sum_{j=1}^{y} \left(n_{ij}^{o} + h_{ij}^{o} \right) * B_i}{C_{total}^{o}} \right) \quad (3.8)$$

3.1.4 Network Billing System (NBS).

The NBS charges calls belonging to different classes of service according to the dynamic price computed in the DTM module. The DTM module updates the computed dynamic prices on the NBS at call set-up. The price remains constant for the entire duration of the call. The NBS sends an unstructured supplementary service data (USSD) message at call set-up to the caller informing them of the dynamic price to pay.

3.1.5 User Notification

The computed dynamic price is presented to a caller in the form of an interactive USSD notification. The duration of this notification is set from the DTM module as part of the business rules. A notification duration of 10 seconds is used for the SDBDP scheme. Included in this notification is:

1. The computed dynamic price for the incoming call.
2. An option to accept the presented dynamic price and continue with the call. This is also the default choice when no option is selected within the notification window.
3. Option to decline the offered dynamic price. This will happen when the caller is not willing to pay the presented dynamic price. The user is further given two choices:
    a. An option to be notified when the price falls below a specific amount P. The caller provides the value of P.
    b. An option to abandon the call completely.

3.2 Integration to Heterogeneous Wireless Network

The SDBDP scheme is integrated into different nodes of an HWN for it to perform optimally. Although the SDBDP scheme is not tied to any specific MWN technology, we are going to use an LTE technology to illustrate its applicability to modern day networks.

3.2.1 LTE Architecture

Figure 3-5 shows an LTE schematic with the key nodes and interfaces interconnecting them. These nodes are: eNodeB, mobility management entity (MME), serving gateway (S-GW), packet data network gateway (P-GW), home subscriber server (HSS) and policy and charging rules function (PCRF). The various main interfaces between different nodes shown are described as: Uu between the UE and the eNodeB, S1-MME between the eNodeB and the MME, S1-U between the eNodeB and the S-GW, S11 between MME and S-GW, S5/8 between S-GW and P-GW, Gx between the P-GW and PCRF and SGi connecting the P-GW to the internet.

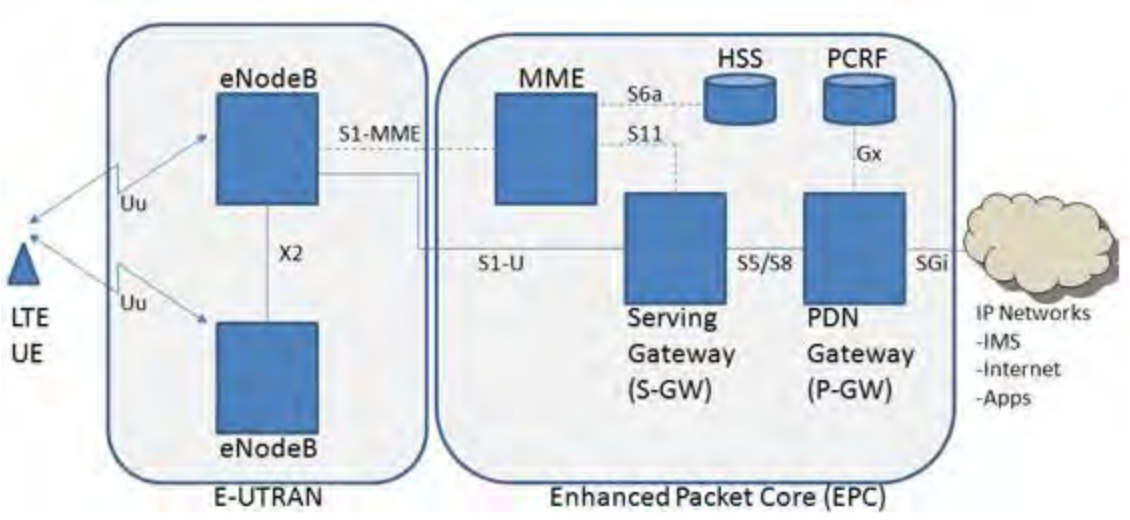


Figure 3-5: LTE Architecture

The key nodes which interface directly with the SDBDP in an LTE implementation are described next.

3.2.1.1 Mobility Management Entity (MME)

The MME supports idle user equipment (UE) location management. It tracks the area update process used in order for the MWN to be able to join UEs in case of incoming calls. The SDBDP scheme obtains the location of the callee from the MME. From the callee location, the DTE is able to fetch the call-destination cell utilization from the PM system

3.2.1.2 Home Subscriber Server (HSS)

The HSS holds dynamic information, for instance the identity of the MME to which a user is attached or registered. This information is crucial to the SDBDP scheme for identifying the call originating and destination cells from the relevant MME. The HSS is also responsible for:

1. User identification and addressing-The HSS stores the IMSI (International Mobile Subscriber Identity) and MSISDN (Mobile Subscriber ISDN Number) or mobile telephone number details.
2. User profile information-The HSS holds information related to the user subscriptions such as the QoS profile.
3. Generating security information and authentication- The HSS is integrated to the authentication center (AUC) which is responsible for generating vectors for

authentication and security keys. As a result, the HSS also provides mutual network-terminal authentication, radio path ciphering and integrity protection to ensure data and signaling transmitted between the network and the terminal is neither eavesdropped nor altered.

3.2.2 Integration of SDBDP to LTE Network

Figure 3-6 shows the positioning of the SDBDP scheme in a functional LTE network. The key integration points which are crucial for the successful operation of the scheme are shown and are numbered from 1 to 7.

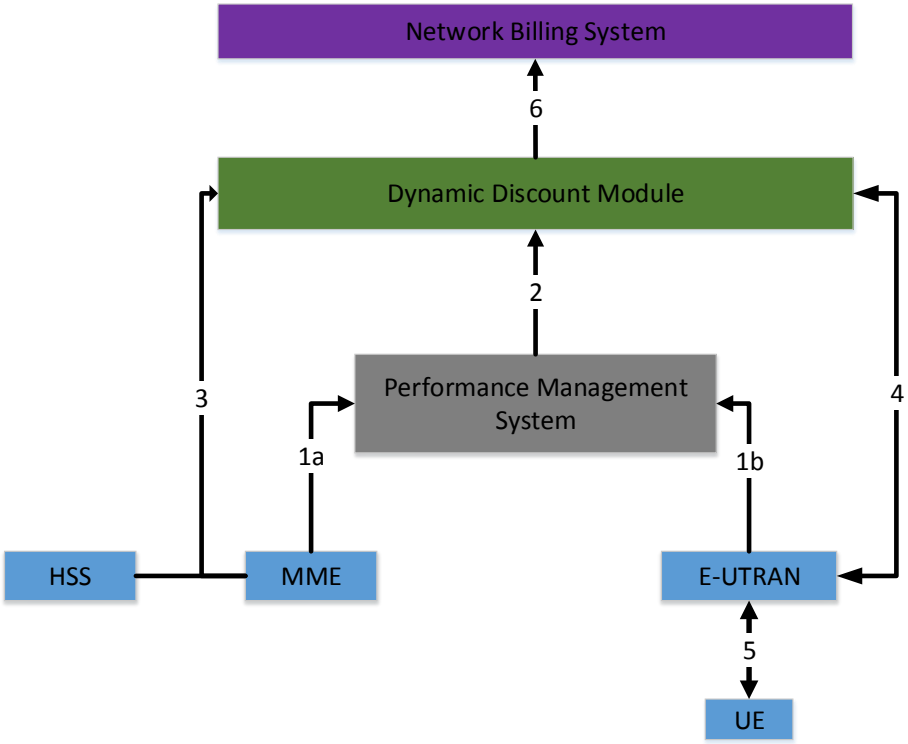


Figure 3-6: SDBDP Scheme Integration Points to an LTE Network

The integration interfaces are described as follows:

1. Interface 1a and 1b: This is integration of the PM system to the NEs for fetching of the traffic and utilization information.
2. Interface 2: This is integration of the DDM to the PM system. This integration allows for fetching of radio network traffic details for use in the computation of the dynamic price.

3. Interface 3: This is the integration of the DDM to the MME and HSS which enables the DDS to determine the call-originating and destination cells. Once the call-originating and destination cells are identified, DDM module fetches the corresponding network traffic from the PM system.
4. Interface 4: This interface allows the sending of computed dynamic price to the call-originating eNodeB for further forwarding to the UE. It also allows for sending of acceptance or rejection responses to the DDM module from the caller.
5. Interface 5: This is the air interface used to send the USSD notification from the eNodeB to the UE. The USSD notification is interactive allowing the user to accept or reject the offered dynamic price.
6. Interface 6: This interface links the DDM to the NBS for billing of services offered.

3.3 Network Evolution Support

There are constant changes in an MNO environment. Cells can be added, removed or re-parented and the existing cell capacity can be expanded or decreased. The SDBDP scheme factors these changes through:

1. Detection of new cells by the DDM module automatically once deployed from the statistics reported by the PM system. The PM system discovers any changes and maintains an up to date view of all the radio resources in the network.
2. Detection of changes in the cell capacity (capacity expansion or capacity decrease) by the DDM module based on the statistics received from the PM system.

3.4 Call Flow Path

The caller is the user initiating a call while a callee is the user receiving the call. Calls are usually charged at the call set-up, thus, a caller pays the total cost of a call. The caller will utilize MWN resources from the call-originating cell while a callee will utilize MWN resources from the call-destination cell.

3.4.1 New Call

The flow chart in figure 3-5 shows the flow for both new and handoff calls. Once a caller initiates a new call, the DTM module gathers the performance statistics from the RAN. The performance statistics gathered from the network are used to compute the congestion levels in the call-originating and the call-destination cells. Depending on the computed congestion levels, a dynamic price is calculated and presented to the caller through a USSD notification message.

The USSD message is interactive and prompts the caller to either "accept" or "decline" the presented dynamic price. When the caller selects "accept", the JCAC module is invoked to check if there is sufficient bandwidth to support the call from any of the available RATs. The default choice is "accept" if no option is selected within the notification duration set from the business rules in the DTM module. The incoming call is admitted into the network if there is sufficient bandwidth in one of the RATs, and the call is completed when the callee receives it. The NBS module starts billing once the call set-up is complete. The incoming call is blocked if there is no sufficient bandwidth in any of the MWN's RATs to support it.

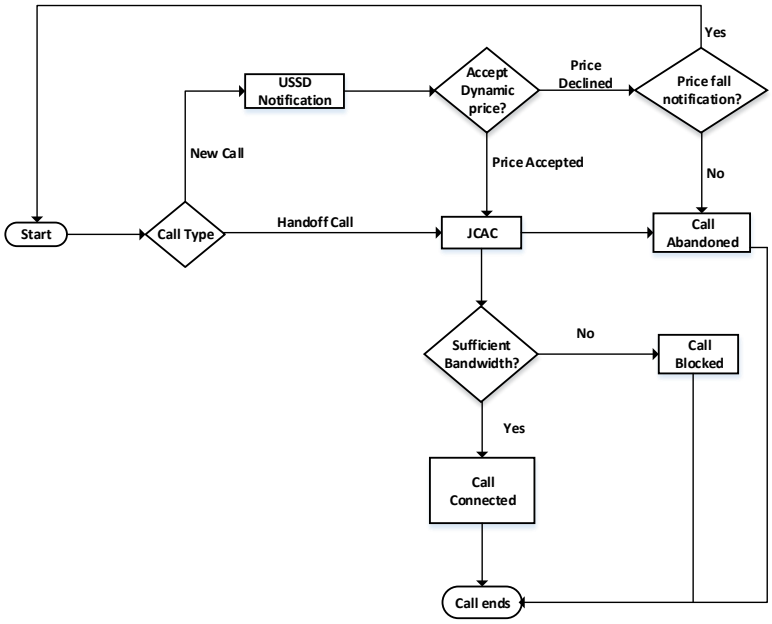


Figure 3-5: Call Flow Chart

When the caller selects “decline call”, they are further prompted to enter an amount they would like to pay so that they will be notified when the dynamic price falls below the stated amount. Otherwise, the user “completely abandons” the call. The amount entered by the user is then saved in the NBS. When the dynamic price falls below the entered amount, the NBS sends a notification to the user. The call attempt is terminated immediately when the user selects “abandon” the call.

3.4.2 Handoff Call

The bandwidth reservation technique gives more priority to handoff calls than new calls. The JCAC algorithm checks if there is sufficient bandwidth in any of the co-located cells to support the class *i* handoff call. The handoff call is admitted into the network if there is sufficient bandwidth available in one of the MWN’s RATs. When there is insufficient bandwidth in any of the MWN’s RATs, the handoff call is dropped. The handoff call flow is also illustrated in Figure 3-5.

3.5 Chapter Summary

This chapter has presented the architecture of the SDBDP scheme and discussed the building blocks of the scheme namely: DTM module, JCAC module, dynamic price notification module and the NBS. The operation of each of the module has been examined in detail. The DTM module monitors the congestion level in the RAN. At call setup, the DTM module computes the congestion level in both the call-originating and the call-destination cells. Based on the congestion level in both cells, the DTM module further computes the dynamic price for the incoming new call and presents this to the NBS. The NBS is responsible for billing usage of the MWN resources and sending USSD notification to the callers at call set-up.

Integration of the SDBDP scheme to the HWN has also been discussed. The DTM module obtains statistical data from the call-originating and the call destination cells in the RAN, from which it computes the dynamic price. All the call details are recorded in the CDR system. The CDR system generates reports, important in setting business rules in the DTM module so as to maximize resource utilization.

Chapter 4 Analytical System Model and Assumptions

This chapter presents the analytical model applied in this work and the performance metrics adopted in studying both SDP and SDBDP schemes in an HWN. We present a Markov decision chain based on M/M/m/m queuing model to evaluate the performance of both SDP and SDBDP schemes. These two schemes are integrated to a load-based JCAC algorithm.

4.1 Heterogeneous Network

A heterogeneous wireless network consisting of 2 RATs with collocated cells is considered. Figure 4-1 shows a heterogeneous network composed 2 groups of co-located cells, with each group having 2 RATs.

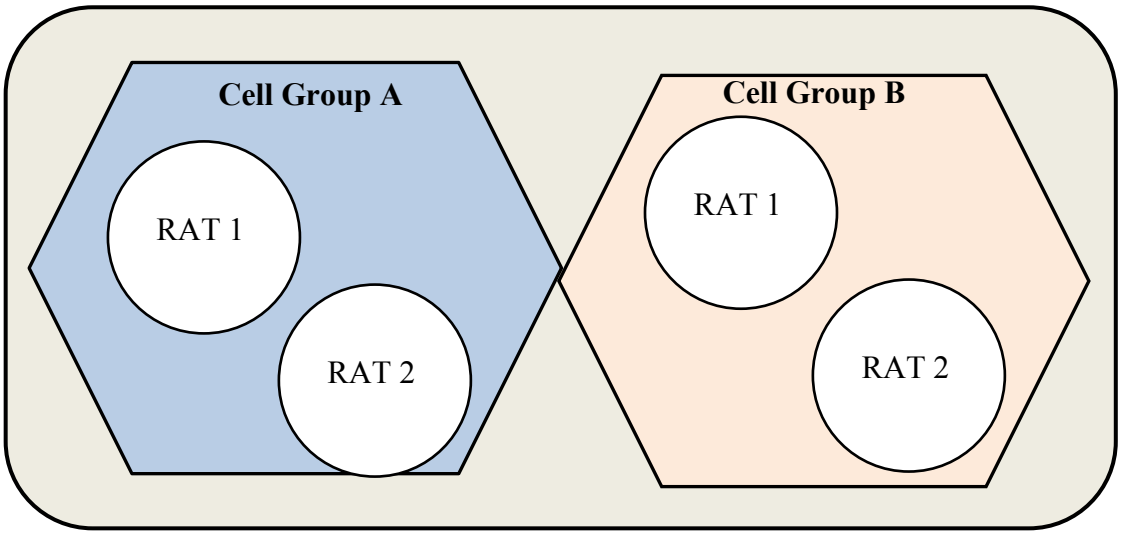


Figure 4-1: Heterogeneous wireless network

Each cell of RAT *j (j=1, 2)* has a maximum amount of radio resources, which is its capacity, denoted by C_j. The radio resources could be frequency channels, code sequence or timeslots depending on the multiple access technology implemented in the air interface. However, these radio resources can always be interpreted in terms of effective or equivalent bandwidth. In our model, we have used basic bandwidth unit (bbu) to represent the various radio resources. The total

capacity in a given group of collocated cells of the heterogeneous network, C_{total} is given by Equations (4.1) and (4.2).

$$C_{total} = C_1 + C_2 \tag{4.1}$$

$$C_{total} = \sum_{j=1}^{2} C_j \tag{4.2}$$

4.2 Markov Decision Chain

A Markov decision chain consists of a set of states and labelled transitions between the states. The states summarize the effects of the past occurrences on the future while the transitions represent the probability of changing from one state to the next [55]. It is therefore ideal in the design of analytical systems which have changing states over time [56]. The following assumptions are made [55] [56]:

1. If the current state is *i*, the time until the next transition is exponentially distributed with a given parameter *t*, independent of the past history of the process and of the next state.
2. If the current state is *i*, the next state will be *j* with a given probability P_{ij}, independent of the past history of the process and of the time until the next transition.

The key attributes of markov decision chain employed in modelling the SDP and SDBDP schemes are the average service time and traffic intensity.

4.2.1 Average Service Time

The average service time is also referred to as the call residence time (CRT). It is related to the duration of a call and is assumed to follow an exponential distribution [57]. The average service time (*t*) is given by Equation 4.3 where μ denotes the mean service rate.

$$t = \frac{1}{\mu} \tag{4.3}$$

4.2.2 Traffic Intensity

The traffic intensity, denoted by ρ, is the ratio of the mean arrival rate to the mean service rate. This is given by Equation (4.4), where λ denotes the mean arrival rate while μ denotes the

mean service rate. The traffic intensity is an indication of user demand for MWN resources. Traffic intensity is also referred to as utilization factor and is measured in Erlangs.

$$\rho = \frac{\lambda}{\mu} \tag{4.4}$$

From Equation (4.4), the traffic intensity for new class-*i* calls is given by Equation (4.5).

$$\rho_{ij}^{n} = \frac{\lambda_{ij}^{n}}{\mu_i} \tag{4.5}$$

Similarly, the traffic intensity of class-*i* handoff calls is given by Equation (4.6)

$$\rho_{ij}^{h} = \frac{\lambda_{ij}^{h}}{\mu_i} \tag{4.6}$$

4.3 *M/M/m/m* Queuing Model

According to the M/M/m/m queue, the mean arrival rate λ follows a Poisson distribution while the mean service rate μ follows an exponential distribution [10] [51]. The M/M/m/m queuing model is used to model call arrival in telecommunication systems [58] [59]. There are *y* RATs, and users who arrive when all RATs are busy leave without being served. New calls are blocked while handoff calls are dropped [60]. This queue is also called Blocked Customers Cleared (BCC) queue, because new users arriving when there are no free RATs are cleared [61].

4.3.1 One Dimensional M/M/m/m Queuing Model

Let *C* denote the total capacity of the heterogeneous network, *R* the bandwidth reserved for handoff calls, λ_n the arrival rate for new calls, λ_h the arrival rate for handoff calls and μ_c the average service time for both new and handoff calls. Figure 4-2 shows the state space diagram for one dimensional M/M/m/m queuing model with bandwidth reservation for handoff calls [52].

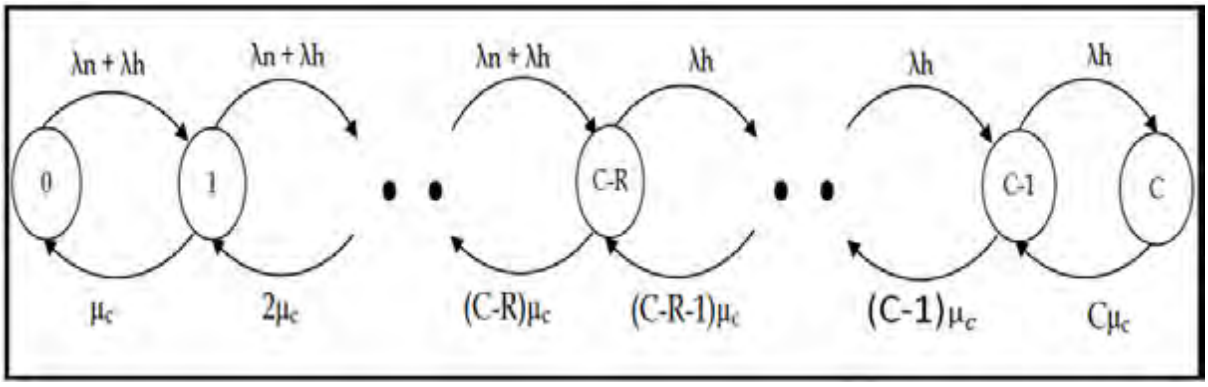


Figure 4-2: One-dimensional *M/M/m/m* [52]

4.3.2 Multi-dimensional Markov Model

A multi-dimensional Markov model supports different mean service times for new and handoff calls. It is therefore preferred over the one-dimensional Markov model in the evaluation of the SDBDP scheme integrated with a load-based JCAC [10] [51]. The state space diagram for the multi-dimensional Markov model is shown in Figure 4-3.

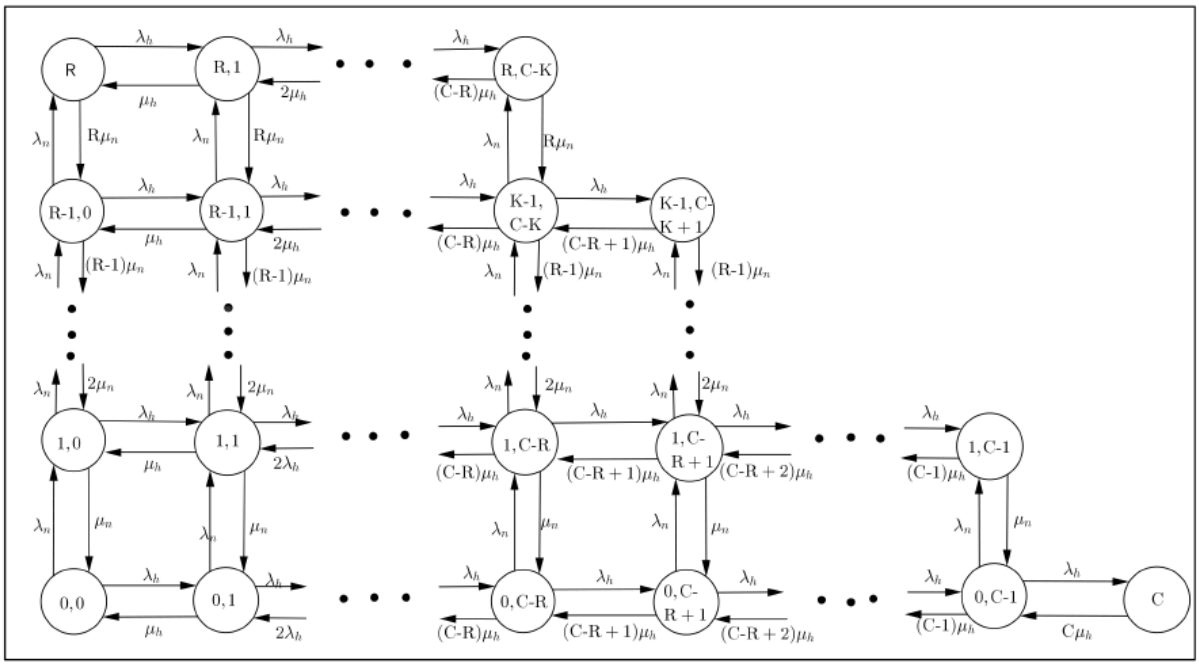


Figure 4-3: Two-dimensional Markov model [51] [52]**.**

4.4 Number of Users in the Heterogeneous Network

The number of users in the network, both new and handoff, is obtained using little's law. Little's law shows the relationship between the average number of users in a queuing system *N*, the average arrival rate of users into the system λ and the mean service time *t* [57]. This relationship is shown in Equation (4.7).

$$N = \lambda * t \tag{4.7}$$

Let n_{ij} denote the number of new class *i* calls in RAT *j* and h_{ij} the number of handoff class *i* calls in RAT *j*. The state space of all possible values of an HWN, with *x* classes of service and *y* RATs is given by vector Ω, as shown in Equation 4.8

$$\Omega = (n_{ij}, h_{ij} : i = 1,2.....x, j = 1,2...y) \tag{4.8}$$

The state space of all admissible states in the HWN is denoted by *S* and is given by Equation 4.9. T_{ij}^{h} denotes the threshold bandwidth capacity for handoff calls while T_{ij}^{n} denotes the threshold bandwidth capacity for new calls. B_i denotes the bandwidth required for a class-*i* call.

$$S \in \Omega \mid \sum_{j=1}^{y}\sum_{i=1}^{x}(n_{ij} * B_i) \le T_{ij}^{n} \wedge \sum_{j=1}^{y}\sum_{i=1}^{x}(h_{ij} * B_i) \le T_{ij}^{h} \wedge \sum_{j=1}^{y}\sum_{i=1}^{x}\left((n_{ij} * B_i) + (h_{ij} * B_i)\right) \le C_j \tag{4.9}$$

An admissible state is a combination of the number of users (both new and handoff calls of different service classes), that an HWN can support while maintaining the required QoS and meeting all the resource constraints.

The steady state probability P_k of the system in state *k* is given by Equation (4.10).

$$P_k = \prod_{j=1}^{y}\prod_{i=1}^{x}\frac{(\rho_{ij}^{n})^{n_{ij}}}{n_{ij}!} * \frac{(\rho_{ij}^{h})^{h_{ij}}}{h_{ij}!} * P_o \qquad k \in S \tag{4.10}$$

Where P_o is normalization constant given by Equation (4.11).

$$P_o = \sum_{k \in S}\prod_{j=1}^{y}\prod_{i=1}^{x}\frac{(\rho_{ij}^{n})^{n_{ij}}}{n_{ij}!} * \frac{(\rho_{ij}^{h})^{h_{ij}}}{h_{ij}!} \tag{4.11}$$

4.5 Performance Metrics

We consider the call-blocking probability, call-dropping probability, and the MWN resource utilization to evaluate the performance of an HWN under different caller-callee distribution. These performance metrics are discussed next:

4.5.1 New Call-Blocking Probability

A new class-*i* call will be blocked when either of the following two conditions is met:

a) If there is no sufficient bandwidth available in any group of co-located cells to meet the QoS requirements of a particular class of service. This condition is shown in Equation (4.12)

$$\left(B_i + \sum_{i=1}^{x} \left(\left(n_{ij} + h_{ij} \right) * B_i \right) \right) > C_j \tag{4.12}$$

b) If the threshold capacity for the new class-*i* call is exceeded as described by Equation (4.13).

$$\left(B_i + \sum_{i=1}^{x} \left(\left(n_{ij} \right) * B_i \right) \right) > T_{ij}^{n} \tag{4.13}$$

The set of states S_b for which a new call is blocked is given by Equation (4.14).

$$S_b = k \in S \mid \left(\left(B_i + \sum_{i=1}^{x} \left(\left(n_{ij} + h_{ij} \right) * B_i \right) \right) > C_j \vee \left(B_i + \sum_{i=1}^{x} \left(\left(n_{ij} \right) * B_i \right) \right) > T_{ij}^{n} \right) \quad \forall j \tag{4.14}$$

The new class-*i* call-blocking probability Pb_i, is then obtained as shown in Equation (4.15).

$$Pb_i = \sum_{k \in S_b} P_k \tag{4.15}$$

4.5.2 Handoff Call-Dropping Probability

A handoff class-*i* call will be dropped when either of the following two conditions is met:

a) When the capacity of the heterogeneous network is exceeded as shown in Equation (4.16).

$$\left(B_i + \sum_{i=1}^{x} \left(\left(n_{ij} + h_{ij} \right) * B_i \right) \right) > C_j \tag{4.16}$$

b) When the threshold capacity for the handoff call of a particular class of service is exceeded as described in Equation (4.17)

$$\left(B_i + \sum_{i=1}^{x} \left(\left(h_{ij} \right) * B_i \right) \right) > T_{ij}^{h} \tag{4.17}$$

The set of states S_d for which a handoff call is dropped is given by Equation (4.18).

$$S_d = k \in S \mid \left(\left(B_i + \sum_{i=1}^{x} \left((n_{ij} + h_{ij}) * B_i \right) \right) > C_j \vee \left(B_i + \sum_{i=1}^{x} \left((h_{ij}) * B_i \right) \right) > T_{ij}^{h} \right) \quad \forall j \tag{4.18}$$

The class-*i* call-dropping probability Pd_i, is then obtained as shown in Equation (4.19).

$$Pd_i = \sum_{k \in S_d} P_k \tag{4.19}$$

4.5.3 Average System utilization

The average utilization (*U)* of a group of co-located cells can be obtained by taking the ratio of the MWN resources in usage to the total capacity of the group. This is as shown in Equation (4.20).

$$U = \frac{\sum_{j}^{y} \sum_{i}^{x} \left(\lambda_{ij}^{n} * t_i * B_i + \lambda_{ij}^{h} * t_i * B_i \right)}{C_{total}} \tag{4.20}$$

4.6 Demand Model

Let ρ_i^n denote the mean traffic intensity of class-*i* calls when the current dynamic price is P_i^c, α_i^n the demand shift factor and β_i the price elasticity of demand of new class-*i* calls. The effect of price on demand for MWN resources is shown in Equation (4.21) [34] [62].

$$\rho_i^n = \alpha_i^n * e^{(-\beta_i * p_i^c)} \quad \forall i \tag{4.21}$$

The different parameters in the demand model are explained below:

4.6.1 Demand Shift Factor (α)

The demand shift factor shows how the demand curve position changes with non-price determinants such as income, user preferences, and time of a day. The demand for MWN resources varies at different times of the day [62]. For instance, the demand for voice MWN resources at 1 am is very different from that at 2 pm. At 1 am, a majority of the people may be sleeping and the probability of making or receiving voice calls is very low. On the other hand, the probability of making or receiving voice calls is higher at 2 pm since a majority of users are awake at this time.

It is important to note that time of a day is not the only factor that affects demand shift of MWN resources, but one among the many.

Figure 4-4 shows the demand curve position changing from D_1 to D_2, and the corresponding change in demand shift factor Δα.

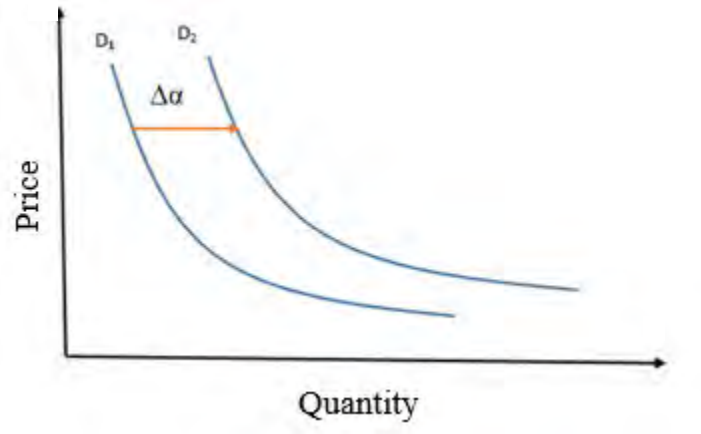


Figure 4-4: Demand shift factor illustration

From the demand distribution of MWN resources obtained from an MNO in [62], we have derived α for 24 hours in a typical day as shown in Table 4-1. Therefore, α varies at different hours of the day to reflect the varying demand for resources [17].

Table 4-1: Demand shift factor (α) for a 24 hour period

Demand Shift Factor	Values
α(t): t=0,1,2…23	[1.2,1.2,1.2,1.2,1.2,1.2,4,8,12,20,24,32,24,18,24,24,28,20,16,12,8,2,2,1.2]

4.6.2 Price Elasticity of Demand (β)

The price elasticity of demand (β), is a measure of the relationship between change in demand of MWN resources, of a particular class of service, and the corresponding change in price. It is always a negative value because of the inverse relationship between price and demand as per the law of demand [63]. Let P_i and Q_i denote the price and demand for class-i calls respectively. The price elasticity of demand is illustrated in Figure 4-5 and shown in Equation (4.22).

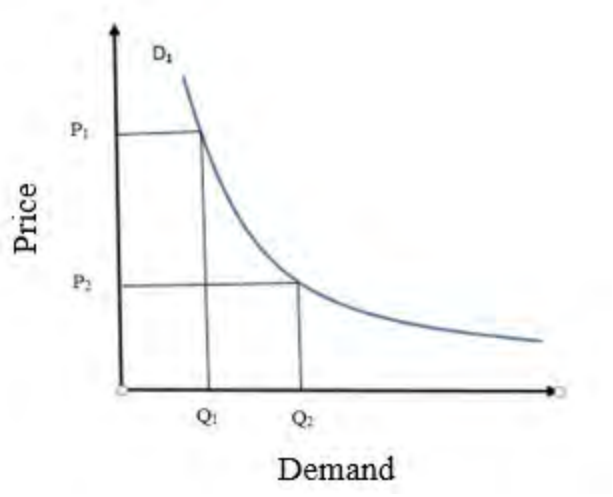


Figure 4-5: Illustration of price elasticity of demand

$$\beta_i^n = \frac{\Delta Q_i / Q_i}{\Delta p_i / p_i} = \frac{(Q_2 - Q_1)/Q_2}{(p_2 - p_1)/p_2} \tag{4.22}$$

4.7 Chapter Summary

This chapter has presented an analytical model of the current SDP schemes and the developed SDBDP scheme using a multi-dimensional Markov chain based on the *M/M/m/m* queuing model. The average service time and the traffic intensity attributes of the M/M/m/m queuing model are used. The metrics used to evaluate the performance of the schemes are call-blocking probability, call-dropping probability, and the MWN resource utilization. An exponential demand model has been presented to evaulate the response of users to dynamic pricing. Key components of the demand model adopted are demand shift constant and price elasticity of demand. The demand shift constant shows the change in the demand as a result of non-price determinants such as the time of day. The price elasticity of demand shows the relationship between the change in price of a call and the corresponding change in demand.

Chapter 5 Numerical Example and Results

A numerical example and simulation results are presented in this chapter. The performance metrics used to evaluate the current SDP and the developed SDBDP schemes are:

1. Call blocking probability.
2. Call dropping probability.
3. MWN resource utilization.

The call blocking and the call dropping probabilities are used to evaluate the congestion level in the MWN during the peak hours. During the off-peak hours, the resource utilization metric is used to evaluate the improvement in MWN resource utilization due to the application of dynamic pricing.

5.1 Evaluation Scenario

The evaluation scenario consists of four groups of co-located cells, as shown in figure 5-1. Each cell group is heterogeneous in nature and has two RATs. The two RATs have different capacity, measured in basic bandwidth units (BBUs). RAT1 has a capacity of 50 BBUs and RAT2 has a capacity of 40 BBUs. We consider call group A as our cell of interest. The call-blocking probability, call-dropping probability, and the MWN resource utilization performance metrics will be evaluated for the cell group of interest, under different caller-callee distributions.

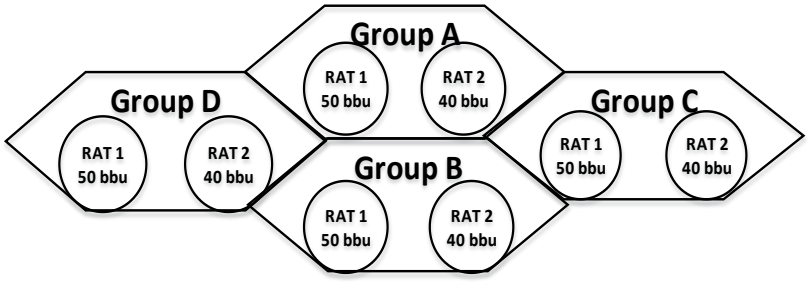


Figure 5-1: Evaluation scenario

5.2 Traffic in a Typical Day

User traffic in a typical day is shown in Figure 5-2, where a day is divided into 24 hours [10]. The various hours of the day are assigned different arrival rates based on the arrival rates in

a typical business day. The user traffic in Figure 5-2 shows the traffic in a network before dynamic pricing is implemented. When dynamic pricing is implemented the user traffic varies depending on the prices [1].

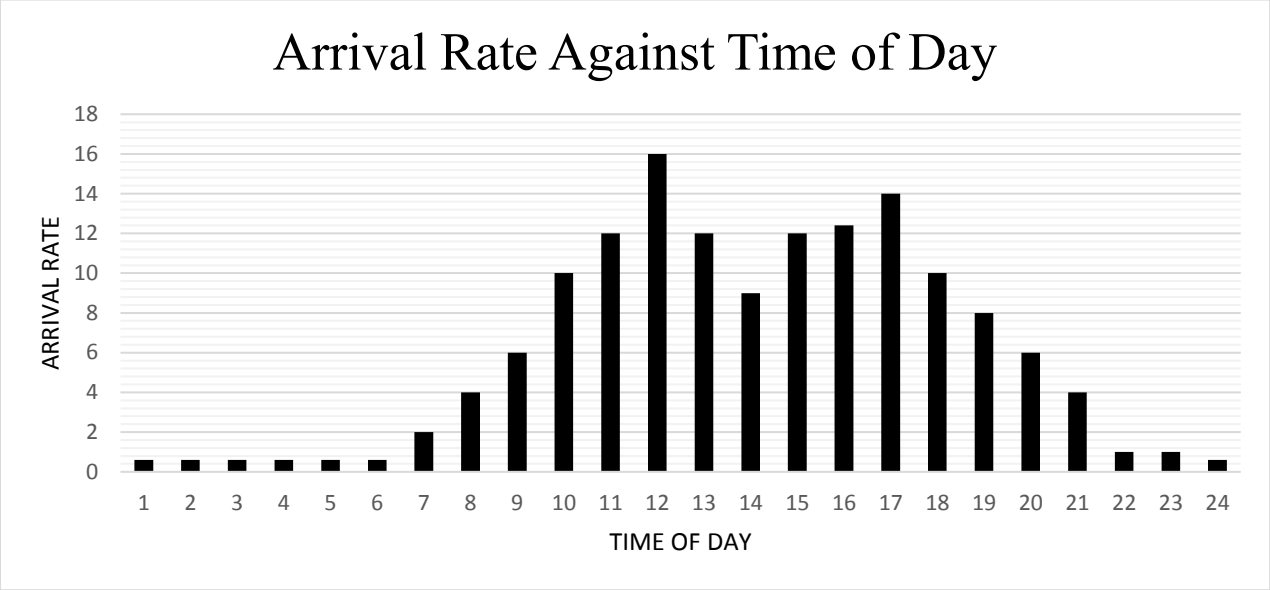


Figure 5-2: User traffic in a typical Day

From Figure 5-2, it is observed that peak period (maximum call arrival rate) occur at 1200 hrs and 1700 hrs. Congestion in a heterogeneous network is usually experienced during the peak periods [2]. Off-peak periods, when the network is under-utilized, are observed in the early morning hours (0000 hrs-0700 hrs) and at late night (2100 hrs-2300 hrs).

5.3 System Parameters

Two RATs and two classes of calls are also considered. Therefore, $x=2$ and $y=2$. The other system parameters used are shown in Table 5-1. It was noted that the trends for class 1 and 2 calls were similar albeit with different values for call blocking probability, call dropping probability and MWN resource utilization. Therefore, the results for one class (class 1) calls are presented.

Table 5-1: System parameters used in the simulation

Parameter	Value(s)
α(t): t=0,1…23	[1.2,1.2,1.2,1.2,1.2,1.2,4,8,12,20,24,32,24,18,24,24,28,20,16,12,8,2,2,1.2]
β	-1
C_j	$C_1 = 50,\ C_2 = 40$
B_i	$B_1 = 2,\ B_2 = 1$
T_{ij}	$T_{11}^n = 40, T_{12}^n = 30, T_{21}^n = 35, T_{22}^n = 30, T_{11}^h = 50, T_{12}^h = 40, T_{21}^h = 45, T_{22}^h = 40$
μ_i	$\mu_1^n = 1, \mu_2^n = 1, \mu_1^h = 1, \mu_2^h = 1$
θ	0.5
Φ	[0, 0.25, 0.5, 0.75, 1]
P_i^n	$P_1^n(0000hrs) = 0.10H, P_2^n(0000hrs) = 0.05H$

5.4 SDP Scheme.

The SDP scheme is obtained when ϕ=0 in Equation (3.5). In this scheme, only the call-originating cell congestion level is considered in computing the dynamic price paid by the caller. Figures 5-3 and 5-4 show the traffic intensity in this scheme under different caller-callee distributions.

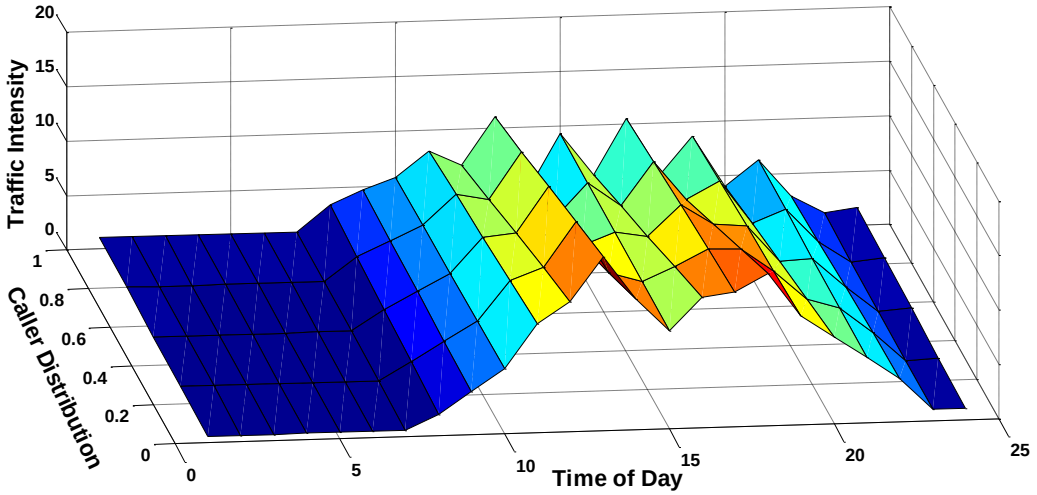


Figure 5-3: Traffic intensity under different caller-callee distribution, SDP scheme (3D)

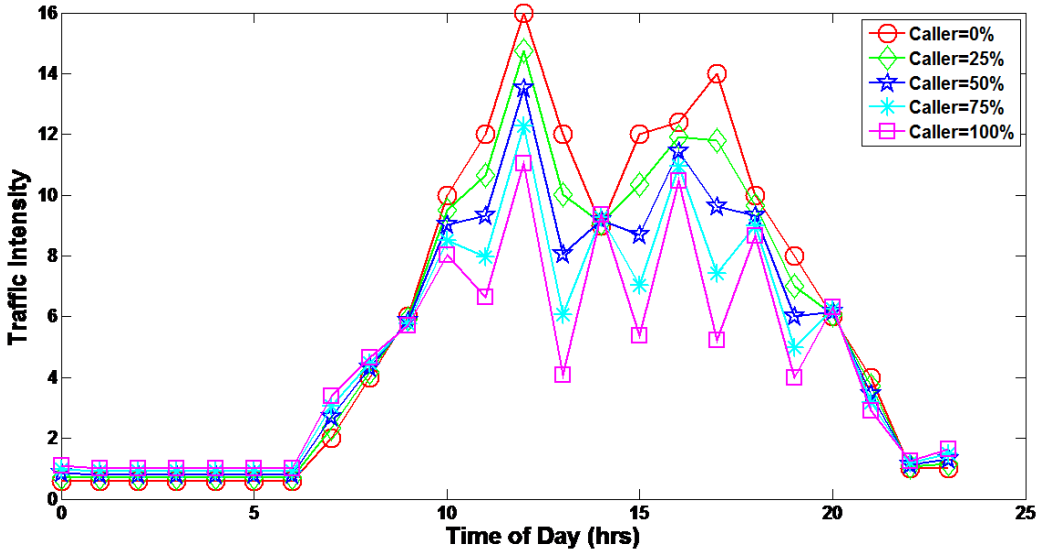


Figure 5-4: Traffic intensity under different caller- callee distribution, SDP scheme

During the peak period, maximum traffic intensity is observed with a caller distribution of 0% (100% callee distribution). Callees do not pay to receive calls and hence dynamic pricing does not discourage them from receiving calls during the congested peak periods. The traffic intensity decreases with increase in caller distribution and a minimum value is obtained with 100% caller distribution. Callers pay to make calls hence dynamic pricing discourages usage during the congested peak hours when the dynamic price is high, thereby reducing the traffic intensity.

During off-peak hours, the traffic intensity is minimum with a caller distribution of 0% and maximum for a caller distribution of 100%. Application of SDP scheme to an under-utilized cell entices prospective callers to make calls but has no effect on prospective callees. Callees do not receive an incentive to increase the utilization of MWN resources as they do not pay for the calls.

5.4.1 Call-Blocking and call-dropping Probability

Figures 5-5 and 5-6 show the call-blocking probability and the call-dropping probability for different hours of the day.

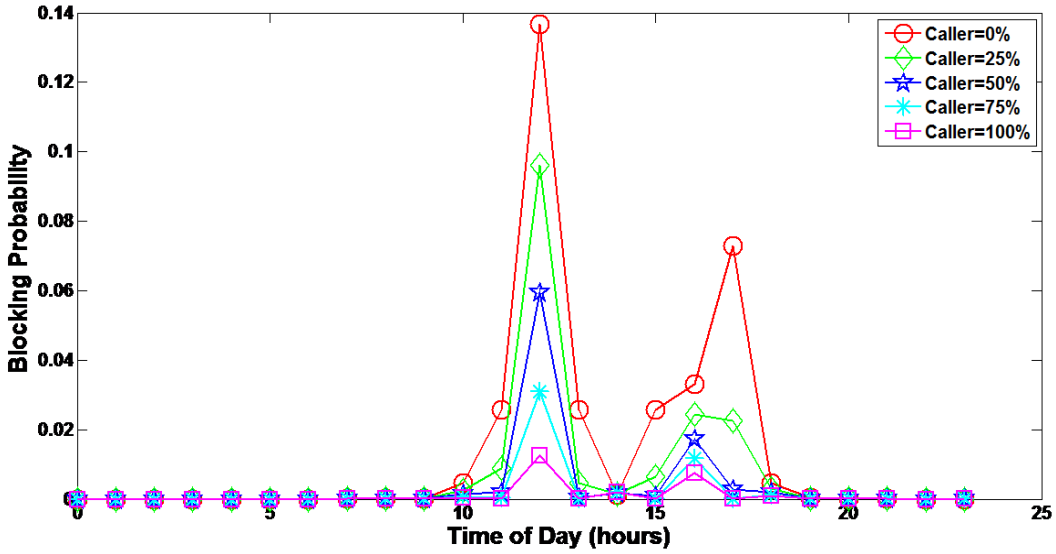


Figure 5-5: Call-blocking probability under different caller-callee distribution, SDP scheme

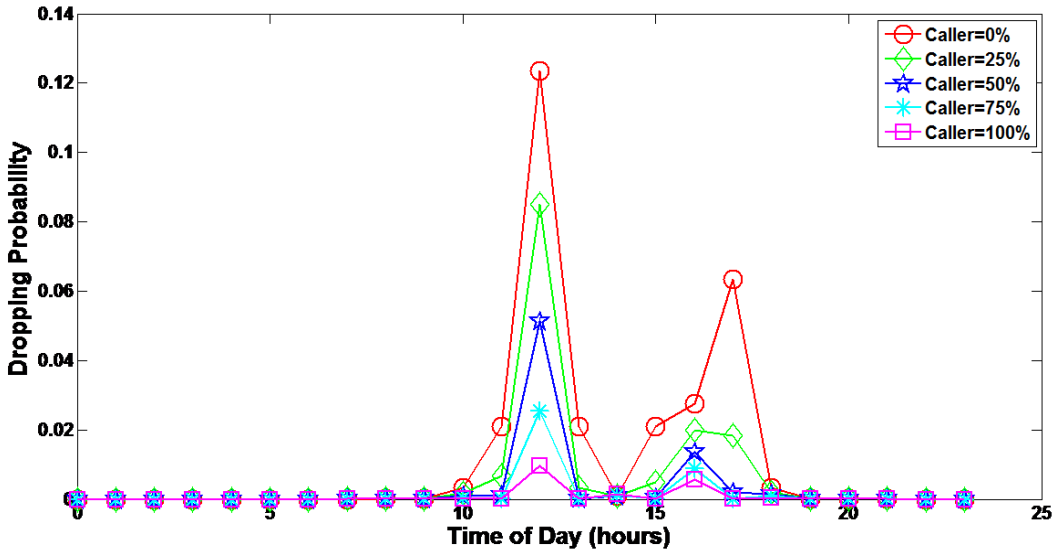


Figure 5-6: Call-dropping under different caller-callee distribution, SDP scheme

At the peak hour (1200 hours), the highest call-blocking and call-dropping probabilities are obtained for a caller distribution of 0%. A steady decrease of both probabilities is observed as the caller distribution increases with the minimum value obtained with a caller distribution of 100%.

Low call-blocking and call-dropping probabilities indicate that a low number of new and handoff calls are blocked and dropped respectively. This further translates to low congestion in the network. Therefore, SDP scheme achieves the lowest congestion levels when the caller distribution is 100%. In any other caller-callee distribution, the congestion level in the MWN is relatively higher.

5.5 SDBDP Scheme.

The SDBDP scheme is obtained when the destination cell congestion factor Φ in Equation (3.5), is greater than 0 ($\Phi>0$). In our evaluation, the value of Φ is varied in steps of 0.25, up to 1. This ensures that the whole call originating cell's network load and a fraction of the call-destination cell's network load is considered in computing the discount advanced to callers. By considering network load in both call originating and call destination cells, the SDBDP scheme gives a more accurate representation of congestion in the RAN.

5.5.1 SDBDP Scheme when Φ=0.25

When Φ is 0.25, the whole call-originating cell's network load and 0.25 of the call-destination cell's network is considered for the computation of the dynamic price paid by the caller. The call-blocking and dropping probabilities obtained when Φ= 0.25 shown in Figures 5-7 and 5-8 respectively.

The maximum call-blocking and call-dropping probabilities are noted at the peak hour (1200 hours). A call-blocking probability maximum value of 0.095 is obtained at 1200 hours with a caller distribution of 0%. This is lower, compared to the call-blocking probability of 0.135 obtained with SDP scheme (when $\Phi=0$) for the same time. A maximum value of 0.085 for call-dropping probability is obtained at 1200 hours. This is lower compared to 0.12 obtained in the SDP scheme at the same time. Therefore, SDBDP scheme with Φ=0.25 obtains lower call-blocking and call-dropping probabilities than the SDP scheme for the same caller distribution and time of day.

The call-blocking and call-dropping probabilities values decrease with increasing caller distribution with the lowest values obtained when the caller distribution is 100%. Therefore,

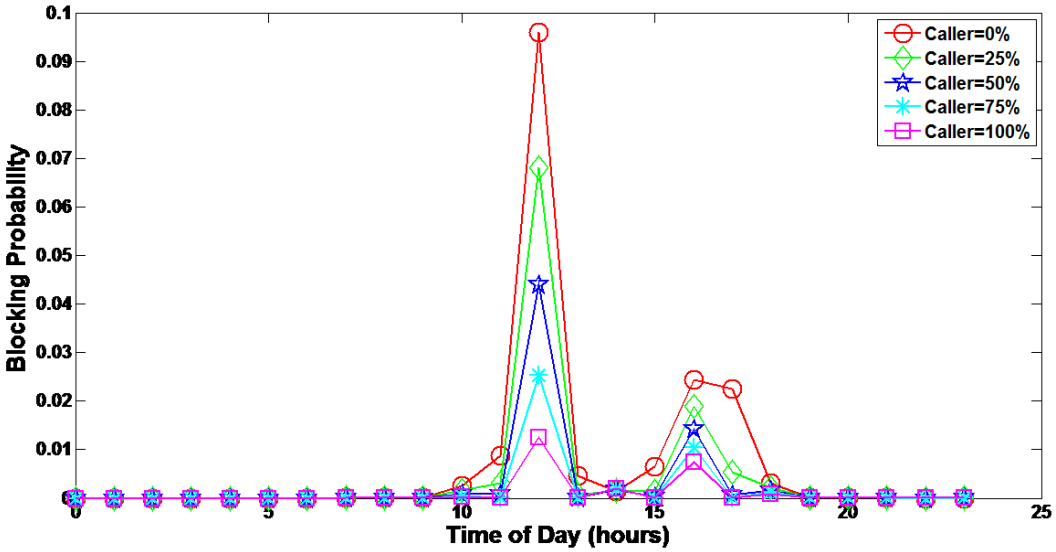


Figure 5-7: Call-blocking probability against time of day with *Φ=0.25*

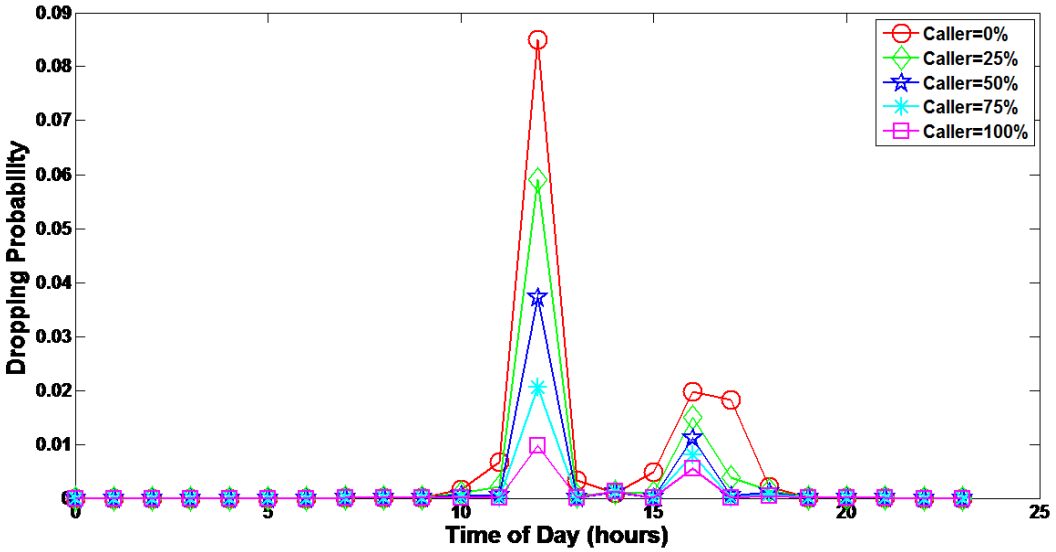


Figure 5-8: Call-dropping probability against time of day with *Φ=0.25*

although this scheme obtains better congestion control than the SDP scheme discussed previously, it still fails in effectively controlling congestion for all caller-callee distributions.

5.5.2 SDBDP Scheme when Φ=0.50

In this scheme, all the call-originating cell network load and half of the network load in the call-destination cell is used to compute the overall dynamic price advanced to the callers. Figures 5-9 and 5-10 show the call-blocking and call-dropping probability respectively under various caller-callee distribution for a period of 24 hours, with Φ=0.5.

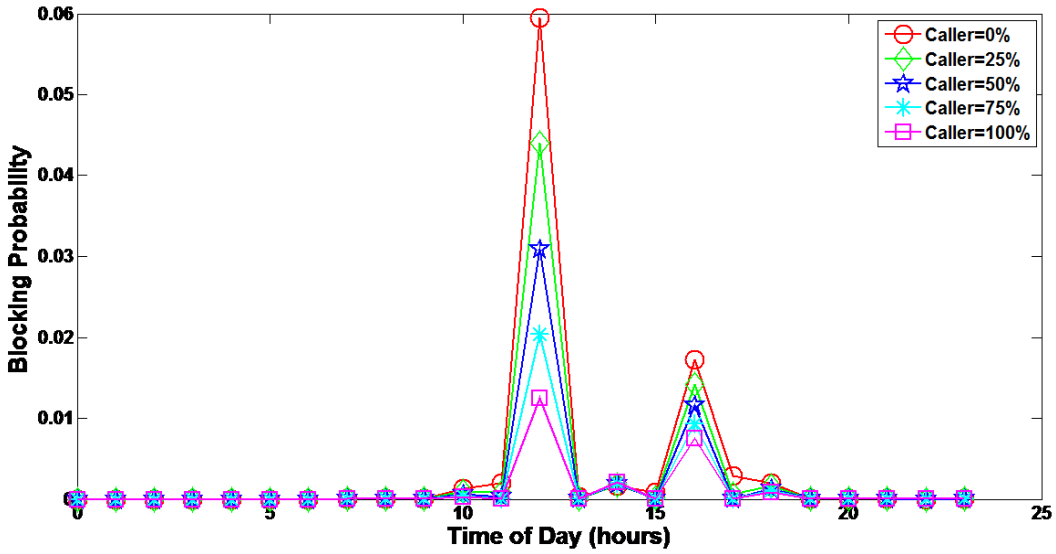


Figure 5-9: Call-blocking probability against time of day with *Φ=0.50*

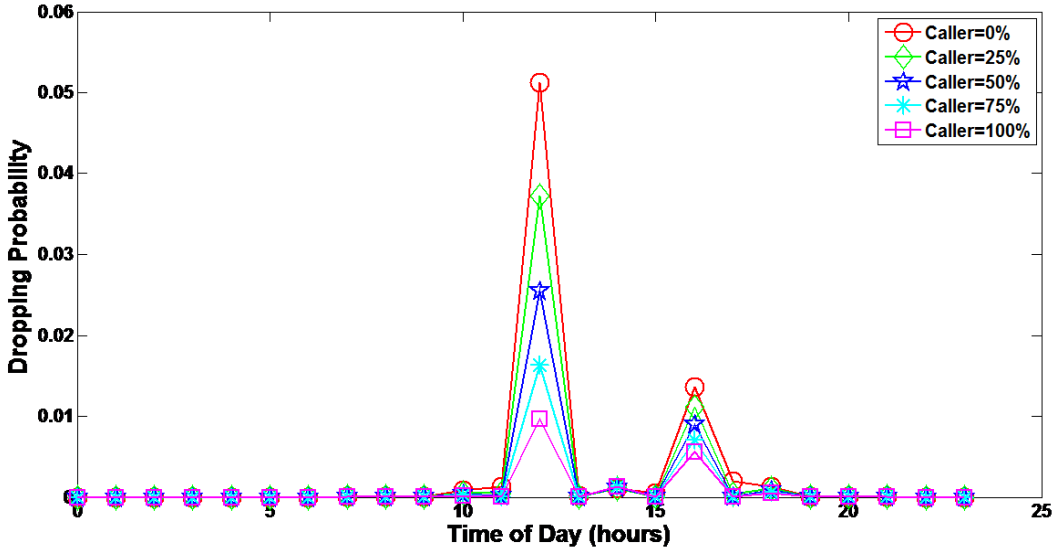


Figure 5-10: Call-dropping probability against time of day with *Φ=0.50*

The maximum values of both new call-blocking and handoff call-dropping probabilities are obtained at the peak-hour when the caller distribution is 0%. The maximum value for call-blocking probability is 0.06 while that of call-dropping probability is 0.05. Both probabilities decrease as the caller distribution increases with a minimum value obtained when the caller distribution is 100%. Compared to the SDBDP scheme with Φ=0.25 and the SDP scheme, the SDBP scheme with Φ=0.5 achieves lower new call-blocking and handoff dropping-probability during the peak periods for all caller-callee distribution. However, for low caller distribution, this scheme is still ineffective in controlling congestion in an HWN.

5.5.3 SDBDP Scheme when Φ=0.75

For this scheme, the entire call-originating cell network load and a fraction (0.75) of the call-destination network load are used in the overall dynamic price computation. Figures 5-11 and 5-12 show the call-blocking and the call-dropping probability for the SDBDP scheme when Φ=0.75, under various caller-callee distributions for a period of 24 hours.

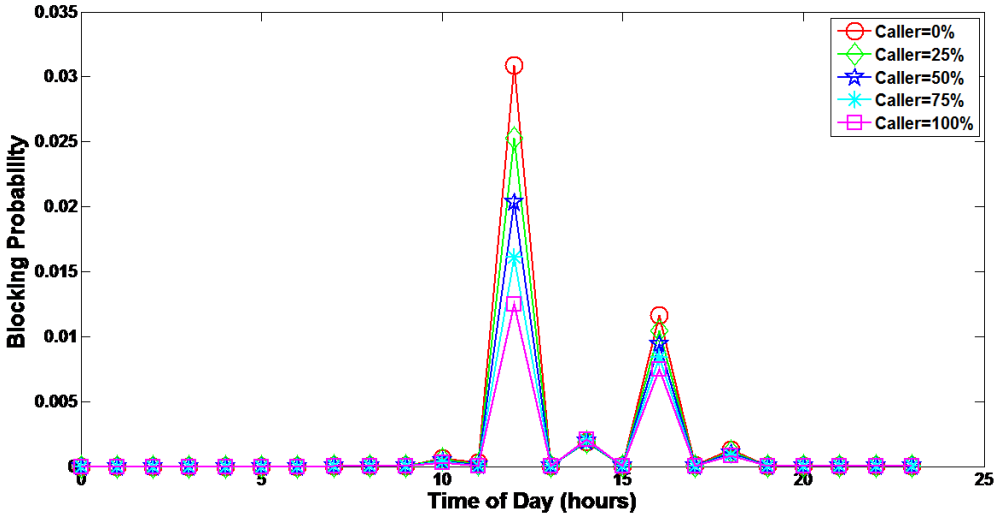


Figure 5-11: Call-blocking probability against time of day with Φ=0.75

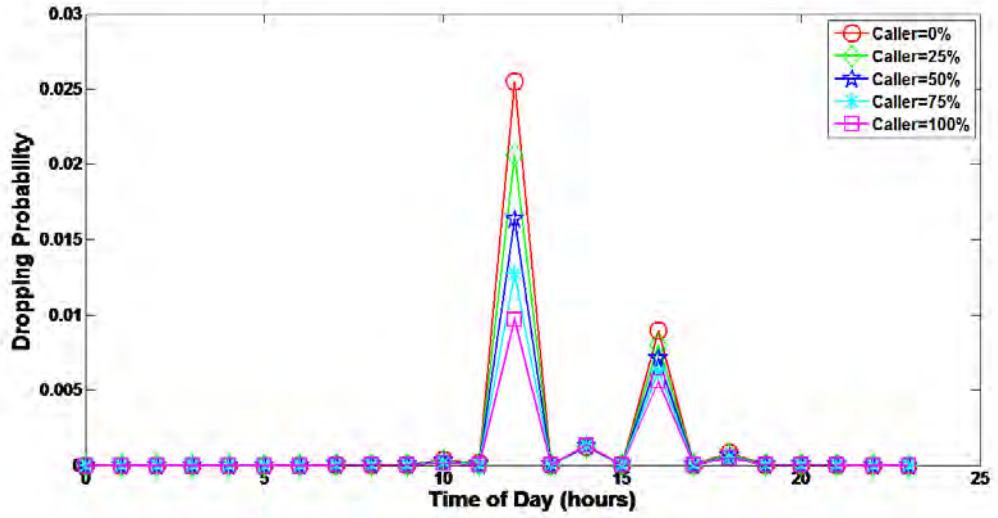


Figure 5-12: Call-dropping probability against time of day with Φ=0.75

From Figure 5-11, a call-blocking probability maximum value of 0.03 is observed with a caller distribution of 0%. This is lower, compared to the call-blocking probability obtained in the preceding SDBDP schemes (Φ=0.5, 0.25 and 0), for similar parameters. From Figure 5-12, the maximum call-dropping probability is observed to be 0.025, a lower value also compared to the call-dropping probability obtained in the preceding SDP and SDBDP schemes, for the same parameters. The lower values obtained in this scheme indicates lower congestion levels in comparison with the preceding schemes during the peak periods. Therefore, the SDBDP scheme with Φ=0.5, achieves better congestion control than all the schemes discussed previously.

The call-blocking and call-dropping probabilities values decrease with increasing caller distribution percentage with the lowest values obtained when the caller distribution is 100%. Therefore, although this scheme obtains better congestion control than the previous three schemes (Φ=0.5, 0.25 and 0), it still fails in controlling congestion for all caller-callee distributions.

5.5.4 SDBDP Scheme when Φ=1.0

In this scheme, the entire network loads in the call-originating and call-destination cells are used in computing the dynamic price to be paid by callers. Figures 5-13 and 5-14 show the call-blocking and call-dropping probabilities under different caller-callee distributions for a 24 hour period.

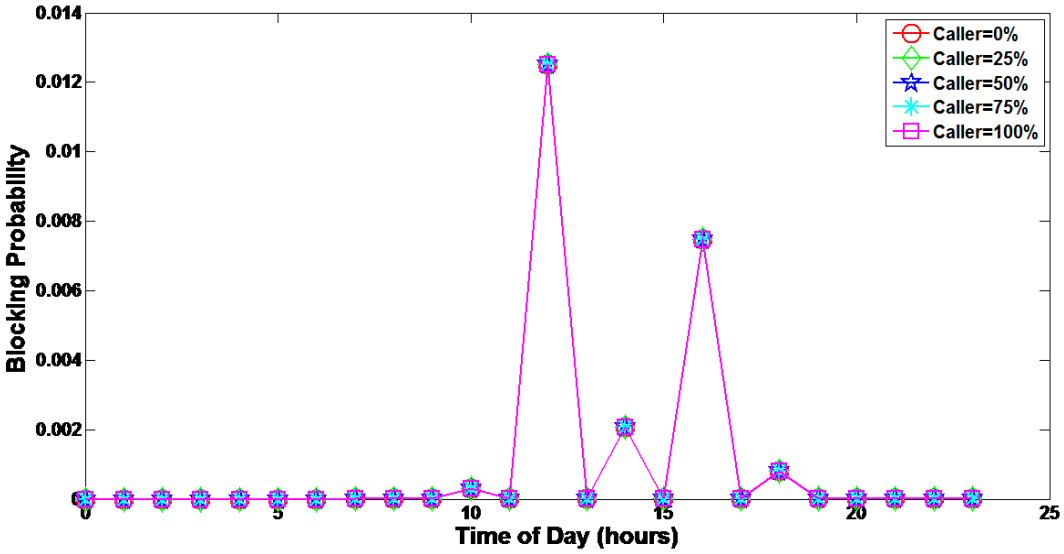


Figure 5-13: Call-blocking probability against time of day with Φ=1

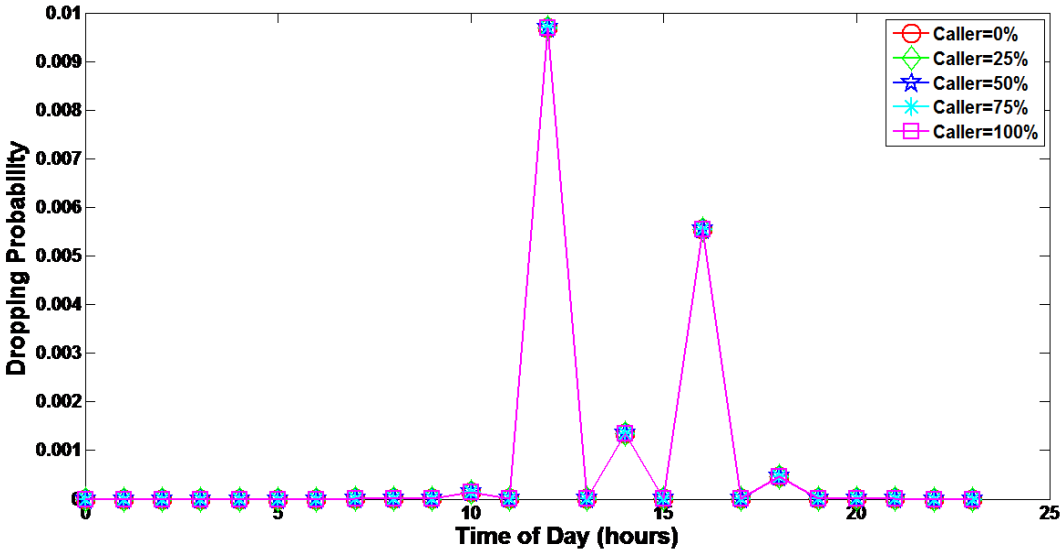


Figure 5-14: Call-dropping probability against time of day with Φ=1

Figure 5-13 and 5-14 show uniform call-blocking and call-dropping probabilities for all the caller-callee distributions, respectively, for the different hours of the day. The call-blocking and dropping probabilities are observed to be highest at 1200 hours. However, the call-blocking and call-dropping probabilities obtained under this scheme at the peak hours are the lowest when

compared to schemes examined previously. Therefore, the SDBDP scheme with Φ=0 achieves the best congestion control for all caller-callee distributions in an HWN.

5.6 Peak Hour Performance Evaluation

At the peak hour (1200 hrs.), we are going to evaluate the following performance metrics for different values of Φ, under different caller-callee distributions:

1. Call-blocking probability
2. Call-dropping probability

In Figures 5-15 and 5-16, when Φ=0 (SDP scheme), the call-blocking and call-dropping probabilities are observed to be maximum for a caller distribution of 0%. As the percentage caller distribution increases, both the call-blocking and call-dropping probabilities decrease with a minimum value being obtained at 100%. Low values of call-blocking and call-dropping probability indicate a reduction in congestion since fewer new and handoff calls are blocked and dropped respectively.

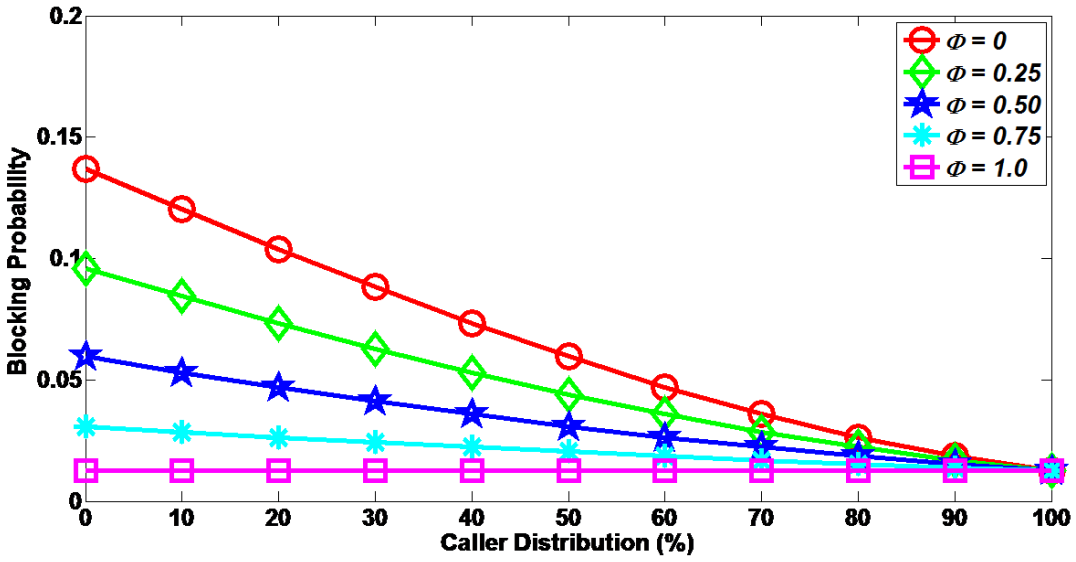


Figure 5-15: Call-blocking probability under different user distributions varying Φ

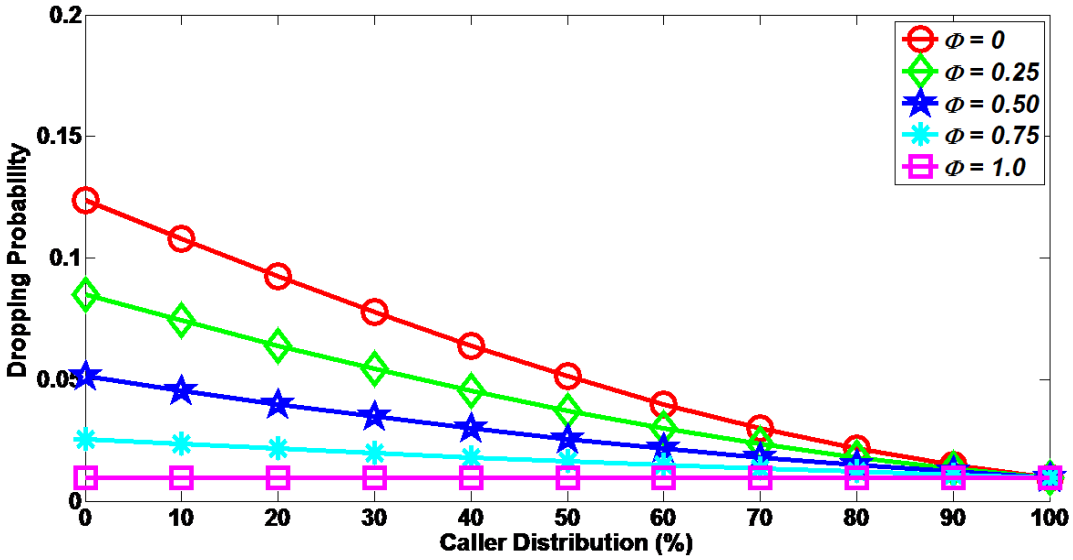


Figure 5-16: Call-dropping probability under different distributions varying Φ

For SDBDP scheme, obtained when $\Phi > 0$, Figures 5-15 and 5-16 show that the call-blocking and dropping probabilities for a particular caller-callee distribution decrease with increasing values of Φ. The minimum values, for all caller distributions, are obtained when Φ=1. In this case, the congestion levels in both call-originating and call-destination group cells are given equal weight in computing the dynamic price. Therefore, Φ=1 is the ideal value for the SDBDP scheme where it achieves lowest call-blocking and call-dropping probabilities compared to all other schemes (SDBDP when $\Phi<1$ and SDP) for arbitrary caller-callee distribution, at the peak hour.

5.7 Off-Peak Hour Performance Evaluation

At the off-peak periods, an MWN is usually under-utilized and MNOs strive to increase the utilization so as to maximize their revenues. Therefore, we shall evaluate the percentage MWN resource utilization during the off-peak period so as to examine the effectiveness of both SDP and SDBDP schemes in increasing MWN resources utilization.

5.7.1 MWN Resource Utilization

The percentage MWN resource utilization at 0700 hrs is shown in Figure 5.17 for both SDP and SDBP. The SDP scheme is shown by the line graph where Φ=0 while the SDBDP

schemes are shown by the line graphs where $\Phi > 0$.

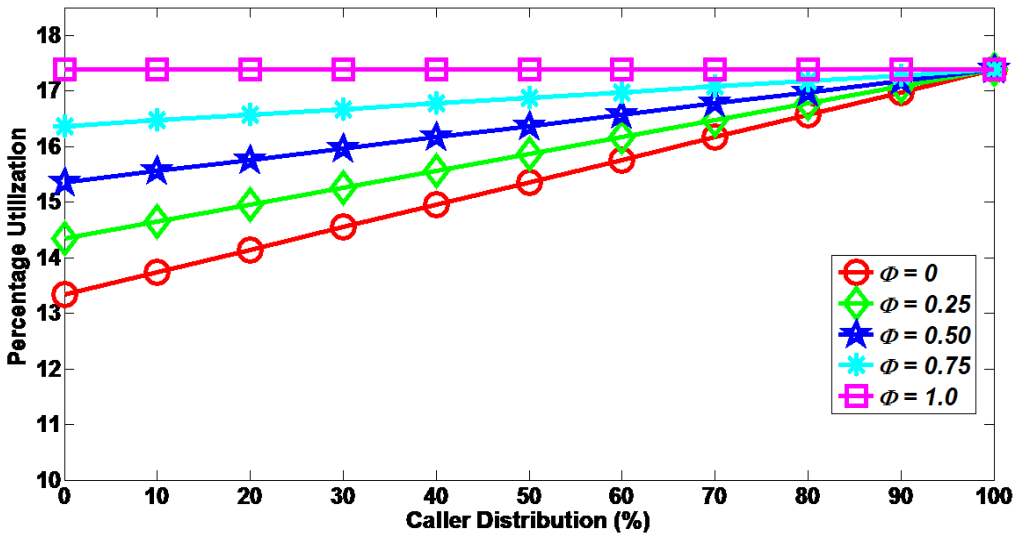


Figure 5-17: Percentage MWN resources utilization during off-peak hour

In the SDP scheme ($\Phi=0$), the percent utilization increases with increase in the caller distribution. The maximum utilization is obtained when the caller distribution is 100%. This scheme is therefore most effective in increasing the system utilization when the caller distribution is 100%. For any other caller-callee distribution, the percentage MWN resource utilization is relatively lower. Therefore, the SDP scheme is unsuitable for improving resource utilization during the off-peak hours.

The SDBDP schemes where $0<\Phi<1$ achieve a higher resource utilization compared to SDP scheme for the same caller-callee distribution. However, under low caller distribution, the percentage MWN resource utilization remains low. Therefore, the SDBDP scheme with $0<\Phi<1$ do not improve the utilization of resources for arbitrary user distributions.

For SDBDP scheme with $\Phi=1$, uniform resource utilization is achieved for all the callee-caller distributions. When compared to SDBDP scheme with $0<\Phi<1$ and SDP scheme, the SDBDP scheme with $\Phi=1$ achieves higher percentage resource utilization for all caller-callee distributions. Therefore, the SDBDP scheme with $\Phi=1$ is the most effective for improving the MWN resource utilization during the off-peak hour.

5.8 Chapter Summary

Both SDP and SDBDP schemes have been compared under different caller-callee distributions. Their efficiency in controlling congestion during peak hours and improving the MWN resource utilization during off-peak hours has been examined.

In the SDP scheme, the minimum call-blocking and call-dropping probabilities are achieved with a caller distribution of 100%. The minimum call-blocking and call-dropping probabilities indicate the least congestion in the HWN. Therefore, SDP scheme achieves the best congestion control when the caller distribution is 100%. During the off-peak period, the percentage MWN resource utilization increases with increase in caller distribution, with a maximum value obtained when the caller distribution is 100%. For all other caller-callee distributions, the MWN resource utilization is relatively low. Therefore, the SDP scheme does not effectively maximize the system utilization during off-peak hours.

Four variants of the SDBDP scheme have been examined. These schemes are identified by the destination cell congestion factor (Φ), which ranges from 0.25 to 1. These schemes achieve better congestion control during the peak hour than the SDP scheme. When Φ=1, effective congestion control for all caller-callee distribution is achieved. The SDBDP scheme with Φ=1 also achieves better congestion control than the other variants of the SDBDP scheme. During off-peak hours, the four variants of the SDBDP scheme also perform better in improving the MWN resource utilization compared to the SDP scheme. The SDBDP scheme with Φ=1 achieves the best MWN resource utilization for all caller-callee distributions compared to the SDP and SDBDP schemes where $0<\Phi<1$.

Chapter 6 Conclusion, Recommendation and Future Work

6.1 Conclusion

Mobile network operators are faced with a challenge of controlling MWN congestion during the peak periods and increasing the utilization during off-peak hours. Various congestion control methods have been proposed. These include increasing the network capacity, splitting cells and even deployment of small cells. However, the infrastructure investments have not translated to increased revenues or decrease in MWN congestion. Economic and user behavior mechanisms have also been used with dynamic pricing of cellular services being the most common. In dynamic pricing, the price of service requests changes according to the network load or the duration. However, the current dynamic pricing schemes have only considered call-originating cell's network load while ignoring the call-destination cell's network load. Due to this property, we have called these current dynamic pricing schemes source-based dynamic pricing (SDP) schemes. The SDP schemes cause poor congestion control during peak hours and low MWN resource utilization during off peak hours.

In this work, we have developed a source-destination pricing (SDBDP) scheme, which considers both the call-originating cell and call-destination cell network load in computing the dynamic price to be paid by the caller. The SDBDP scheme is integrated to a load-based JCAC algorithm which admits new and handoff calls to the least loaded RAT. The load-based JCAC algorithm also ensures that QoS of ongoing calls is not affected by incoming calls. An analytical model based on multi-dimensional Markov chain has been developed. An HWN is evaluated using call-blocking probability, call-dropping probability and percentage MWN resource utilization performance parameters.

Simulation results have shown very high blocking probabilities and dropping probabilities for the SDP schemes when the caller distribution is low. It has been established that SDP schemes achieve the best congestion control under high caller distribution. With high callee distribution many new calls and handoff calls are blocked and dropped respectively.

The percentage MWN resource utilization for SDP schemes is high under high caller distribution. Under low caller distribution, the percentage utilization remains relatively low. This indicates that SDP schemes are ineffective in increasing the MWN resource utilization, during off peak hours, under low caller distribution. These SDP schemes perform best under high caller distribution.

The variants of the SDBDP scheme achieve lower call blocking and call dropping probability for all caller-callee distributions compared to the SDP scheme, during the congested peak periods. The SDBDP scheme with Φ=1 achieves the lowest call blocking and dropping probabilities. Therefore, the SDBDP schemes are more efficient in controlling congestion during the peak hours, than SDP schemes. The SDBDP scheme with Φ=1 is the most efficient in controlling congestion. Also, unlike the SDP schemes, the SDBDP scheme with Φ=1 controls congestion for all caller-callee distributions.

During the under-utilized off-peak periods, all the variants of the SDBDP scheme obtain higher MWN resource utilization than the SDP scheme. The SDBDP scheme with Φ=1 obtains the highest resource utilization for all caller-callee distributions.

Overall, the SDBDP scheme with Φ=1 has been found to be more effective than the SDP schemes in controlling congestion during the peak periods and in improving the MWN resource utilization during the under-utilized off-peak periods.

6.2 Recommendations and Future Work

We have considered connection-level QoS in this work. However, for comprehensive QoS support, packet-level QoS on top of connection level QoS support should also be considered. Therefore, more research can be done on incorporating both connection and packet level QoS support in dynamic pricing to improve further the overall user experience.

Price sensitive users are bound to respond to high prices by either shifting their service request to a cheaper period or shortening the length of their service request. In this work, we have considered the shifting the non-sensitive traffic from the peak hours when the network is congested and prices are high to off-peak hours when the MWN is under-utilized and the prices are low. Therefore, this work can be extended to consider the effect of dynamic pricing on the duration of

the service requests.

The key principle in dynamic pricing for congestion control in HWN is the increased prices of service requests during the congested period. The high prices during the peak periods may not be affordable to low-income users. Therefore, dynamic pricing can be seen to create social unfairness. Further research can be done on how to apply dynamic pricing in HWNs while maintaining social fairness for all the users.

Design of charging system to enable real-time relaying of dynamic prices to the user before call set-up should be investigated further. New mobile services and software applications need to be developed for the user's mobile terminals so as to display the dynamic prices to the users in real-time. Also, these end-user applications should help in information gathering to help anticipate the user's utility demand for wireless resources.

The developed SDBDP scheme is based on a monopolistic environment. In a multi-operator environment, users can call across different networks. Therefore, further investigation should be conducted on how to dynamically charge requests destined to a different operator's network while ensuring optimal use of the MWN resources.

The SDBDP scheme can be analysed further to determine the impact on the operator revenue. This is important since MNOs are interested in maximizing the income from their deployed infrastructure.

References

[1] Cisco Visual Networking Index: "Forecast and Methodology, 2014-2019"

[2] B. Labs, “Metro Network Traffic Growth: an Architecture Impact Study: Strategic White Paper,” p. 12, 2013.

[3] O. E. Falowo and Ha. Chan, “Adaptive Bandwidth Management and Joint Call Admission Control to Enhance System Utilization and QoS in Heterogeneous Wireless Networks,” *EURASIP J. Wirel. Commun. Netw.*, vol. 2007, no. 1, p. 034378, 2007.

[4] G. Wu, M. Mizuno and P. J. M. Havinga, “Networks and Navigation Services Mirai Architecture for Heterogeneous Network” *IEEE Commun. Mag.*, no. February 2002, pp. 126–134, 2005.

[5] E. Bodanese and P. Qhwzrun,“A Brief introduction to Heteorgeneous Networks and its Challenges”, IET International Conference, vol. 7, PP605-609, 2011.

[6] A. Damnajovic, J. Montojo, “A Survey on 3GPP Heterogenous Networks,” *IEEE Wireless communications,* Vol 18, Issue no3, pp. 10–21, June 2011.

[7] T. V. K. Reddy and P. S. S. S. Teja, “Joint Call Admission Control for Multi-Mode Terminals in Heterogeneous Cellular Networks,” no. 1, pp. 240–248, 2013.

[8] O. E. Falowo, S. Zeadally, and H. A. Chan, “Dynamic pricing for load-balancing in user-centric joint call admission control of next-generation wireless networks,” no. October 2009, pp. 335–368, 2010.

[9] N. Verma and I.Chen, “Admission Control Algorithms Integrated with Pricing for Revenue Optimization with QoS Guarantees in Mobile Wireless Networks”, Proceedings of the Parallel and Distributed Systems

[10] S. Kabahuma and O. E. Falowo “Analysis of Network Operators ’ Revenue with a Dynamic Pricing Model Based on User Behaviour in NGWN Using JCAC.”

[11] O. E. Falowo and H. A. Chan, “Analysis of Joint Call Admission Control Strategies for Heterogeneous Cellular Networks,” 12th IEEE Symposium on Computers and

Communications (ISCC) pp. 775–780, Portugal, July 2007.

[12] B. M. Epstein, M. Schwartz, and L. Fellow, "Predictive QoS-Based Admission Control for Multiclass Traffic in Cellular Wireless Networks," vol. 18, no. 3, pp. 523–534, 2000.

[13] J. Bigham, H. Pervaiz, J. Bigham, P. Jiang, and M. P. Chan, "A game theoretic based Call Admission Control scheme for competing WiMAX networks A Game Theoretic based Call Admission Control Scheme for Competing WiMAX Networks," no. November 2016, 2009.

[14] Dusit Niyato and Ekram Hossain "Radio resource management games in wireless networks : an approach to bandwidth allocation and admission control for polling service in ieee 802 . 16", *IEEE Wireless Communications,* pp. 27–35, February 2007.

[15] J. Antoniou, A. Pitsillides, and S. Member, "4G Converged Environment : Modeling Network Selection as a Game," 2007.

[16] G. Parr and A. Bashar, "Novel distributed call admission control solution based on machine learning approach A Novel Distributed Call Admission Control Solution based on Machine Learning Approach," no. May, 2011.

[17] B. Al-Manthari, N. Nasser, and H. Hassanein, "Congestion Pricing in Wireless Cellular Networks," *IEEE Commun. Surv. Tutorials*, vol. 13, no. 3, pp. 358–371, 2011.

[18] C. A. MacK, "Fifty years of Moore's law," in *IEEE Transactions on Semiconductor Manufacturing*, 2011, vol. 24, no. 2, pp. 202–207.

[19] "Apple's Voice Recognition Siri Doubles IPhone Data Volumes - Bloomberg." [Online]. Available:http://www.bloomberg.com/news/articles/2012-01-06/apple-s-voice-recognition-siri-doubles-iphone-data-volumes. [Accessed: 02-May-2016].

[20] S. Sen, C. Joe-Wong, S. Ha, and M. Chiang, "Incentivizing time-shifting of data: A survey of time-dependent pricing for internet access," *IEEE Commun. Mag.*, vol. 50, no. 11, pp. 91–99, 2012.

[21] S. Sen, C. Joe-Wong, S. Ha, and M. Chiang, "Smart Data Pricing (SDP): Economic Solutions to Network Congestion," *Recent Adv. Netw.*, pp. 221–274, 2013.

[22] R. Thanawala, M. El-sayed, T. Morawski, A. Mukhopadhyay, J. Zhao, and C. Urrutia-

valdés, “The mobile data explosion and new approaches to network planning and monetization,” *Bell Labs Tech. J.*, vol. 16, no. 2, pp. 79–99, 2011.

[23] Real Wireless, “Techniques for increasing the capacity of wireless broadband networks : UK , 2012-2030 Produced by Real Wireless on behalf of Ofcom,” pp. 2012–2030, 2012.

[24] C. a. Gizelis and D. D. Vergados, “A survey of pricing schemes in wireless networks,” *IEEE Commun. Surv. Tutorials*, vol. 13, no. 1, pp. 126–145, 2011.

[25] H. Sangtae, S. Soumya, C. Joe-Wong, Y. Im, and M. Chiang, “TUBE : Time-Dependent Pricing for Mobile Data,” *SIGCOMM ’12 Proc. ACM SIGCOMM 2012 Conf. Appl. Technol. Archit. Protoc. Comput. Commun.*, pp. 247–258, 2012.

[26] L. Jiang, S. Parekh, and J. Walrand, “Time-dependent network pricing and bandwidth trading,” *2008 IEEE Netw. Oper. Manag. Symp. Work. - NOMS 08*, pp. 193–200, 2008.

[27] M. J. Sheng, C. Joe-Wong, S. Ha, F. M. F. Wong, and S. Sen, “Smart data pricing: Lessons from trial planning,” *Proc. - IEEE INFOCOM*, pp. 3327–3332, 2013.

[28] L. Zhang, W. Wu, and D. Wang, “Time dependent pricing in wireless data networks: Flat-rate vs. usage-based schemes,” *INFOCOM, 2014 Proc. IEEE*, pp. 700–708, 2014.

[29] H. R. Varian, *Intermediate Microeconomics: A Modern Approach*, no. 8. 2010.

[30] S. Papavassiliou, “Integration of pricing with call admission control to meet QoS requirements in cellular networks,” *IEEE Trans. Parallel Distrib. Syst.*, vol. 13, no. 9, pp. 898–910, 2002.

[31] P. Camp, “Van Westendorp ’ s Price Sensitivity Meter Scientific Pricing : Research is the Key,” 2012.

[32] P. C. Fishburn and A. M. Odlyzko, “Dynamic Behavior of Differential Pricing and Quality of Service Options for the Internet,” *Proc. First Int. Conf. Inf. Comput. Econ.*, pp. 128–139, 1998.

[33] S. L. Hew and L. B. White, “Optimal integrated call admission control and dynamic pricing with handoffs and price-affected arrivals,” *2005 Asia-Pacific Conf. Commun. (Apcc), Vols 1& 2*, no. October, pp. 396–400\r1116, 2005.

[34] S. Mandal, D. Saha, and A. Mahanti, “A technique to support dynamic pricing strategy for differentiated cellular mobile services,” *IEEE Globecom 2005*, pp. 3388–3392, 2005.

[35] S. Yaipairoj, “Auction-based Congestion Pricing for Wireless Data Services,” vol. 00, no. c, pp. 1059–1064, 2006.

[36] J. K. MacKie-Mason and H. R. Varian, “Pricing congestible network resources,” *IEEE J. Sel. Areas Commun.*, vol. 13, no. 7, pp. 1141–1149, 1995.

[37] S. Mandal, D. Saha, and M. Chatterjee, “Pricing wireless network services using smart market models,” *CCNC 2006. 2006 3rd IEEE Consum. Commun. Netw. Conf. 2006.*, vol. 1, pp. 574–578, 2006.

[38] S. Mandal, D. Saha, and M. Chatterjee, “Dynamic price discovering models for differentiated wireless services,” *J. Commun.*, vol. 1, no. 5, pp. 50–56, 2006.

[39] C. Courcoubetis and R. Weber, *Pricing communication networks*, vol. 2. 2003.

[40] J. W. Lee, R. R. Mazumdar, and N. B. Shroff, “Downlink Power Allocation for Multi-class CDMA Wireless Networks,” vol. 00, no. c, pp. 1480–1489, 2002.

[41] V. Siris, “Resource Control for Elastic Traffic in CDMA Networks,” *Proc. Conf. Mob. Comput. Netw.*, 2002.

[42] S. W. Han and Y. Han, “A simple congestion pricing in wireless communication,” *IEEE Veh. Technol. Conf.*, vol. 2, pp. 795–798, 2015.

[43] “Axon Partners Group Consulting Global Insights Mobile Data Pricing strategies around the World,” no. April, 2014.

[44] C. Parris, S. Keshav, and D. Ferrari, “A framework for the study of pricing in integrated networks,” 1992.

[45] J. K. MacKie-Mason and H. R. Varian, “Pricing the internet,” *Public access to Internet*, pp. 269–314, 1995.

[46] “The mother of invention | The Economist.” [Online]. Available: http://www.economist.com/node/14483880. [Accessed: 10-May-2016].

[47] “Telus Tuneage: Now powered by Rdio.” [Online]. Available: http://thenextweb.com/ca/2011/08/03/telus-tuneage-now-powered-by-rdio/. [Accessed: 22-Apr-2016].

[48] “Orange refreshes Panther tariff with Swapables benefits | Mobile Today.” [Online]. Available: http://www.mobiletoday.co.uk/News/12573/Orange_refreshes_Panther_tariff_with_Swapables_benefits_.aspx. [Accessed: 22-Apr-2016].

[49] “Dean Bubley’s Disruptive Wireless: Belgian MNO tries app-specific zero rating - bad idea in my view.” [Online]. Available: http://disruptivewireless.blogspot.co.za/2011/11/belgian-mno-tries-app-specific-zero.html. [Accessed: 22-Apr-2016].

[50] S. Sen, C. Joe-Wong, S. Ha, and M. Chiang, “Pricing data: A look at past proposals, current plans, and future trends,” *arXiv Prepr. arXiv1201.4197*, vol. 46, no. 2, pp. 1–37, 2012.

[51] A. D. Mafuta, “Congestion Control based on Dynamic Pricing Scheme and Service Class-based Joint Call Admission Control in Heterogeneous Wireless Networks,” Masters Thesis, School of Mathematics, Statistics and Computer Science, Univeristy of KwaZulu Natal, South Africa, December 2013.

[52] S. Kabahuma, “Joint Call Admission Control Incorporating Pricing for Congestion Control to Enhance QoS and Ensure Revenue for Network Operators in Next Generation Wireless Networks”, Masters Thesis, Electrical Dept, Universisty of Cape Town, South Africa, November 2010.

[53] “Quality of Service Overview - Technical Documentation - Support - Juniper Networks.” [Online]. Available: http://www.juniper.net/techpubs/en_US/junos-mobility12.1/topics/concept/service-parameters-mobility-overview.html. [Accessed: 14-Mar-2016].

[54] J. D. C. Little and S. C. Graves, “Chapter 5 Little ’ s Law,” *Oper. Manag.*, vol. 115, pp. 81–100, 2008.

[55] D Bertsekas and J Tsitsiklis “Introduction to Probability", 2nd Edition, MIT, 2008

[56] Finite State Markov Chains, Available: http://www.rle.mit/edu/rgallager/documents/6.262-4vawa.pdf

[57] L. Kleinrock, "Queueing Systems, Volume 1: Theory," *John Wiley Sons, Inc.*, vol. 1, no. 1, p. 417, 1975.

[58] H. Zeng and I. Chlamtac, "Adaptive guard channel allocation and blocking probability estimation in PCS networks," *Comput. Networks*, vol. 43, no. 2, pp. 163–176, 2003.

[59] G. Bolch, S. Greiner, H. de Meer, and K. S. Trivedi, *Queueing Networks and Markov Chains*. 2006.

[60] M. Veeraraghavan, "Applications of queueing theory to circuit switched networks," vol. 1, pp. 1–11, 2004.

[61] "Blocked customers cleared", Available:http://ocw.mit.edu/courses/sloan-school-of-management/15-763j-manufacturing-system-and-supply-chai-design-spring-005/recitations/review-of-qmole.pdf.

[62] B. Al-Manthari, N. Nasser, N. Abu Ali, and H. Hassanein, "Efficient Bandwidth Management in Broadband Wireless Access Systems Using CAC-based Dynamic Pricing," *2008 Ieee 33rd Conf. Local Comput. Networks, Vols 1 2*, vol. 2008, pp. 473–480\r947, 2008.

[63] J. K. H. Quah, "The Law of Demand and Risk Aversion," *Econometrica*, vol. 71, no. 2, pp. 713–721, 2003.