Simulation-Assisted QoS-Aware VHO in Wireless Heterogeneous Networks

Ismaeel Al Ridhawi

Thesis submitted to the
Faculty of Graduate and Postdoctoral Studies
in partial fulfillment of the requirements
for the Doctorate in Philosophy degree in Electrical and Computer Engineering

School of Electrical Engineering and Computer Science (EECS)
Faculty of Engineering
University of Ottawa

ABSTRACT

The main goal of today's wireless Service Providers (SPs) is to provide optimum and ubiquitous service for roaming users while maximizing the SPs own monetary profits. The fundamental objective is to support such requirements by providing solutions that are adaptive to varying conditions in highly mobile and heterogeneous, as well as dynamically changing wireless network infrastructures. This can only be achieved through well-designed management systems. Most techniques fail to utilize the knowledge gained from previously tested reconfiguration strategies on system and network behaviour.

This dissertation presents a novel framework that automates the cooperation among a number of wireless SPs facing the challenge of meeting strict service demands for a large number of mobile users. The proposed work employs a novel policy-based system configuration model to automate the process of adapting new network policies. The proposed framework relies on the assistance of a real-time simulator that runs as a constant background process in order to continuously find optimal policy configurations for the SPs' networks. To minimize the computational time needed to find these configurations, a modified tabu-search scheme is proposed. An objective is to efficiently explore the space of network configurations in order to find optimal network decisions and provide a service performance that adheres to contracted service level agreements.

This framework also relies on a distributed Quality of Service (QoS) monitoring scheme. The proposed scheme relies on the efficient identification of candidate QoS monitoring users that can efficiently submit QoS related measurements on behalf of their neighbors. These candidate users are chosen according to their devices' residual power and transmission capabilities and their estimated remaining service lifetime. Service monitoring users are then selected from these candidates using a novel user-to-user semantic similarity matching algorithm. This step ensures that the monitoring users are reporting on behalf of other users that are highly similar to them in terms of their mobility, used services and device profiles.

Experimental results demonstrate the significant gains achieved in terms of the reduced traffic overhead and overall consumed users' devices power while achieving a high monitoring accuracy, adaptation time speedup, base station load balancing, and individual providers' payoffs.

Dedication

To my loving parents and wonderful wife.
From the bottom of my heart.

Acknowledgements

I would like to express my deepest gratitude to my supervisor, Prof. Ahmed Karmouch, for his guidance, encouragement and support throughout my graduate studies. It would not have been possible to complete this work without his continuous assistance.

I would also like to extend appreciation to my co-supervisor, Prof. Nancy Samaan, for her effective supervision and helpful discussions, both of which significantly contributed to the completion of this thesis.

I am delighted to have had the opportunity to work with exceptional colleagues and friends who provided me with encouragement throughout my research. I would like to thank Imad Abdeljouad, Heli Amarasinghe, Bassim Victor, and all my lab mates at the '*Intelligence for Mobile Autonomic and Cognitive Networks Laboratory*' for their cooperation and advice.

I would like to thank Cistech Limited for their support and the constructive feedback that helped improve the quality of this thesis.

My thanks also extend to my brother, Yousif Al Ridhawi, for his help, guidance and mentoring. Without his assistance it would have been impossible to complete my research.

Words are not enough to thank my wife, Doaa, for her patience, love and emotional support which made it all possible.

Last but not least, my deepest thanks go to my parents for their on-going love, support, and encouragement. Their positive influence provided me with the determination to persevere with my studies. They have always been a source of inspiration throughout my life and their selfless sacrifices shall never go unacknowledged.

Contents

List of Figures

List of Tables

List of Algorithms

Acronyms

1G	First Generation
2G	Second Generation
3G	Third Generation
4G	Fourth Generation
AHP	Analytic Hierarchy Process
AMPS	Advanced Mobile Phone System
AP	Access Point
BER	Bit Error Rate
BS	Base Station
CBR	Constant Bit Rate
CDMA	Code Division Multiple Access
CHOP	Configurable, Healable, Optimizable, protectable
CoA	Care-of-Address
CoS	Class of Service
CPU	Central Processing Unit
DCD	Downlink Channel Descriptor
DMTF	Distributed Management Task Force
DSSS	Direct Sequence Spread
E2E	End-to-End
EDGE	Enhanced Data Rates for Global Evolution
EV-DO	Evolution Data Optimized
EV-DV	Evolution Data Voice
FA	Foreign Agent
FIFO	First In First Out
FMIP	Fast Mobile Internet Protocol
GloMoSim	Global Mobile Information Systems Simulation Library
GLS	Guided Local Search
GMM	Guass Markov Model
GPRS	General Packet Radio Service

GRA	Grey Relational analysis
GSM	Global System for Mobile Communications
HA	Home Agent
HAWAII	Handoff Aware Wireless Access Internet Infrastructure
HIP	Host Identity Protocol
HMIP	Hierarchical MIP
HoA	Home Address
HSDPA	High Speed Downlink Packet Access
HSPA	High Speed Packet Access
HSUPA	High Speed Uplink Packet Access
IDMP	Intra-Domain Mobility Management Protocol
IEEE 802.11	Wi-Fi
IEEE 802.15.1	Bluetooth
IEEE 802.16	WiMAX
IETF	Internet Engineering Task Force
IGL-TS	Iterated Local and Global-Tabu Search
IP	Internet Protocol
IPTV	Internet Protocol Television
IS	Interim Standard
ISP	Internet Service Provider
LDAP	Lightweight Directory Access Protocol
LTE	Long-Term Evolution
MADM	Multiple attribute decision making
MANET	Mobile Adhoc Network
MIMO	Multiple Input Multiple Output
MIP	Mobile Internet Protocol
MN	Mobile Node
MOBIKE	Internet Key Exchange Protocol Mobility and Multi-homing
NIC	Network Interface Card
NIST	National Institute of Standards and Technology

NMT	Nordic Mobile Telephone
NS	Network Simulator
OFDM	Orthogonal Frequency Division Multiplexing
OFDMA	Orthogonal Frequency Division Multiple Access
OntoCAT	Ontology Common API Tasks
OWL	Web Ontology Language
PBM	Policy-Based Management
PBNM	Policy-Based Network Management
PCS	Personal Communications Service
PDP	Policy Decision Point
PEP	Policy Enforcement Point
PM	Potential Monitor
PMIP	Proxy MIP
PMT	Policy Management Tool
PR	Policy Repository
QoE	Quality of Experience
QoS	Quality of Service
RAM	Random Access Memory
ROM	Read-Only Memory
RSS	Received Signal Strength
RSSI	Received Signal Strength Indication
RTP	Real-time Transport Protocol
SAW	Simple Additive Weighting
SC-FDMA	Single Carrier Frequency Division Multiple Access
SDMA	Space Division Multiple Access
SIP	Session Initiation Protocol
SLA	Service-Level Agreement
SLO	Service Level Objectives
SLS	Service-Level-Specification
SM	Simulator Manager
SMS	Short Message Service

SOFDMA	Orthogonal Frequency Division Multiple Access
SP	Service Provider
SS	Service Subscriber
SSON	Service Specific Overlay Network
TACS	Total Access Communication System
TCP	Transmission Control Protocol
TDMA	Time Division Multiple Access
TS	Tabu-Search
UCD	Uplink Channel Descriptor
UDP	User Datagram Protocol
UMB	Ultra-Mobile Broadband
UMTS	Universal Mobile Telecommunications Service
UWB	Ultra Wideband
VBR	Variable Bit Rate
VC	Variable Configuration
VCO	Variable Configuration Optimizer
VHO	Vertical Handover
VNS	Variable Neighborhood Search
VoD	Video on Demand
VoIP	Voice over Internet Protocol
WCDMA	Wideband Code Division Multiple Access
Wi-Fi	Wireless-Fidelity
WiMAX	Worldwide Interoperability for Microwave Access
WLAN	Wireless Local Area Network
WMAN	Wireless Metropolitan Area network
WPAN	Wireless Personal Area Network
WSN	Wireless Sensor Network

CHAPTER 1

INTRODUCTION

1.1. OVERVIEW

Today's wireless telecommunication market has shown tremendous advancement in technology employing an abundant variety of access network types. This trend will most likely continue in the near future, resulting in an unprecedented increase in the number of mobile users. Network connectivity has become an essential part of mobile users' personal lives and work. Although most users have only recently become acquainted with 'all-time' network connection, some are starting to pose more and more stringent service quality demands on wireless telecommunication networks.

Wireless networks have been emphasized due to their capability of providing internet connection regardless of the geographic location and mobility status. The evolution of wireless technology has led to different generations of wireless cellular systems such as 2G, 3G, 4G and 4G LTE. The 4G wireless network is of a heterogeneous nature and consists of different wired and wireless access networks. These wireless access technologies vary not only in their physical media, but also in the span of their coverage, their location suitability (e.g., indoor vs. outdoor) as well as their ability to inter-operate and coexist. Moreover, each wireless access technology provides mobile users with a unique mix of features with respect to its connection's monetary costs, offered capacity, transmission rates and service quality. For example, IEEE 802.11 networks can support high data rate services in hot spots. On the other hand, IEEE 802.16 and cellular networks offer wireless access over longer distances.

To take full advantage of these technologies, current mobile devices have been enhanced by utilizing powerful processors and batteries resulting in longer life spans, as well as multiple radio interfaces. Such features allow users to run advanced media applications and connect, possibly simultaneously, to more than one network technology. These advances

provide roaming users with the opportunity to have optimum and ubiquitous services.

Satisfaction of users' increasing service demands is dependent upon the efficient utilization of the underlying access networks' resources. For sure network and service providers are always eager towards maximizing their profit. However, it is necessary to take into consideration users' satisfaction of the provided service while provisioning network and service profit maximization solutions. Nonetheless, in such heterogeneous environments, it has been shown [1] that this objective can be achieved through cooperation of service providers. Cooperation requires resource pooling in heterogeneous networks and is mainly achieved by switching a mobile user from one network to another, possibly with a different link-layer technology. This functionality is generally known as the vertical-handover (VHO) process [2].

Unfortunately, achieving a satisfactory VHO is becoming one of the most challenging issues with respect to user mobility management. Such a challenge is a result of the diverse requirements of current and emerging applications, as defined in terms of tolerance to service delays, along with variations in bandwidth availability and/or abrupt changes in link quality. Nonetheless, network operators strive to maintain a seamless and uninterrupted service to the users along the various steps of the VHO process.

In general, network and service providers should be able to offer much more flexibility to users in selecting Quality of Service (QoS) facilities for services and tuning them to the capabilities of the network. This is carried out while adhering to service agreements between service/network providers and consumers. In terms of QoS and Mobility, advances in network management will require new policy management architectures that autonomously fine-tune network and mobility configurations. Such solutions will be an important concept for the future Internet infrastructures.

With increasing stringent quality demands from users, the problem of QoS monitoring and performance evaluation must be restructured to consider the massive quantity and the high dynamicity of mobile users. Thus, it is necessary to focus on monitoring solutions that are energy efficient in regard to battery power usage and require less network traffic in terms of data transmission while taking into account redeveloped node clustering mechanisms.

This dissertation, addresses the problems of managing QoS and mobility issues in wireless heterogeneous networks and issues related to improving policy management and

QoS monitoring mechanisms. This chapter briefly discusses the different aspects of QoS, mobility and policy-based network management in fourth generation wireless networks. The system design process and architecture are also presented. Furthermore, the contribution of this work in reference to current research is addressed. Finally, the organization of the remainder of the thesis is presented.

1.2. POLICY-BASED NETWORK MANAGEMENT PROBLEM

The progression of both wired and wireless networks has amplified a diverse selection of network management systems [3] [4] [5] [6] [7] [8] [9] [10] [11] [12] [13]. By separating the rules that govern the behaviour of a network from its functionality, Policy-Based Network Management (PBNM) provides a possibility for reducing maintenance costs while improving QoS and dynamic adaptability. PBNM is conceptualized as a set of mechanisms that are used to fine-tune different network services, such as controlling traffic flow through a network. PBNM is attractive due to its nature and ability to allow a single command to be chosen to implement what previously consisted of a set of commands. However, such a powerful solution for network management does have its limitations. Although Policy-based management (PBM) systems introduce the notion of predefined policies that prescribe a set of rules that guide the behaviour of network components, policies would lend themselves to be of a static nature and thereby introduce an immediate burden on network administrators. New and different policies must be developed in order to maintain the changing environment to sustain its integrity and achieve the requested objectives. With the increasing magnitude and complexity of current network components, this task places excessive demands on network administrators.

Network management is further complicated by a second problem. Network administrators rely on QoS network monitoring systems to estimate network traffic and user requirements when configuring network policies. These estimates can be a major source of inflexibility. It must be acknowledged that these limitations must be solved, while maintaining an acceptable solution time. Fortunately, advances in autonomous self-configurable systems [14] [15] [16] [17], such as reinforcement learning [18], have facilitated valuable reconfiguration tools for PBNM systems. Nonetheless, in order to provide the necessary up-to-date policy and network configurations to maintain a continuous

smooth operation, research in this area is still considered an ongoing research. The challenge specifically lies in the domain of variable reconfiguration accuracy and time requirements. Furthermore, selecting the appropriate system configurations in order to maintain an acceptable level of delivered service quality relies on several criteria. These include the degree of mobility of the serviced users, the variations of the traffic and the Class of Service (CoS) requested by the users.

Therefore, it is necessary that any intelligent and efficient algorithm be evaluated prior to deployment, due to the cost incurred in case of unnecessary resource depletion or unnecessary network handoff in network cooperation situations. A crucial issue that contributes to the problem is having contradicting demands for service subscribers and providers. Subscribers' demands require optimum service quality with the least possible cost. On the contrary, the providers' objectives are to provide subscribers with desired service in a timely manner with the agreed upon performance guarantees such that resources are efficiently utilized and maximize the total revenue. With this increasing number of mobile service subscribers, management systems may not have sufficient resources to deliver the required service quality to all subscribers. Hence, fast and efficient decisions must be made to sufficiently optimize these resources.

Even though there exists many state-of-the-art QoS management systems, e.g. [19], [20], [21], [22], [23], [24], [25], which provide configurations that maximize the global objectives, these systems provide temporary solutions such that the estimated configurations best optimize network resources at a certain time. Thus, the selection of the right strategy is not a trivial task in the presence of the dynamic network components' reconfiguration requirements, and the high subscriber mobility in the environment. Thus, well-constructed management systems must be equipped with state-of-the-art monitoring and simulation components. The latter must be dedicated to providing a feedback mechanism based on QoS measurements and network configuration evaluation, and validation prior to deployment.

An important aspect of network management, namely mobility management and QoS measurements collection issues, is expanded upon in the following section.

1.3. THE MOBILITY MANAGEMENT PROBLEM

Mobility management in 4G not only aids the user in the sense of supporting horizontal and vertical handoff [26], where users can roam freely between different networks such that it increases the geographical span of service coverage, but also enables the networks to keep track of the subscriber's status and location to better serve its customers. Mobility management in wireless networks consists of a set of functions supported by the networks to facilitate service subscriber mobility. It generally deals with automatic roaming, authentication, and inter and intra-system handoff [27]. Mobility management provides the serving networks with the ability to locate a mobile subscriber's point of attachment for service delivery; this concept is referred to as Location Management. Mobility management is also responsible for maintaining a mobile subscriber's connection as it continues to change its point of attachment. This concept is referred to as Handoff or Handover Management. Network handover requires certain configurations to fulfill the initiation, selection, and execution phases [28]. Moreover, as in the case of PBNM, the behaviour of handoff/mobility management mechanisms is controlled by policies.

VHO has been the target of a number of research efforts (e.g., [29], [30], [31], [32], [33], [34], [35], [36], [37]). A critical constraint that most of these solutions inhibit is their inability to continuously fine-tune critical VHO configuration parameters (e.g., received signal strength (RSS) threshold for policy triggered handoffs, dwell time between initiation and execution of VHO [38], use of appropriate network cost functions, and frequency vs. periodicity of VHO initiation). To achieve self-configuration in mobility management, triggered policy and configured handover mechanisms need to provide a closed-loop feedback such that advantages of the performance indications of previously made handover decisions should be considered. As a result, this provides the ability to adapt and fine-tune the VHO operation. Such solutions are closely linked to system self-configuration in PBM and will provide an added benefit to QoS and network management systems.

1.4. QOS MONITORING PROBLEM

Network and mobility management problems are all linked to the main issue of accurately measuring network and service quality in order to provide mobile subscribers with

their required services. This is according to the quality guarantees that have been agreed upon between service subscribers and providers. QoS monitoring is receiving widespread attention in network and service management research. Typically, most multimedia streaming services and applications in existing systems have strict QoS requirements and depend on performance feedback from the network and mobile devices to determine if agreed upon service guarantees are being met. Indications of service quality degradations prompt the service providers to solve such problem either by increasing the service providers' resources or switching the user from one network to another. This is carried out utilizing different data link layer technology in situations where heterogeneous networks are available in the environment.

Earlier network management solutions depended on network resources over-provisioning to avoid the incorporation of QoS measurement collection mechanisms. With the increasing number of mobile service subscribers, and demands of mobility services today, such solutions are unpopular because over provisioning results in wasting of network resources. This results in reduced profit for network and service operators. Furthermore, resource overprovisioning acts as a poor management strategy. Therefore accurate and efficient network and service quality measurement collection schemes must be incorporated within PBNM systems to fulfill the goal of increased service provider profits and service subscriber satisfaction.

1.5. MOTIVATION

Management systems have received much attention over the past decade and much progress has been made since the augmentation of both mobile and web services. Most if not all solutions aim to provide self-configurable, self-healable, self-optimizable and self-protectable (self-CHOP) systems in the management of pervasive environments [39]. The fundamental idea is to support such requirements by providing solutions that are adaptive to varying conditions in highly mobile and heterogeneous, dynamically changing wireless network infrastructures. Today's exponential growth and variety in the number of networks, services and mobile service subscribers necessitates the automation of the management systems that abide QoS guarantees. With the availability of formal notations, standards and models (e.g. [40], [41], [42]), adaptive management systems represent an obvious solution to

the problem of network management.

Existing frameworks have certain limitations in the degree of flexibility, such that, much of the decisions taken are conducted by network administrators. These techniques fail to reinforce past results gained from previously experimented reconfiguration strategies on system and network behaviour. Most of the existing frameworks are built upon previous contextual and temporal configurations which lack the flexibility to handle changes in the underlying environments.

A great challenge for these management systems is to continuously provide adaptive and dynamic configurations, both at the management level, in terms of service agreements in the form of policies, and at the network level, in terms of component configurations. The latter can also be formulated in terms of network policies. To add more complexity to the problem, mobile wireless environments require cooperation between network and service providers to allow handoff of mobile service subscribers between different heterogeneous networks so service providers can meet the requested and agreed upon service quality guarantees. Such cooperation requires management at the mobility level and more precisely at the handoff level.

A critical constraint that most of the existing handover solutions inhibit is their inability to continuously fine-tune critical VHO configuration parameters. Thus, SPs may find many incentives to support VHO. These incentives are measured in terms of the improved access-point (AP) workloads and increased payoffs, reduction in network expansion costs by allowing providers to deliver desired coverage and service rate guarantees, while deploying fewer base stations. From the service subscriber's perspective, the benefit of SP cooperation is an 'always-best-connection' availability [31].

Since SPs are always voracious towards maximizing their profit, the main objective out of cooperation should be to increase providers' individual incomes. Research in the literature has focused on developing efficient mechanisms to ensure the accurate transformation of Service Level Agreements (SLAs) into lower configurations or policies [43] but lack the mechanisms in self-re-configurability of SLAs in dynamically changing mobile environments. Static long-term SLAs, where performance guarantees for service subscribers are fixed over time, may not be suitable for the short relationship between the users and the SPs in today's wireless environments. With today's technological advances, users must be

capable of roaming freely between different networks. Clearly, accepting more users with the same SLAs as provided to existing ones, may result in experienced service degradation and SLA violations for new users. This, in turn, will impose more penalties on the SP and reduce its profit. Hence, short-term SLAs for mobile users must be employed, in which new SLA contracts are created and configured according to current network conditions.

The optimal choice of policy and variable configurations requires a method to continuously explore the space of all possible configurations. Today's advanced meta-heuristic search strategies [44] and real-time simulators [45] provide an incentive to adapt such solutions to self-configurable management frameworks. A challenge that arises from having cooperation is the time sensitivity of the handover problem, where optimal configurations must be found within a limited time frame. Hence, a fast local search procedure must be considered. Over the last decade, the tabu-search (TS) [46] heuristic approach has been one of the most used meta-heuristics for solving optimization problems due to its fast search performance. Nonetheless, modifications must be applied to such search algorithms in areas of discrete variable selection problems to acquire the optimal value of each policy configuration variable. Moreover, the configuration learning process must be decoupled from that of the actual configuration adaptation step to avoid unnecessary time delays in the management process.

Finally, the instability of mobile environments and the high dynamicity of networks where nodes are regularly added and removed at short notice, add more challenge to the already comprehensive task of QoS monitoring. This compels us to devise algorithms for management that select nodes for monitoring purposes that are stable and responsive. Monitors must be selected to reduce resulting traffic and decrease overall transmission power cost. With the current trend of node clustering algorithms in different scopes (e.g. [47], [48], [49], [50], [51], [52]), monitoring mechanisms are emerging to better contribute to network performance evaluation in circumstances where network administrators recourse to using estimates of network traffic when configuring different components.

The aforementioned limitations of current management frameworks exemplify a set of strong motivators for developing a novel simulator-assisted multi-component network management architecture. The management framework must include a distinct variable

configuration selection mechanism and a mobility-aware monitoring system. Moreover, the mechanism must perform the above responsibilities in a continuously changing environment with minimum human intervention.

1.6. DISSERTATION OVERVIEW

This dissertation approaches the issue of network management from two perspectives; the QoS monitoring and performance evaluation perspective, and the adaptive management of dynamically changing configurations perspective.

QoS Monitoring and performance evaluation has the task of efficiently collecting end-user service quality measurements. The mechanism must consider and adapt to the dynamicity and instability of mobile wireless networks. Furthermore, mobile node semantic similarity, residual power levels and session duration must be considered when developing the quality collection algorithm. This will provide low overhead which reduces the total network traffic and provides power efficiency by reducing the total transmission power. An ontology-based semantic similarity decision method can evaluate the sameness between mobile nodes in order to group similar nodes together. Node relative velocity is also considered when constructing node clusters. Evaluation of the collected data is performed at the service provider side where service and network specific reconfigurations are adapted in response to the provided service and network quality level changes.

Management of dynamically changing configurations in the form of network policy variables is approached in a dynamic process where such variables are configured using mechanisms which learn from previously applied adaptation strategies. Using tabu-search meta-heuristics in conjunction with real-time network simulators provides a time-efficient mechanism to modify policy configurations for managing subscriber handover and service level agreements.

Incorporating these aspects is achieved through a multi-component framework. In the first component, the necessary data pertaining to user-specific service quality is collected and evaluated. This information is reported to the second component in which configurations and specifications are observed using search and simulation methods. These new configurations are provided to the third component in which service level specifications and handover policies are dynamically reconfigured to provide a smooth service delivery to the mobile

service subscribers. The objectives of the presented research work can be summarized as follows:

- ***Minimize Human Intervention***: human intervention should be kept at a minimum. This increases the system's management agility and guards network administrators from unnecessary details.
- ***Handling Mobility***: the system must react to high mobility issues such that services will be provided seamlessly to users.
- ***History Repository***: the system must process the ability to use and apply solutions from past experiences of previously applied decisions.
- ***Maximize Service Provider Profit***: the key goal for any profit-seeking enterprise is to maximize revenue of the underlying service providers. This objective is subject to adhering to the service level requirements requested by service subscribers.
- ***Efficiency***: the management system must minimize total transmission power of mobile nodes and minimize the traffic load experience from mobile node quality feedback reports.

1.7. SUMMARY OF CONTRIBUTIONS

The goal of this dissertation is to investigate new principles and design new models for service and network management architectures. The major contributions of this dissertation can be summarized as follows:

- **A Simulator-Assisted Multi-Component Network and Service Management System Architecture**

  A simulator-assisted framework is presented to determine the optimal decisions for the cooperating providers. The work introduces a novel architecture that incorporates the advantages of real-time simulation tools to continuously achieve optimal network decisions in the form of adaptive policy configurations. In policy-based management systems, the events that occur, the conditions, and the actions that must be performed are denoted into policy terms. Thus, the simulator-assisted architecture estimates the impact of different adaptation decisions and guides the decision making process of adapting the behaviour of network components, and different management systems such as handoff and service level agreements. It will provide a solution that is capable of reacting

efficiently to contextual changes in heterogeneous wireless mobile environments.

- **A Novel Tabu-Search-Enhanced Variable Configuration Selection Strategy Mechanism**

  The optimal choice of policy configurations requires a method of continuously exploring the space of all possible configurations. To perform this task, a fast local search procedure based on the tabu-search heuristic approach is, hence, proposed and applied to a discrete variable selection problem to acquire the optimal value of each variable. The new modified tabu-based search algorithm is called the "Iterated Local and Global-Tabu Search (IGL-TS)". The configuration learning process is decoupled from that of the actual configuration adaptation step. Evaluation results demonstrate that using local search techniques is considered an effective method to find near-optimal solutions to service provider profit maximization problems. The modifications applied to the search algorithm are influential on the performance, speed of convergence and running time.

- **A Mobility-Aware QoS Monitoring Mechanism**

  To monitor QoS in mobile service-oriented systems requires design decisions focusing on where and when to measure, evaluate and store the performance related QoS properties, such that the monitoring solution considers and adheres to various technical and infrastructure requirements. QoS monitoring should have minimum performance overhead that acknowledges network traffic and power consumption in mobile environments. A novel QoS monitoring system has been designed that considers a highly mobile and power-limited environment. Semantic similarity and velocity relativity have been introduced into the methodology. Node residual power availability and transmission power capability, in addition to session lifetime identification methods are considered in the developed solution. The process establishes a node clustering algorithm that guarantees power efficiency and upload-traffic minimization. Simulation tests illustrate the validity and usefulness of the QoS performance monitoring and evaluation mechanisms.

1.8. ORGANIZATION OF THE DISSERTATION

The remainder of the dissertation is organized into the following chapters:

Chapter 2 presents some of the related work that forms the background history of this work and presents state of the art research currently adopted. Some of the issues addressed by various research groups are presented. The limitations of their work in the management of wireless networks are also addressed.

Chapter 3 outlines the design of the proposed adaptive management framework. A hierarchal approach is employed to facilitate adding needed functionalities to the framework in a phased manner. Responsibilities of the different components along with their interactions are specified.

Chapter 4 presents a novel scheme for searching the network variable space for configurations that provide the optimal solution in terms of service provider profit maximization, base station load balancing and adaptation time.

Chapter 5 describes the robust scheme used for QoS monitor selection. The method is divided into different specifications required to achieve the aforementioned objectives. Simulation results are also presented to demonstrate the performance of the proposed scheme.

Finally, Chapter 6 summarizes the research work and presented contributions, and discusses directions for the future research plan.

CHAPTER 2

BACKGROUND AND STATE-OF-THE-ART WORK

The advances in network connectivity and wireless technologies have made a huge impact on our daily lifestyles globally, and along with other factors, these advances have dramatically changed the way people communicate and work. Service and network management is a process that requires the collocation of a number of QoS-driven management architectures to provide quality guaranteed services that meet mobile users' requirements. A considerable amount of research has been conducted in the field of network and service management and other related topics.

This chapter presents a background overview of wireless mobile communication systems from the first generation communication networks to the fourth generation broadband networks. The work presented in this dissertation, builds and extends upon such technology to provide an efficient solution to adaptively and dynamically automate network and service related configurations. Moreover, a focus is placed on a number of other related topics including mobility management and the basic handover procedure including various topic-related definitions. This chapter also provides readers with an overview of the active research initiatives in the areas of policy-based and QoS management in heterogeneous networks. Terminology related to the aforementioned subjects is looked at along with the different levels of abstractions for both policies and QoS hierarchy. Specification languages and translation mechanisms are also discussed in this chapter. Finally, the chapter summarizes the limitations of the presented related work and the novelty of the proposed work.

2.1. MOBILE COMMUNICATION SYSTEMS

2.1.1. Cellular Technologies

An increase in demand for wireless services has been observed with the introduction of

cellular communications. The fast increase of network traffic globally is also caused by the increase in the bandwidth requirements of mobile service subscribers. Motivated by this increase in demand, the past decades have witnessed rapid evolution of wireless networks. The first generation (1G) systems were the original analog mobile voice networks. The second generation (2G) systems that emerged almost a decade later were based on digital technologies for voice and data traffic for mobile users. A decade later, the third generation (3G) systems provided hi-speed digital mobile voice, data, and multimedia services. Finally came the fourth generation (4G) systems designed to provide higher data-rate and lower latency. The advances of the mobile communication technologies are illustrated in Figure 2.1.

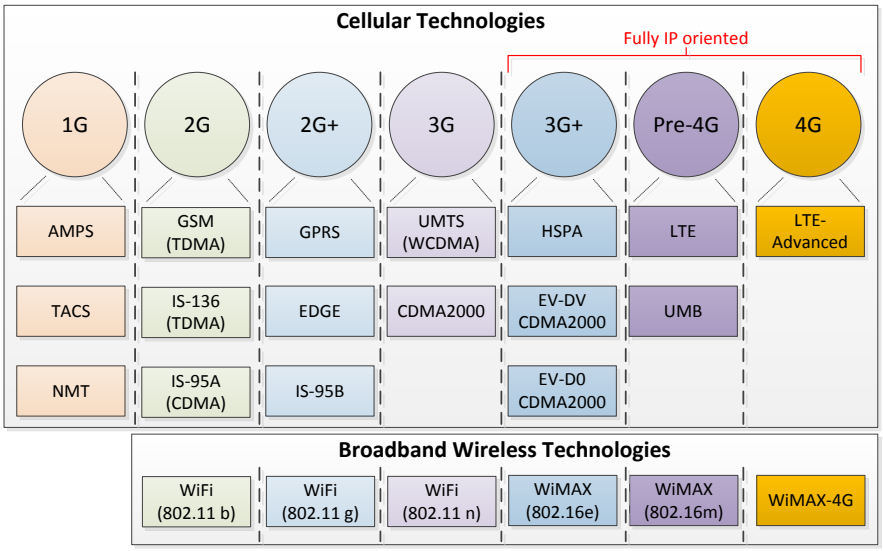

Figure 2.1 Wireless technology Advances

First Generation (1G)

The main 1G standards which were debuted in the late 70's and early 80's included *Advanced Mobile Phone System* (AMPS), *Total Access Communication System* (TACS), and *Nordic Mobile Telephone* (NMT) [53]. Although 1G was confronted with problems such as poor quality of transmission, insufficient security and inefficient capacity of the available frequencies to perform in a more cost effective way, it was a breakthrough in the telecommunication world which provided a revolution towards the more advanced systems seen today.

Second Generation (2G)

Digital system capabilities were introduced for cellular mobile systems in the early 90's. New services other than voice delivery were boasted to subscribers within a variety of technology platforms which used the spectrum in a more efficient way. The main issues that initiated the deployment of 2G were capacity, spectrum utilization, and the use of digital techniques for transporting either voice or data content. Some of the benefits of 2G included increased capacity over analog, reduced infrastructure and subscriber costs and encryption. A few of the main cellular standards developed in this generation were:

- GSM (*Global System for Mobile Communications*) is a European standard that has achieved success all over the world. It uses *Time Division Multiple Access* (TDMA) which enables multiple users to occupy the same channel.
- IS-136 was developed in the United States using TDMA technology which allowed up to six users to operate on the same physical radio channel at the same time. It provided increased system capacity, improved protection for channel interference and authentication over its predecessors. Data rates of up to 30 Kbps were allowed on IS-136.
- IS-95, also known as *Code Division Multiple Access* (CDMA), which was adopted by many system operators in the U.S. and Asia, enables multiple users to occupy the same frequency spectrum concurrently. Each subscriber utilizes their own code to differentiate themselves from other subscribers. Two categories are available in IS-95, namely, IS-95A and IS-95B. The difference between the two is that IS-95B supports data rates of up to 115 Kbps using up to eight channels, while IS-95A supports data rates of up to 14.4 Kbps only. Due to its faster data rates, IS-95B is categorized as a 2.5G or 2G+ technology, which is the next generation transitional technology between 2G and 3G. 2.5G enabled operators to deploy digital packet services prior to the availability of 3G platforms.
- *General Packet Radio Service* (GPRS), which is also considered a 2.5G technology, is a packet-switched solution that provides a more efficient data service at higher speeds than the early 2G circuit-switched solutions such as GSM, with theoretical data rates providing support of up to 171 Kbps.
- *Enhanced Data Rates for Global Evolution* (EDGE) is another 2G+ technology used

to enhance the data throughput capabilities of a GSM/GPRS network. EDGE can theoretically support speeds of up to 384 Kbps and operators can support three times more subscribers than with GPRS, thus categorizing it as a 2.75G technology.

Third Generation (3G)

With the new millennium, a rapid increase in demand of IP data services introduced new challenges for the wireless industry. The requisite to use a multitude of technologies covering many frequency bands, channel bandwidths and higher uplink and downlink data rate speeds was proposed in 3G infrastructure platforms. With the high-speed data utilization accessibility in 3G, true multimedia capabilities were offered for wireless mobile subscribers. Some of the 3G applications that were available to subscribers include wireless internet and e-mail, wireless telecommuting and commerce, and location-based services. Some of the platforms developed in this generation were:

- *Universal Mobile Telecommunications Service* (UMTS) which uses *Wideband Code Division Multiple Access* (WCDMA) as the underlying air radio interface. It was first launched in Japan in 2001 and provides a high data rate of up to 2 Mbps. It specifies services to be defined into four different classes: 1) *Conversational*, characterized by low delay tolerance, low jitter and low error rate, 2) *Interactive*, which consists of a request/response type transaction such as web-browsing. This class is characterized by low error rate tolerance, large delay tolerance and jitter. 3) *Streaming*, which is a real-time type transaction and has low error tolerance and high tolerance for delay and jitter. Applications use buffers to store incoming data so that it can be played to the subscriber in a synchronized manner. Audio and video streaming are typical examples of this category. 4) *Background*, which is characterized by having little or no delay constraint. A typical example of this category is *Short Message Service* (SMS).
- *CDMA-2000* is an enhancement and extension of IS-95 CDMA which is capable of transporting wireless services supporting multimedia applications in both fixed and mobile environments. Earlier releases of CDMA-2000 are deployed on existing IS-95 channels with numerous enhancements such as 144 Kbps data rates, twice the increase in voice capacity and improved handoff mechanisms. Later releases of

CDMA-2000 which are categorized under 3G+ systems include *Evolution Data Voice* (EV-DV) and *Evolution Data Optimized* (EV-DO) which use separate channel frequency carriers for data and offers up to 2.4 Mbps downstream and 153 Kbps upstream for EV-DO and 4.8 Mbps downstream and 307 Kbps upstream for EV-DV.

- *High Speed Packet Access (HSPA)* was developed to improve the performance of UMTS and is categorized as a 3.5G technology. HSPA is a packet-based data service feature of the WCDMA standard and is an amalgamation of two protocols, namely, *High Speed Downlink Packet Access* (HSDPA) and *High Speed Uplink Packet Access* (HSUPA). HSDPA provides improved downlink data rates reaching a theoretical peak rate of 14.4 Mbps. On the contrary, HSUPA provides improved uplink data rate speeds of up to 5.8 Mbps and enhanced QoS features.

While new systems are being deployed such as pre-4G and 4G, it is widely believed that as was the case for 2G networks, 3G networks will continue to be deployed globally and will probably exist in most major telecommunication markets within the next decade.

Fourth Generation (4G)

The evolution of 3G networks towards 4G continued with the introduction of the *Long-Term Evolution* (LTE) radio interface and *Ultra-Mobile Broadband* (UMB). This evolution was enabled by advancements in technologies and services available for mobile systems. New regulations of spectrum use and market aspects of mobile systems have also played a big role in this evolution. Three of the key service-related design parameters for the progression towards 4G systems include data rate, delay and capacity. The ever increasing demand for higher peak data rates of close to 1 Gbps and more is due to services such as media streaming and file transfer. Interactive services such as real-time gaming require very low delay. Spectral efficiency as it relates to total data rate provided from each deployed base station, is a key factor to supplying increased subscriber QoS. These main design patterns influenced the development of UMB and LTE.

- UMB also known as EV-DO Revision C is clearly an extension to CDMA-2000 EV-DO which incorporates advanced antenna techniques such as *Space Division Multiple Access* (SDMA) to provide the 4G requirements described earlier. UMB supports downlink and uplink data rate speed of up to 275 Mbps and 75 Mbps.

- LTE uses *Orthogonal Frequency Division Multiple Access* (OFDMA) for its downlink transmission which provides a broadcast/multicast transmission with flexible bandwidth and a high degree of robustness against channel frequency selectivity. On the contrary, LTE uses Single Carrier Frequency Division Multiple Access (SC-FDMA) for its uplink which avoids interference between uplink transmissions from different terminals within a cell. LTE provides peak theoretical downlink and uplink throughput rates of up to 300 Mbps and 75 Mbps respectively. It provides improved support for mobility of terminal speeds up to 500 km/h, cell sizes ranging from a radius of 10 meters to 100 km, and supports more than 200 active data mobile subscribers simultaneously. LTE-Advanced is an evolutionary step in the development of LTE which provides wider bandwidth through the use of advanced antenna techniques and the aggregation of multiple carriers. It introduces support for relaying [54] and provides improved inter-cell interference coordination in heterogeneous network environments. Self-CHOP capabilities are present in LTE-Advanced networks to support automatic and autonomous network configuration and operation.

Fourth Generation networks, and more precisely LTE, is a very flexible platform and will most definitely evolve towards better solutions to meet the future needs of wireless communication. During the time of writing this dissertation, there is already talk of the development of the next major phase of mobile telecommunications' standards that are beyond 4G standards. No official name has been specified for such platforms, but most likely, it will be named 5G.

2.1.2. Wireless Broadband Technologies

As is the case for the latest cellular technologies available in 4G, wireless broadband technologies which are part of the mobile wireless technology structure (i.e. 1G, 2G, 3G and 4G), provide ubiquitous mobile access to service subscribers and mobile users. These technologies provide users with a wide range of mobility-enabled multimedia services. As the Internet expands to offer services to billions of subscribers, the development and structure of broadband technology has dramatically advanced to offer high speed Internet access with a multitude of services.

Wired broadband access was the first to emerge into the networking world, and has dominated in providing rich performance applications such as video on demand at speeds of Gigabits per second. However, with the enormous availability of mobile devices, such solutions are not fully acceptable. Wireless access technologies have evolved rapidly over the decades to provide the same capabilities that wired connections have, with the advantage of having wireless coverage to allow mobile devices to roam freely within the environment.

There are various broadband wireless access technologies that are, in most cases, classified according to their coverage area and performance capabilities towards user mobility. A *Wireless Personal Area Network* (WPAN) is a wireless data network limited to a few meters and used for communication among devices in proximity. Examples of WPAN technologies include *Bluetooth* (IEEE 802.15.1) [55] and *Ultra Wideband* (UWB) [56]. A *Wireless Local Area Network* (WLAN) is a wireless data network limited to small proximities such as home or office environments. The most well-known WLAN technology is *Wi-Fi* (IEEE 802.11). Finally, *Wireless Metropolitan Area network* (WMAN) connects a number of wired and wireless LANs and mobile users allocated over a large geographical coverage area. WiMAX (IEEE 802.16) is a famous example of WMAN technology. Below, is a synopsis of some of the main wireless broadband standards developed, and their evolution:

- *Wireless-Fidelity* (Wi-Fi), also used nowadays as a synonym for WLAN since almost all WLAN solutions conform to IEEE 802.11 standard, provides a broadband connection in hot-spot locations that include airports, restaurants, malls and hotels. In addition, almost all homes and office buildings have a WiFi connection. Following in the steps of cellular technologies, WiFi has gone through a number of phases to become one of the most widely used wireless communication technologies globally. Initially introduced in the late 90's, the IEEE 802.11b relied on *Direct Sequence Spread Spectrum* (DSSS) transmission technology with a data transmission rate of up to 11 Mbps operating in the unlicensed 2.4 GHz band. Many devices operate in the 2.4 GHz range such as Bluetooth devices, cordless telephones, radio equipment, microwave ovens and wireless keyboards, causing interference to 802.11b devices. In the same year, IEEE 802.11a was also introduced, relying on *Orthogonal Frequency Division Multiplexing* (OFDM) transmission that operates in the 5 GHz band, thus

providing increased theoretical speeds of up to 54 Mbps. The disadvantage of using IEEE 802.11a which uses the higher carrier frequency over IEEE 802.11b, is the shorter transmission range. Thus, IEEE 802.11b has a higher transmission range at low speeds, while 802.11a has higher transmission speeds at a shorter transmission range. A third modulation standard was introduced with the new millennium, IEEE 802.11g, which uses the advantages of both OFDM and the unlicensed 2.4 GHz band. It provides theoretical speeds of up to 54 Mbps with transmission range of up to 140 meters. With the deployment of 3G networks, an improvement to the previous standards was introduced by using *Multiple Input Multiple Output* (MIMO) antennas, named IEEE 802.11n. This technology provides a maximum theoretical throughput of up to 540 Mbps with a broader transmission range of up to 250 meters. The capabilities of WiFi are being reinforced day by day to support higher data rates, longer transmission range and better QoS support. New standards are waiting to be deployed such as the IEEE 802.11ad that uses a tri-band solution operating at 2.4/5/60 GHz bands to achieve theoretical speeds of up to 7 Gbps.

- *Worldwide Interoperability for Microwave Access* (WiMAX) or IEEE 802.16 is a metropolitan area network standard, with important differences from WLAN. These include its capability to operate in point-to-point and point-to-multipoint to build bridges between two locations in the first case and deliver internet access and telephony services in the latter case. Also, 802.16 defines in detail how to ensure QoS mechanisms for applications like *Voice over Internet Protocol* (VoIP). The IEEE 802.16e standard was released in late 2005 with enhancements on all layers of the protocol stack including true mobility functionality for wireless devices in between networks, enabling devices to roam through the network with handover capabilities. IEEE 802.16e allows for advanced antenna diversity schemes with MIMO technology that use *Scalable Orthogonal Frequency Division Multiple Access* (SOFDMA). This results in selectable channels' bandwidths ranging between 1.25 MHz and 20 MHz. Further refinements led to the release of 802.16m (also called IEEE 802.16 4G) to satisfy the needs of 4G data networks in providing data rates of up to 1 Gbps for low mobility users. This is implemented by aggregating multiple channels.

2.1.3. Heterogeneity in Wireless Access Technologies

Based on the brief outline of different wireless access technologies presented earlier in this dissertation, it is clear that 4G mobile networks consist of multiple heterogeneous networks. This heterogeneity will provide ubiquitous coverage and seamless mobility by allowing mobile users to switch their connection among the different access technologies to provide services according to users' preferences. Each access technology has its own advantages and disadvantages, such that not a single access technology can or will ever replace all other existing and upcoming technologies, thus requiring handover between heterogeneous networks. Therefore, combining the advantages of diverse network technologies to get the best service possible is achieved through *interworking*, which connects multiple distinct access networks. This will allow mobile users to access the interworked networks to maintain service continuity and sustainability. Many other benefits result from using multiple heterogeneous networks in the environment. This can be implemented by extending the radio range or allowing users to connect to other access technologies when one is highly loaded to have the equivalent services and achieve higher QoS levels. This happens when the current access technology cannot support such requests. Such motivations have provoked many researchers to introduce efficient mobility management schemes. These schemes will be addressed later in this dissertation.

2.2. MOBILITY MANAGEMENT IN HETEROGENEOUS NETWORKS

Due to the advances and diversity of network technologies and the rapid growth in the number and sophistication of service subscribers' mobile devices, mobility management is one of the major challenging and complex issues in the wireless mobile communication domain. Two sub-management problems are classified under the mobility hierarchy: *Location Management* and *Handoff Management*. The former's responsibility is to track locations of mobile users and is out of the scope of this dissertation. It involves two major subtasks: *location registration* and *call delivery*. In the location registration approach, mobile subscribers periodically inform cell stations of their current location. Today's mobile service subscribers' requirements have stemmed from new mobile location identification mechanisms (e.g. [57], [58], [59], [60], [61], [62], [63]). Most importantly are the fast and accurate mobility prediction techniques which have become an important topic for current

researchers. There have been several attempts to address the issue of mobility prediction (e.g. [64], [65], [66], [67], [68], [69]). Such approaches are based on the use of historical movement patterns, environment and user contextual information to calculate possible future locations. The latter subtask of location management is invoked after localizing or identifying the location of a mobile subscriber. Thus call delivery, which is also known as paging, is responsible for successfully delivering a call or a particular service. Location management responsibilities also include minimization of signaling overhead [70] and meeting the guaranteed QoS level.

On the contrary, handoff management, the other sub-management problem under the mobility management hierarchy, is the task of providing an active connection for a mobile service subscriber when moving or switching from one access point to another. This problem will be discussed in detail in the following sections of this chapter, and a novel simulator-assisted handover scheme is introduced in Chapter 3.

2.2.1 Handover Terminologies

Handover is a key component to maintaining a seamless and uninterrupted service to mobile users. Handover management has been the focus for many researchers as an attempt to tackle all relevant technical issues in this field. A large number of solutions have been proposed to provide seamless and automated handover procedures (e.g. [31], [32], [71], [72], [73], [74], [75]). Before discussing the handover procedure and the different layers involved in the management procedure, it is necessary to define the following concepts:

- *Horizontal vs. Vertical Handover*: Handover is the process of changing the mobile node's point of connection from one access point to another. Horizontal handoff takes place when a mobile device's point of attachment is switched within the same access technology. On the contrary, vertical handoff occurs when the origin and target access points have different access technologies.
- *Upward vs. Downward Vertical Handover*: Upward handoff occurs when a mobile device switches to an access network with a wider coverage (e.g. Wi-Fi to Cellular). Conversely, downward handoff occurs when the target access network has a smaller coverage.

- *Hard vs. Soft Handover*: In soft handoff, also known as make-before-break handover, the mobile device is allowed two or more simultaneous connections, such that the connection of the serving cell is kept in parallel with the target's cell connection. Hard handover, which is also known as break-before-make handoff, does not allow multiple simultaneous connections on a mobile device. Thus, at most one wireless connection is kept at a time, such that the connection in the serving cell is released and then the connection in the target cell is engaged.
- *Personal vs. Terminal Mobility*: Personal mobility refers to the user's ability to access mobile services from anywhere, anytime, using any terminal. Terminal mobility, on the other hand, refers to the user's ability to use his/her terminal to move across heterogeneous networks while having access to the same set of subscribed services.

2.2.2 Handover Procedure

There are three stages in the handoff process: 1) handover initiation, which includes cell discovery and measurement 2) handover decision, and 3) handover execution. Details of each stage are discussed further in the following sub-sections.

2.2.2.1. Handover Initiation

The initiation of a handover can be triggered by many factors such as users' devices, network conditions, network agents, and so on. Thus, it is crucial to first identify the need for a handover by taking both user and network measurements. Examples of network measurements are *Bit Error Rate* (BER), throughput, delay, jitter, packet loss, etc. On the contrary, user perceived measurements are application specific. A well-known user measurement that many handover decision algorithms use is the *received signal strength* (RSS) (e.g. [76], [77], [78], [79], [80]). Signal strength is averaged over time so that fluctuations due to radio propagation are eliminated. Other information such as terminal capabilities, context and surrounding environment information are also gathered and reported to the network in an effort to assist in the handover decision. Neighboring or available cells are discovered by the mobile device by scanning different channels or through provisioning information sent by the device's current base station. Once this information is gathered, the

next stage involves making a network selection.

2.2.2.2. Handover Decision

As stated earlier, the handover decision stage is responsible for determining when and how to perform the handover and to which access network the user will switch to. The measurement results from the first stage will indicate if a handover is required. If so, the next step is to determine the best access network to switch to among the available networks. Different criteria must be evaluated either by the user or network before selecting the best or optimal one. Various selection strategies exist [81], including solutions that consider user preferences and roaming or interworking agreements stated in an SLA. Once a decision has been approved, a handover will be initiated either vertically or horizontally. Certain pre-registration information of the mobile subscriber will be relayed to the target BS for handover preparation. The next step is to execute the handover of the mobile subscriber.

2.2.2.3. Handover Execution

At this stage, establishing a connection with the new access point is accomplished by authenticating the mobile subscriber with the target network. This is accomplished using proper user credentials through valid encryption keys. Delivery of the data from the old connection path to the new connection path is achieved using a variety of methods [82] [83]. *Mobile IP* (MIP) [84] enables such packet delivery from one network to another. MIP introduces two entities: 1) *Home Agent* (HA), which controls the movements of subscribers' mobile nodes and 2) *Foreign Agent* (FA), responsible of handling a mobile node that has transferred to its network. When a mobile node moves to a foreign network, a *care-of-address* (CoA) is provided to the mobile node by the foreign agent, which is an IP address. The mobile node will then register the CoA with its home agent. It will now receive all packets destined to the mobile subscriber at its home address instead of the mobile node's address. The HA will encapsulate and tunnel [85] the packets to the mobile node's CoA. The FA will then receive and encapsulate the packets and send them to the mobile subscriber's node. Further details are illustrated in Figure 2.2.

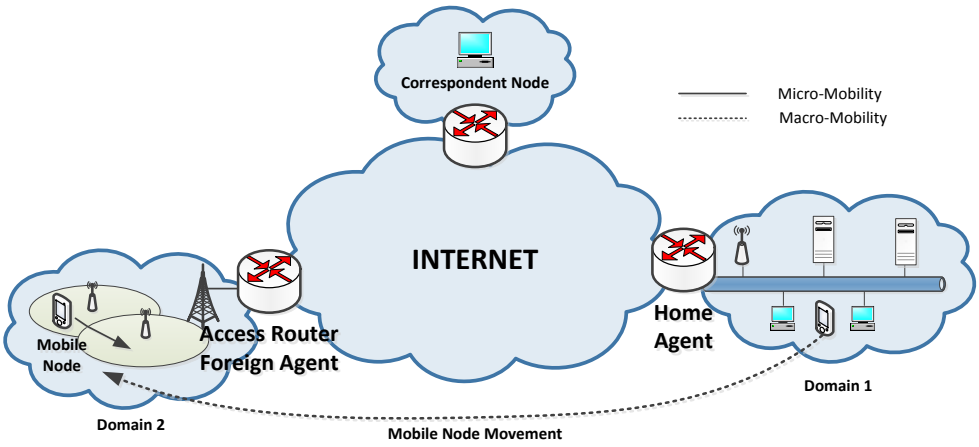

Figure 2.2 Mobile IP Architecture.

2.2.3 Mobility Management at Different Layers

Mobility management in heterogeneous networks is more complex than in homogeneous networks. Management mechanisms involve different layers of the TCP/IP protocol stack. Several mobility management protocols have been presented for next generation IP networks. These mechanisms are classified according to the communication protocol layer involved in the solution.

- *Link Layer Mobility Management* – these types of solutions are wireless technology specific, which allow mobile roaming among different physical points of attachment while keeping the point of attachment to the IP network unchanged. Thus, no IP sub-network configuration is required when moving from one point of attachment location to another.
- *Network Layer Mobility Management* – these types of protocols communicate using IP layer messages and are not specific to the underlying wireless access technologies. This category is further classified into two sub-categories: *Macro- and Micro-Mobility*, illustrated in Figure 2.2. A local movement of a mobile terminal within a single administrative domain is referred to as micro-mobility. While large-step movements of mobile terminals across different domains can be referred to as macro-mobility and are independent of the underlying routing protocols, access techniques and other mechanisms. Under macro-mobility, when a mobile node moves between different domains, its IP address will change as well. Thus, as described earlier, to

maintain node connectivity, mobile IP is used to solve this problem. MIPv4 and MIPv6 [86] [87] [88] are two well-known mechanisms, in which MIPv6 includes features for mobility support such as the exclusion of FAs and address translators which are a must in MIPv4. In addition to the CoA, MIPv6 includes a *Home Address* (HoA) used as the mobile node's identifier. Besides MIP new mobility management, protocols have been proposed which are beyond the scope of this dissertation. These include, as examples, *Host Identity Protocol* (HIP) [89] and *Internet Key Exchange Protocol version 2 Mobility and Multi-homing* (MOBIKE) [90] [91]. To manage mobile handoffs within a single domain, a number of micro-mobility protocols have been developed. Such mechanisms have the added benefit of signaling overhead and handover latency reduction over the MIP protocol. Examples of micro-mobility protocols are *Hierarchical MIP* (HMIP) [92] [93], *Fast MIP* (FMIP) [94], *Intra-Domain Mobility Management Protocol* (IDMP) [95] and *Handoff Aware Wireless Access Internet Infrastructure* (HAWAII) [96]. Mobility protocols are classified as either *host-based solutions* that necessitate host involvement at the IP layer, such as FMIP and HMIP, or *network-based solutions*, that neglect a host's involvement in the handover decision. An example of a network-based mobility protocol is Proxy MIPv6 (PMIPv6) [97].

- *Application Layer Mobility Management* - MIP solutions are a network-layer mechanism which only supports seamless connectivity across subnet changes. More and more mobile internet multimedia applications require application-layer mobility support which is beyond the capability of MIP. The *Session Initiation Protocol* (SIP) [98] is a novel application-layer control (signaling) protocol used for creating, modifying, and terminating sessions consisting of multiple media streams with one or more participants. Application layer mobility management mechanisms are application-specific and do not cover a wide spectrum of applications.
- *Cross-Layer Mobility Management* – several state of the art cross-layer solutions that combine network- and application-layer protocols have been investigated in [99] [100] [101] [102]. These protocols provide network layer handoff with the aid of communication and signaling from the link layer. The benefits of such protocols include the reduction of both packet loss and latency.

2.3. POLICY MANAGEMENT IN HETEROGENEOUS NETWORKS

2.3.1. Policy Definition

Policy management is usually used to define rules for different user roles such as administrators, guests, etc. to restrict their access to network or system resources. Lately, policies have played a significant role in providing ways to simplify the management of many different areas in both networking- and non-networking related fields. Many of the operations required to enforce and manage such systems use similar techniques. Therefore, it is necessary to examine policies as a mechanism that can be used for different disciplines. When managing heterogeneous networks, policies must satisfy the following requirements to achieve a smooth performance:

- *Precision*: all policies have to be specified to a level of detail that can be inferred by the managed system in which no ambiguity can be interpreted for a specific policy. To enforce a certain policy, all requisite details must be specified within it.
- *Consistency*: all policies provided to manage a system must be consistent such that the managed system elements must be configured in such a way that the enforced policies are properly interpreted.
- *Compatibility*: the set of policies enforced on a system must be compatible with the capabilities that the system can support.
- *Mutual consistency*: when multiple policies are specified for a system, each must be consistent with the other in a way that the system is able to use these policies in an explicit manner.
- *Specification easiness and intuitiveness*: policies must be simple, easy to specify and defined in terms that are user-friendly in terms of the business process at the 'high-level'.

2.3.2. High- vs. Low-Level Policies

Policies in general are described in the form of one or more rules that specify actions that have to be performed in response to certain conditions [103].

$$\boldsymbol{if} \ < condition\ 1 > \ \boldsymbol{and} \ < condition\ 2 > \ldots \ \boldsymbol{and} \ < condition\ n >$$
$$\boldsymbol{then} \ < action 1 > \ \boldsymbol{and} \ < action\ 2 > \ldots < action\ m >$$

A policy rule will often comprise of other rules, such that polices can contain other sets of polices within it. Based on the adopted definition of policies, the requirements stated earlier cannot be satisfied together simultaneously. Thus, policies are described at different hierarchical levels of abstraction. Specifically, four main abstraction levels were used in this research, and these abstractions will be further discussed in Chapter 3. However, at its simplest form, policies are divided into *high*- and *low-level* policies. Each of these levels imposes different sets of requirements on the way of defining and specifying policies. High-level policies must be expressed in terms of users' perspectives (i.e. administrators, operators, subscribers, etc.). Policies at this level must be easy to specify and in a format that is intuitive to the user. The definition of high-level policies depends on the end goal of the business. For example, if a service provider wants the network to operate so as to satisfy specific *service level objectives* (SLO), then those objectives must be defined as high-level policies. This specification will be independent of the underlying technology deployed in the heterogeneous network. High-level policies are usually static compared to the state of a system, as seen in many state-of-the-art works [104] [105] [106] [107]. With today's highly dynamic network environment, this static state of business-driven service-level policies is undesirable. In this research, a novel simulator-assisted policy-based network adaptation algorithm is introduced that dynamically reconfigures policies to achieve an optimal state to provide services to mobile subscribers according service level agreements.

Policies are specified in a language that is processed and interpreted by computing devices. Many policy specification languages exist [103] [108] [109] [110] [111] [112] [113] [114] [115] [116] [117] [118]. PONDER [119] is a state-of-the-art specification language that supports policies that utilize event triggered condition-action rules, as seen above, for policy based management of networks and distributed systems. PONDER is declarative and object-oriented which makes the language flexible and adaptable to a wide range of management requirements.

2.3.3. Policy Architecture

According to the specifications of IETF/DMTF [120] [121], the policy architecture consists of four main components as illustrated in Figure 2.3:

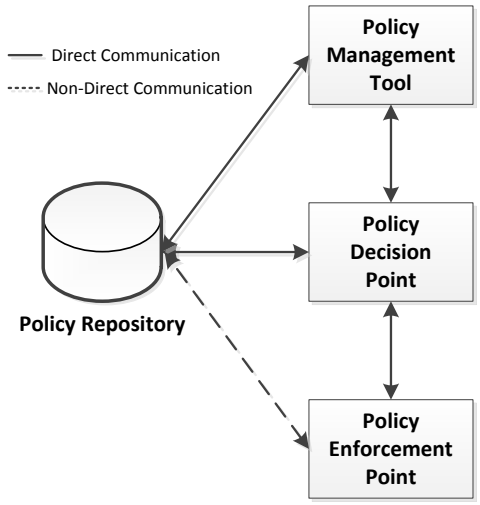

Figure 2.3 Basic components of a policy architecture.

- *Policy Management Tool (PMT)* – takes as input the high-level policies that a user enters in the system through some form of user interface for further translation into precise low-level component-specific policy description. PMT is responsible of validating the syntactic and semantic correctness of such policy inputs including policy parameters relationship validation. The validation process includes ensuring that the policies can be satisfied given the system's resources and constraints. The management tool is also responsible for storing and retrieving policies from the *policy repository*.
- *Policy Repository (PR)* – communication of components of the policy architecture are done through a policy repository which stores all the generated policies by the management tool. Both high- and low-level policies are stored at this repository. A well-known example of a policy repository is the *Lightweight Directory Access Protocol* (LDAP) directory [122] used to store policies required for network operations. Regardless of the type of repository used, all stored policies must conform to specific policy specification language rules.
- *Policy Decision Point (PDP)* – responsible of interpreting the policies stored in the repository for further communication to the *policy Enforcement Point (PEP)*. It will transform the set of rules retrieved from the PR to a syntax that is convenient for the PEP. It is also responsible of validating that any changes within a policy's conditions required for the policy enforcement are kept up to date.

- *Policy Enforcement Point (PEP)* – responsible for executing the policies defined by the PDP. It also has the responsibility of monitoring the system's state and reporting it to the appropriate system's components, including the PDP.

2.3.4. Policy Translation and SLA Support

High-level policies are translated into lower level policies that are interpreted by each system's components and devices. At this level policies are specified precisely and consistently for each component in the system. For example, the low-level policy view for a heterogeneous network with handover capabilities describes the amount of time for service description required before initiating an actual handoff. To bridge the gap between the business needs and the technology being deployed in a system, policy at the SLA-level must be translated into network-level and device-specific parameters. To perform such translations, it is necessary to have a representation of the low-level polices that are required at each PEP within the system. Many authors [123] [124] [125] [126] have provided policy translation solutions to domain-/application-specific systems which are beyond the scope of this dissertation.

A *Service-level Agreement* (SLA) is a statement or agreement between the service provider and the service customers that states the obligations and expectations that exist between both parties to assure that the provided services fulfill the requirements. It contains details such as the type of service provided, the expected performance level of the service, the time frame in which a problem is resolved if a problem occurs, and the process for monitoring the service level. An SLA is an aggregation of multiple performance objectives, such that an SLA refers to the overall business agreement and consists of several *service-level objectives* (SLO) which require satisfaction. SLAs are translated into high-level policies which are further translated into low-level policies, and so on. Policy translation is meant to derive low-level policies from high-level requirements. It is therefore important to monitor a SLA's performance-related service properties and enforce the service's quality during a SLA's validity.

2.4. QoS Management in Heterogeneous Networks

Quality of Service is defined as a measurement of the network and computing a system's

ability to provide different levels of services to selected applications [127]. Effective management of QoS requires an understanding of the roles of the different actors involved in both managing and acquiring resources from a network or a service provider. Different groups of service subscribers have different quality requirements which result in varying rankings and levels of performance. Resource optimization requires not only knowledge of such service class distinction but rather further contextual knowledge of the subscribers' devices and surrounding environment. To manage cooperative heterogeneous networks efficiently, it is necessary to understand the characteristics of QoS that are defined according to the level of abstraction in a system. It is also necessary to know the mechanisms involved to achieve an automated and managed network such as network performance monitoring and analysis.

2.4.1. QoS Layering

Traditional QoS was defined specifically for the network layer of the OSI model. But with advanced improvement in deliverable services, the QoS concept was further enhanced to induce QoS at multiple layers. Similar to policy classification levels, QoS can be characterized into different classes: user QoS, application QoS, network QoS, transport QoS, etc. Many QoS models have been introduced in the literature [5] [7] [8] [41], in which these models change according to users, applications, system and individual device components, such that individual layers (e.g. network-layer QoS) can have other sub-layers. In this dissertation, QoS specifications are classified into upper-level and network-level QoS.

2.4.1.1. Upper-Level QoS

This layer is further subdivided into three levels:

- *Perceptual QoS* – this layer represents user-level QoS specifications that describe the perceptive qualities in terms expressed to the user's perceivable effects, such as media qualities (in terms of smoothness, color depth, brightness), response time (in terms of interactivity), security level, and pricing choices. Today, the theories of *quality of experience* (QoE) [128] and quality of security [129] have become commonly used to represent user perception. It is important for the network-provider or service-provider to be aware of the influence of each network's factor on the user

perception. In this dissertation, the focus will only be on quantitative user-level QoS specifications. QoE and quality of security will not be considered.

- *Application QoS* – parameters at this level describe requirements for application services in terms of media characteristics (e.g. frame rate, frame resolution), transmission characteristics (e.g. application-level end-to-end latency, delay variation, jitter, packet loss), and media relations, such as media transformations and frame synchronization skews.
- *Device/system QoS* – given the variety of multimedia devices, device QoS parameters specify timing and throughput demands for media data units. Such parameters include sampling size and rate, compression details, etc. Device or system-level QoS considers node or end-system resources that are required when executing a service. Such resources include CPU, memory, system bus bandwidth, coding and decoding capabilities and input/output capabilities.

Later in this dissertation, attention will be given to a monitoring technique in a heterogeneous wireless environment that considers upper-layer QoS parameters for node clustering. It is important to note that words such as user-perceived QoS, upper-layers QoS or device-specific QoS will be used alternatively when referring to QoS parameters at the user-, application- and device-level.

2.4.1.2. Transport/Network-Level QoS

Parameters at this level describe the low level network services in terms of network load and performance. Network load describes the ongoing network traffic within an average or minimal interval time on the network connection. Transport/network performance describes the network service guarantees and provides key information on performance of transport and network protocols (IP, UDP, RTP, etc.). Examples of transport/network-level parameters include bandwidth, latency, delay, jitter, bit-error or packet loss rate and connection establishment/release failure rate. Performance at this level directly affects service quality. This information can be measured either at the client-side or at several points within the network such as routers and access points.

2.4.2. QoS Monitoring

With the growth of mobile services and devices, it has become crucial for a network or service provider to receive accurate QoS measurements of its network and customers. These measurements are used to analyze faulty and non-optimal resource provisioning to improve the performance of both the network and services in a cost-efficient way. The results will help boost the provider's competitiveness and customer loyalty.

Monitoring, maintaining, and predicting a network behavior is crucial to ensure that a contracted QoS is sustained. Monitoring is considered the first step towards full network management after launching a service. Any self-managed system requires monitoring to achieve a dimension of the system's resources that will best meet the expected demands. This monitoring process will provide a determination of the number of resources that should be allocated for the class of service provided. Further analysis of the monitored network might require network cooperation [130] when the provider's resources cannot meet the requested demands. Other solutions might consider traffic classification [34] [76] [131], scheduling [132] and buffering [133] to support increased service subscriber demands. Thus, it is necessary to regularly monitor a system's QoS to maintain and provide optimal performance to service subscribers. In this dissertation a novel state-of-the-art QoS monitoring mechanism is introduced in highly mobile network environments.

2.4.3. QoS Performance Analysis

When monitoring a system, it is not just mandatory to only collect data, but to also conduct an analysis of the current system behavior. Therefore, an investigation of how traffic and resource management mechanisms deployed in the system can affect the system's performance should be performed. Self-management requires new mechanisms to analyze data and make changes to system configurations rather than being heavily dependent on decisions by human operators. Such reliance on human operators is a major contributor to costly services and significant service delays. There have been group efforts that tackle the management problem through state-of-the-art performance analysis mechanisms [134] [135] [136] [137] [138] [139]. A survey of some of the outstanding solutions will be discussed in the following section. Although such solutions have contributed to the self-management of today's heterogeneous networks, neglect in real-time

network simulation analysis has led to a lack of real self-manageable networks. For this reason, in this research a simulator-assisted framework has been developed to determine the optimal decisions for cooperating service providers in heterogeneous mobile network environments. This will be discussed in detail in Chapter 3.

2.5. STATE OF THE ART IN POLICY-BASED SYSTEM MANAGEMENT

The previous sections introduced an overview of the adopted frameworks and infrastructure for mobile service delivery. The need for a generic, integrated, and dynamic service and network management framework to provide a higher level of delivered QoS and optimize service provider resources were also shown. This section of the chapter discusses some of the related work that exists in terms of business-driven policy-based system management architectures which focus on maximizing the profit of the provider. Also addressed is cooperative profit sharing which considers user mobility. Section 2.5.1 presents some existing policy-based management frameworks that aim at maximizing service provider profit through network cooperation. Section 2.5.2 discusses mobility management architectures that incorporate vertical handover algorithms. Section 2.5.3 presents SLA management frameworks and some of the existing formal languages to define service agreements. Section 2.5.4 presents some existing simulator tools used in operating different systems. This section also considers related work in the area of variable search methods that employ meta-heuristic algorithms. Finally, Section 2.5.5 discusses the necessary background and similar approaches for QoS monitoring, node clustering and packet forwarding.

2.5.1. Business-Driven Policy-based Management and Network Cooperation Solutions

In the literature, there have been several proposals for designing systems that maximize the profit of wireless communication network providers. Such systems mainly focus on increasing the number of clients and the fulfillment of their business-driven optimization problem.

Yet, most of these solutions dismiss the surrounding rigid competition by other network providers. For example, in [140], the authors consider the problem of pricing and transmission scheduling for a service provided in a wireless network environment for mobile

users. The ambition of the SP is to maximize its individual profit. Huang et al. develop an online method that jointly solves the pricing and transmission scheduling problem in a dynamic environment to regulate incoming traffic and maximize revenue. The scheme can attain a solution that is close to the optimum, with a tradeoff in the average delay. This scheme, suffers from inefficient utilization of resources [141] and the lack of revenue increase within the competitive domain.

On the contrary, several proposals consider the concept of cooperative profit sharing between various service and network providers, while acknowledging the reality of competition between enterprises. Singh et al. [1] propose a solution to demonstrate how cooperation may enhance throughput through better utilization of resources while lowering the overall energy consumption of mobile nodes. This cooperation results in higher customer satisfaction and payoffs for the providers. The authors model the cooperation using the theory of transferable payoff coalitional games, such that the optimum cooperation strategy is computed as a solution of an integer-programming problem. The proposal provides a solution using distributed computations and limited exchange of confidential information among the service providers.

Recently, policy-based management (PBM) tools have been employed to effectively control the behavior of various systems. For example, in [142], Chai et al. employ policies to effectively configure military networks. Samaan et al. [12] present a novel paradigm to approach the issue of autonomous PBM of wired/wireless differentiated communication systems. The proposed architecture addresses the management issue by learning from the current system behavior, while creating new policies at runtime. A policy model is used to acquire users and administrators' higher-level goals into lower level objectives. Policies are created in response to changing requirements and are assembled at runtime.

A main controversy that is not addressed in most state-of-the-art profit maximization methods is the use of management policies. All network management systems require policies to control and administer their behaviour, including VHO and SLA management. Nonetheless, a demand for policy automation is neglected in most cooperation-based systems. Jia et al. [143] develop a method to optimally compute budget allocations for policy improvement to maximize a lower bound of the probability for correctly selecting the best action. The authors focus on improving the system performance from a given set of policies

instead of dynamically creating policies. This method developed by Jia et al. differs from other solutions such as reinforcement learning [18], approximate dynamic programming [144], and potential-based methods [145], which are time-consuming to run and provide noisy performance estimations.

2.5.2. Handover Management Solutions

This research focuses on the first two stages of vertical handover, which are typically approached as a multiple attribute decision making (MADM) problem [146] that deals with selecting an optimal decision among several alternative proposals; each proposal is characterized in terms of values of various attributes. VHO schemes differ in the methods used to solve the MADM problem and the set of used attributes. For example, in [78], the authors propose a method that combines two MADM techniques, namely, the Analytic Hierarchy Process (AHP) and the Grey Relational Analysis (GRA) to select the best access network. This is carried out according to the perceived QoS as defined in terms of network availability, throughput, timeliness, etc. The main advantages of their scheme are its simplicity and low computational cost. However, the user must explicitly rank the various quality attributes. To overcome this limitation, a fuzzy-logic strategy is applied to network handover in [37], [147] to deal with user preferences. This strategy also tackles the configuration of an appropriate dwell timer which must elapse before the execution of the VHO to guarantee the stability of the handover. The system provides a method to calculate weighting factors for each application. A decision matrix is then constructed for multiple triggering factors of different networks.

Various research efforts have also leveraged advantages of policies for VHO. For example, Vidales et al. [73] incorporate high-level user, application and network policies during the execution of various VHO stages. Likewise, a policy-enabled VHO model is proposed in [148]. The model considers the preference of the user and the trade-off between different characteristics of the networks (e.g., bandwidth, access cost, and power consumption). In contrast to the approach of this thesis, these schemes rely on statically defining a set of VHO configuration policies. For a comprehensive review of various VHO schemes the reader is referred to [2], [81].

The aforementioned research outlined in the previous section disregards today's heterogeneity of wireless networks and the requirement for mobile node handoff both vertically and horizontally [149] in cooperative profit sharing cases. In [150], the authors consider an environment where WLAN access points have overlapping coverage. They investigate a solution to account for incentives that can trigger handovers between the APs to improve performance for all wireless networks. A model is presented that can approximate the aggregate performance for each AP to predict when handovers between WLANs are beneficial for the parties, giving all APs the incentive to support handover. The work neglects profit gains, which are the SPs main concern in this regard.

With today's abundant collection of handover schemes [151] [152], it is becoming increasingly difficult to distinguish between outstanding handover methods, such that a wide selection of VHO algorithms focus only on when to trigger VHO to improve a connection's QoS. Those solutions neglect the problem of how one can consider the optimal choice of network for VHO from all available candidates. Lee et al. [29] provide a highlight of the metrics best suited for decision making in VHO algorithms. The authors consider metrics such as bandwidth, available mobile node battery, power consumption of Network Interface Cards (NIC) and received signal strength (RSS) for handoff decisions. Their solution seeks to optimize battery lifetime of the mobile nodes and AP load balancing. Liu et al. [153] present an analytical framework to evaluate VHO algorithms that can be used to provide guidelines for the optimization of handoff in heterogeneous wireless networks. The authors also propose a general handoff decision algorithm to trigger both horizontal and vertical handoffs in heterogeneous wireless networks at the appropriate time.

2.5.3. SLA Management Solutions

SLA management requires a robust language so that an SP can obtain, monitor and enforce QoS guarantees for its customers. Ludwig et al. [40] propose a novel formal language to define an SLA. It is used to automate the process of supervision and management of SLAs and the provisioning of corresponding systems. The language provides a large degree of flexibility by utilizing an XML-based representation and a runtime system for SLAs. Parties involved in an SLA can describe their desired guarantees with respect to QoS parameters. In [43] the authors present a systematic approach to business and policy

driven refinement. A use-case is presented in which they derive a low-level policy-based Service-Level-Specification (SLS) from a general SLA. This research provided insight on how to bridge the gap between high-level business goals and low-level management actions. Their work also aims at optimizing the business profit of an SP.

2.5.4. Simulator Tools and Solution Search Mechanisms

The selection of configuration policies necessitates the continuous use of variable configuration search methods. Well respected meta-heuristic-based search algorithms such as tabu-search [46] presented the incentive in this research to find a solution to policy and QoS management. In [154], the authors propose a solution using tabu-search to refine existing policies used in reinforcement learning. The use of tabu search provides a method for speeding the convergence of on-policy reinforcement learning [155] and provides a mechanism to escape local minimum states. Other areas in which tabu-search is used involve logistic regression classification models. Since the goal of classification problems is to classify cases that are characterized by variables, the authors in [156] use tabu search to find, from a set of variables, a smaller subset to be placed in the same category.

It should be noted that with the advances in and usage of multi-core processors, real-time simulators have become an integral part of the decision making process in many areas including manufacturing, gaming and military [157], [158]. These systems focus on run-time operations to estimate optimal decisions based on a finite number of simulation runs under different configurations. In [159], the authors present a novel approach to increase the scalability and efficiency of parallel network simulations. The model partitions the networks into domains and separates simulation time into intervals, where each domain is independently and concurrently simulated with the others over the same simulation time interval. Data is exchanged between domain simulators at the end of each interval. This real-time simulator can support simulations of large scale networks in nodes with limited resources.

Another proposal is developed by Aib et al. [160], in which the authors develop a policy simulator tool for validation and performance evaluation purposes of policy-based management solutions. The simulation environment is discrete and allows the specification of management policies, events, metrics, probes, service level objectives, SLAs, as well as

business-level objectives. However, the feedback from the simulator is not directly provided to the management entity of the network.

2.5.5. Monitoring, Clustering and Packet Forwarding Mechanisms

The problem of designing distributed QoS monitoring systems has been addressed extensively in the literature. Existing approaches differ in their target environments (e.g., wired versus wireless networks), their target measurement layers (e.g, application and network layers versus a cross layer approach). The approaches also differ in their objectives (e.g., enhancing specific aspects of QoS such as service delay and throughput versus more generic service qualities), and the network components selected to perform the monitoring process (e.g., end-users, proxies and routers).

For example, several E2E QoS monitoring frameworks have been presented in [161], [162] and [163]. These distributed approaches promote the collaboration among different entities (e.g., network routers and end-users) and domains. To reduce the incurred traffic overhead of these methods, the work developed in [164] focuses on selecting a subset of virtual nodes to perform the monitoring task in virtualized networks environments. These nodes continuously communicate with nearby nodes to collect various measurements such as CPU usage, memory utilization, queue size and service waiting times. Similarly, the authors in [165] use clustering mechanisms to group similar cells and nodes, respectively. Network performance is, then, gathered from these cells and nodes and aggregated for future retrieval by system administrators.

Unfortunately, in addition to not being designed specifically for cellular networks, those solutions did not consider the huge amount of traffic generated when sending measurements to the selected subset of nodes responsible for the measurements collection. Furthermore, the majority of these approaches did not consider the power depletion effects of the developed schemes on the users' devices.

In [166] and [167] the authors propose a mobile service management architecture that solely relies on data collected from the mobile devices, without the need to insert measurement probes in the network environment. Mobile nodes gather and aggregate data related to the monitored service performance and forward it to a QoS broker. The broker is responsible for sending the collected data to the SP and providing the clients with the

threshold values required to force a report back about QoS measurements. This method uses a historical service measurement database to archive all submitted measurements and then tunes the reporting threshold values. The work disregards issues related to network mobility, consumed node power, and traffic overhead.

Jurca et al. [168] propose a QoS monitoring mechanism based on client feedback, where each client runs a monitoring code and periodically reports its service ratings to a trusted center. The center aggregates the reports and estimates the delivered service quality for each provider. The authors focus on ensuring honest feedback from the clients. Since clients may tamper with the default monitoring code, special incentives are given to guarantee reliable QoS reporting. On the other hand, the developed scheme neglects the issue of the excess traffic overhead introduced on the network when reporting about the experienced clients' service quality.

In [169] the authors present a broker-based architecture for monitoring web service quality. The broker is responsible for generating many instances of replicated clients that invoke the web service. After monitoring the clients' behaviour, the web services broker assists clients in selecting web services that best match their demands. The drawback of such a solution is the cost of adopting the broker within network management architectures and its single point of failure.

Michlmayr et al. [170] use a combination of both client- and server-side QoS monitoring. Clients use a tool that sends probe requests periodically to the services available in the server and react to any service quality violations by informing the service provider. The provider, in essence, compares its measurements to that of the clients. Corrective actions are then taken if the results are similar. The work focuses more on the design of QoS measurement tools to be installed on the clients' devices rather than on developing QoS monitor selection techniques. The work also disregards the effects of the incurred traffic overhead and client mobility.

Clegg et al. [171] address the problem of provisioning monitoring nodes within highly dynamic virtual network environments. Their research focuses on selecting nodes which reduce monitoring and management traffic on the network. In the proposed scheme, traffic is regularly sent from all nodes to a small subset of nodes chosen according to two criteria: placement and longevity. The selected nodes are placed close to the nodes they communicate with. The algorithm calculates the amount of traffic which would be removed from the

network when a certain node is elected as a leader (i.e. monitoring node). In addition, selected nodes need to be stable and so as a result, they must remain connected to their neighbors. The work presents an algorithm named Pressure which places monitoring nodes in a dynamic network. Node longevity estimation is achieved using an improved Kaplan-Meier lifetime estimator method.

The authors in [135] present an architecture for large scale monitoring in distributed virtualized environments. The authors propose an event based monitoring framework used to measure system performance such as the physical and virtual CPU usage, memory utilization, queue size and waiting time, throughput and power consumption. Monitoring sensors are deployed on each node to report the system's performance at different periodic sampling times.

Node clustering has been used in many areas to overcome certain contextual limitations. For example, much state-of-the-art research [47] [48] [49] [50] [51] [52] uses clustering to provide data aggregation, traffic overhead reduction between neighboring clusters, traffic overhead reduction between clusters and sink nodes, along with many other objectives. Clustering has been widely applied in both *wireless sensor networks* (WSN) and *mobile ad-hoc networks* (MANET). Different clustering techniques have been used, in which nodes are grouped within the same cluster according to objectives such as: load balancing, energy efficiency [172], and data similarity.

In [173] the authors presented a distributed clustering method used to reduce energy consumption in multi-hop data delivery within WSNs. The algorithm determines the suitable size for clusters depending on the hop distance to the sink node. Performance results demonstrate the effectiveness of the clustering algorithm in extending the network lifetime and providing equalization of node energy levels. Although the presented clustering algorithm saves energy for the inter-cluster communications, the adopted method produces a significant amount of summary packets which result in heavy traffic loads. The cause of such an issue is that the authors adopt an unequal clustering mechanism that produces smaller sizes of clusters for nodes that are closer to the sink node.

In [174] the authors present a clustering method used for the management of self-monitoring distributed systems. The algorithm requires nodes to periodically publish their status in terms of events. Node clusters are created based on the relative density of points

within the environment. The system's cluster size is directly proportional to the point density of an environment. The algorithm also considers when nodes join and leave a cluster and anomalies. Evaluations were conducted on network devices to monitor and report device usage levels and to find a relationship between devices and users.

Packet forwarding is adopted within this research's QoS monitoring architecture to reduce transmission power consumption when sending data from the cluster-head (i.e. monitor) to the BS. A large number of forwarding node or relay node selection strategies have emerged lately, especially with the introduction of LTE networks [175] [176] [177] [178] [179] [180] [181] [182] [183] [184] [185]. Many of these mechanisms aim to extend the network coverage area, reduce the total consumed power of mobile nodes, and relay data back to the sink node when a mobile node is out of reach, given the limited power supply of the devices. For example, the authors in [186] consider a wireless cooperative cellular network with a BS and subscribers. The subscribers have the ability to relay information to each other to improve the overall network performance. The authors propose a solution that selects the best relay node when transmitting data from the subscriber to the BS. At the same time, consideration is given to user traffic demands and physical channel realizations in order to allocate power and bandwidth optimally for each user. Both amplify-and-forward and decode-and-forward relaying strategies are considered when forwarding data to the BS.

Despite the abundant existing studies on service performance measurement techniques, to the best of the authors' knowledge, the presented work is the first to employ node stability and similarity mechanisms to select QoS performance monitors. The mechanism is utilized for wireless network environments employing highly mobile and heterogeneous users. The algorithm aims at reducing the overhead traffic incurred from client-side performance reporting through clustering. Moreover, the solution also reduces the overall transmission power consumption since mobile nodes have limited power availability.

2.6. SUMMARY

In this chapter, a global view of the mobile network evolution has been addressed from the first generation towards today's 4G systems characterized by interworking of heterogeneous access networks. One of the key factors for the success of such a convergence

is the introduction of policy-based mobility- and QoS-management. An overview of the features and necessary background information required to enhance such self-management architectures, was presented. A clear trend is emerging in the form of wireless mobile location-aware service availability that requires interworking of heterogeneous networks and presence of state-of-the-art self-manageable network architectures. The policy-based simulator-assisted network management approach provides an improvement to existing solutions. In the following chapter, the focus will be on the design of the proposed simulator-assisted management framework and on research initiatives in the area of self-management in heterogeneous networks that led to the design of the proposed framework.

Despite the abundant existing studies on policy-based management, handover mechanisms, heuristics-based search algorithms for variable configuration selection and real-time simulators. To the best of the authors' knowledge, the presented work is the first to employ tabu search for variable selection to automate the policy management process jointly with the assistance of real-time simulators. This is done with the objective of offering improved performance and profit gains for SPs. This work also incorporates an efficient QoS monitor selection mechanism in an environment that is totally independent of network-inserted monitoring probes.

CHAPTER 3

A SIMULATOR-ASSISTED MULTI-COMPONENT NETWORK AND SERVICE MANAGEMENT ARCHITECTURE

The previous chapter introduced some of the related work that exists in regard to policy-based network management, in general. The need for a generic and dynamic network and SLA policy parameter configuring system was shown previously. Such system guarantees that both service providers and subscribers acquire maximum satisfaction in terms of providers' profit maximization, resource optimization, and subscribers' service quality satisfaction. This chapter gives an overview of the simulator-assisted multi-component network system architecture designed to aid in network management. Section 3.1 presents a general overview that discusses some of the challenges and limitations of existing systems. Section 3.2 discusses the architectural design requirements, models the problem and architecture specifications, and illustrates the different components of the architecture. Three sets of experiments are conducted in Section 3.3 to present and analyze the results and performance of the architecture. Finally, Section 3.4 concludes the chapter.

3.1. OVERVIEW

The multi-component network management architecture used in this work stems from rigorous growth in wireless communication technologies coupled with rapid advances in offered services. Advanced services in real-time applications such as video-on-demand, *Voice over IP* (VoIP), and interactive gaming with stringent time and network resource constraints, renders the management of heterogeneous wireless networks a major challenge.

The continuous struggle of management system paradigms to keep up with continuous demands of optimal and adaptive configurations of network policies and service agreement parameters, stipulates a necessity for adapting rejuvenated solutions. Such adopted system paradigms might introduce challenges that have neither been seen nor foreseen in the past.

Given such advances and growth in wireless networks, it is necessary to consider network cooperation in the design of a QoS management system. A benefit of such cooperation is the satisfaction for both service providers and subscribers in terms of augmented profit gains and improved QoS guarantees. Another clear incentive for cooperation is the reduction in network expansion costs by allowing providers to deliver desired coverage and service rate guarantees while deploying fewer base stations. Cooperating networks require service subscribers to switch between the available access networks to achieve maximum user satisfaction, SP profit, and access point load balancing. Thus, VHO mechanisms must be recognized within the network-cooperative-aware management system.

In addition to the aforementioned issues, various research challenges must be identified for provider cooperation network management mechanisms to succeed. The first challenge is to find an incentive for commercial SPs to cooperate and share their resources with other providers. Since SPs are always voracious towards maximizing their profit, the main goal of cooperation should be to increase providers' total and individual incomes. To provide a realistic means for profit calculation, this work relies on an SLA template that is employed to derive different SLA instances to be used by the SPs.

Research in the literature has focused on developing efficient mechanisms to ensure the accurate transformation of SLAs into lower configurations or policies [43]. Unfortunately, static long-term SLAs, where performance guarantees for each class of service (CoS) subscribers are fixed over time, may not be suitable for the short relationship between the users and the SPs in today's wireless environments. With today's technological advances, users must be capable of roaming freely between different networks. Hence, short-term SLAs for mobile users are employed, in which new SLA contracts are created and configured according to current network conditions. Clearly, accepting more users with the same SLAs provided to existing ones may result in experienced service degradation and SLA violations for new users. This, in turn, will impose more penalties on the SP and reduces its

profit. In the proposed SLA template, VHO is also included to provide an added-service-value that can increase the users' satisfaction.

To maximize the profit obtained from these SLAs, a framework is presented to determine the optimal network configurations to be applied by the providers using a real-time simulator and a meta-heuristic search strategy. The second challenge of having cooperation is the time sensitivity of the handover problem, where optimal configurations must be found within a limited time manner. To address this challenge, a new modified tabu-based search algorithm is introduced called Iterated Local and Global-Tabu Search (IGL-TS) and the configuration learning process is decoupled from that of the actual configuration adaptation step.

To this end, an architecture is proposed that is built upon the behavior of five main components: a simulator manager (SM), a VHO manager, an SLA manager, a variable configuration (VC) optimizer, and a monitoring and performance evaluation component. The SM is responsible for adapting candidate VHO and SLA configurations through enacting a set of simulation scenarios and then updating those configurations according to the obtained results. The VHO and SLA managers are responsible for the actual execution of the configured VHO and SLA management functionalities. The VC optmizer is responsible for discovering optimal policy configurations in order to ensure an increased profit and AP load balancing for SPs. The presented work jointly adapts SLA offerings and VHO operations with the objective of maximizing the user satisfaction and the SP monetary profit.

3.2. THE MANAGEMENT ARCHITECTURE

3.2.1. Management System Architecture Design Requirements

The emergence of the network management paradigm due to the advent of a plethora of new access technologies and services, multimedia services in particular, has raised questions related to traditional ways of optimizing network resources. Due to the heterogeneity in network infrastructure, services must go through a variety of transmission media for an End-to-End (E2E) service delivery. Optimizing the performance of a network requires an enormous amount of work configuring and coordinating different entities to achieve the desired objective, such as in the case of this research, optimizing resource usage and maximizing a SP's profit. Thus far, the SPs problem has been introduced. In this section, the

design requirement of the architecture which promotes the utilization of run-time simulators for parameter configuration in VHO policies and SLAs is presented.

In the design of a network management architecture, network operation is optimized by adapting configurable policies. The policy configuration process is based on some type of model that requires information about the current and previous state of the network, and thus information must be provided by the monitoring and measurement component of the architecture. In order to foster extensibility and adaptability of the network management system, the following requirements are considered within the architecture design:

- ***Ability to integrate existing simulation tools***: the networking community has produced an enormous amount of network simulators capable of observing and predicting the behaviour of a network, without the presence of an actual network. In contrast to existing, statically configured schemes that lack the ability to quickly adapt to changing network loads, the proposed framework continuously derives optimal network configuration policies. To achieve this goal, the proposed work relies on a policy-based real-time simulator such that based on the outcome of the simulation and the monitored network performance, a set of VHO and SLA configuration policies are communicated to the VHO and SLA modules to reconfigure and/or fine-tune the running VHO scheme and SLAs.

- ***Ability to integrate existing handover mechanisms***: the dynamicity and mobility of user devices within the available heterogeneous networks enforces management systems to integrate and manage mobility models. With the wide variety of available handover algorithms, there is a critical need for developing a management architecture. Such an architecture must integrate handover algorithms for connection management and optimal resource allocation. The architecture must be able to integrate handover mechanisms that are managed by policies, such that the integrated simulator within the system evaluates a set of possible new VHO configurations before actually applying them to the real network. Based on the outcome of the simulation and the monitored network performance, a set of VHO configuration policies are used by the adopted policy-triggered handover mechanism to reconfigure and/or fine-tune the running VHO scheme.

- ***Ability to incorporate a monitoring and measurement tool***: the process of reconfiguration is not self-contained, but distributed through the whole system. It does not solely detect the necessity for new policy configurations, but rather is notified by another component of the management system. Hence, the system's architecture design requires a notification module responsible for detecting the events occurring in the network utilizing monitoring and measurement tools. When monitoring and executing measurements, alarms should be generated as soon as a measured value within the enforced policies exceeds its defined threshold. The monitoring module should be adaptable in wireless mobile environments within heterogeneous networks.

- ***Ability to incorporate an SLA refinement tool:*** network management systems are developed to optimize resource usage with the goal of increasing provider profit and user satisfaction. Thus, the design of a management architecture should incorporate SLA management capabilities to refine and offer new short-term SLAs which utilize network resources and maximizes the business profit for the service provider. The management system must satisfy the providers' business requirements as imposed by network administrators.

- ***Persistent storage:*** there are several entities that the management system has to store persistently, including simulator, handover, SLA, monitoring, and ontology-related specifications. Base stations, access points, and user devices must have the capability of accessing the temporary and ontology repositories used to store information, such as monitored and measured data, user-related data, and so on.

- ***Cooperation of components within the architecture:*** the adopted and integrated architecture components and modules must be capable of coordinating and adapting to changes enforced by other components. For example, the simulation component will communicate new policy configurations to both the handover and SLA management components.

- ***Ability to integrate meta-heuristic search methods***: The selection of configuration policies necessitates the continuous use of variable configuration search methods. Well respected meta-heuristic-based search algorithms such as tabu-search [46] should be incorporated within the management architecture design to find optimal solutions to policy and QoS management.

3.2.2. Modeling the Problem and Design Requirements

In this research, an environment comprised of multiple heterogeneous wireless networks with overlapping coverage is considered. The networks are owned by N service providers, SP_i, $i = 1, \ldots, N$. A new mobile client u_j negotiates an SLA with a provider i for a choice of one of three CoS offerings (CoS_1: high, CoS_2: intermediate, and CoS_3: low) for the service provided. Each client may be attached to at most one network at a given time. In the considered system model, a mobile client can subscribe to a single CoS with service quality offerings described by the vector $\boldsymbol{q_{ij}} = \left(\underline{\boldsymbol{q}}_{\boldsymbol{ij}}, \overline{\boldsymbol{q}}_{\boldsymbol{ij}}, \boldsymbol{p_{ij}}\right)$, where $\underline{\boldsymbol{q}}_{\boldsymbol{ij}} = \left(\underline{q}_{ij}^{1}, \underline{q}_{ij}^{2}, \ldots, \underline{q}_{ij}^{|K|}\right), \overline{\boldsymbol{q}}_{\boldsymbol{ij}} = \left(\overline{q}_{ij}^{1}, \overline{q}_{ij}^{2}, \ldots, \overline{q}_{ij}^{|K|}\right)$ and $\boldsymbol{p_{ij}} = \left(p_{ij}^{1}, p_{ij}^{2}, \ldots, p_{ij}^{|K|}\right)$, K is the set of service features (e.g., service rate, expected delay, handover latency and error rate). The values $\underline{q}_{ij}^{k}$ and $\overline{q}_{ij}^{k}$ represent the upper and lower bounds of the offered service feature range, $k \in K$. The corresponding value p_{ij}^{k} represents the monetary penalty imposed on a SP if it does not meet the minimum requirements for the QoS features outlined in the SLA. For example, if SP_i offers a certain service rate (i.e. $k = 1$) guarantee range, the SLA will state the range of the acceptable service rates, $\underline{q}_{ij}^{1} = \underline{R}_{ij}$ and $\overline{q}_{ij}^{1} = \overline{R}_{ij}$, respectively, and a penalty $p_{ij}^{1} = p_{ij}^{R}$ if it fails to provide the user with this rate.

The subscription class is determined when the mobile client initiates a connection with an SP. The ongoing connection must remain in the same class until the contract terminates. It is assumed that all providers in the considered model are willing to cooperate and share resources [214]. Thus, client u_j which is provided a service by provider i, will benefit from the same service at the same quality levels determined by the original SP with other cooperating providers. To illustrate this, Figure 3.1 provides an example of several SPs, in which each $\boldsymbol{SP_i}$ deploys base stations or access points that belong to a certain access

technology. Customers belong to one SP but are able to roam in the coverage areas of other SPs with the same benefits agreed upon with its provider in the negotiated SLA. All SPs provide the same service to different mobile users $\{m_1, m_2, \ldots, m_9\}$ with different CoS. Results of the provided service are measured in terms of the observed service quality readings $\tilde{\boldsymbol{q}}_{\boldsymbol{ij}}(t)$, where $\tilde{\boldsymbol{q}}_{\boldsymbol{ij}}(t) = \left(\tilde{q}_{ij}^{1}(t), \tilde{q}_{ij}^{2}(t), \ldots, \tilde{q}_{ij}^{|K|}(t)\right)$ is the observed QoS reading of service feature $k \in K$ at time t. The observed service quality may also include other performance parameters such as the number of handovers and the total handover latency.

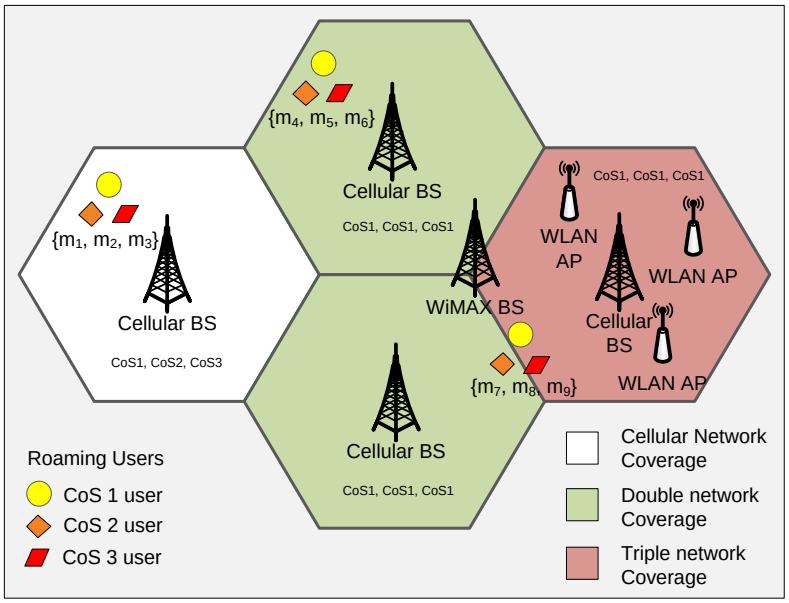

Figure 3.1 Example of cooperation between SPs.

Evidently, a SP's intention is to maximize its profit through increasing its revenue $\mathcal{R}$ and reducing its cost $\mathcal{C}$. The business profit function provides a measure of the profitability of a SP. Assuming discrete time intervals $t = t_0, t_1, \ldots, T$, where t_0 is the time marking the initiation of the current service session readings for u_j. The profit function $\mathcal{P}\left(\boldsymbol{q}_{\boldsymbol{ij}}, \tilde{\boldsymbol{q}}_{\boldsymbol{ij}}(\boldsymbol{t})\right)$ is defined as the sum of the total profit gained from a contracted service agreement based on the observed QoS readings $\tilde{\boldsymbol{q}}_{\boldsymbol{ij}}(t)$. The profit from one service agreement is the net revenue gained from client u_j minus any costs incurred on the SP. Net revenue is the revenue generated from the service price paid by the client and alternative sources such as SP rewards, minus any penalties suffered by the provider. Such penalties are incurred from the

service level agreement instance paid out to the client due to a service violation. For a SP at time t serving client u_j for $j = \{1, \dots, U(t)\}$, profit is calculated as follows:

$$\mathcal{P}\left(\boldsymbol{q_{ij}}, \tilde{\boldsymbol{q}}_{\boldsymbol{ij}}(t)\right) = \mathcal{R}\left(\boldsymbol{q_{ij}}, \tilde{\boldsymbol{q}}_{\boldsymbol{ij}}(t)\right) - \mathcal{C}\left(\boldsymbol{q_{ij}}, \tilde{\boldsymbol{q}}_{\boldsymbol{ij}}(t)\right) \tag{3.1}$$

Using the example outlined in Listing 1 (see Section 3.2.3.4), given the observed QoS reading $\tilde{\boldsymbol{q}}_{\boldsymbol{ij}}(t) = \left(\tilde{r}_{ij}(t), \tilde{c}_{ij}(t), \tilde{e}_{ij}(t), \tilde{l}_{ij}(t)\right)$ provided to a client up to T, where $\tilde{r}_{ij}(t), \tilde{c}_{ij}(t), \tilde{e}_{ij}(t)$ are the effective service rate, the used capacity and the experienced error rate, respectively, by u_j at time t and $\tilde{l}_{ij}(t)$ is the amount of time of service disruption at time t, the revenue and cost imposed on a SP is calculated as follows:

$$\mathcal{R}\left(\boldsymbol{q_{ij}}, \tilde{\boldsymbol{q}}_{\boldsymbol{ij}}(t)\right) = \sum_{\substack{t=t_0 \\ \underline{R}_{ij} \le \tilde{r}_{ij}(t) \le \overline{R}_{ij}}}^{T} c_i \tilde{r}_{ij}(t) + \sum_{\substack{t=t_0 \\ \tilde{r}_{ij}(t) > \overline{R}_{ij}}}^{T} c_i \overline{R}_{ij} + c_i^R\left(\tilde{r}_{ij}(t) - \overline{R}_{ij}\right)$$

$$+ \sum_{\substack{t=t_0 \\ \tilde{c}_{ij}(t) > C_{ij}}}^{T} c_i^C \max\left(0, \tilde{c}_{ij}(t) - C_{ij}\right) \tag{3.2}$$

$$\mathcal{C}\left(\boldsymbol{q_{ij}}, \tilde{\boldsymbol{q}}_{\boldsymbol{ij}}(t)\right)$$

$$= \sum_{t=t_0}^{T} p_i^L \max\left(0, \tilde{l}_{ij}(t) - L_{ij}\right) + \sum_{\substack{t=t_0 \\ \tilde{r}_{ij}(t) < \underline{R}_{ij}}}^{T} p_i^R \, max\left(0, \underline{R}_{ij} - \tilde{r}_{ij}(t)\right)$$

$$+ \sum_{\substack{t=t_0 \\ \tilde{e}_{ij}(t) > \varepsilon_{ij}}}^{T} p_i^{\varepsilon} \, max\left(0, \tilde{e}_{ij}(t) - \varepsilon_{ij}\right) + \left(p_i^T \middle| (t_0 + \Delta t) > T_{ij}\right) \tag{3.3}$$

where c_i is the service charge in cents/Mbit, $\overline{R}_{ij}$ is the maximum guaranteed service rate in Mbps, c_i^R is the extra cost incurred on user u_j for requesting an increase in the service rate beyond $\overline{R}_{ij}$. c_i^C is the extra cost incurred on u_j when exceeding the assigned capacity C_{ij}. p_i^L is the penalty incurred on the service provider for providing a disrupted service beyond the agreed upon normal disruption time L_{ij}. p_i^R is the penalty incurred on the service provider for decreasing the service rate. p_i^{ε} is the penalty incurred on the service provider for exceeding the agreed upon bit error rate ε_{ij}. p_i^T is the penalty incurred on the service provider for forcefully handing off a user to a different network. A description of an SLA template is presented in Section 3.2.3.4.

In addition to the revenue and cost details outlined, another reduction in the fixed cost expenditures also results from SP network cooperation, which this research does not consider but needs to be clarified for the reader. Cooperation may reduce network expansion costs by allowing providers to deliver desired coverage and service rate guarantees while deploying fewer base stations and hot spots.

To attain monetary profit maximization for SPs, the framework provides a cooperation strategy to achieve the aforementioned objective while achieving additional benefits such as base station load balancing and user QoS guarantees fulfillment. Thus, through SP cooperation, clients will either be provided the same service guarantees by a different SP through handoff, or a new instance of the original SLA template (discussed in Section 3.2.3.4) will be created with modifications to the original service guarantees. This will be done by finding equilibrium between customer satisfaction and reduced SP penalty costs.

3.2.3. Proposed Management Architecture

Inspired by the aforementioned system design requirements imposed by the service provider and subscriber, specifically profit maximization and SLA adherence, the proposed architecture design is driven by the following main goals:

1. Guarantee the QoS levels requested by the mobile clients. This guarantee might require a VHO of the user to other heterogeneous networks by dynamically configuring new short-term SLAs with the objective of maximizing both user satisfaction (i.e. QoS) and the SP profit.
2. Continuously derive and adapt optimal VHO and SLA configurations through the use of meta-heuristic approaches. The adaptation process must be decoupled from the parameter configuration search process due to the time sensitivity of the VHO process.
3. Adopt a VHO algorithm capable of executing the three stages of handover (initiation, decision and execution). The algorithm must be capable of applying optimized configurations and adapting them to complex heterogeneous wireless environments.
4. Monitor the QoS of the network in an accurate and efficient approach. Such measurements are used to analyze and improve the performance of the network and offered services.

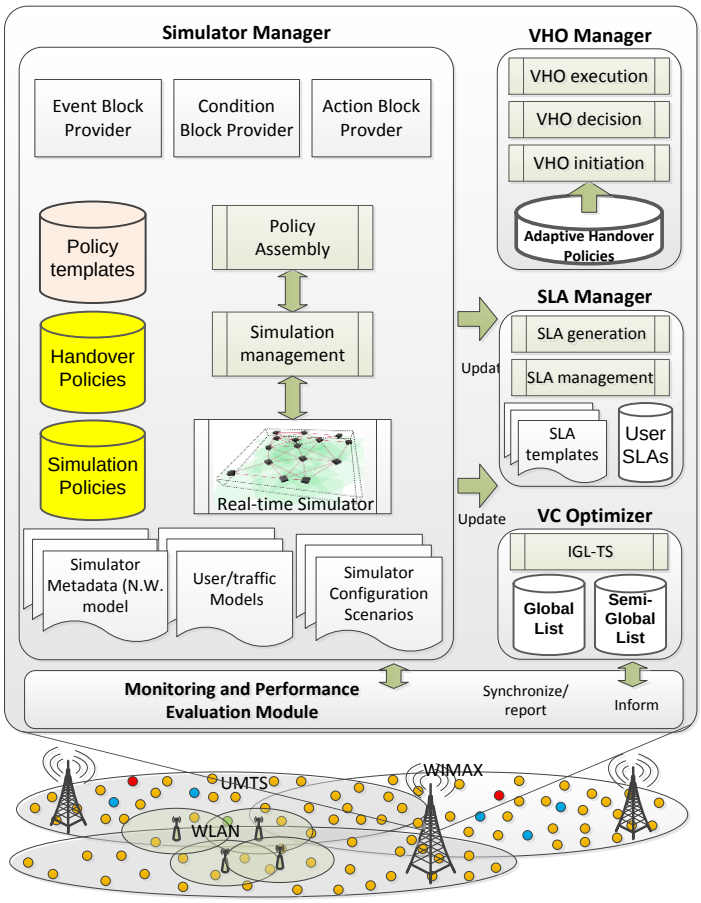

Figure 3.2 Network management architecture.

The proposed architecture, illustrated in Figure 3.2, realizes these goals through the unification of advances in simulation tools and an heuristic optimization algorithm. The architecture relies on the cooperation among five components. Separating the architecture into components reflects the flexibility and scalability of adapting such a framework into network entities. The first two components are the SLA and VHO managers, which are typically responsible for offering and managing the appropriate SLAs to the mobile users and executing the various VHO operations, respectively. Additionally, the Simulator Manager (SM) receives user and network related monitoring measurements from the '*monitoring and performance evaluation module*', and analyzes and employs this data to control the behavior of a real-time simulator. The variable configuration optimizer (VCO) will search for good policy configurations, communicate them to the SM, and based on the simulator results, the SM submits the obtained configurations to both the VHO and SLA managers. The VHO and SLA managers are responsible for the actual execution of the configured VHO and SLA

management functionalities. Further details of each component will be discussed in the following sections and chapters.

3.2.3.1. The Policy Hierarchy Model

This section describes the methodology for using policies in system management. In policy-based network management, the events that occur, the conditions, and the actions that must be performed are denoted into policy terms for an automatic management of all entities. Policies can help to automate the different stages of VHO (initiation, decision and execution) and manage and control the behaviour of the simulator. Scalability, flexibility and simplicity are all advantages of policy-based approaches. The ability to separate polices from the underlying managed system reflects the flexibility and scalability of using policies for network management.

Different stages of VHO require different policy sets to control the VHO process. For example, in the first phase, user and network related data is collected and employed to test some triggering policy conditions that are specific to the VHO scheme (e.g., received signal strength (RSS), user velocity, battery level, or a newly detected network with a cheaper cost). Thus, a handover initiation policy would require a condition (e.g. $RSS < RSS_{Threshold}$ && $UserVelocity > Velocity_{Threshold}$) for an action (e.g. $initiate\ VHO$) to be triggered as an example. The same concept is applied to other policy types used for network and simulator management.

Three classes of handover policies have been adopted: *handover initiation, network elimination/recommendation and VHO configuration policies*. In addition, three classes of simulator policies have been employed to control the behavior of network simulators: *performance evaluation, simulator configuration and scenario configuration policies*. Further details of each policy type are explained in the following sections.

The policy hierarchy, illustrated in Figure 3.3, consists of four layers, i.e., abstract, VHO/NW/simulator, intermediate and low level policies. *Layer 1* policies are defined from the service agreements negotiated between the user and the SP. *Layer 2* policies are constructed using a set of events, conditions and actions. *Layer 3* policies express QoS parameters for specific users and networks which are then translated into technology dependent configurations in *layer 4*. The task of mapping different layers (i.e. translating

parameters of QoS at each layer to a lower level) is not trivial due to the dynamics and complexity of the underlying wireless environment. This task is done through different mapping methodologies [188] [189] and is beyond the scope of this dissertation.

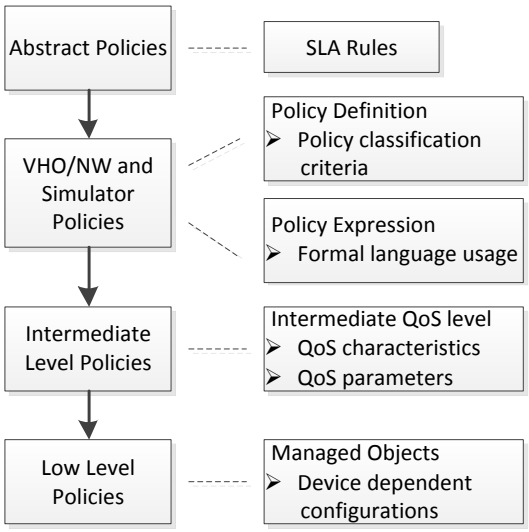

Figure 3.3 . Policy hierarchy model.

Let $\boldsymbol{x} = (x_1, x_2, \dots, x_M)$ be a vector of policy configurations comprised of M variables. Each variable x_m, $m \in 1,2, \dots, M$, may represent a variable in a policy event, condition or action and is associated with a domain $D(x_m)$ that contains all permissible values that x_m can take. Thus, a configuration space X which is used to manage the network is defined by the product $D(x_1) \times D(x_2) \dots \times D(x_M)$. For example, the VHO configuration policy given by:

$$\boldsymbol{if}\ NW_{Load} > WorkLoad_{Thshld}\ \boldsymbol{then}\ set\ Dwell_{Timer} = Simulator_{MeauredTimer}$$

The policy is represented by the policy variables x_1 and x_2, where x_1 represents the policy variable corresponding to $WorkLoad_{Thshld}$ and x_2 represents the variable corresponding to $Dwell_{Timer}$. The objective is to automate the process for finding an optimal policy configuration $\boldsymbol{x}^* \in X$ such that when it is applied to the network, it will result in maximizing service provider profit.

3.2.3.2. The Simulator Manager

To manage the SP's resources optimally, it is crucial to test network and handover decisions prior to the adaptation process. Since handover execution includes connection

establishment and release of resources by an SP, the manager uses a run-time simulator to enact a set of simulation scenarios and then employs the results to modify the VHO and network operations. Since the effect of mapping $\boldsymbol{x}$ to the network environment cannot be directly measured, a real-time simulator is used to achieve such tasks. The simulator manager (SM) is tasked with two functionalities: the first is to adapt the VHO model maintained by the VHO manager to measure and minimize the difference between the desired VHO and network performance objectives and the actual measurements. The second functionality is concerned with determining appropriate configurations of newly offered SLAs to maximize the SP profit. Additionally, the SM is also responsible for maintaining an accurate network-simulated model that closely resembles the physical network and its performance. The VC optimizer provides the simulator with the appropriate configurations for newly offered SLAs, network and handover policies according to the profit maximization problem which will be discussed in the following sections.

As shown in Figure 3.2, the manager maintains the necessary metadata to describe the static information related to the underlying physical network, as well as the models of the users and their generated traffic. No restrictions are imposed on the simulator, but due to the large size and scale of networks, one challenging concern is having the simulator know information for the entire network. Thus, it is wise to decompose a large network into parts, such that each part is simulated independently and simultaneously. Each part represents a subdomain of the entire network, where each part comprises metadata describing the physical networks, users, and the traffic generated. Each domain is simulated separately when needed, thus reducing the enormous computational power needed to execute all events that take place in the network. Another benefit of dividing the simulated network into domains is the reduction in the amount of memory size required in large scale network simulations.

The manager maintains the necessary metadata to describe the static information related to the underlying physical network, as well as models of users and their generated traffic. All necessary network and service quality parameters are collected through the performance evaluation module. This information is represented by two models M^{NW} and M^{Users} where the former is the vector of current network settings (e.g., the network communication graph, capacities of the links and assumed traffic classes) and the latter models the users in the

network along with their behavior (e.g., number of users, device capabilities, used bandwidth, velocity and trajectories in the network). The SM also maintains scenario templates for various scenarios to be run by the simulator. A scenario, S, is defined by the tuple $(M^{NW}, M^{Users}, \boldsymbol{x})$. A single simulation run produces an evaluation of a single scenario. This scenario includes the new parameter configurations selected through the optimization heuristic, discussed later in this dissertation. Since a single simulation run for a specific network model with a reasonable number of nodes can take an excess amount of time, it is necessary to decouple the simulation process from the adaptation process. It is also necessary to minimize the number of simulation runs through a state-of-the-art variable configuration optimization algorithm, discussed later.

To maintain scalability and provide a variety of service applications to users, the manager estimates the network and handover performance using five traffic classes: *conversational, streaming, emergency, interactive, and background*. Each traffic class is further divided into three different subclasses each with a different priority: *high, medium, and low*. The SM also maintains the active instances of the SLA and VHO as well as templates for various scenarios to be run by the simulator. Indications of performance degradation, received from the performance evaluation module, trigger the SM to instruct the run-time simulator to run a set of simulation scenarios while testing one or more new configuration policies. This is done for either the VHO scheme or the SLA configurations, depending on the type of the performance degradation. Obtained simulation results are used as indicators for either performing further simulations or communicating the new policies to the SLA and VHO managers. Based on the results obtained from the simulation runs, the SM may decide to adapt the VHO scheme or reconfigure the SLA template used in offering services to new users. To decouple the process of updating these configurations from their runtime operation, the SM communicates these updates in the form of VHO and SLA management policies. As briefly discussed earlier, the overall behavior of the SM is controlled by a set of predefined policies, outlined below:

- ***Performance evaluation policies***. These policies are triggered by any performance degradation during the VHO process, as well as by SLA violations. These policies usually indicate that once degradation events are experienced, the SM should be

notified. The aim of such policies is to achieve an accurate network-simulated model. Any difference that exceeds a certain threshold will evoke the SM to synchronize the simulator. For example, performing a handover without checking whether it has achieved a better performance, results in an undesired VHO. Performance evaluation policies aim at correcting handover decision algorithms by including an evaluation function that inspects various parameters such as the VHO delay, VHO failure probability, throughput, ping-pong phenomenon (unnecessary frequent fluctuation between two or more networks) [190], user satisfaction, etc. The action part of these policies notifies the simulation manager, triggers more adequate monitoring events or reconfigures one or more parameters of the simulator (e.g., simulation frequency or duration).

Example: $\boldsymbol{if}\ PerformanceParameter < TargetValue$

$\boldsymbol{then}\ notify\ (SimulatorManager, PerformanceParameter)$

- ***Simulator configuration policies***. The SM employs this set of policies to directly control the behavior of the simulator. For example, one critical factor is controlling the degree of synchronization [159] between the simulator and the actual network. Obviously, there is a trade-off between the simulator accuracy and the frequency of network status updates. Hence, these policies are needed to tie these two factors together along with other business-level objectives. Other issues that are controlled through these policies are: whether to allow concurrent simulations, and how these policies control the frequency of simulation runs or the simulation length. Examples of policy control include: indicating how many simulation seconds are considered for future look-ahead performance and the number of different simulation runs considered for each scenario as a function of the achieved performance. For example, several simulation runs must be employed to reach a certain confidence degree when generating a VHO configuration policy for calculating the optimal dwell time [38]. An example follows:

  Example: $\boldsymbol{if}\ VHOPolicyParameter = DwelTimer$

  $\boldsymbol{then}\ Set\ MaxSimRuns = 100\ \&\&\ ConfidenceInterval = 90\%$

- ***Scenario configuration policies***. Based on the performance evaluation policies, the SM selects a set of operations to perform for a certain scenario. A scenario includes configuration parameters such as: the number of MNs, location of MNs in proportion to the base stations, link delay, jitter, bit error rate, bandwidth, price, and MN velocity. The fine tuning of these scenario scripts is achieved through the scenario configuration policies. The following are two examples of these policies:

  Example: $\boldsymbol{if}\ VHOFrequency < VHOFequencyThreshold$

  $\boldsymbol{then}\ set\ TestRangeForSimulationParameters(\ DwellTimer = [c1, c2] * UserVelocity)$

  Example: $\boldsymbol{if}\ VHOFrequency < VHOFrequencyThreshold\ \&\&\ Application = "streaming"\ \boldsymbol{then}\ set\ simulationParameters(\ AvailabilityWeight = 6, DelayWeight = 4, PriceWeight\ = 4, AppType = streaming\ \&\&\ AppPriority = High\ )$

These policies are constructed with the assistance of the policy templates repository. The policy assembly module selects the appropriate events, conditions and actions that apply new parameter values, depending on the results of the network performance. The process of optimal policy configuration selection starts with indications of performance degradation. This is received from the performance evaluation module, which in turn, triggers the SM to instruct the run-time simulator to run a set of simulation scenarios (with parameter configurations provided from the VCO). The SM will test one or more new configuration policies for SP resource reallocation, VHO scheme reconfiguration or SLA configurations, depending on the type of performance degradation. Obtained simulation results are used as indicators to either perform further simulations or to communicate the new policies to the SLA and VHO managers. The operation of the simulator can be regarded as the mapping of a scenario S running over a simulation period t to behavior values $\tilde{\boldsymbol{q}}_{\boldsymbol{ij}}(t)$.

Figure 3.4 depicts this mapping process, where the simulator starts with a scenario S_0 reflecting the current (initial) behavior of the network and the current models (network, users, VHO, SLA). The search process for a new scenario S is initiated by the SM with the aid of the VC optimizer by reconfiguring one or more of the parameters in the vector $\boldsymbol{x}$. Further details concerning the search algorithm are provided in Chapter 4.

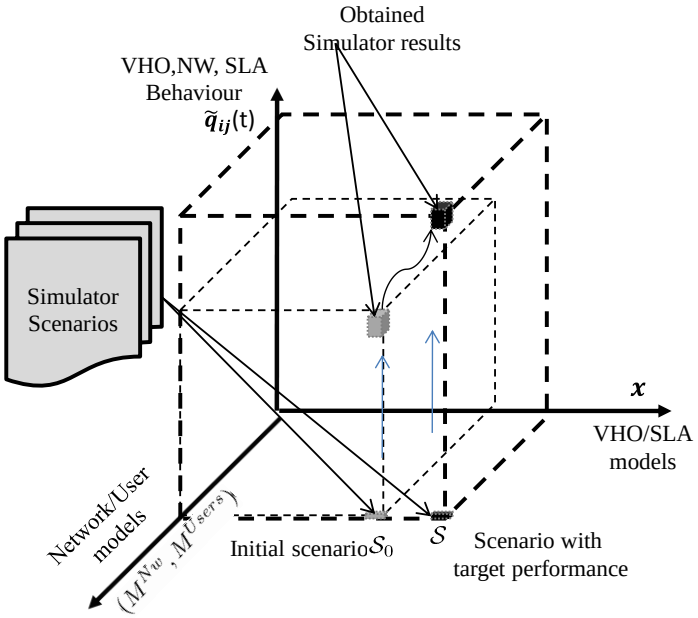

Figure 3.4. Scenario search space consists of inputs to the simulator (network, user, VHO and SLA models) and behavioral outputs in terms of QoS performance.

3.2.3.3. The Variable Configuration Optimizer (VCO)

The VCO is the backbone of the management architecture, such that the SM employs exhaustive responsibilities on this component. It is important to note here that while each provider SP_i can explicitly control the service offerings $\boldsymbol{q_{ij}}$ to each client u_j, it is not easy to determine or predict $\tilde{\boldsymbol{q}}_{\boldsymbol{ij}}(t)$ a priori. In fact, $\tilde{\boldsymbol{q}}_{\boldsymbol{ij}}(t)$ depends on a number of factors such as the number of users serviced by the provider as well as the behavior of the current network management functionalities. Therefore, it is necessary to find the ultimate set of network and handover policy configurations that will provide the optimal observed QoS performance $\tilde{\boldsymbol{q}}_{\boldsymbol{ij}}(t)$ and maximize the SP monetary profits $\mathcal{P}\left(\boldsymbol{q_{ij}}, \tilde{\boldsymbol{q}}_{\boldsymbol{ij}}(t)\right)$. Searching for the configurations that produce optimal results is the responsibility of VCO. A modified tabu-search algorithm [46] called the *Iterated Global and Local - Tabu Search* (IGL-TS) was developed. IGL-TS is a meta-heuristic search strategy that provides a method to speed up the process of finding the optimal set of policy configurations and provides a mechanism to escape local states and follow previous optimal solutions. Details of the VCO algorithm are discussed in Chapter 4.

3.2.3.4. The SLA Manager

A service relationship between the provider and the subscriber constitutes a business relationship defined in a contract. The most crucial aspects of a contract are the set of QoS guarantees a SP gives to the service subscriber, the costs governed on the subscriber, and the penalties imposed on the provider for failing to deliver such promises. SLAs are also used within corporations to negotiate their relationship with each other. However, having SLAs being manually provisioned and monitored would translate into an increased cost burden to the service provider, inability to provide the optimal service quality, and increased time delays in providing updated versions of SLA offers. For economic and autonomic management reasons, an automated mechanism of SLA management is desired to provide both service and system provisioning. In addition, static long-termed SLAs, where performance guarantees for each CoS subscribers are fixed over time, may not be suitable to the short relationship between the users and the SPs in today's wireless environments. One way to approach this issue is to use SLA templates that include a few automatically processed fields, as illustrated in Listing 1.

SLA management requires a robust language so that an SP can obtain, monitor and enforce QoS guarantees for its customers. Ludwig et al. [40] proposes a novel formal language to define and express SLAs and assist in SLA provisioning and management to aid in SLA refinement [40]. It is used to automate the process of supervision and management of SLAs and the provisioning of corresponding systems. The language provides a large degree of flexibility by means of an XML-based representation and a runtime system for SLAs. Parties involved in an SLA can describe their desired guarantees with respect to QoS parameters.

The SLA management module acquires user requests to assist and ease the class of service selection process. It is also responsible for finding the best user service [191] by selecting the appropriate network capabilities, given the available options. Additionally, the SLA generation module adapts the newly configured parameters and generates either short-term SLAs for roaming clients or new SLAs for potential clients, according to the SP capabilities.

An SLA must be comprised of the following parts, while taking into account the specifics of an SLA environment:

- **Parties**: the SLA must describe the parties involved in both delivering and acquiring the service. This includes the signatory parties as well as the supporting parties (i.e. third-parties) that are brought into the agreement to act on behalf or support the service provider or customer. Third-parties cannot be held responsible for any SLA violation.
- **Service Definitions:** the SLA must also describe the service properties onto which obligations are defined. Such properties include the type of service delivered to the subscriber, the class of service, and so on. The task of establishing a common definition of a service and service qualities defined within the SLA can be represented in an ontology; see Chapter 5.
- **Obligations:** most importantly, the SLA must define the service level that is guaranteed in terms of SLA parameters. Obligations also include the actions taken under particular conditions.

Below, an outline is given with an example of a contracted SLA between a client u_j and a provider SP_i: when u_j requests a service from SP_i, an SLA instance is advertised to the client as outlined in Listing 1. The advertised SLA instance is represented by the service quality offerings vector $\boldsymbol{q_{ij}} = \left(\underline{\boldsymbol{q}}_{\boldsymbol{ij}}, \overline{\boldsymbol{q}}_{\boldsymbol{ij}}, \boldsymbol{p_{ij}}\right)$, where $\underline{\boldsymbol{q}}_{\boldsymbol{ij}} = \left(\underline{R}_{ij}, \underline{L}_{ij}, \underline{\varepsilon}_{ij}, \underline{C}_{ij}\right)$, $\overline{\boldsymbol{q}}_{\boldsymbol{ij}} = \left(\overline{R}_{ij}, \overline{L}_{ij}, \overline{\varepsilon}_{ij}, \overline{C}_{ij}\right)$ and $\boldsymbol{p_{ij}} = (p_i^R, p_i^L, p_i^\varepsilon, p_i^T)$.

Listing 1 SLA Template Example

1) Service charge is c_i cents/Mbit with expected session duration of T_{ij} min.
2) The guaranteed service rate is in the range of $\left[\underline{R}_{ij},\ \overline{R}_{ij}\right]$ Mbps.
   a) A rate downgrade below the guaranteed rate at time t incurs the SP a penalty p_i^R cents/Mbps.
   b) u_j can increase its service rate more than $\overline{R}_{ij}$ at an extra cost of c_i^R cents/Mbps of excess rate.
3) The guaranteed bit error rate is in the range of $\left[\underline{\varepsilon}_{ij},\ \overline{\varepsilon}_{ij}\right]$.
   a) An error rate violation at time t incurs a penalty of p_i^ε cents on the SP.
4) Maximum used capacity during T_{ij} is $\underline{C}_{ij}$ MB.
   a) m_j can use up to $\overline{C}_{ij}$ MB at extra cost of c_i^C cents/MB.
5) Time of service disruption or QoS degradation must be within $[\underline{L}_{ij}, \overline{L}_{ij}]$ seconds, otherwise SP incurs a penalty of p_i^L cents.
6) A forced handover before T_{ij} min. incurs a penalty p_i^T cents on the SP.

The first clause of the SLA instance states that SP_i offers a service to u_j using one of the three offered CoSs with a fixed service cost of c_i cents/Mbps for a maximum session duration of T_{ij} as long as u_j is within its coverage area. The second rule indicates that provider i will guarantee client u_j a minimum service rate $\underline{R}_{ij}$ and a maximum service rate $\overline{R}_{ij}$. If the SP fails to achieve this, a penalty p_i^R will be induced on the SP. Clauses 3 and 4 indicate that the guaranteed bit error rate is in the range of $[\underline{\varepsilon}_{ij}, \overline{\varepsilon}_{ij}]$ and the allowed capacity during the session lifetime is $\underline{C}_{ij}$. Any violation will incur penalties and extra charges of p_i^{ε} and c_i^C on the SP and user, respectively. The fifth rule states that a service disruption of time in the range of $[\underline{L}_{ij}, \overline{L}_{ij}]$ is acceptable to the client; otherwise, the SP will incur a penalty of p_i^L. Finally, the last rule states that if SP_i switches u_j to another network before T_{ij} while he/she is still in its coverage area, the SP will incur a penalty p_i^T.

The outcome of the contract is measured in terms of the revenue and costs incurred on the SP from the provider's perspective. From the client's perspective, the outcome of the contract is measured in terms of the observed service quality $\tilde{\boldsymbol{q}}_{\boldsymbol{ij}}$, which may also include other performance parameters such as the number of handovers and the total handover latency.

In order to meet the agreed upon commitments outlined in the SLA, a number of service level objectives are extracted and considered. For example, if a SP's goal is to achieve high levels of throughput for mobile clients, then the service level objectives that are considered in this case in an SLA are related to the service rate. To meet such service level objectives, policies are considered to define the action that must take place in case an objective has not been respected [192]. The management component does not implement the business logic to decide whether the management action is in the best interest of the service provider. Rather, the VCO will find the ultimate set of configurations that maximize the SP's profit while considering the subscriber's service quality requests. Upon receipt of a notification that a term of the SLA has been violated, the action part of these policies will trigger a message to the SM to modify the new SLAs such that the same problem is avoided in the future. Such a fine-tuning of the values of SLA parameters in new SLA instances (i.e., $\boldsymbol{q}_{\boldsymbol{ij}}$ for a new client u_j) is crucial for maximizing the SP profit. Any new short-term SLA created through the SLA manager will be in the form outlined in Listing 1.

3.2.3.5. The VHO Manager

The VHO manager is responsible for the execution of the various steps for the VHO process. A specific scheme is not assumed in this research; rather any VHO algorithm is adopted which can be controlled by a set of policies that drives the VHO mechanism to control various aspects of the VHO process. Thus, a set of policy templates is adopted and acts as the interface between the VHO manager and the actual employed VHO algorithm by linking their parameters with the actual scheme's parameters.

Before describing the model, the common functionalities for any VHO scheme are presented. As depicted in Figure 3.5, the process involves three major phases: VHO initiation, network selection, and VHO execution. In the first phase, user data and network related data is collected and employed to test a few of the triggering conditions that are specific to the VHO scheme (e.g., received signal strength (RSS), user velocity, battery level, or a newly detected network with a cheaper cost). Once the VHO process has been initiated, the next phase selects one or more candidate networks, collects offered SLAs from these networks and employs a specific decision technique (e.g., AHP [29]) for the selection of the new network to which the user will be handed over to. The final step is concerned with the actual execution of the VHO process which involves various operations at different layers.

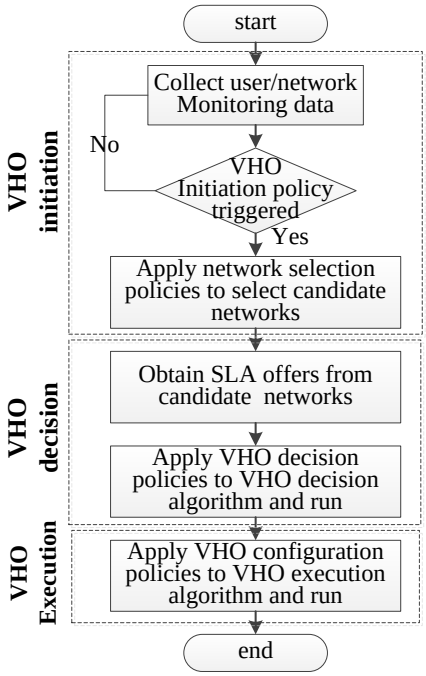

Figure 3.5 Policy-based model for VHO schemes.

Handover policies help to automate and simplify the process of handover management. The proposed VHO scheme relies on configuration policies to fine-tune execution of various steps of the handover process. A sample of these policies ordered according to handover stages is presented below:

- ***Handover initiation policies.*** This type of policy template controls when to invoke a VHO. Several QoS parameters are used to trigger such an event. Those parameters can either be network- (e.g., RSS, bandwidth, cost, coverage and network), terminal- (e.g., velocity, battery power and active interfaces) or user- (e.g., preference for a certain network) related. This type of policy also controls the frequency of VHO initiation (e.g., event-triggered vs. periodic). As will be shown later, the configuration of the policy parameters is performed by the SM with the aid of the VCO. A sample of this type of policy is as follows:

  Example: $\boldsymbol{if}\ RSSCurrent < RSSThreshold\ \&\&\ UserVelocity > VelocityThreshold$ $\boldsymbol{then}\ initiate\ VHO$

  Example: $\boldsymbol{if}\ NumberCurrentUsers < UsersThreshold$ $\boldsymbol{then}\ increment\ NewUserScanFrequency$

- ***Network elimination/recommendation Policies.*** Rather than specifying exactly which network to choose, recommendation/elimination policies specify networks that do/do not meet user/device requirements. These sets of policies can also recommend access networks based on the satisfaction of business objectives (e.g., prioritize cellular access for streaming applications and manage access privileges according to predefined users' preferences).

  Example: $\boldsymbol{if}\ MNPower < PowerThreshold\ \boldsymbol{Then}\ eliminate\ WLANNWx$

- ***VHO configuration policies.*** These types of policies are the core of the VHO process as they define various configurations of the VHO procedure as a function of the network status, user preferences or device characteristics. One example is a policy that determines the usage of a dwell timer [30] as a function of the network workload. In this situation the handover is performed only after the time period equal to the value of the dwell timer has elapsed. This will guarantee the stability of the connection after a VHO process.

  Example: $\boldsymbol{if}\ NetworkWorkload > WorkloadThreshold$ $\boldsymbol{then}\ set\ DwellTimer = SimulatorCalculatedTimer$

- ***VHO performance guarantees policies.*** To achieve stability in the state of the network, performance policies provide a guarantee that a network will not overload or provide services for mobile users that, according to its status, it is not capable of handling. One example of such policy is:

  Example:

  $\boldsymbol{if}\ Probability\,(VHOFailure) > FailureThreshold\ \boldsymbol{then}\ suspend\ VHO\ initiation$

The scheme also incorporates traffic classes and priorities in the condition clause of the above policies, as mentioned earlier in Section 3.2.3.2. One policy can be defined for emergency traffic to allow a high priority class MN to be connected to the same network unless the link quality has deteriorated to a point that a disconnection to the access point will occur shortly; and if it is necessary to initiate a handover, the one with the least handover delay is selected. On the other hand, another policy can be defined for interactive traffic, indicating that low bandwidth and moderate handover delays are acceptable. These requirements are described as various handover initiation policies. Clearly, these policies provide great flexibility for adapting the employed VHO scheme.

The VHO manager maintains a list of the currently active VHO policies communicated by the SM, and is automatically mapped [193] and used to configure the VHO scheme. By using the policy templates discussed in [76] (examples shown above), various policy instances that are specific to controlling the employed VHO scheme are derived. At any point in time t, the collection of these instances can be modeled as a tuple $\left(\boldsymbol{x}(t), \tilde{\boldsymbol{q}}_{\boldsymbol{ij}}(t)\ \forall ij\right)$. Here, $\boldsymbol{x}(t)$ represents the configurations of the variables for the set of active policies including all of the VHO phases at time t, as explained earlier. On the other hand, $\tilde{\boldsymbol{q}}_{\boldsymbol{ij}}(t)$ represents the behavior or performance of the network in terms of service quality readings for all users in the system.

It is obvious that statically defining a network and handover policy is not realistic, given the dynamic nature of the network and users. In Chapter 4, a novel scheme is developed for the dynamic management of these policies. Once the new configurations are communicated to the VHO manager, it will maintain a list of the currently active VHO policies to configure the VHO scheme. The simulator manager, with the aid of the VC optimizer, communicates a new set of configuration policies $\boldsymbol{x}$ to other components of the architecture in which the new policies are automatically mapped to reconfigure the running network entities.

The SM will communicate a new set of VHO configuration policies to the VHO manager. The policies are mapped to configure the VHO scheme. In the current implementation, a simple VHO procedure was adopted based on the analytic hierarchy process (AHP), which is used in both the simulator and the network to determine the weight of each QoS factor. The overall score of each candidate network is calculated using various cost functions such as the *simple additive weighting* (SAW) and the *multiplicative weighting* functions [78]. The best available network will be chosen for mobile users. For the purpose of achieving the best network performance and supporting the best possible QoS for mobile users, several QoS factors were considered. Those factors define the condition of the network and are divided as such: availability, timeliness, reliability, bandwidth, and price. Received signal strength (RSS) is selected as an element of availability, delay and jitter as a sub-factor of timeliness, and packet loss, an element of reliability. For clarity, the steps of the adopted VHO scheme are outlined below:

- *Step 1*: Active VHO initiation policies may trigger a VHO as well as select candidate networks.

- *Step 2*: A VHO decision policy determines the appropriate QoS factors and their weights for the current MN according to the class of traffic and/or network conditions.

- *Step 3*: The VHO scheme employs the AHP to compute the relative weight matrix used to calculate the priority of each QoS factor when selecting the new network. For instance, the following matrix example computes the relative weights for three parameters {availability (A), delay (D), and price (P)} by dividing the weights of step 2. The matrix is normalized, using (3.4) and (3.5), then averaged to obtain the final weights of each factor. An example is provided as follows:

$$\begin{array}{c} \\ A \\ D \\ J \\ L \\ B \\ P \end{array}\begin{array}{c} \begin{array}{cccccc} A & D & J & L & B & P \end{array} \\ \begin{bmatrix} 1 & rv_{AD} & rv_{AJ} & rv_{AL} & rv_{AB} & rv_{AP} \\ \frac{1}{rv_{AD}} & 1 & rv_{DJ} & rv_{DL} & rv_{DB} & rv_{DP} \\ \frac{1}{rv_{AJ}} & \frac{1}{rv_{DL}} & 1 & rv_{JL} & rv_{JB} & rv_{JP} \\ \frac{1}{rv_{AL}} & \frac{1}{rv_{DL}} & \frac{1}{rv_{JL}} & 1 & rv_{LB} & rv_{LP} \\ \frac{1}{rv_{AB}} & \frac{1}{rv_{DB}} & \frac{1}{rv_{JB}} & \frac{1}{rv_{LB}} & 1 & rv_{BP} \\ \frac{1}{rv_{AP}} & \frac{1}{rv_{DP}} & \frac{1}{rv_{JP}} & \frac{1}{rv_{LP}} & \frac{1}{rv_{BP}} & 1 \end{bmatrix} \end{array} \quad (3.4) \qquad \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} & a_{15} & a_{16} \\ a_{21} & a_{22} & a_{23} & a_{24} & a_{25} & a_{26} \\ a_{31} & a_{32} & a_{33} & a_{34} & a_{35} & a_{36} \\ a_{41} & a_{42} & a_{43} & a_{44} & a_{45} & a_{46} \\ a_{51} & a_{52} & a_{53} & a_{54} & a_{55} & a_{56} \\ a_{61} & a_{62} & a_{63} & a_{64} & a_{65} & a_{66} \end{bmatrix} \quad (3.5)$$

Example:

$$\begin{pmatrix} & A & D & P \\ A & 1 & 3/2 & 5/3 \\ D & 2/3 & 1 & 9/7 \\ P & 3/5 & 7/9 & 1 \end{pmatrix} \xRightarrow[normalize]{} \begin{pmatrix} 0.375 & 0.85 & 1.00 \\ 0.063 & 0.375 & 0.64 \\ 0 & 0.167 & 0.375 \end{pmatrix} \xRightarrow[clac.\ weights]{} \begin{pmatrix} 0.7396 \\ 0.3601 \\ 0.1806 \end{pmatrix}$$

- *Step 4*: A decision VHO policy defines the appropriate score function (e.g., additive weight vs. multiplicative weight aggregation) to calculate the score of each of the candidate networks, given the monitored QoS (i.e., network availability, delay and price in the previous example). A relative weight value $w_{factor\,i}$ indicates how much more important QoS factor i is than factor j [194]. The average across each row i is calculated (equation 3.6) to obtain the relative weights of the QoS factors.

$$w_{factor\,i} = \frac{a_{i1} + a_{i2} + a_{i3} + a_{i4} + a_{i5} + a_{i6}}{6} \quad (3.6)$$

The overall score of each candidate network is determined by the weighted sum of all the QoS attribute values. The score of a candidate network i is obtained using equation (3.7):

$$NW_{best} = \arg\max_{i \in M} \sum_{j=1}^{N} w_{factor\,i} r_{ij} \quad (3.7)$$

where N is the number of QoS parameters, M is the number of candidate networks, r_{ij} is the attribute value of metric j, $w_{factor\,i}$ denotes the weight factor of each QoS metric obtained from equation (3.6).

- *Step 5*: VHO policies control the remaining execution of the VHO procedure (e.g., waiting for a dwell time and calculating VHO failure before performing link up to session transfer).

3.2.3.6. Monitoring and Performance Evaluation Module

QoS monitoring is used to evaluate and improve the performance of the network or services for current and future customers to help maintain customer loyalty and increase profit accumulation for service providers. User-perceived QoS performance metrics are used to measure the integrity of the provided service and confirm the SLA honorability from the SP's side. Network and service management architecture designs must provide end-to-end network monitoring to provide quality guarantees and insure that the system operates with desirable configuration parameters, specifically policy configurations. A novel network monitoring approach is incorporated that aims at providing end-user quality measurements in an energy-efficient and traffic overhead minimization technique. The algorithm retrieves QoS performance data from mobile terminals, generates QoS reports, and analyzes the collected parameters against expected values. The algorithm is targeted specifically for mobile environments with limited power availability and absence of SP-provided monitoring probes. The designed self-monitoring module provides the following capabilities: node clustering, node similarity identification, and packet forwarding capabilities. The algorithm provides a similarity-based and stability-based QoS monitor selection strategy. Residual and transmission power levels, in addition to service request durations within a service session, indicate how stable a mobile node is within a wireless environment. Moreover, a comparison is made between nodes both semantically and in terms of relative velocity, to identify their similarity for embracing the QoS monitor selection strategy. The proposed algorithm's details are outlined in Chapter 5.

3.2.4. Architecture Component Interaction Overview

Having discussed the details of each component and briefly looked at the interaction between the components, subsequently a flowchart is provided, illustrated in Figure 3.6, of the interactions between the architecture components from Figure 3.2. Based on the measured network and user behavior using the monitoring and performance evaluation

module, the SM may be triggered to examine the system behavior through simulation in three different states. These include: a current SLA guarantee is violated (QoS degradation), or a degradation in the VHO operation is detected through various parameters (e.g., VHO latency, handover frequency) using the mobile node monitoring capabilities, or a deviation of the real-time simulator model from that of the real network. In the first two states, the SM aims at detecting the causes of violations and service degradation and decides whether to modify the VHO or the SLA manager behavior or not after crediting the customer with the incurred penalties. This decision is taken after the VC optimizer provides the adequate set of optimal policy configurations to the SM. On the contrary, when the third state is encountered, the SM employs the collected measurements to modify the current configuration of the simulator, and hence, synchronizing its behavior with that of the actual network.

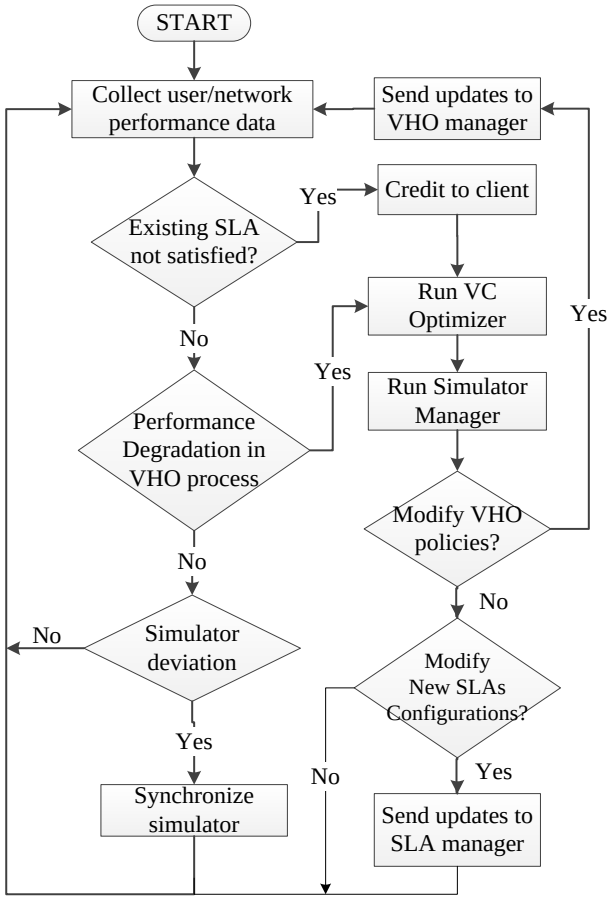

Figure 3.6 Flowchart diagram of the interaction between the five architecture components.

3.3. PERFORMANCE EVALUATION

In this section, performance results obtained from simulating the proposed system using NS-2 [187] are presented. The features provided by the NIST mobility patch [195] are extended to include a handover decision engine that implements the handover scheme using AHP and SAW [34]. This framework is compared with the statically configured AHP/SAW scheme [34], [196]. Two simulation scenarios are adopted: the main focus of the first is to provide some insight with regard to the performance of the simulator manager, while the second scenario demonstrates the performance of the overall architecture with respect to the increase in the overall network revenue and the achieved load balancing. Finally, it is significant to note that in this research's current implementation of the proposed framework, a simple VHO procedure was adopted based on AHP [197]. This procedure is used in both the simulator and network to determine the weight of each QoS factor. The overall score of each candidate network is calculated using various cost functions such as the simple additive or multiplicative weighting functions.

3.3.1. Simulation Results with a Single Service Provider

As shown in Figure 3.7, the simulated scenario consists of a mobile node (MN) that is equipped with UMTS, WiMAX and WLAN interfaces. The simulation starts with UMTS as the only available network where the MN is first connected. As the MN moves, it enters WiMAX coverage (at t = 23 sec). At this point, there are two possibilities, either the MN stays in the same network or traffic is redirected to WiMAX. The MN continues to move until it enters WLAN coverage (at t = 60 sec) where three possibilities arise: traffic is directed to UMTS, to WiMAX, or to WLAN. Finally, the MN leaves the WLAN coverage (at t = 73 sec) creating two possibilities: traffic direction via UMTS, or via WiMAX.

In the simulated scenario, the packet size is set to 1240 bytes [153] using variable bit rate (VBR) traffic with 0.1 sec intervals originating at the server. The settings for WiMAX are as follows: coverage has a 500 m radius, the connection broadcast messages (the Downlink Channel Descriptor (DCD) and Uplink Channel Descriptor (UCD)) are transmitted every 5 sec with 4 msec frame duration. On the other hand, the WLAN coverage has a 20 m radius, data rate of 11 Mbps, and beacon interval at t = 0.1 sec. The UMTS base station has a coverage of 1km with a data rate of 384Kbps.

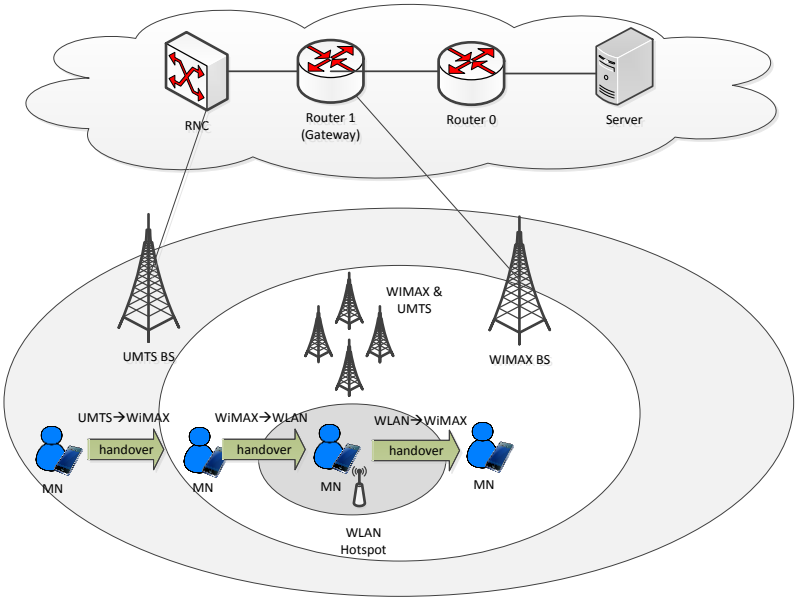

Figure 3.7 Simulated network topology in scenario 1.

Figure 3.8 shows the achieved throughput for the MN with the AHP/SAW scheme. Handover occurs at the following time intervals: t = 23 sec (UMTS to WiMAX), t = 60 sec (WiMAX to WLAN), t = 73 sec (WLAN to WIMAX), and t = 113 sec thereafter, there is a ping-pong handover effect occurring between WiMAX and UMTS. Hence, this results in degrading the average achieved throughput.

Figure 3.9 shows the throughput achieved when employing the proposed framework. As is shown, the first fluctuation between the two networks triggers the simulator manager to test four scenarios and apply a new set of policies to reconfigure the VHO policies. Thereby, resulting in better throughput and smaller delay.

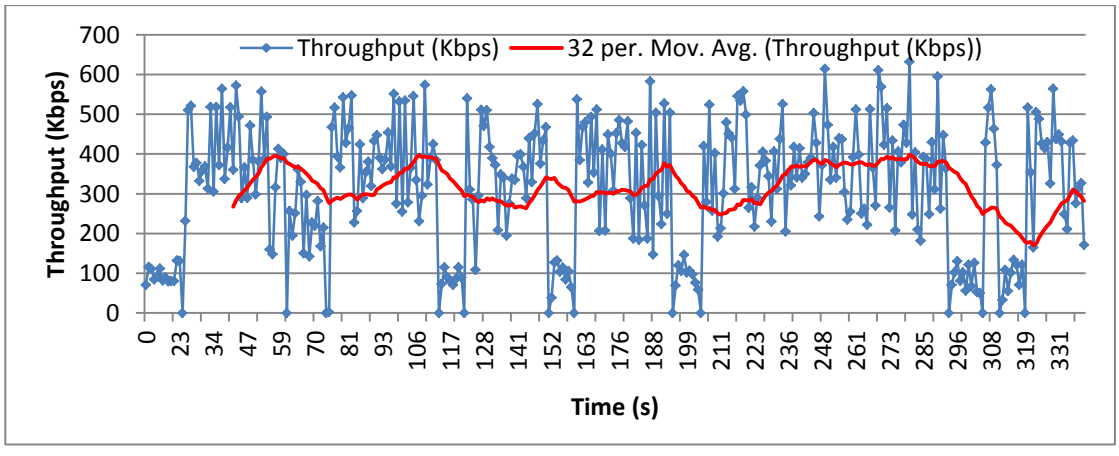

Figure 3.8 Achieved mobile node throughput with static AHP/SAW scheme.

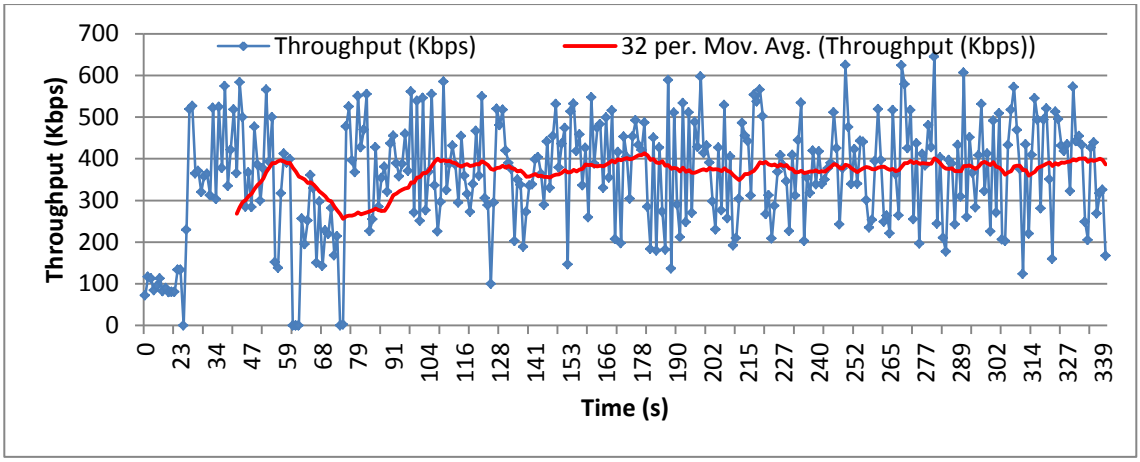

Figure 3.9 Achieved mobile node throughput with the proposed scheme.

Figure 3.10 compares the delay resulting from the proposed approach to that arising from the AHP/SAW scheme. The figure shows a substantial improvement in the delay upon the application of new policies introduced by the SM.

To provide some insight on the performance of the SM, Figures 3.11 and 3.12 depict the throughput and delay, respectively, for the simulation runs that were triggered, where different QoS performance measurements are analyzed to discover the best VHO configuration policies. For example, Figure 3.12 depicts that the smallest experienced delay is achieved with the VHO configurations in 'simulation run 2'.

Since 'simulation run 2' provides the best QoS performance results, the policies and configurations of this simulation are applied to the actual VHO scheme.

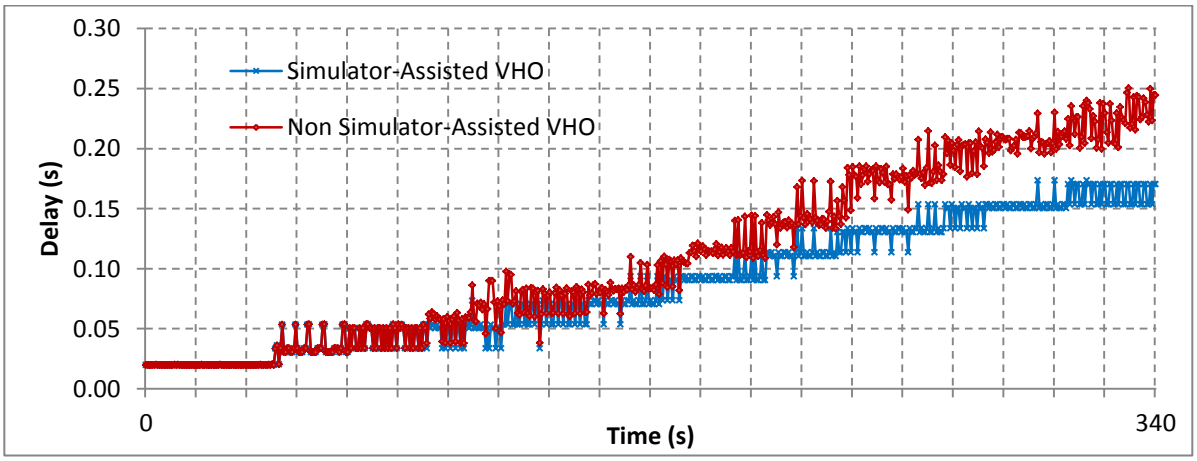

Figure 3.10 The achieved delay for the proposed framework.

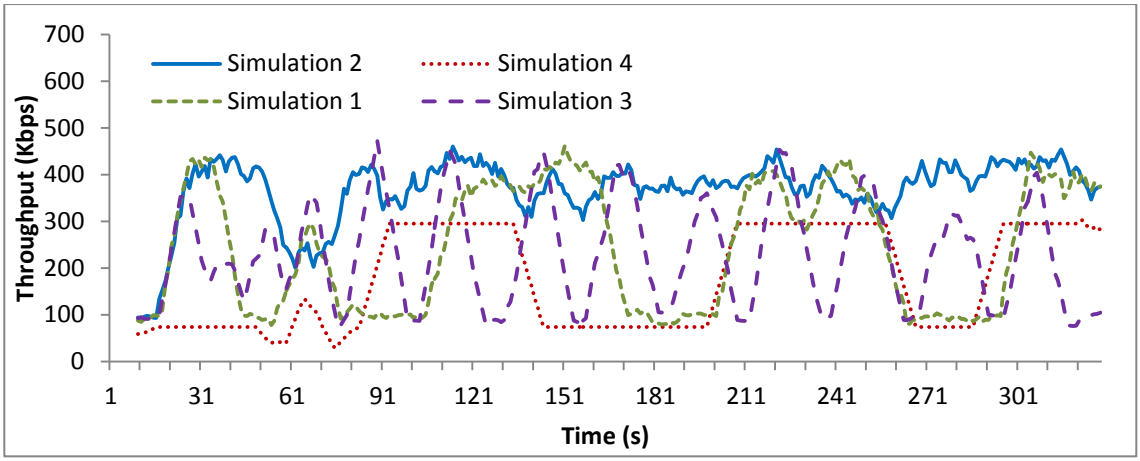

Figure 3.11 Throughput performance for four different simulation scenarios with different applied policy configurations.

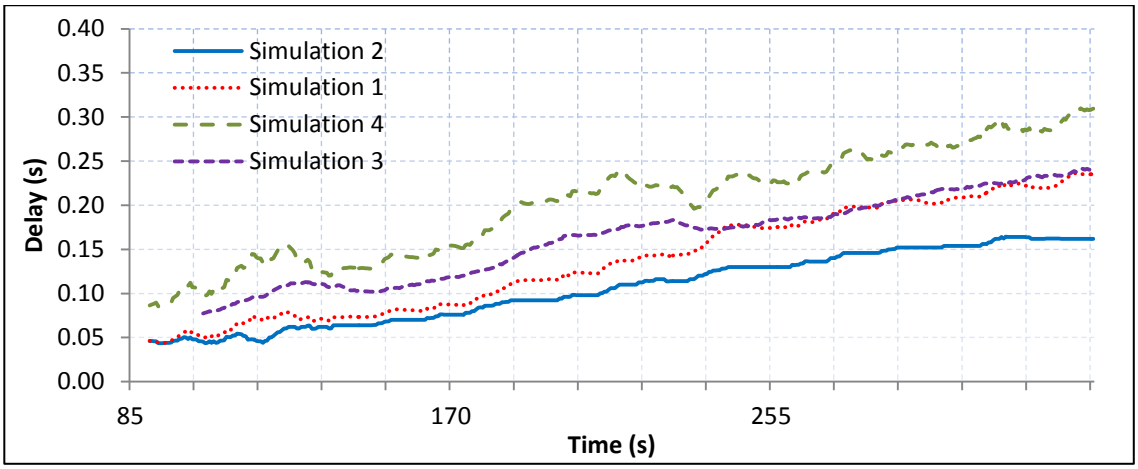

Figure 3.12 Delay performance for four different simulation scenarios with different applied policy configurations.

3.3.2. Simulation Results with Multiple Service Providers

In this set of experiments, the observed QoS performance $\tilde{\boldsymbol{q}}_{\boldsymbol{ij}}(t)$ is measured through the NS-2 simulator in a network scenario, illustrated in Figure 3.13. The performance of the architecture is illustrated with respect to the increase in the SP profit and the achieved load balance among the access points. The experiment consists of clients that have the option to acquire a service by connecting to one of the SPs available. Thus, a client can connect to a UMTS network through BS1 or to a WIMAX network through BS2, owned by two service providers SP_1 and SP_2, respectively. Also, there are three WLANs managed by three different SPs, namely, SP_3, SP_4 and SP_5. The first has two access points, namely AP1 and

AP2; the second includes AP3 and AP4, while the third SP has a single AP, namely, AP5. The packet size is set to 1240 bytes using constant bit rate (CBR) traffic with 0.1 sec. intervals originating at the server. The WiMAX network has a coverage area of 500 m radius with a 2 Mbps bandwidth service rate. On the other hand, each of the access points AP1-AP5 has a coverage radius of 20 m, a data rate of 11 Mbps, and beacon interval of t = 0.1 sec. The UMTS base station has a coverage of 1km and a data rate of 384 Kbps. Three user classes of service (CoS) are available in which each class is identified by the offered service data rate of a MN. Thus the services are available according to the following rates: 64 Kbps, 128 Kbps, and 256 Kbps. When a new MN arrives, the requested CoS is randomly selected from the three allowed data rates.

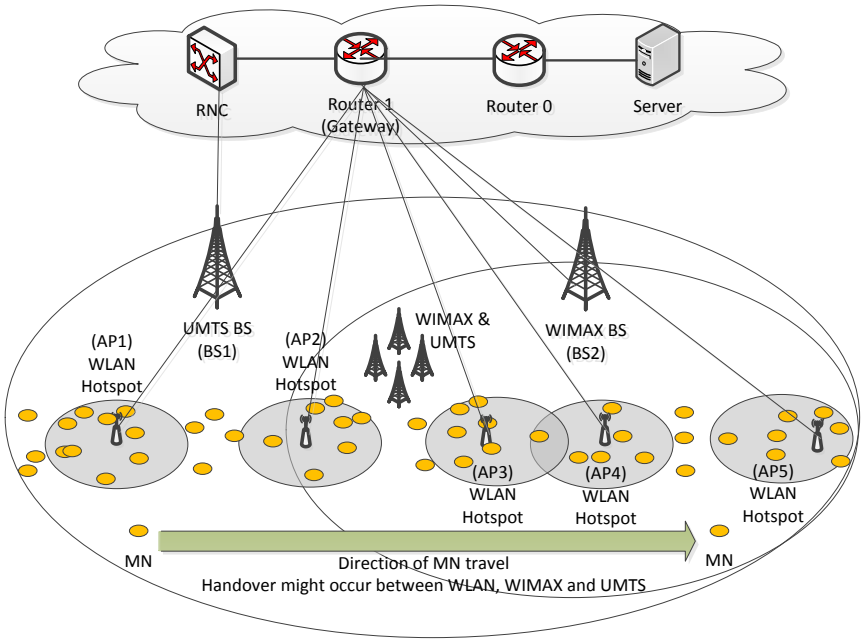

Figure 3.13 Simulated network topology 2.

3.3.2.1. Experiment 1

In this experiment, the arrival rate of the MNs was set at a constant rate of 0.1 user/sec and moved randomly in the network. All MNs enter the network from one end and stop at different locations for the remainder of the experiment. Looking at Figure 3.14. it can be seen that within three time units most MNs join BS1, or more precisely most MNs keep their connection with BS1, where at 300 s of the simulation run the load on BS1 reaches 1472 Kbps. This is already surpassing the limit for the downlink channel of the UMTS access technology (384 Kbps). A large delay will be experienced by most of the MNs, particularly

the ones that requested a data rate of 128 Kbps or higher. On the fourth time unit it is evident that most of the MNs have left BS1 and either joined the WiMAX network or other WLAN AP. Almost all new MNs will join either BS2 or APs 1-5 to avoid delay and retransmissions due to the high byte error rate.

Unfortunately, the same issue will be repeated but this time with BS2, where from Figure 3.14, it is evident that beyond the seventh time unit, BS2 surpasses the limit of the WiMAX downlink channel (2 Mbps). Therefore, it is clear that without adapting new sets of policies according to the current network conditions, nodes might be clustered into a small set of networks while keeping others vacant. This situation will cause a negative effect on both the SP profit and users satisfaction. Penalties will be incurred on the SP because it is not able to meet the SLAs requirements and users will experience delay and degradation in the QoS requested.

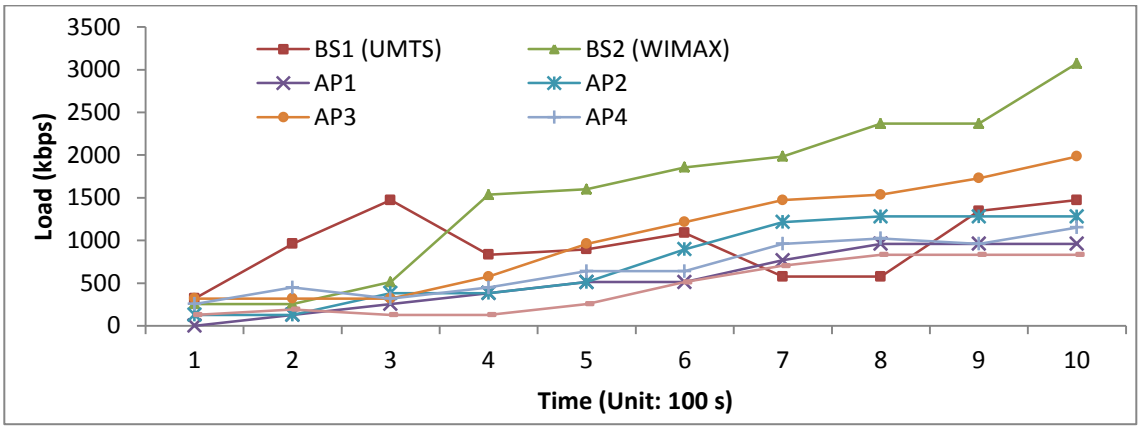

Figure 3.14 Achieved load versus time with 100 mobile nodes without using the proposed scheme.

To illustrate the advantage of the proposed architecture, Figure 3.15 illustrates that MNs do not join networks that will not meet their demands at the moment of arrival and in the future. A balance is kept at each BS and AP, where the right number of users is provided with the required service without incurring penalties on the SP. From the Figure, it is evident that BS1's load is kept around 384 Kbps, BS2's load does not exceed the limit for the WiMAX downlink channel (2 Mbps) and the load is distributed among the five APs so that potential customers can freely join the different APs. The detailed load distribution data corresponding to these figures is given in Table I.

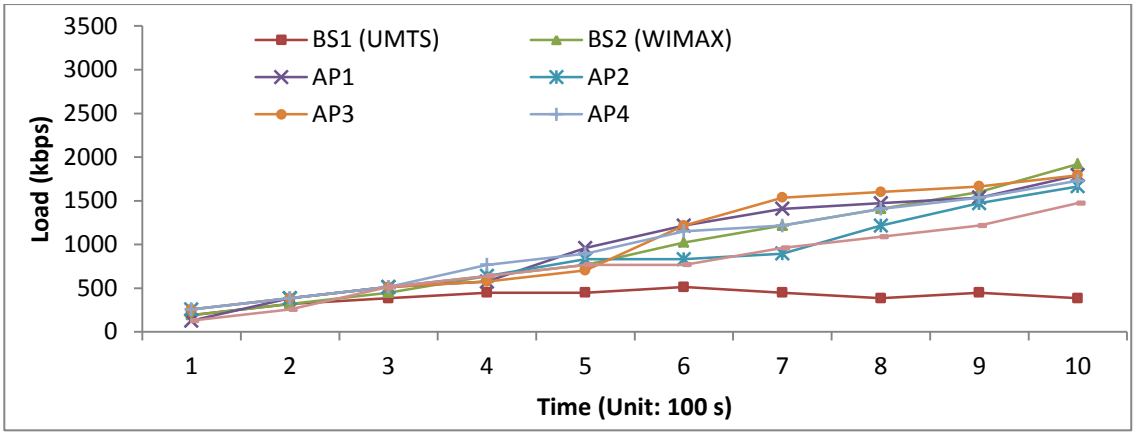

Figure 3.15 Achieved load versus time with 100 mobile nodes when using the proposed scheme.

TABLE I
APS AND BSS LOADS (IN KBPS) FOR EXPERIMENT 1 (TIME UNIT: 100 S)

	TRADITIONAL SCHEME							PROPOSED ARCHITECTURE						
Time	BS1	BS2	AP1	AP2	AP3	AP4	AP5	BS1	BS2	AP1	AP2	AP3	AP4	AP5
1	320	256	0	128	320	256	128	192	192	128	256	256	256	128
2	960	256	128	128	320	448	192	320	320	384	384	384	384	256
3	1472	512	256	384	320	320	128	384	448	512	512	512	512	512
4	832	1536	384	384	576	448	128	448	640	576	640	576	768	640
5	896	1600	512	512	960	640	256	448	768	960	832	704	896	768
6	1088	1856	512	896	1216	640	512	512	1024	1216	832	1216	1152	768
7	576	1984	768	1216	1472	960	704	448	1216	1408	896	1536	1216	960
8	576	2368	960	1280	1536	1024	832	384	1408	1472	1216	1600	1408	1088
9	1344	2368	960	1280	1728	960	832	448	1600	1536	1472	1664	1536	1216
10	1472	3072	960	1280	1984	1152	832	384	1920	1792	1664	1792	1728	1472

3.3.2.2. Experiment 2

In the second experiment, 100 MNs are randomly placed simultaneously in the environment at t = 0. The MNs are then allowed to randomly move within the network. Simulation runs with duration of 1000 sec each were performed, comparing the performance of the proposed architecture with the fixed AHP/SAW scheme; measurements were taken at intervals of 100 s. In Figures 3.16 and 3.17, the overall load is plotted at each AP and BS versus time. Comparing these two figures, it can be shown that the load is fairly distributed among the APs and BSs using the proposed architecture. The detailed load distribution data corresponding to these two figures is given in Table II. Clearly, BS1's load starts off with a large number of MNs associated with it, causing the QoS to deteriorate and provoking many of the MNs to handover to other available networks at the same time. This leads to an

increase in the number of handovers and total handover delay. Not only will users be affected, but also the SP will have profit drawbacks. Any QoS level which was not met according to the SLA signed between the customer and the provider will incur penalties on the SP, hence decreasing the total profit of the SP as depicted in Figure 3.19. For that reason, to achieve stability in the overall environment, the system used in this particular experiment will monitor the performance of the network and based on the results, will simulate a new set of handover configuration policies. Policies which show a positive effect in the overall performance are chosen and are applied to the actual network handover scheme.

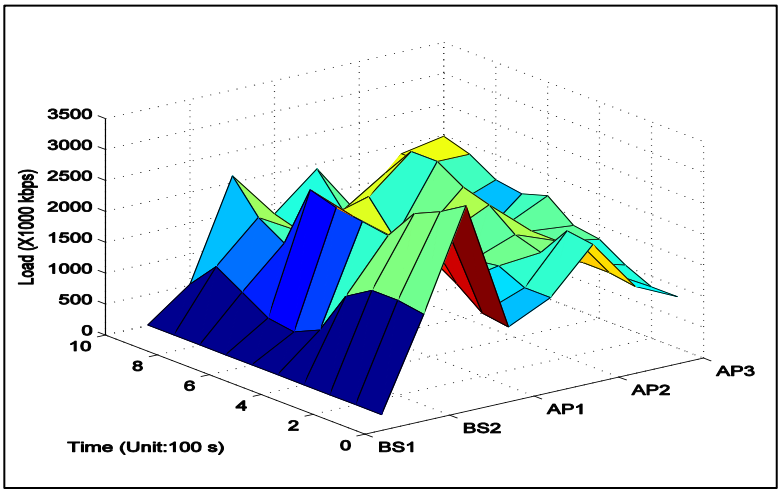

Figure 3.16 AP and BS load at different time instants without using the proposed scheme.

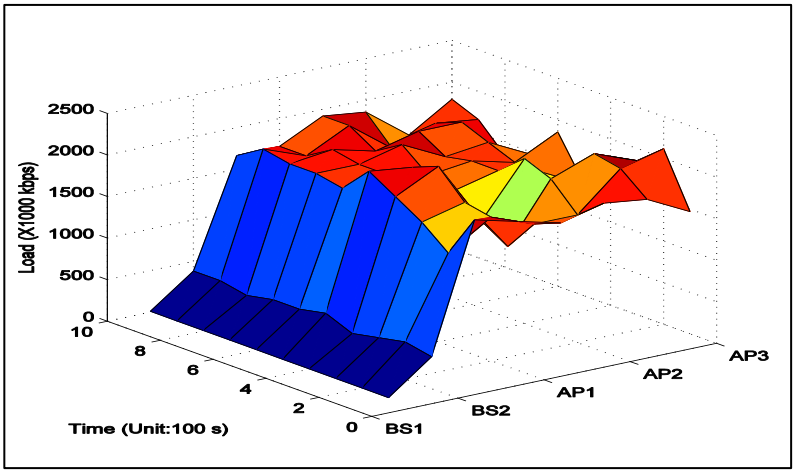

Figure 3.17 AP and BS load at different time instants using the proposed scheme.

Results recorded in Figure 3.17 and Table II show that the load is distributed among the APs and BSs, causing a decrease in the number of unwanted handovers and handover delay as illustrated in Figure 3.18.

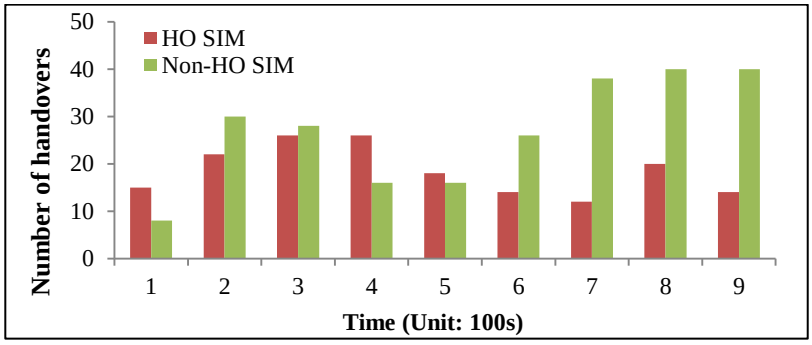

Figure 3.18 Number of handovers versus time with 100 mobile nodes.

TABLE II
APS AND BSS LOADS (IN KBPS) FOR EXPERIMENT 2 (TIME UNIT: 100 S)

	TRADITIONAL SCHEME							PROPOSED ARCHITECTURE						
Time	BS1	BS2	AP1	AP2	AP3	AP4	AP5	BS1	BS2	AP1	AP2	AP3	AP4	AP5
1	1472	3072	960	1280	1984	1152	832	192	192	128	256	256	256	128
2	1536	2816	1024	1280	2048	1216	832	320	320	384	384	384	384	256
3	1536	2624	1344	1472	1536	1216	1024	384	448	512	512	512	512	512
4	1280	2112	1600	1408	1664	1408	1280	448	640	576	640	576	768	640
5	576	2176	1600	1664	1984	1408	1344	448	768	960	832	704	896	768
6	384	2240	1536	1472	2112	1344	1664	512	1024	1216	832	1216	1152	768
7	448	2368	1792	1280	2368	960	1536	448	1216	1408	896	1536	1216	960
8	704	1344	1600	1792	2368	1344	1600	384	1408	1472	1216	1600	1408	1088
9	960	1600	1216	1408	1856	1856	1856	448	1600	1536	1472	1664	1536	1216
10	512	2112	1344	1920	1024	1856	1984	384	1920	1792	1664	1792	1728	1472

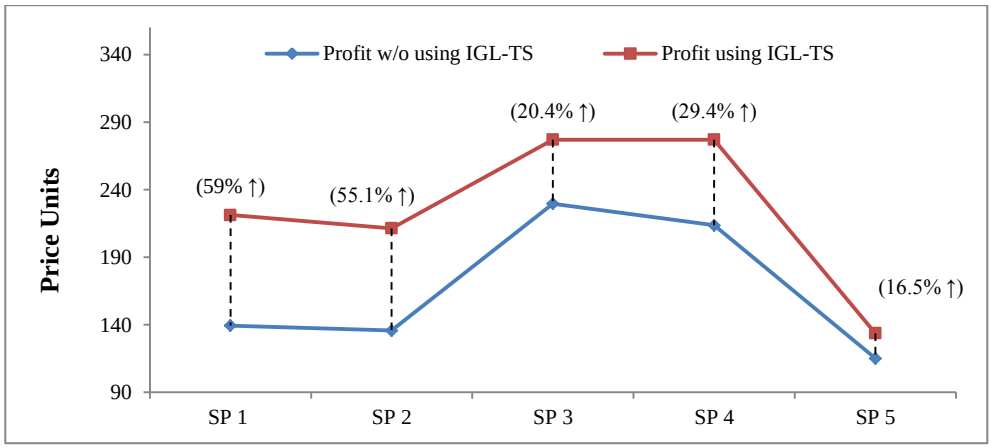

Figure 3.19 A comparison of the total gained profit for each SP between the proposed and traditional schemes.

Inability to meet the requirements of users, and failure to abide to guarantees provided to the clients through SLAs will dissatisfy mobile users, as well as have profit drawbacks on the SP. Any QoS level which was not met according to the SLA signed between the customer and the provider will incur penalties on the SP, hence decreasing the total profit of the SP, as depicted in Figure 3.19. The average cooperative profit for all SPs is maximized by 39.6%, and specifically SP1 has gained the most profit increase i.e. 59%, compared to SP5 which gained the least profit increase at 16.5%.

3.4. SUMMARY

This chapter presented the research's multi-component policy-based network and service management architecture in wireless heterogeneous network environments. The benefit of the proposed work lies in that given sets of objectives, constraints and goals, network policies are applied to govern the behavior of network entities, specifically user behavior. The use of policies that are functions of different parameters have been proven to give more freedom to users and service providers to describe their requirements, in a continuously changing manner. The design of the framework proposed a separation between five major correlated components: the simulation manager, the handover manager, the SLA manager, the variable configuration optimizer, and the network monitoring and performance evaluation module. The proposed mechanism provides proactive response to possible deteriorations in the delivered QoS, aiming to achieve user-accepted services. The algorithm also allows for utilization of knowledge of past experiences learnt from previous management decisions to adapt to new policy and SLA configurations to best suit current environment status. The search process for new configurations is conducted through the use of a modified meta-heuristic search mechanism, which will be discussed in detail in the next chapter.

CHAPTER 4

A TABU-SEARCH ASSISTED VARIABLE CONFIGURATION OPTIMIZER

Many problems are continuous or abundant in reality, and in some cases finding solutions for these problems is quite challenging. Finding a solution is often computationally too expensive in terms of processing power capabilities and time sensitivity. Today's advanced development in computer technologies has achieved a substantial increase in computational speed. Nonetheless, this advancement is not adequate alone, particularly if the size of the problem is large. Several new techniques that rely on heuristic approaches have been contemplated. Heuristic solutions are popular for the reasonable computational time required to tackle the problem, and the approximate and near-accurate solution provided.

In the previous chapter the multi-component policy-based network management architecture was introduced. It was shown that the behaviour of all the components and the Simulator Manager in particular, depend on the results of the Variable Configuration Optimizer. This chapter presents a robust scheme to reliably obtain policy configuration variables for near-optimal network behaviour. The scheme relies on an enhanced version of the tabu-search algorithm that not only finds the best solution within the scenario search space, but also minimizes the computational time needed to find these configurations.

This chapter proceeds as follows. Section 4.1 introduces concepts of heuristic search schemes and highlights the motivation behind the adapted scheme. Section 4.2 introduces the necessary background information about tabu-search and the key ideas behind the utilization of tabu-search for policy variable configuration selection. The proposed scheme is described in Section 4.3. In Section 4.4, linear regression and its utilization in the proposed scheme is

discussed. In Section 4.5, performance evaluation results of the adapted VCO module are presented. Finally, Section 4.6 concludes the chapter.

4.1. OPTIMAL VARIABLE RECONFIGURATION

4.1.1. Overview of the Search Problem

In the previous chapter, it was brought to the reader's attention that the process of optimal configuration selection starts with indications of performance degradation. This is received from the performance evaluation module, which in turn, triggers the simulator manager to simulate a set of scenarios composed of a set of parameter configurations. These configurations must be dynamically reconfigured using the variable configuration optimizer to cope with network changes. Since a single simulation run for a specific network model with a reasonable number of nodes can take an excess amount of time, it is necessary to minimize the number of simulation runs. This minimization process is achieved using a search technique that selects the optimal solution from a large search space. Provided with the optimal configurations from the VCO, the SM will test one or more new configuration policies for SP user handoff, VHO scheme or SLA reconfigurations. Finally, the SM will communicate the new policies to the SLA and VHO managers.

The objective that must be considered in the search problem is to maximize service provider profit and user satisfaction while adhering to obligations provided by the contracted SLA. Hence, at each time epoch t, the SPs face the following problem:

$$\boldsymbol{P}\text{: } \left(\text{maximize} \sum_{i=1}^{N} \sum_{j=1}^{U(t)} \mathcal{P}\left(\boldsymbol{q_{ij}}, \tilde{\boldsymbol{q}}_{\boldsymbol{ij}}(t)\right) \right) \tag{4.1}$$

The SPs must minimize the number of times an SLA is violated. Strict constraints are posed on the time taken to reconfigure the underlying network, ensuring that the time taken to maximize the profit is less than the time entailed to incur SP penalties. In addition, the solution must satisfy the resource constraints:

$$\tilde{r}_{ij}(t) \leq \boldsymbol{R}_{\mathrm{i}} \qquad for\ i = 1, \dots, N, j = 1, \dots, U(t). \tag{4.2}$$

i.e. the sum of the service rate allocated to all the users in a particular network must not exceed the provider's total resources $\boldsymbol{R}_{\mathrm{i}}$. Later in this dissertation, it will be shown how the proposed management framework, aided with the help of a run-time simulator, can

continuously fine tune the configurations of the underlying networks in order to achieve a maximized profit.

4.1.2. Solution Search Methods

Faced with the challenge of solving a hard optimization problem, classical methods often encounter great difficulty, even when equipped with a theoretical guarantee of finding an optimal solution. Most of the time, using efficient heuristic algorithms will outperform other methods with regard to computation time. However, there is no guarantee that such methods perform well for all optimization problems.

4.1.2.1. Classical Heuristics

Classical heuristic methods can be categorized according to their application area [198]. These include the following:

- **Construction techniques**. These types of algorithms generate a solution through adding components such as variables one at a time and individually until a feasible solution is obtained. Most construction algorithms do not reach a feasible solution till the end of the search. The most commonly used construction techniques are the *greedy algorithms*. Such methods seek to maximize the improvement at each step. They start from a given solution in which the best move to improve the solution is chosen at each iteration. Another well-known search method is the *look-ahead algorithm*. It estimates the sequence of candidate solutions at each iteration. Candidate solutions which may lead to a bad final solution are discarded.

- **Improvement techniques**. These types of heuristics are also known as *local search methods*. Improvement heuristics start with a feasible solution and are improved through a sequence of exchanges in a local search. A neighborhood of candidate solutions is defined for each solution, hence the name *local search*. A move to a neighboring solution is selected if the new solution is better than the current solution, until the local optimum is found. Combinations of construction and improvement techniques have been used to obtain an initial starting solution using construction

techniques, while improvement techniques are used to search and find the local optimum to improve the solution.

- **Mathematical programming techniques**. Such approaches use mathematical optimization models and exact solution procedures together to modify the solution procedure and obtain efficient heuristic to solve the problem. Mathematical programming techniques are more application-specific and their success depends greatly on the design process, thus giving developers more opportunities to blend in different procedures together.

- **Decomposition techniques**. In this type of approach, the problem is solved by dividing it into a sequence of smaller problems, the output of one being the input to the other, and then inductively merging the solutions together. Many scheduling heuristics use decomposition techniques.

- **Partitioning techniques.** These types of heuristics break or partition the problem into smaller sub-problems, similar to decomposition. However, each sub-problem is solved independently and then merged together into a single solution to the overall problem.

- **Solution space restriction techniques.** Such methods restrict the set of solutions to simplify the problem. These particular restrictions will allow a search only among a set of solutions with specific properties. A drawback of such techniques is that an optimal solution to a restricted problem will most likely not be a global optimum to the problem.

- **Relaxation techniques.** Such approaches are the exact opposite of restriction techniques. The solution space is expanded to obtain a manageable problem.

4.1.2.2. Meta-heuristics

Heuristic methods were introduced more than half a century ago to solve specific problems. A more general heuristic methodology was introduced called *meta-heuristics*. Meta-heuristics is defined by Osman et al. [199] as follows: "*A meta-heuristic is formally*

defined as an iterative generation process which guides a subordinate heuristic by combining intelligently different concepts for exploring and exploiting the search space. Learning strategies are used to structure information in order to find efficiently near-optimal solutions." Given this definition, it can be deduced that meta-heuristics are strategies that guide the search process. They explore the search space in order to find optimal or near-optimal solutions. Many meta-heuristic algorithms exist that range from simple local search procedures to complex learning processes. It is important to note that meta-heuristics are approximate and are not problem-specific. Today, more advanced methods use search experience learned previously to guide the search process. Similar to heuristic techniques, meta-heuristics can be classified as follows [200]:

- **Population-based vs. single point search techniques**. Population-based approaches perform search processes on a set of points as a group in the search space. On the contrary, single point search techniques are usually called trajectory methods and encompass local search-based meta-heuristics.

- **Static vs. dynamic objective function search techniques**. Some meta-heuristics keep the objective function fixed during the search. While other approaches such as the *Guided Local Search* (GLS) modify the objective function to escape from local minima. The techniques are called static and dynamic objective function search methods, respectively.

- **Memory usage vs**. **memory-less search techniques.** Other methods of classifying meta-heuristics are according to the use of search history (i.e. memory). When the current state of the search process is used to determine the next location, algorithms of such type are called memory-less techniques and are said to perform a Markov process [203]. On the contrary, techniques that depend on search history can be classified either as short-term or long-term memory. Short-term memory keeps track of recently performed visited solutions (i.e. moves) or decisions taken. Long-term memory, on the other hand, is an accumulation of parameters about the search. Most powerful meta-heuristics nowadays make use of memory to provide more accurate decision-making.

- **Single vs. various neighborhood search techniques.** Most search algorithms have a single neighborhood structure, such that the topology does not change in the course of the algorithm. On the contrary, there exists some meta-heuristic algorithms such as *Variable Neighborhood Search* (VNS) [204], that use a set of neighborhood structures to diversify the search process.

4.1.3. Variable Configuration Search Method Requirements

Given that a policy's parameters require dynamic reconfiguration in accordance to network or service quality deterioration, it is necessary that the search technique provide optimal solutions. A challenge that arises from the dynamicity of a mobile environment and having service provider cooperation is the time sensitivity due to user mobility and the handover problem. Thus optimal configurations must be found within a limited time frame, hence a fast local search procedure must be considered. The search technique must also be capable of using memory to keep track of visited solutions. The search rules must be flexible to allow previously optimal solutions to be chosen or looked at more frequently than other solutions. Since users overall share similar mobility descriptions, previously chosen solutions for a particular set of users can be applied to other sets of users under certain circumstances. Hence, the search method should provide a fast and accurate solution.

Over the last decade, the *Tabu-Search* (TS) [46] heuristic approach has been one of the most used meta-heuristics for solving optimization problems, due to its fast search performance. Nonetheless, modifications must be applied to such search algorithms in areas of discrete variable selection problems to acquire the optimal values for configuration variables at the network, handover, policy and SLA level. Moreover, the configuration learning process must be decoupled from that of the actual configuration adaptation step, so as to avoid unnecessary time delays in the management process. In the next section, background information about Tabu-Search is presented for the purpose of providing a better understanding for the reader of the modified variable selection method.

4.2. TABU SEARCH

One of the most utilized and cited meta-heuristics employed for solving optimization problems is the Tabu Search (TS) method. It was first introduced in [201] and originated as a

device for implementing oscillating assignment strategies [202]. TS is a meta-heuristic that guides a local heuristic search procedure to explore a solution space beyond a local optima. The local search procedure uses the *move* operation to define the neighborhood of a given solution. To provide search behaviour flexibility, TS uses adaptive memory. The adaptive memory feature of TS allows the implementation of procedures that are capable of searching the solution space efficiently, such that previously visited solutions are used to analyze current alternatives. The term *tabu* refers to something that is forbidden or banned, hence in TS, *tabus* are used to prevent cycling and going back to local optima. Solutions which are marked as tabu are stored in a short-term memory referred to as a *tabu list*. Most recent visited solutions stored in the tabu list are prevented from being revisited in the future. This will prevent moves from being repeated or reversed and restricts the neighborhood of the current solution to solutions that do not belong to the tabu list. The latter is referred to as the *allowed set*, and thus the size of the solution space is reduced.

The process of tabu search in its simplest form is as follows: at each iteration, the best solution from the allowed set is chosen as the new solution and is also added to the tabu list. This is carried out while removing one from the list, according to specific techniques, usually in a *first in first out* (FIFO) order. The length of the tabu list is referred to as *tabu tenure*. It is used to control the memory of the search process and can either be dynamic (varying tabu tenure) or static (fixed tabu tenure). Large tabu tenures allow for a larger search space scope and prevent solution cycling.

A problem with incorporating tabus is that they may lead to the loss of some unvisited good solutions. Thus, additional precautions must be taken to avoid missing good solutions. Strategies that incorporate such features are referred to as *aspiration criteria*. Aspiration criteria allow moves considered tabu to be overridden and consider the solutions to be admissible and allowed to be visited.

Two other important components of TS that make it fully effective are *intensification* and *diversification* strategies. *Intensification* forces the search to explore more portions of the search space. It allows for modification of choice rules to encourage moves to solutions historically found to be good or promising. This will ensure that the best solution is indeed found. In addition to considering solutions that are close to unvisited solutions by means of the standard move mechanism, intensification will generate neighbors from previously

visited good solutions. On the contrary, *Diversification* forces the search process to examine unvisited regions of the exploration space. It generates solutions that differ in various significant ways from those seen before. It is based on a long-term memory strategy that records the number of iterations that various solution elements have been involved within the current solution. Different types of diversification strategies exist. For example, *restart diversification* forces the algorithm to restart the search at certain rarely used points.

As in any meta-heuristic search strategy, different termination conditions can be applied. Some examples of termination conditions are: terminate after a certain number of iterations, or after a certain time threshold, or after a number of iterations where no improvement to the objective function has been visible, or finally, terminate if the objective function reaches a certain threshold value. The algorithm is illustrated in Algorithm 1 for further clarity.

Algorithm 1: Tabu Search Algorithm

```
Set S is the search space
Choose an initial solution s
Define the neighborhood of s: N(s) ⊆ S
Store best solution thus far s′
Define initial tabu list TL and allowed set AL if applicable
begin
    TL ← ∅
    replace s′ ← s, where s′ϵ N(s)
    update(TL)
    do
        AL ← N(s) − TL
        //choose best solutionfrom AL
        if f(s) < f(s′) then
            replace s′ ← s
        update TL
    until termination condition
    return s′
end
```

4.3. THE ITERATIVE GLOBAL AND LOCAL TABU-SEARCH (IGL-TS) ALGORITHM

Suppose that up to time point t, the network performance was within the normal range. Then at time $t + \Delta t$, the performance evaluation module detected a service qualtiy degradation. To maximize the SPs profits $\mathcal{P}\left(\boldsymbol{q_{ij}}, \widetilde{\boldsymbol{q}}_{\boldsymbol{ij}}(t)\right)$ for each user as outlined in (8), it is

necessary to apply an optimal or near-optimal set of policy configurations. In order to re-optimize the network behavior, this step is realized through the support of a scenario search mechanism. Indications of performance degradation, received from the performance evaluation module, trigger the SM to instruct the run-time simulator to run a set of simulation scenarios. Configurations are provided by the VC optimizer to test new policy configurations.

To automate the process of policy reconfiguration represented by a policy vector $\boldsymbol{x}$, as explained earlier, a modified tabu-search algorithm called *Iterated Global and Local - Tabu Search* (IGL-TS) is developed. The policy vector $\boldsymbol{x}$ represents policy configurations that need to be evaluated in terms of the observed QoS performance $\tilde{\boldsymbol{q}}_{\boldsymbol{ij}}(t)$ through the simulator summarized by the behavior function $B\big(\boldsymbol{x}(t)\big)$ at time t:

$$B\big(\boldsymbol{x}(t)\big) = \sum_{i=1}^{N} \sum_{j=1}^{U(t)} \sum_{k=1}^{|K|} \frac{score\left(\tilde{q}_{ij}^{k}(t)\right)}{N \times U(t) \times |K|} \tag{4.3}$$

where $score\left(\tilde{q}_{ij}^{k}(t)\right)$ is a score function obtained for all QoS aspects K. If $k \in K$ has a positive effect on the total outcome (i.e., QoS aspects that the system wishes to maximize such as throughput), then the score is calculated as follows:

$$score\left(\tilde{q}_{ij}^{k}(t)\right) = \begin{cases} 0, & if\ \tilde{q}_{ij}^{k} < \underline{q}_{ij}^{k} \\ \frac{\tilde{q}_{ij}^{k} - \underline{q}_{ij}^{k}}{\overline{q}_{ij}^{k} - \underline{q}_{ij}^{k}}, & if\ \underline{q}_{ij}^{k} \le \tilde{q}_{ij}^{k} \le \overline{q}_{ij}^{k} \\ 1, & if\ \tilde{q}_{ij}^{k} > \overline{q}_{ij}^{k} \end{cases} \tag{4.4}$$

On the contrary, if k has a negative effect (i.e., QoS aspects that the system wishes to minimize such as delay and jitter), then $score\left(\tilde{q}_{ij}^{k}(t)\right)$ is calculated as follows:

$$score\left(\tilde{q}_{ij}^{k}(t)\right) = \begin{cases} 0, & if\ \tilde{q}_{ij}^{k} > \overline{q}_{ij}^{k} \\ \frac{\overline{q}_{ij}^{k} - \tilde{q}_{ij}^{k}}{\overline{q}_{ij}^{k} - \underline{q}_{ij}^{k}}, & if\ \underline{q}_{ij}^{k} \le \tilde{q}_{ij}^{k} \le \overline{q}_{ij}^{k} \\ 1, & if\ \tilde{q}_{ij}^{k} < \underline{q}_{ij}^{k} \end{cases} \tag{4.5}$$

where $\underline{q}_{ij}^{k}$ is the minimum acceptable QoS value and $\overline{q}_{ij}^{k}$ is the maximum acceptable value for a service provided by SP_i, $\tilde{q}_{ij}^{k}$ represents the observed QoS levels for the set of concerned QoS aspects. Using this function, the SM can calculate a score for the current configuration

of the network $\boldsymbol{x_0}$, both at the time of normal operation $\boldsymbol{x_0}(t), B\big(\boldsymbol{x_0}(t)\big)$, and when a service quality degradation is observed, $B\big(\boldsymbol{x_0}(t+\Delta t)\big)$.

Due to time restrictions in mobile environments, network configurations need to be updated for current clients. This will require a list of the best, recently found, previous configurations, or solutions to $\boldsymbol{P}$ (4.1) (i.e., policy configurations $\boldsymbol{x}$) to be created. This list, which offers the optimal performance results are stored in the lists X_G and X_{sG}, where X_G is a *Global Candidate List* $(\boldsymbol{x}_G^1, \boldsymbol{x}_G^2, \dots, |\boldsymbol{x}_G|)$ and X_{sG} is a semi-*Global Candidate List* $(\boldsymbol{x}_{sG}^1, \boldsymbol{x}_{sG}^2, \dots, |\boldsymbol{x}_{sG}|)$, such that $X_G \subset X$ and $X_{sG} \subset X$. The rank (equation 4.6) of a solution $\boldsymbol{x}_G^z \in X_G$ increases and decreases according to the number of times solution $\boldsymbol{x}_G^z$ has been accepted as the final solution over the number of times the solution has been visited. The configurations are ordered in X_G and X_{sG} in a descending order according to their rank.

$$Rank(\boldsymbol{x}_G^z) = \frac{\#\ of\ times\ \boldsymbol{x}_G^z\ is\ chosen}{\#\ of\ times\ \boldsymbol{x}_G^z\ is\ visited}, \quad \forall \boldsymbol{x}_G^z \subset X_G \tag{4.6}$$

Updating the network configurations entails changing one or more variables in the current configuration $\boldsymbol{x_0}$. The problem with selecting this subset of variables that can achieve a better network performance can be solved using the modified tabu-search approach (IGL-TS) and is described as follows:

Step 1:

Starting from the initial set of the current policy variable configurations $\boldsymbol{x_0}$, evaluate the observed QoS performance of the solution using the behavior function $B(\boldsymbol{x})$. The IGL-TS method performs a fast online search procedure followed by a more enhanced in depth offline search. A solution that provides a better performance to the SM, VHO and SLA manager than the current configuration is first found by the former. The latter continues to search for more optimal configurations and update both the Global and semi-Global candidate lists. These solutions will be available for later use by other mobile clients in similar network scenarios.

Step2:

If the global candidate list X_G is not empty, perform an online search in X_G, where each solution is run by the simulator and evaluated using (4.3). When a solution is found that

outperforms $\boldsymbol{x_0}$, the new configurations are adopted. If a better solution is not found from X_G, then a list of neighbors for $\boldsymbol{x_0}$ is generated according to Algorithm 2. When determining how many neighbors for $\boldsymbol{x_0}$ to consider, there are two tradeoffs at stake. If a large number of neighbors are chosen, it will provoke the TS method to carry out a deep search in a local region, leading to increase in the calculation burden. On the other hand, a small number of neighbors will omit potential solutions in the local region, but will reduce the calculation burden. Hence, the number of neighbors must be chosen so that a balance is kept between the calculation burden and the search depth. The definition of neighborhood depends on the problem being considered. In this case, a neighborhood $N(\boldsymbol{x})$ of a vector $\boldsymbol{x}$ is controlled by two parameters, θ and β. The value of θ determines the size of the area to be searched, while β determines the distance between the two closest configurations in the policy configuration search space. More precisely:

$$N\big(\boldsymbol{x_0}(t)\big) = \{\boldsymbol{x} \in X: x_0^k \leq x^k \leq \theta x_0^k, k = 1, \dots, |k|\} \tag{4.7}$$

$$\theta = \frac{B\big(\boldsymbol{x_0}(t)\big) - B\big(\boldsymbol{x_0}(t + \Delta t)\big)}{B\big(\boldsymbol{x_0}(t)\big)} \tag{4.8}$$

β is in the range of (0,1) and is chosen experimentally. A large value of β, means a greater probability for the neighbors to overlap local optima and will result in a convergence to a faster solution by reducing the size of the list. On the contrary, a small value means the neighbors will follow the best previous solutions from the Global and semi-Global lists (X_G, X_{sG}).

Algorithm 2: Find $\mathbf{x}' \in \mathbf{N}\big(\mathbf{x}(\mathbf{t})\big)$ that results in better $\tilde{\mathbf{q}}_{\mathbf{ij}}(\mathbf{t})$

1: construct $\mathrm{N}\big(\mathbf{x_0}(\mathrm{t})\big)$ with θ and β
2: **Repeat**
3: remove $\mathbf{x}'$ from $\mathrm{N}\big(\mathbf{x_0}(\mathrm{t})\big)$
4: simulate $\mathbf{x}'$ and compute $\mathrm{B}(\mathbf{x}')$
5: **until** $\big(\mathrm{B}(\mathbf{x}') > \mathrm{B}\big(\mathbf{x_0}(\mathrm{t})\big)$ or $\mathrm{N}\big(\mathbf{x_0}(\mathrm{t})\big)$ is empty$\big)$

Step 3:

Perform a more in depth search using the space $X_G \cup X_{sG} \cup N\big(\boldsymbol{x}(t)\big)$ with the general tabu search procedure. Exchange an element from the examined online candidate list $X_{candidate}$ with that of an element of the new search space according to Algorithm 3. If the

neighboring solution is not in the tabu list, pick it to be the new current solution. If the best of these neighbors is found to lead to a better performance with respect to the current optimum, override the tabu status and pick it to be the new current solution. The performance function of the new solution is then evaluated. To avoid repetitive looping when a move is performed, an element in the solution space is prevented from returning for a certain number of iterations (TS_{tenure}). Update X_G and X_{sG} by swapping a solution of a higher rank with the lowest ranking member of X_G. Swap the lowest ranking member of X_G with the best ranking member of X_{sG}. This update process for the global and semi-global candidate lists will produce a faster optimal solution search for upcoming iterations. The search process terminates when it reaches the time threshold T_{IGL}. The search process and exchange of messages between the components of the architecture is summarized in Figure 4.1.

Algorithm 3: Find a new configuration x' that results in better $\tilde{\mathbf{q}}_{\mathbf{ij}}(\mathbf{t})$ and update $\mathrm{X_G}$ and $\mathrm{X_{sG}}$

construct $\mathrm{N}(\mathbf{x}(\mathrm{t}))$ with a smaller step size β
$\forall\, \mathbf{x}' \in \mathrm{X_G} \cup \mathrm{X_{sG}} \cup \mathrm{N}(\mathbf{x}(\mathrm{t}))$
  $\mathrm{tabu}(\mathbf{x}') = 0;$
  $\mathrm{X_{candidate}} = 0;$
  Repeat
    select $\mathbf{x}' \in \mathrm{X_G} \cup \mathrm{X_{sG}} \cup \mathrm{N}(\mathbf{x}(\mathrm{t}))$ when $\mathbf{x}'$ is not tabu
   $\mathrm{tabu}(\mathbf{x}') = \mathrm{tabu}(\mathbf{x}') + 1;$
  if $(\mathrm{TS_{tenure}} - \mathrm{tabu}(\mathbf{x}') \geq 0)$ **then**
      simulate and calculate $\mathrm{B}(\mathbf{x}')$
        if $\mathrm{size}(\mathrm{X_{candidate}} \leq \mathrm{MaxSize})$
          add $\mathbf{x}'$
        else $\left(\exists\, \mathbf{x} \in \mathrm{X_{candidate}}\ \mathrm{s.t.}\ \mathrm{B}(\mathbf{x}) < \mathrm{B}(\mathbf{x}')\right)$ **then**
          replace $\mathbf{x}$ with $\mathbf{x}'$ in $\mathrm{X_{candidate}}$
          calculate $\mathrm{Rank}(\mathbf{x}')$;
            if $\mathrm{Rank}(\mathbf{x}') > \mathrm{Rank}(|\mathbf{x}_G|)$**then**
              $\mathrm{X_{sG}} = \mathrm{X_{sG}} \cup |\mathbf{x}_G| - |\mathbf{x}_{sG}|;$
              $\mathrm{X_G} = \mathrm{X_G} \cup \mathbf{x}' - |\mathbf{x}_G|;$
    else
      $\mathrm{X_{candidate}} = \mathrm{X_{candidate}} - \mathbf{x}';$
      mark $\mathbf{x}'$ as tabu
  until duration $= \mathrm{T_{IGL}}$

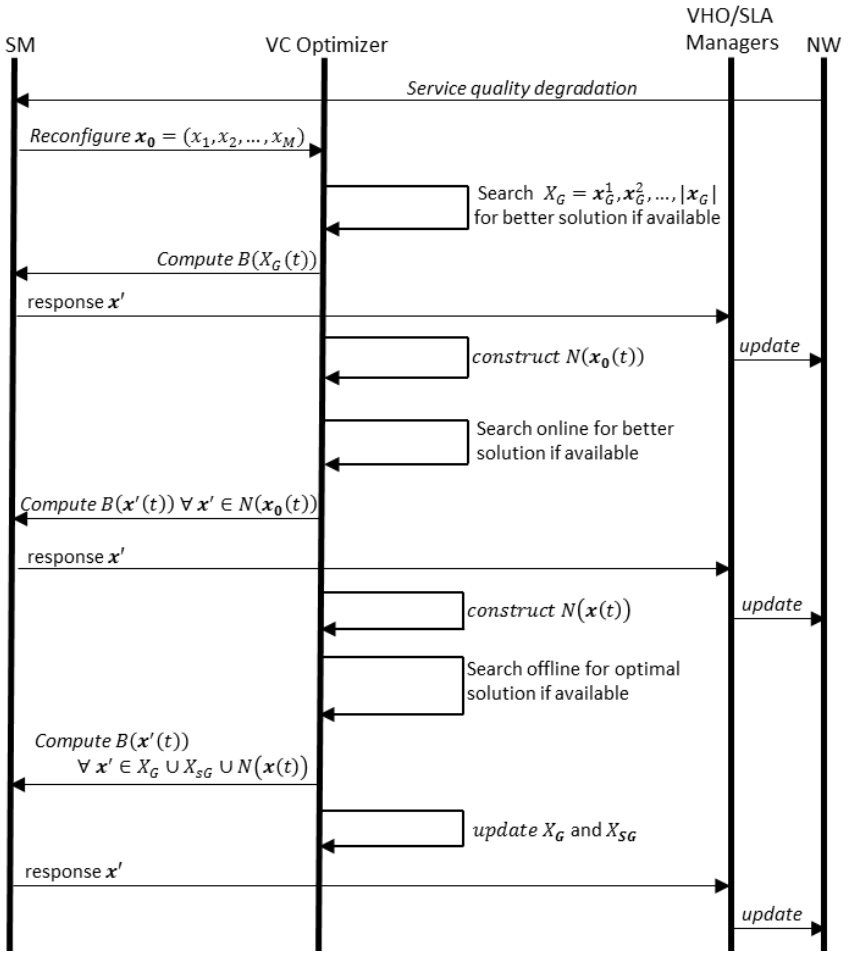

Figure 4.1 Timeline exchange of messages between Simulator Manager, Variable Configuration Optimizer, VHO and SLA managers, and the mobile client.

4.4. SIMPLIFYING THE VARIABLE CONFIGURATION SEARCH SPACE USING LINEAR REGRESSION ANALYSIS

4.4.1. Correlation between Policy Parameters and Scenario Settings

As in most optimization problems, the search space must be discretized to limit the search region and simplify the problem for the adopted meta-heuristic algorithm. Hence, by discretizing continuous domains, continuous variables are also discretized, and thus meta-heuristics in general. In addition, tabu-search particularly can be applied to find an optimal solution over the scenario search space. It is also undesirable to simulate the extensive number of different scenario configurations. Instead it is necessary to find a mathematical function that will give us the appropriate scenario configurations to produce the desired

behavior according to the VHO used. In this case, a VHO mechanism that adopts the AHP method is used. The mathematical function must accept current network performance parameters, and provide the appropriate scenario configurations, (i.e. AHP settings). Therefore a scenario is simplified to a set of dynamic and static set of variables. Static variables do not need to be changed to optimize network setting. For example, the number of mobile users, the network devices, the device settings, etc. is omitted from the variable reconfiguration problem. Thus, only policy variables need to be reconfigured. To minimize the set of variable reconfiguration, a focus is put on policies that reconfigure the AHP settings. The following policy example illustrates the correlation between AHP parameter settings and the received signal strength, delay, and bandwidth:

$$\boldsymbol{if}\ RSSI < RSSI_{threshold}\ \&\&\ Delay > Delay_{threshold}\ \&\&\ BW < BW_{threshold}\ \boldsymbol{then}\ set\ simulationParameters(\ AvailabilityWeight = 6, DelayWeight = 4, PriceWeight\ = 4\ )$$

From the simulation experiments conducted it was found that the scenario search space can be limited to a discrete but large number of variable configurations. From this set of variables, a prediction can be made of the policy parameter configurations (i.e. AHP settings) that will make up the scenario search space. Based on the results of the simulation experiments, it is evident that the number and location of the users in the network affect the configurations of the independent variables. These independent variables have a direct impact on the AHP settings. Therefore, it is logical to adapt regression analysis models to find the correlation between delay, bandwidth, RSSI and AHP configurations. Given a correlation, it is then possible to know the policy configurations that may produce optimal network performance using the IGL-TS algorithm.

4.4.2. Regression Analysis

Statistical techniques provide mechanisms for answering questions about possible patterns in empirical data. *Regression analysis* is a statistical technique that attempts to predict the values for one variable with the aid of the values from one or more other variables. Technically, the predicted variable is called the *dependent variable* and the other variables being used as predictors of the variable are called *independent variables*. Regression analysis provides a mechanism for estimating the form of the relationship

between variables as well as a mechanism for assessing how an accurate and independent variable predicts a dependent variable. A large set of techniques for performing regression analysis has been developed. Examples of such techniques include: linear, ordinary least squares, non-linear, and nonparametric regression.

Multiple regression analysis employs a linear function of two or more independent variables to provide the variation in the dependent variable. In linear multiple regression analysis the observed values of the dependent variable is predicted using a linear function of the observed values of the multiple independent variables. The relationship of the variables can be formulated as follows:

$$\boldsymbol{y} = \boldsymbol{u}a + \boldsymbol{x_1}b_1 + \boldsymbol{x_2}b_2 + \cdots + \boldsymbol{x_n}b_n + \boldsymbol{e} \tag{4.9}$$

where $\boldsymbol{y}$ is the dependent variable, $\boldsymbol{x_1}$ to $\boldsymbol{x_n}$ are the independent variables, a is the intercept (the value of $\boldsymbol{y}$ when the independent variables are set to zero), and $\boldsymbol{e}$ is the prediction error.

Most applications of linear regression aim at either prediction or forecasting, such that linear regression is used to fit a predictive model to an observed data set of $\boldsymbol{x_1}$ to $\boldsymbol{x_n}$. It is also used to quantify the strength of the relationship between $\boldsymbol{y}$ and $\boldsymbol{x_i}$ to assess which $\boldsymbol{x_i}$ may have no relationship with $\boldsymbol{y}$ or to identify which subsets of $\boldsymbol{x_i}$ contain redundant information about $\boldsymbol{y}$. Linear regression models are usually fitted using the least-squares approach.

4.4.3. Applying Regression Analysis to a Wireless Network Environment

The variable configuration search problem is simplified and minimized to a three variable selection problem in the form of a linear regression function:

$$L_j = \mu + \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3 \tag{4.10}$$

where L_j are the action variables of a policy that must be reconfigured, such that $j = \{Availability, Delay, Jitter, Loss, Bandwidth, Price\}$; α_1, α_2, α_3 are the current average performance of the network in terms of RSS, delay and bandwidth provided to a mobile client with a specific CoS. Thus, the values of variables x_1, x_2, x_3 in the linear regression function need to be updated through the IGL-TS method.

Experimental tests were conducted on three types of network: WiMAX, WiFi, and UMTS to provide the necessary linear regression functions for the AHP settings. All tests were conducted using the NS-2 simulator. The packet size is set to 1240 bytes using constant

bit rate (CBR) traffic with 0.1 sec intervals originating at the server. The WiMAX network has a coverage area of 500 m radius with a 2 Mbps bandwidth service rate. WiFi access points have a coverage radius of 20 m, a data rate of 11 Mbps, and beacon interval of t = 0.1 sec. The UMTS base station has coverage of 1km and a data rate of 384 Kbps. Three user classes of service (CoS) are available in which each class is identified by the requested service data rate of a MN. Thus the services are available according to the following rates: 64 Kbps, 128 Kbps, 256 Kbps.

4.4.3.1. Regression Analysis Tests in a WiMAX Network

When varying the number and location of CoS-1 users (i.e. 256 Kbps data rate) in a WiMAX network the following observations were reported:

$$L_A = u\,(10.1253) + x1\,(0.0073) - x2\,(0.0133) - x3\,(0.0089)$$
$$L_D = u\,(7.5587) - x1\,(0.0211) - x2\,(0.0119) - x3\,(0.0072)$$
$$L_J = u\,(7.0651) - x1\,(0.0009) + x2\,(0.0025) + x3\,(0.0005)$$
$$L_L = u\,(7.5376) + x1\,(0.0012) - x2\,(0.0313) - x3\,(0.0005)$$
$$L_B = u\,(7.6002) - x1\,(0.0044) - x2\,(0.0036) - x3\,(0.0002)$$
$$L_P = u\,(1.6522) - x1\,(0.0012) + x2\,(0.0017) + x3\,(0.0010)$$

Figure 4.2 illustrates a visualization of the correlation between $L_A, x1$ and $x2$.

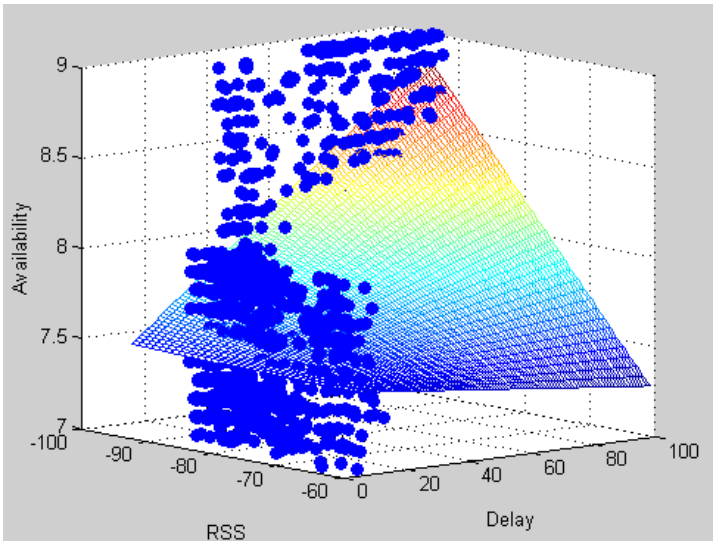

Figure 4.2 Correlation between Availability, RSS and Delay for CoS-1 users in a WiMAX environment.

On the contrary, varying the number and location of CoS-2 users (i.e. 128 Kbps data rate) in a WiMAX network, the following observations were reported:

$$L_A = u\,(3.1304) - x1\,(0.0184) - x2\,(0.0120) - x3\,(0.0019)$$
$$L_D = u\,(4.6563) - x1\,(0.0007) + x2\,(0.0058) + x3\,(0.0026)$$
$$L_J = u\,(5.3313) + x1\,(0.0038) - x2\,(0.0026) - x3\,(0.0009)$$
$$L_L = u\,(5.1407) + x1\,(0.0021) + x2\,(0.0022) + x3\,(0.0006)$$
$$L_B = u\,(4.7804) - x1\,(0.0017) + x2\,(0.0036) + x3\,(0.0010)$$
$$L_P = u\,(5.0032) + x1\,(0.0011) + x2\,(0.0013) + x3\,(0.0006)$$

Figure 4.3 illustrates a visualization of the correlation between $L_A, x1$ and $x2$.

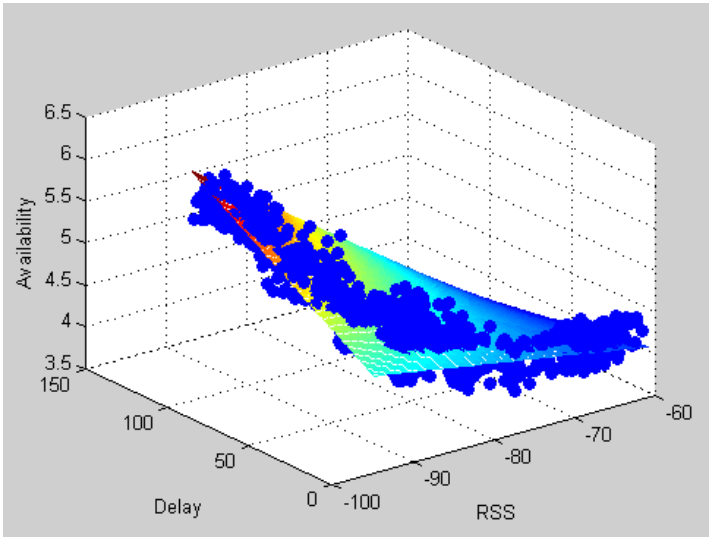

Figure 4.3. Correlation between Availability, RSS and Delay for CoS-2 users in a WiMAX environment.

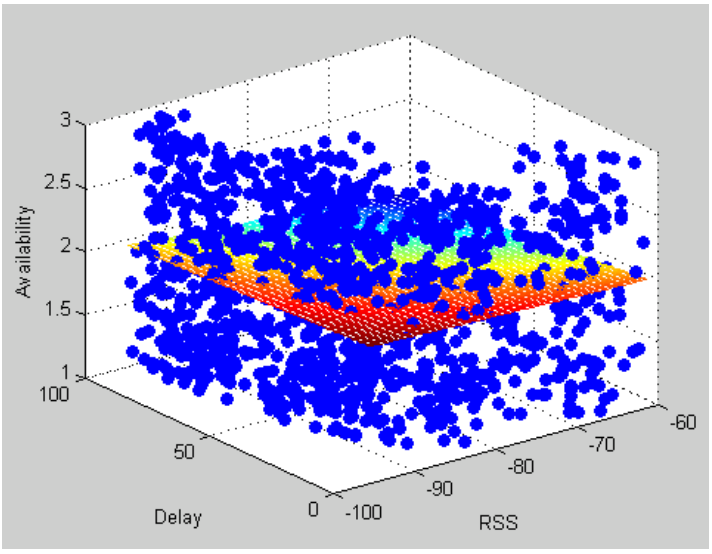

Figure 4.4. Correlation between Availability, RSS and Delay for CoS-3 users in a WiMAX environment.

Finally, varying the number and location of CoS-3 users (i.e. 64 Kbps data rate) in a WiMAX network, the following observations were reported:

$$L_A = u\,(1.6600) - x1\,(0.0029) - x2\,(0.0010) + x3\,(0.0020)$$
$$L_D = u\,(2.2653) + x1\,(0.0006) - x2\,(0.0021) - x3\,(0.0035)$$
$$L_J = u\,(1.7626) - x1\,(0.0026) - x2\,(0.0017) + x3\,(0.0009)$$
$$L_L = u\,(3.0694) + x1\,(0.0072) - x2\,(0.0042) - x3\,(0.0086)$$
$$L_B = u\,(1.5874) - x1\,(0.0032) + x2\,(0.0022) + x3\,(0.0029)$$
$$L_P = u\,(1.8720) - x1\,(0.0026) - x2\,(0.0034) - x3\,(0.0002)$$

Figure 4.4 illustrates a visualization of the correlation between $L_A, x1$ and $x2$.

4.4.3.2. Regression Analysis Tests in a WiFi Network

When varying the number and location of CoS-1 users (i.e. 256 Kbps data rate) in a WiFi network, the following observations were reported:

$$L_A = u\,(0.5467) - x1\,(0.0372) + x2\,(0.0851) + x3\,(0.0203)$$
$$L_D = u\,(3.6982) - x1\,(0.0229) + x2\,(0.0376) - x3\,(0.0111)$$
$$L_J = u\,(7.9846) - x1\,(0.0179) - x2\,(0.0207) - x3\,(0.0046)$$
$$L_L = u\,(7.1426) - x1\,(0.0107) - x2\,(0.0460) - x3\,(0.0001)$$
$$L_B = u\,(17.1567) + x1\,(0.0282) - x2\,(0.0951) - x3\,(0.0293)$$
$$L_P = u\,(1.503) + x1\,(0.0012) + x2\,(0.0289) + x3\,(0.0135)$$

Figure 4.5 illustrates a visualization of the correlation between $L_A, x1$ and $x2$.

On the contrary, varying the number and location of CoS-2 users (i.e. 128 Kbps data rate) in a WiFi network, the following observations were reported:

$$L_A = u\,(6.2458) + x1\,(0.0019) - x2\,(0.0188) + x3\,(0.0000)$$
$$L_D = u\,(4.9024) + x1\,(0.0132) - x2\,(0.0072) - x3\,(0.0000)$$
$$L_J = u\,(4.9538) + x1\,(0.0128) - x2\,(0.0065) + x3\,(0.0000)$$
$$L_L = u\,(5.9442) + x1\,(0.0045) - x2\,(0.0161) + x3\,(0.0000)$$
$$L_B = u\,(5.2820) + x1\,(0.0103) - x2\,(0.0103) + x3\,(0.0000)$$
$$L_P = u\,(6.3767) - x1\,(0.0162) - x2\,(0.0162) + x3\,(0.0000)$$

Figure 4.6 illustrates a visualization of the correlation between $L_A, x1$ and $x2$.

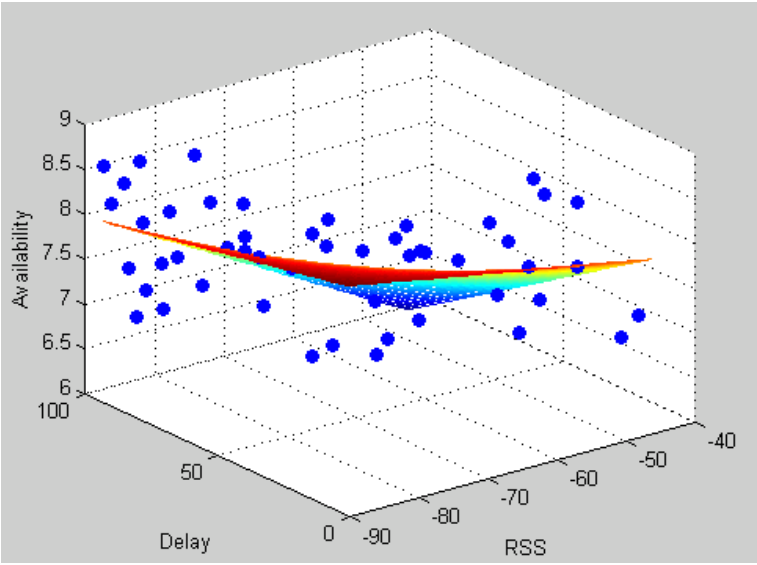

Figure 4.5. Correlation between Availability, RSS and Delay for CoS-1 users in a WiFi environment.

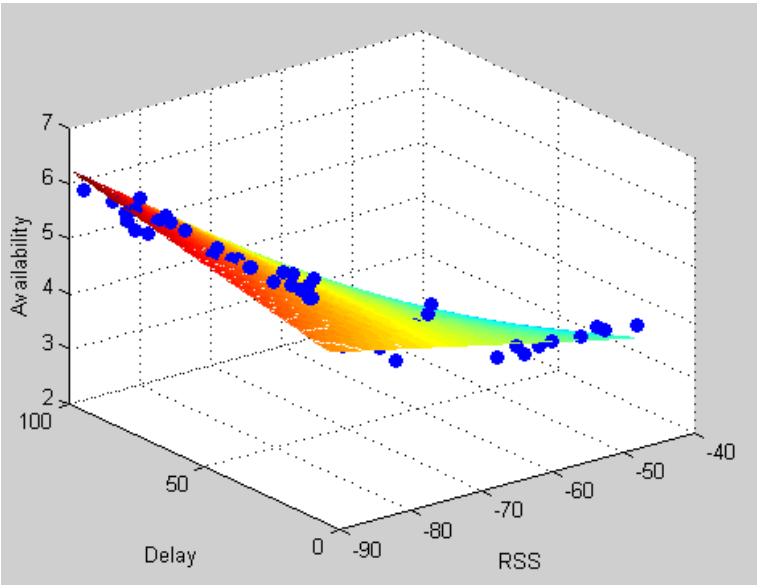

Figure 4.6. Correlation between Availability, RSS and Delay for CoS-2 users in a WiFi environment.

Finally, varying the number and location of CoS-3 users (i.e. 64 Kbps data rate) in a WiFi network, the following observations were reported:

$$L_A = u\,(0.9092) - x1\,(0.0392) + x2\,(0.0024) + x3\,(0.0079)$$
$$L_D = u\,(0.7098) - x1\,(0.0059) + x2\,(0.0082) + x3\,(0.0052)$$
$$L_J = u\,(0.1336) - x1\,(0.0120) + x2\,(0.0059) + x3\,(0.0075)$$
$$L_L = u\,(1.9677) - x1\,(0.0127) - x2\,(0.0253) - x3\,(0.0116)$$
$$L_B = u\,(0.7962) - x1\,(0.0051) + x2\,(0.0081) + x3\,(0.0025)$$
$$L_P = u\,(7.4559) - x1\,(0.0078) + x2\,(0.0070) + x3\,(0.0007)$$

Figure 4.7 illustrates a visualization of the correlation between $L_A, x1$ and $x2$.

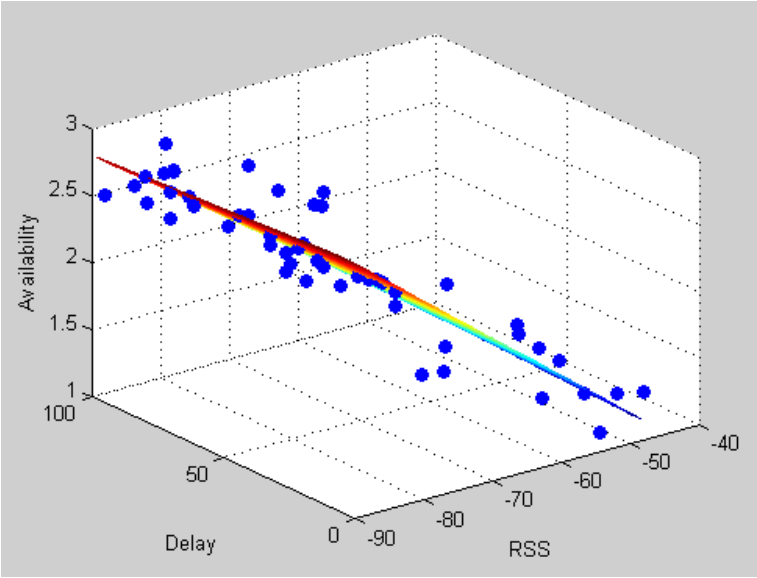

Figure 4.7. Correlation between Availability, RSS and Delay for CoS-3 users in a WiFi environment.

4.4.3.3. Regression Analysis Tests in a UMTS Network

When varying the number and location of CoS-1 users (i.e. 256 Kbps data rate) in a UMTS network, the following observations were reported:

$$L_A = u\,(9.5208) - x1\,(0.0036) - x2\,(0.0060) - x3\,(0.0094)$$
$$L_D = u\,(7.7364) - x1\,(0.0133) + x2\,(0.0093) - x3\,(0.0038)$$
$$L_J = u\,(8.5839) - x1\,(0.0031) - x2\,(0.0021) - x3\,(0.0059)$$
$$L_L = u\,(8.2020) + x1\,(0.0008) - x2\,(0.0036) - x3\,(0.0005)$$
$$L_B = u\,(8.6456) - x1\,(0.0116) - x2\,(0.0031) - x3\,(0.0072)$$
$$L_P = u\,(2.7041) + x1\,(0.0020) - x2\,(0.0063) - x3\,(0.0022)$$

Figure 4.8 illustrates a visualization of the correlation between $L_A, x1$ and $x2$.

On the contrary, varying the number and location of CoS-2 users (i.e. 128 Kbps data rate) in a UMTS network, the following observations were reported:

$$L_A = u\,(5.1657) - x1\,(0.0036) - x2\,(0.0015) - x3\,(0.0094)$$
$$L_D = u\,(5.5307) - x1\,(0.0120) - x2\,(0.0013) - x3\,(0.0106)$$
$$L_J = u\,(5.6505) - x1\,(0.0030) - x2\,(0.0010) - x3\,(0.0095)$$
$$L_L = u\,(5.6546) - x1\,(0.0117) - x2\,(0.0020) - x3\,(0.0112)$$
$$L_B = u\,(5.6786) - x1\,(0.0038) - x2\,(0.0018) - x3\,(0.0097)$$
$$L_P = u\,(4.8059) - x1\,(0.0032) - x2\,(0.0008) + x3\,(0.0001)$$

Figure 4.9 illustrates a visualization of the correlation between $L_A, x1$ and $x2$.

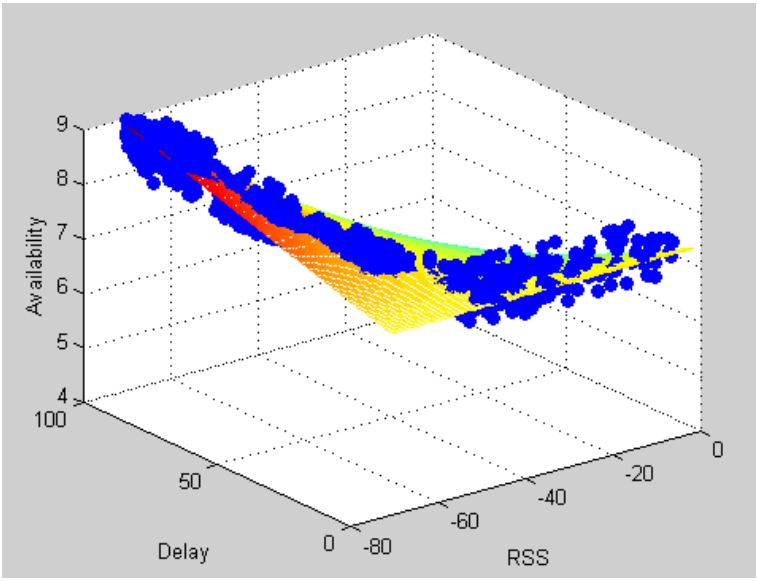

Figure 4.8. Correlation between Availability, RSS and Delay for CoS-1 users in a UMTS environment.

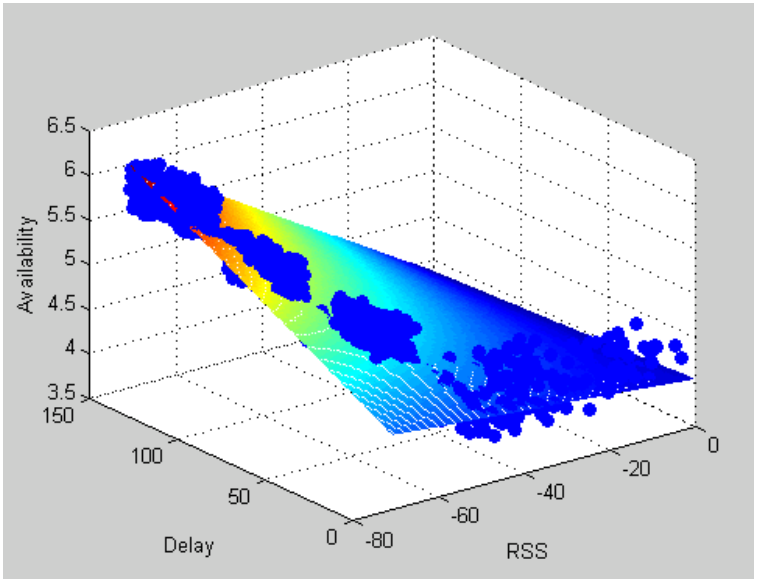

Figure 4.9. Correlation between Availability, RSS and Delay for CoS-2 users in a UMTS environment.

Finally, varying the number and location of CoS-3 users (i.e. 64 Kbps data rate) in a UMTS network, the following observations were reported:

$$L_A = u\,(1.8940) - x1\,(0.0029) + x2\,(0.0011) - x3\,(0.0002)$$
$$L_D = u\,(2.2242) + x1\,(0.0003) - x2\,(0.0056) - x3\,(0.0027)$$
$$L_J = u\,(2.2396) + x1\,(0.0033) - x2\,(0.0034) - x3\,(0.0007)$$
$$L_L = u\,(1.8023) + x1\,(0.0016) + x2\,(0.0024) + x3\,(0.0037)$$

$$L_B = u\,(1.7026) + x1\,(0.0035) + x2\,(0.0054) + x3\,(0.0047)$$
$$L_P = u\,(7.2926) + x1\,(0.0001) + x2\,(0.0108) + x3\,(0.0092)$$

Figure 4.10 illustrates a visualization of the correlation between $L_A, x1$ and $x2$.

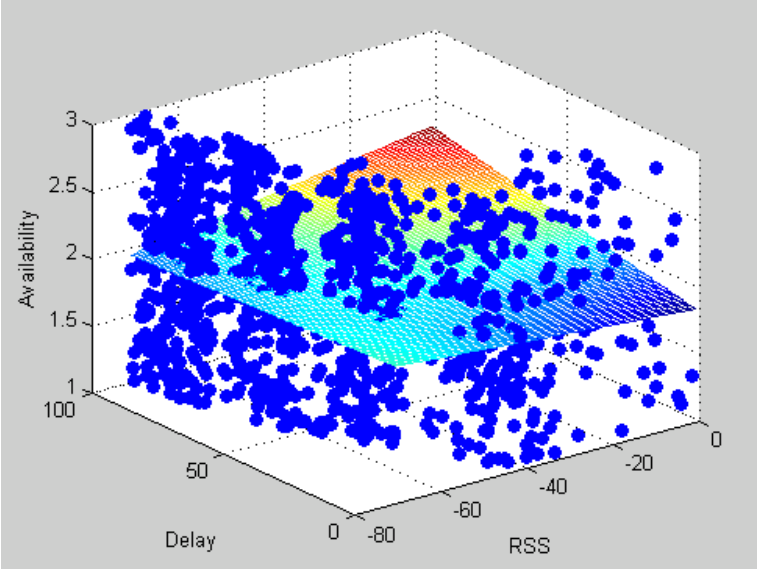

Figure 4.10. Correlation between Availability, RSS and Delay for CoS-3 users in a UMTS environment.

4.5. PERFORMANCE EVALUATION

In this section, initially the effectiveness of adopting the IGL-TS algorithm for the scenario search problem is shown and then the performance results of adapting the proposed framework are presented. Several experiments were conducted to evaluate the impact of the tabu search method parameters. The TS algorithm is implemented using Java and the OpenTS library [205]. The proposed system is simulated using NS-2. The IGL-TS method has to be designed and implemented as the existing tabu search approaches for variable optimization cannot be used in this case, as explained earlier in this dissertation. The algorithms runs on a 3GHz Core 2 Duo Intel processor, equipped with a Linux server and 4-Gbytes of memory.

The observed QoS performance $\tilde{\boldsymbol{q}}_{\boldsymbol{ij}}(t)$ is measured through the NS-2 simulator in a network scenario, illustrated in Figure 4.11, which consists of clients that have the option to acquire a service by connecting to one of the SPs available. Thus, a client can connect to a UMTS network through BS1 or to a WIMAX network through BS2, owned by two service

providers SP_1 and SP_2, respectively. Also, there are three WLANs managed by three different SPs, namely, SP_3, SP_4 and SP_5. The first has two access points AP1 and AP2, the second includes AP3 and AP4, while the third SP has a single AP, namely, AP5. NS-2 simulation settings are similar to what has been discussed in Chapter 3.3.

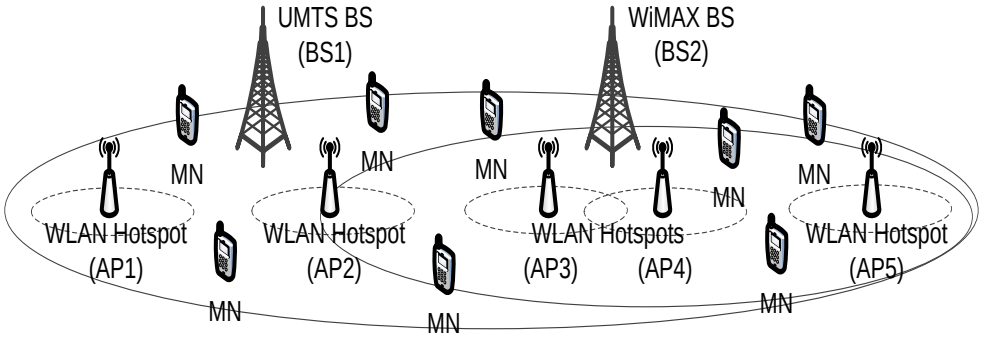

Figure 4.11 Simulated network topology. MNs are able to roam coverage area of other SPs with the same original service rate. All SPs provide the same service to all MNs.

4.5.1. Analysis of IGL-TS and Adaptation Time

The problem of policy configuration selection is simplified to a 3-variable linear regression function such that each policy configuration is affected by three variables. These are in the form of: received signal strength, delay, and bandwidth provided to a mobile client with a specific CoS. Thus, the values of variables x_1, x_2, x_3 in the linear regression function need to be updated. The IGL-TS method is applied to find the optimal values that will maximize both the SP's profit and the mobile client's satisfaction in terms of QoS guarantees.

The size of the tabu list has a direct impact on the quality of the solution. By analyzing the impact of the list's size over QoS performance improvement, it is clear that the best results are obtained using tabu lists whose size is larger than 120. A size of 130 produced a 100% improvement in terms of observed QoS when compared to a previous initial solution. If the size of the list is too small, cycling will occur. On the contrary, if the size is too big, the quality of the solution will deteriorate by forbidding too many moves. Figure. 4.12 illustrates the result of this investigation. When performing an online search (Algorithm 2), β is set to a large value, such that $\beta > 0.45$, to converge to a solution faster by reducing the size of the

list. On the contrary, when conducting an offline search (Algorithm 3), to perform a more in-depth search, $\beta \leq 0.45$.

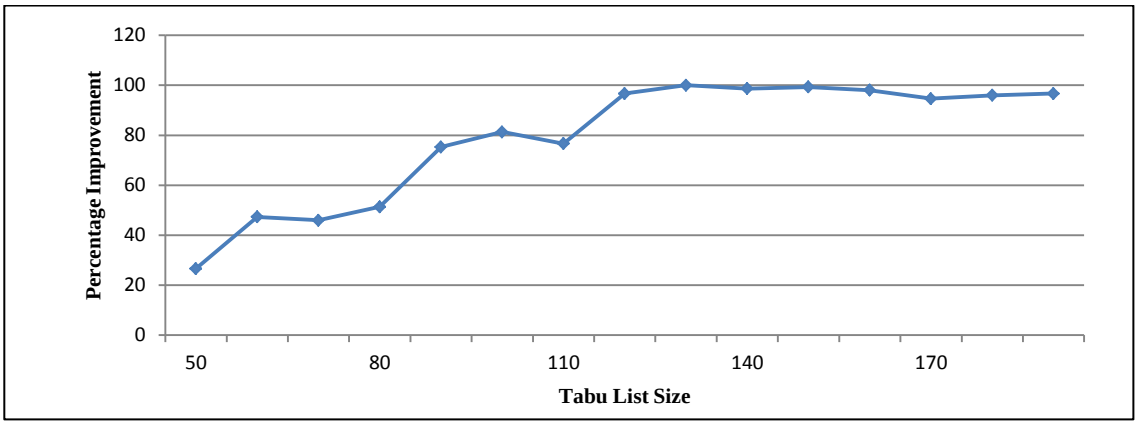

Figure 4.12 Impact of the tabu list size on performance improvement.

The online IGLTS solution will provide service guarantees outlined by the SLA for clients to minimize SP penalties and provide client service satisfaction. The offline search will update both the Global and semi-Global candidate lists with optimal solutions. These solutions will be available for use later by future clients in similar network scenarios. The goal is to decrease the amount of time needed to produce an optimal solution. Starting from an empty set, IGL-TS took 17 iterations to achieve a solution where optimality is found in 10 seconds (i.e. better solution is equal to optimal solution). Thus the online solution provides an optimal solution, which can be applied to the current clients in the network. Furthermore, after 17 iterations, TS achieves an optimal solution in 5800 seconds while a better solution is found in 390 seconds. Figures 4.13-4.14 provide further details.

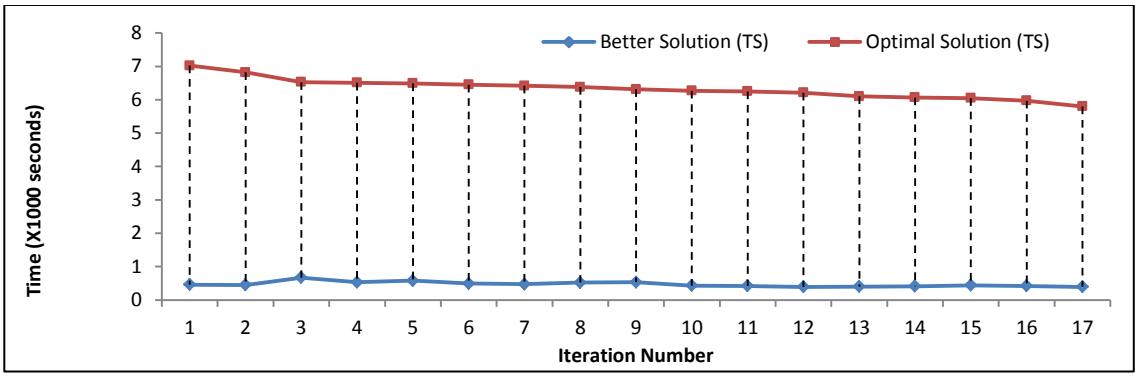

Figure 4. 13 Impact of the number of iterations on the execution time for the non-modified TS method.

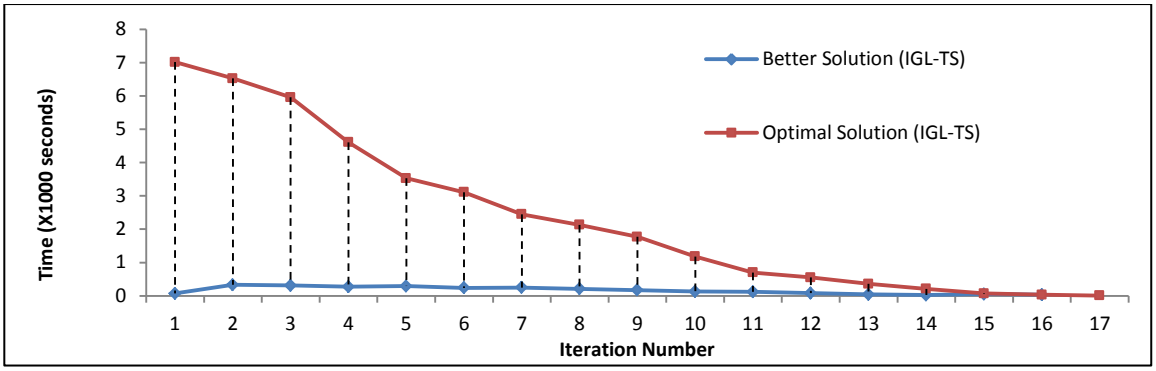

Figure 4. 14 Impact of the number of iterations on the execution time for the IGL-TS method.

4.5.2. QoS Performance Analysis

The analysis regarding the impact of performance difference between better and optimal solutions show that IGL-TS exceeds TS by 39% after 17 iterations. The normalized QoS for IGL-TS is 2.93 while TS is 1.798, as illustrated in Figure 4.15. Thus, it can be deduced that the quality of the solution is enhanced by increasing the number of iterations until a value is reached where the execution time increases without enhancing the solution efficiently, which in this experiment, 17 iterations are shown to be the maximum execution time.

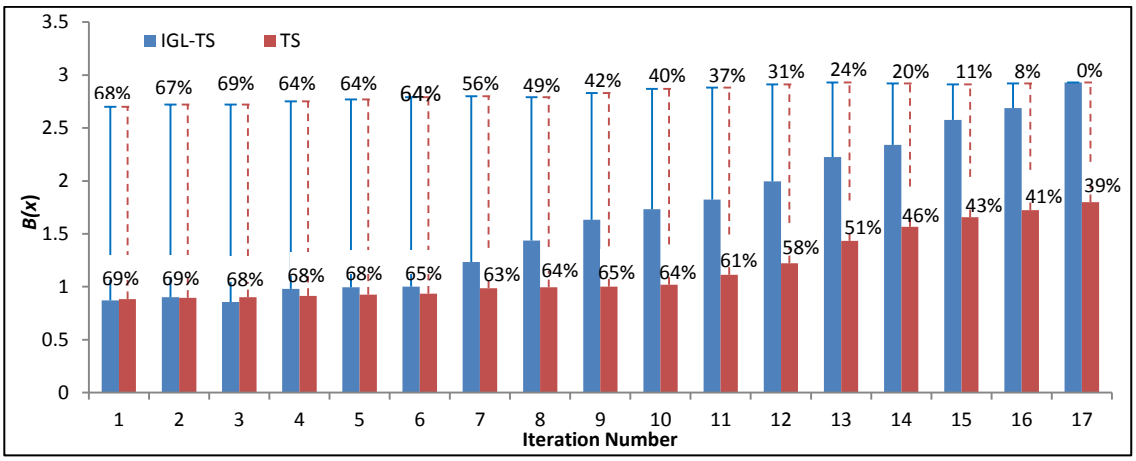

Figure 4.15 Number of iterations required to reach optimality for both TS and IGL-TS methods. The figure also shows the percentage difference between a better solution and the optimal solution.

The IGL-TS approach was devised for the scenario search problem because it was noticed from the conducted experiments that the modified approach ensures an optimal solution. Other experiments were conducted in order to evaluate the difference between the performances of the non-modified TS and IGL-TS approaches. The best performance result

is achieved with a normalized QoS theoretical value of 3. Even though both algorithms achieve an optimal value, the IGL-TS approach shows a capability of achieving a closeness of 2.4% to the best theoretical value achievable, while TS is only capable of achieving a closeness of 23% (Figure 4.16). Also, the difference between the two performance scores emanates in a larger gap, such that the performance difference at the first iteration is roughly 11%, while after 17 iterations the gap exceeds 24 %, which shows that the IGL-TS approach provides a better solution in time, and therefore faster than the original method.

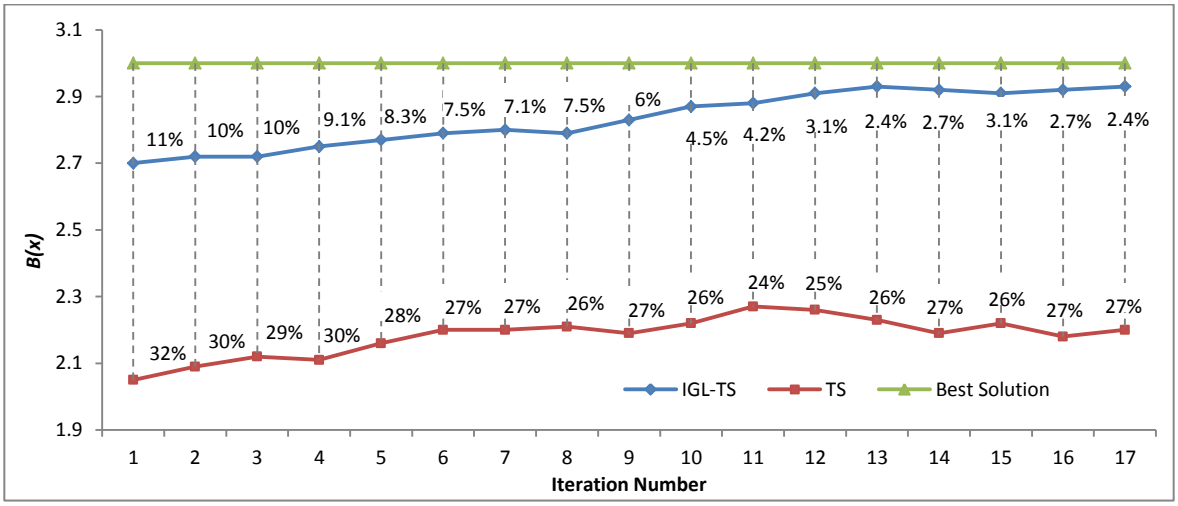

Figure 4.16 Comparing the number of iterations required to reach the best theoretical solution performance.

Other experiments were conducted to study the performance when varying the network size for the two approaches (Figure 4.17). Results show effectively that the average nodes' performance for the IGL-TS approach is better than those using the non-modified TS. For example, the average normalized QoS performance in which a better solution is found for 5 nodes using IGL-TS is 2.92, compared to normalized QoS of 1.78 for the TS approach. Also, the average performance for 20 nodes using IGL-TS is 2.95 while the average optimal solution found is 2.97 (difference of 0.7%). On the contrary, the average better solution found for 20 nodes using TS is 2.05 (30% difference between this solution and the optimal one).

From these experiments it can be concluded that generally, the IGL-TS algorithm performs better than the non-modified TS algorithm. The IGL-TS algorithm is much more efficient than TS, in the sense that the amount of time and iteration numbers needed to achieve optimality is much less. Not only does it achieve outstanding results after a number

of iteration runs, but also results show that an improved solution can be achieved using the 'online search' algorithm.

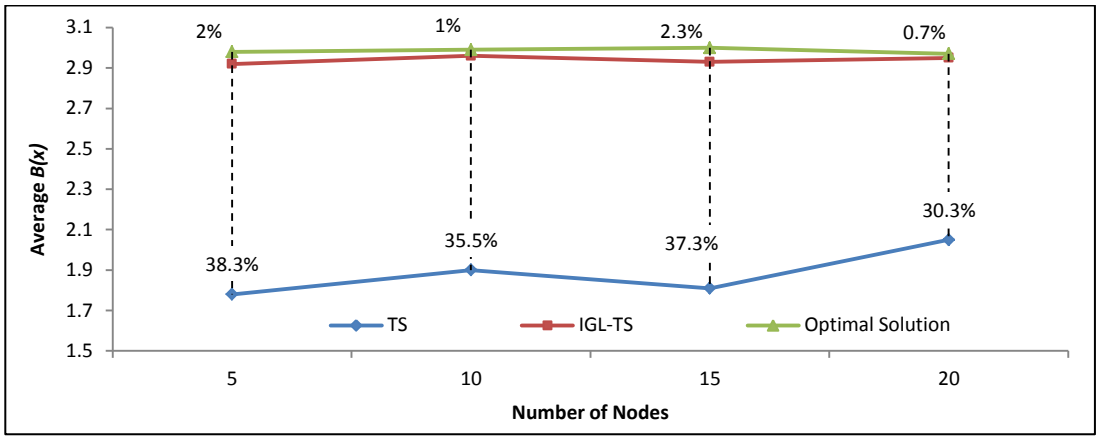

Figure 4.17 Comparing the QoS performance for TS and IGL-TS against the number of mobile clients.

4.6. SUMMARY

This chapter presented details of the variable configuration component of the framework introduced in Chapter 3. The framework is supported by a tabu search-based heuristic approach which was identified as the Iterated Local and Global Tabu-Search (IGL-TS). It is responsible for exploring the scenario search space for an optimal choice of policy configurations. The benefits of using tabu search and the modifications required to achieve the system requirements were illustrated. In addition the need for applying the regression analysis to simplify and provide a more accurate reading of the network behavior was clarified. Applying IGL-TS to the management framework has provided a benefit for reducing the time needed to update policy configurations. Simulation results reveal significant gains in terms of solution time requirements such that an optimal solution is achieved within the time requirements of a handover.

The next step in this dissertation is to provide a clear approach for monitoring the received service quality at the subscriber's side. A stability and similarity node clustering and monitoring mechanism is presented in Chapter 5. Using this approach, service subscribers' properties are semantically and syntactically compared. The highest scoring subscribers are chosen for QoS monitoring purposes.

CHAPTER 5

MONITORING QOS IN WIRELESS NETWORKS

The current mobile users' wide adoption of newly emerging mobile applications, such as mobile gaming, video on demand (VoD) and IPTV has brought along new profit opportunities for service providers (SPs). Yet, one of the main challenges that providers face is the ability to efficiently collect measurements related to the performance of these offered services. Another challenge is to timely evaluate their compliance with the agreed upon QoS levels with the mobile service subscribers (SS). These collected service quality measurements are also employed to analyze and improve the performance of the underlying network or offered services for future SSs. The continuous process of monitoring, analyzing and adapting the offered services helps the SPs in maintaining subscriber loyalty while increasing their profits.

Existing QoS measurement collection methods rely mostly on base stations (BSs) to collect measurements taken either from the SP's side or somewhere along the path between the provider and the end-user (e.g., [138], [206], [207]) in a centralized manner. Unfortunately, these approaches cannot provide timely and accurate per-user measurements and do not scale well in networks with a large number of mobile SSs.

Client-side, or distributed, monitoring approaches report measurements directly from the mobile SSs. Hence, these approaches reduce the BS workload and provide more accurate measurements. Nonetheless, they usually incur a much higher cost in terms of the needed traffic overhead and user-device power consumption.

This chapter presents a novel client-side performance monitoring scheme that relies on measurements collected by a carefully chosen subset of mobile SSs, referred to therein as *monitors.* Due to the reduced number of reporting monitors, the proposed scheme efficiently reduces the overall network traffic overhead caused by QoS performance feedback which is generated by all the SSs in traditional client-side approaches. In turn, the proposed work also minimizes the average per-user transmission power needed to report these measurements

and, hence, prolongs the lifetime of the mobile SSs' device batteries within the environment.

The remainder of this chapter is presented as follows: Section 5.1 provides an overview of the QoS monitoring steps and discusses some of the challenges and limitations of existing systems. Section 5.2 explains the requirements and objectives of the developed QoS monitoring mechanism. Section 5.3 discusses the QoS monitoring model chosen in this system. Section 5.4 provides details of the proposed monitor selection method. Section 5.5 provides results of experimentations conducted on the system to evaluate its robustness. Finally, Section 5.6 provides a summary and discussion of the main concepts discussed in this chapter.

5.1. OVERVIEW

In order to obtain end-to-end QoS measurements, different network elements must be involved in the monitoring process. The QoS monitoring process consists of the following functions:

- Retrieve QoS performance data from network elements including mobile terminals. Quality data is usually stored in a common database for later report generation and analysis.
- Generate QoS reports which can be used for network or service planning and optimization, benchmarking, monitoring of SLAs, and improving sales, marketing and product management.
- Analyze collected QoS parameters against expected values. Alarms are generated in situations of fault and error detection. In self-management systems, QoS performance analysis is done autonomously, such that corrective actions are initiated automatically when the quality levels are considered unsatisfactory.

In this chapter consideration is only given to the first function of the QoS monitoring process (i.e. QoS performance measurement retrieval).

5.1.1. QoS Measurement Collection

Measurement retrieval or collection is concerned with gathering information related to the performance of the provided service. Measurement retrieval schemes differ mainly in the

manner they address the following questions:

- What types of measurement collectors are used: *passive* (i.e. packet sniffing) or *active* (i.e. packet interception)?
- From what point are measurements collected (e.g. provider side, end-user side, or network in between)?
- Who is in charge of collecting the quality measurements?

To simplify the problem, it is assumed that the provision of the service is unilateral, such that only the SP provides a service. Therefore, only the performance of the provider needs to be measured and evaluated. In reality, it is quite possible that applications exist with bilateral service provisioning, where the two parties deliver something to each other. It is emphasized that the performance of the subscriber affects the performance of the provider. For example, it is quite natural that end-user devices have different computing performance characteristics such that high end devices accommodate software applications and hardware components that can correct packet errors and make up for packet loss. Less bandwidth rate can be given to such devices in which the service provider's resources are optimized to handle a larger community of service subscribers. Based on the above stated questions, existing measurement collection techniques can be divided into four categories:

- *End-user measurements retrieval* - measurements are collected by the service subscriber as the service is being used. Quality measurement tools are installed on the clients' devices as illustrated in Figure 5.1. It is usually assumed that monitoring is carried out with the aid of third parties with the expertise in measuring a given list of parameters and storing the collected results in its repositories to ensure that the results are trusted by both the service provider and subscriber. The client is also responsible of storing the collected quality measurements for later retrieval upon periodic requests from the provider. Client-side measurement provides enough knowledge to the SP to cope with the current network conditions to optimize its delivered services.

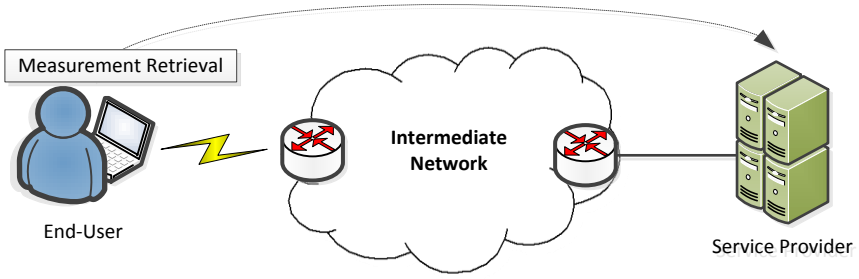

Figure 5.1 Client-side QoS measurements retrieval.

- *Service provider measurements retrieval* – quality measurements are collected by tools installed in the service provider itself as illustrated in Figure 5.2. It is important to note that with this approach, SP performance measurements are taken directly from the SP's resources. This will not consider the intermediate network's condition which plays a major role in the end-user's perceived quality performance. The latter, in effect contributes to the resource allocation solution within an enterprise's resource optimization problem. Provider-side measurement retrieval techniques aim at preventing contractual violation from the first place by monitoring its own resources. In addition, these techniques take corrective measures to deliver the service with the expected quality levels rather than reacting to violation notifications from other sources. This type of monitoring is considered proactive when preventive measures are applied to a system to avoid violations in the future. Since server-side measurement collection is private, there has to be some sort of trust and integrity provided to the client side.

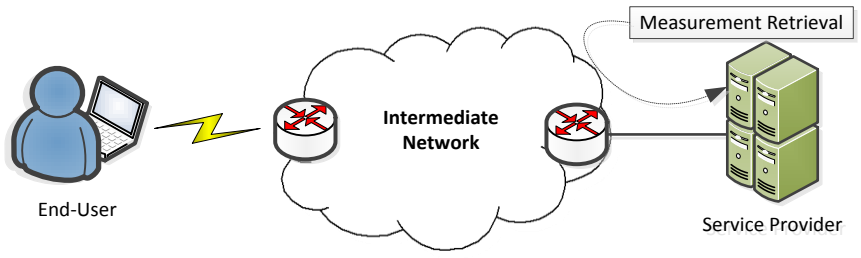

Figure 5.2 Provider-side QoS measurements retrieval.

- *Periodic polling with probe clients* – performance quality measurements are collected neither by the provider nor by the end-users, but rather by third parties trusted by all network elements [192]. Synthetic clients called 'probes' are strategically placed in

hotspots equipped with measurement tools. Such devices periodically probe the provider to measure its response. Figure 5.3 illustrates this technique. Some limitations to this approach include extra costs incurred to deploy the probes and the perception that the provider's performance might be different from that seen by the service subscriber. It is worth noting that measurements collected from the end-users side (i.e. end-user measurement collection and periodic polling with probe clients) do not solve the problems concerning the origin of service degradation. This is because there is not enough information to say whether the degradation is caused by an underperformance of the provider or by the condition of the network. However, client-side measurement collection does provide enough knowledge and notification to the service provider to cope with the current network conditions and optimize its resources according to the current conditions.

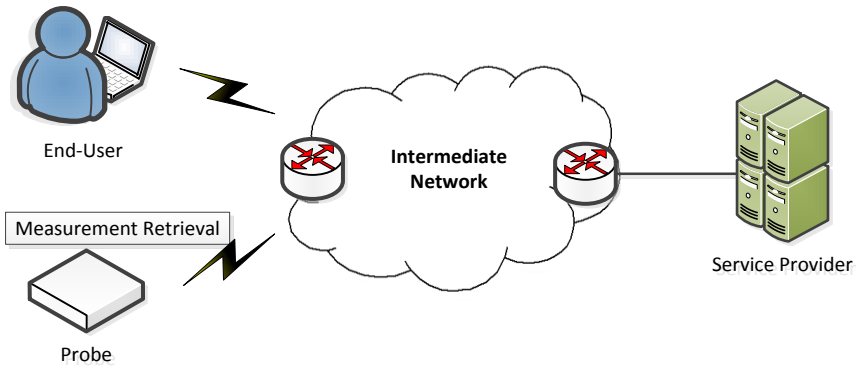

Figure 5.3 Measurements retrieval through periodic polling with probe client.

- *Network packet collection via request and response messages* – measurement tools are installed somewhere in the path between the provider and the end-users to collect all the packets being sent from the SP, as illustrated in Figure 5.4. Packet collection is done either by interception or sniffing. Such measurement tools can be realized by trusted third parties since retrieval of information is done inside neither the provider nor the consumer. This measurement retrieval technique is not a trivial task [207] [208] as it requires specialized hardware and software to be deployed in the communication link between the two ends with a great deal of packet analysis. Much work has been invested in network packet collection and analysis using packet interception techniques [209] [210] [211].

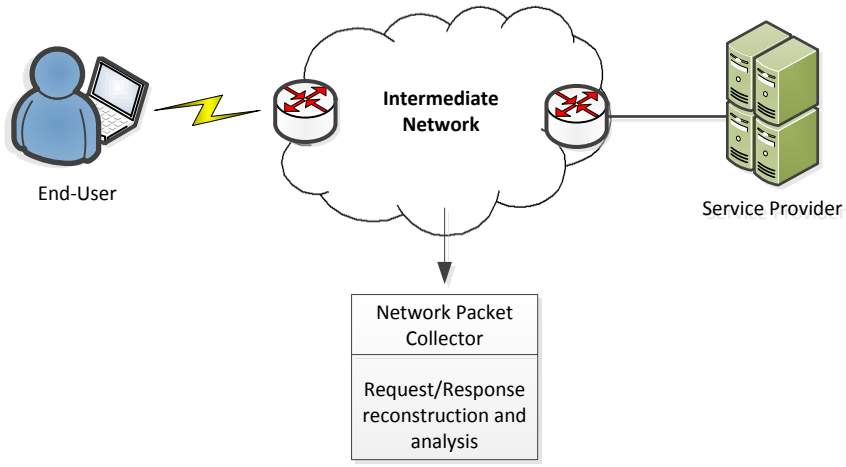

Figure 5.4 Network packet collection via request and response messages.

5.1.2. QoS Measurements Evaluation

Any monitoring system must decide on where to evaluate performance-related QoS measurements. As in any distributed system, the evaluation feature can be either centralized or distributed. In centralized service quality evaluation systems, a central evaluation component is responsible for evaluating all end-users' performance-related measurements. For instance, as depicted in the overall system architecture in Chapter 3, each base station includes an evaluation component used to report QoS results periodically to the service and network providers. Centralized evaluations are performed in a faster timely manner than distributed evaluations because the QoS measurements are mainly not collected from each client. However, end-users have to submit their performance-related QoS measurements to the centralized evaluation component. The mechanism of QoS measurement collection will be discussed further in section 5.2. On the contrary, distributed performance evaluation shifts the responsibility of performing evaluations to the client. The main drawback for such a technique is the performance overhead incurred on the client. Evaluating network performance based on aggregated measurements from a global system perspective is preferred. An advantage of distributed client performance evaluation is the avoidance of a single point of failure.

5.2. SYSTEM-SPECIFIC REQUIREMENTS

In this section, details are given on the criteria driving the monitoring process as well as

the various requirements imposed on the system. System requirements focus on the general requirements of the QoS monitoring architecture such as scalability and efficiency. For instance, requiring a QoS monitoring system that has a minimal data transmission overhead and overall node power efficiency induces a design of a scalable and distributed QoS monitoring mechanism. The requirements and decision criteria are described in the following paragraphs.

The first requirement is to design a measurement collection mechanism that supports a self-monitoring system [174] [212] [213]. The purpose of any self-monitoring system is to detect events of interest rather than respond to events after they occur. A self-monitoring system has the ability to observe and analyze a system's state and behavior, to discover violations and to notify autonomic or human administrators to effectively apply appropriate management actions. Self-monitoring mechanisms are subject to the same failures that occur on the devices being assisted. The robustness of these mechanisms is highly important to ensure overall system reliability.

In addition, with the help of network and Internet Service Providers (ISP) most service providers today guarantee service performance only at the network and transport level. Client-side performance monitoring that relies on data collected from the mobile clients is usually neglected. Reasons for neglecting client-side measurement collection include node integrity (i.e. service providers cannot rely on information provided by any subscriber, since most of the subscribers might provide false information to gain better service quality without extra charges), limited power availability, in-network data overhead and client performance overhead. On the contrary, neglecting client-side monitoring would result in omission of user-perceived performance quality inspection. This omission would result in the inability to optimize the provider's resources efficiently in an effort to cope with the current network conditions.

Since traditional monitoring systems are no longer able to adequately support the highly adaptive networking environments, new monitoring approaches are evolving. These approaches are directed towards an awareness plane [134] that consists of distributed traffic measurement, collection and analysis. This management plane relies on a complete end-to-end cross layer monitoring framework that spans across all system segments including service provider, network provider and customer domains [137]. Such monitoring

capabilities are crucial for service quality assurance, fault detection and system optimization. Therefore, in this dissertation an attempt is not made to solve the problem from all these perspectives. Instead, a client-side performance monitoring and reporting mechanism is adopted which adds additional support to the end-to-end monitoring paradigm. In addition, with the necessity to dynamically fine-tune network configurations, SLAs, and different entities, it is necessary to have the ability to monitor mobile devices to provide the optimal configurations in a wireless heterogeneous network environment. Thus, the monitoring mechanism that has been developed considers the following environmental conditions and requirements:

- Highly dense and mobile wireless environments.
- Limited power of service subscriber mobile devices.
- Absence of service and network performance monitoring probes.

Furthermore, the developed monitoring mechanism assumes the following objectives:

- Construct a client-side performance monitoring technique which supports media services such as video streaming.
- Decrease the overall network traffic overhead caused by QoS performance feedback generated by the service subscriber.
- Minimize the overall network transmission power consumption used towards QoS monitoring to prolong the lifetime of mobile nodes within the environment.
- Consider performance measurement accuracy in order to provide valid evaluation results. Imprecise measurements cause inaccurate evaluation results. As a consequence, this will result in false SLA violation triggers and imprecise resource optimization.

5.3. CLIENT-SIDE QOS MONITORING DESIGN MODEL

Centralized monitoring of service quality for large-scale distributed mobile devices is a task that is rapidly exceeding the central monitor's ability, given its quality feedback dynamics and the large amount of data involved. On the contrary, localized monitoring of QoS is overwhelming the network's capability given the large amount of traffic generated. Hence, a solution that avoids these two extremes is essential to ensure a continued

performance feedback and robust operation of the monitoring mechanism. Node clustering for performance monitoring provides a robust mechanism to achieve low traffic overhead when reporting on service quality. Since only one node will take the responsibility for providing quality feedback, two directions are to be considered. The first is to have all nodes report on service quality performance to the node cluster-head, in which the cluster-head will aggregate or simply forward the data to the SP. The second option is to have a single node within a cluster bear the responsibility of measuring the service quality and reporting to the SP. In the first case, reducing network traffic and extending the life of mobile nodes cannot be achieved since all nodes will send performance measurements; hence energy is consumed to transmit data.

This problem can be eliminated in the second case when only a single node within a cluster bears the responsibility of transmitting data; hence there is a reduction in the total energy consumed. A question that arises is: will that node within a cluster report accurate measurements? It is crucial to understand that this node will only measure service quality on its local device and thus this node within a cluster must reflect other nodes' behaviour. Thus, any solution must consider node similarity when constructing clusters.

Given the requirements and objectives outlined in the previous section, any considered solution must realize three capabilities within a self-monitoring system: *node clustering techniques*, *node similarity identification techniques*, and *packet forwarding capabilities*. The application of clustering techniques to QoS monitoring models provides the benefit of exclusively restricting performance feedback reports to the cluster-head. This will reduce traffic overhead within the network and maximize mobile device lifetime. 'Similarity identification' techniques provide support to node clustering, where the node closest, in terms of sameness, to all other nodes within proximity, is nominated as the cluster-head. This node will thus provide near-accurate or approximated user-perceived service quality measurements. Packet forwarding techniques will give the added advantage of extending a nodes lifetime resulting in a node consuming less energy when transmitting QoS measurements back to the provider. Figure 5.5 illustrates an overview of how monitors are deployed within a wireless environment. The following section describes in detail the proposed monitor selection mechanism.

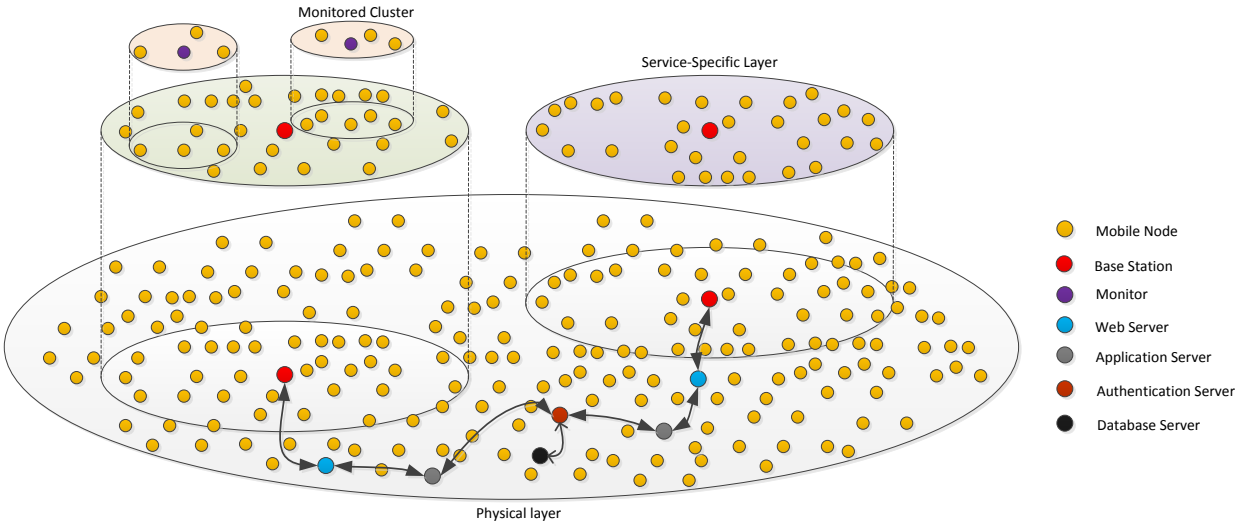

Figure 5.5 Deployment of monitors within a wireless environment

5.4. PROPOSED MONITOR SELECTION MECHANISM

Given that end-user QoS measurement retrieval provides a more accurate reading of the current network and service conditions, selecting a few SSs for monitoring purposes is more preferred. This results in a robust mechanism to achieve low traffic overhead when reporting on service quality. It is assumed that all mobile SSs are willing to cooperate with the BS and are given incentives such as credit or reputation [214] to act as QoS monitors on behalf of other SSs.

The monitor selection process is divided into three phases, as shown in Figure 5.6. In order to ensure a reliable and accurate monitoring service while minimizing frequent triggering of the monitor selection process, candidate monitors must meet the criteria as identified in Phases I and II. Phase I identifies the nodes that are capable of handling the monitoring task. Monitors are required to achieve a minimum acceptable score in terms of their residual power levels and expected remaining service lifetime. In other words, a higher-level of the node's residual power in conjunction with small expected power consumption when transmitting measurements to the BS, results in a higher node score. In addition, the longer a node keeps its service session active, the higher the score is, and the greater the chances are for the node to be nominated as a monitor. The nodes with the highest scores are then selected as candidate monitors.

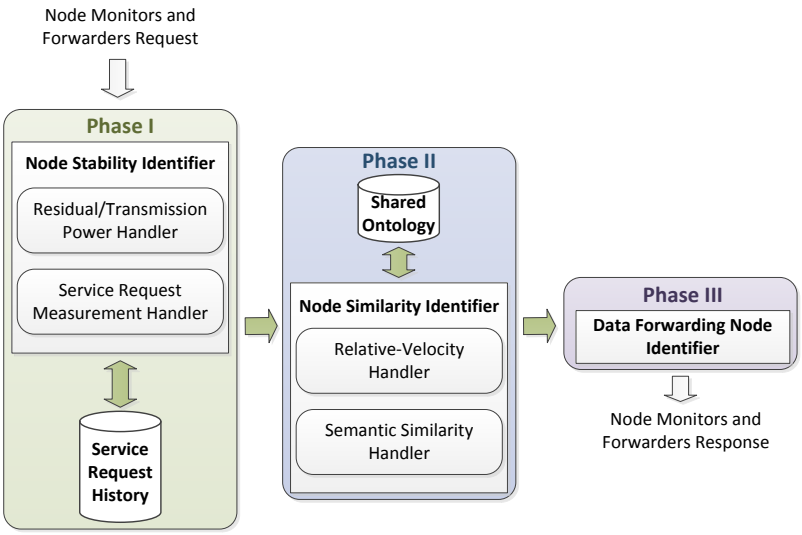

Figure 5.6 Proposed monitor selection mechanism design model.

Once candidate monitors are identified, the next step is to match these nodes to existing SSs which require service monitoring. This matching process is based on the similarity between a monitor and a neighbor SS with respect to their mobility, used services and devices. To better illustrate the importance of this process, consider the profiles of the three users, namely, X, Y and Z, in Figure 5.7. Clearly, X and Y might be similar in terms of their mobility patterns; nonetheless, X and Z may achieve a higher similarity level if their used class of service (CoS), device properties, and device capabilities are used. Obviously, any two nearby nodes with similar devices, having, relatively the same speed and heading direction and using the same service are likely to experience the same service quality and would produce highly similar QoS measurements reports.

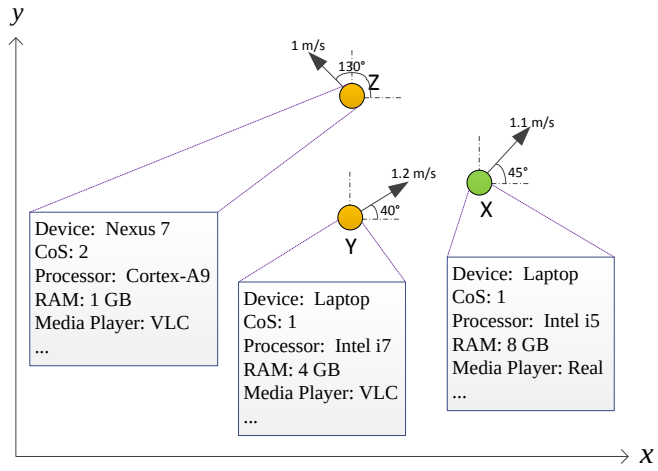

Figure 5.7 Similarity identification between mobile nodes.

To take advantage of this phenomenon, in phase II, a 'similarity matching' technique between candidate monitors and other nearby SSs is incorporated. The algorithm clusters the SSs based on their similarity scores to nearby candidate nodes and assigns the role of a monitor to one candidate monitor within each cluster. This node is responsible for monitoring the service quality and for reporting to the SP service quality measurements on behalf of other nodes in the cluster. The monitors will periodically report on the measured performance-related QoS properties in pre-defined time intervals. The high degree of similarity between the monitors and other SSs ensures that the measurements difference between the monitor and the monitored nodes are kept to a minimal value. Each cluster monitor will contribute to a much smaller traffic overhead and, hence, reduce the average per-user consumed transmission power for QoS measurements reporting.

Since centralized performance evaluation provides a faster and more accurate analysis of the network, service quality measurement analysis and evaluation is performed at the BS.

Feedback sent by monitors to the BS can be either direct, i.e., using a 1-hop, or indirect using multi-hops. In the latter case, measurements are sent from the monitor to one or more intermediate nodes and then to the BS. Based on the monitor's location, the use of intermediate forwarding nodes may further contribute to reducing the needed consumed power by the monitoring node and provide more incentives for these nodes to cooperate. Phase III is responsible for this process of selecting the forwarding path between the monitor and the BS. Details pertaining to each of these phases are discussed in the following sections.

5.4.1. Candidate Monitors Selection (Phase I)

5.4.1.1. An Overview

Identifying candidate monitors requires a measure of service stability and transmission capabilities. Nodes that have maintained their service session with large residual energy and lower transmission power requirements are more likely to be selected as monitors and are said to be stable.

To maintain the goals of transmission power and traffic overhead reduction, adaptive cell sectoring is employed and a threshold is enforced on the allowed distance between users and monitors on the one hand, and monitors and the BS on the other. Since the SS density can be highly non-uniform, as shown in Figure 5.8, partitioning the cell into equal width sectors is

not the best way of assigning nodes fairly to sectors due to user mobility. Adaptive cell sectoring [215] in which sector configurations are adaptive to the load in the cell is used in this developed monitoring scheme. The aim is to provide a fair competition among SSs such that if a network management system requires a certain percentage of the mobile nodes to be assigned as QoS monitors, that percentage will be enforced among each sector of the cell.

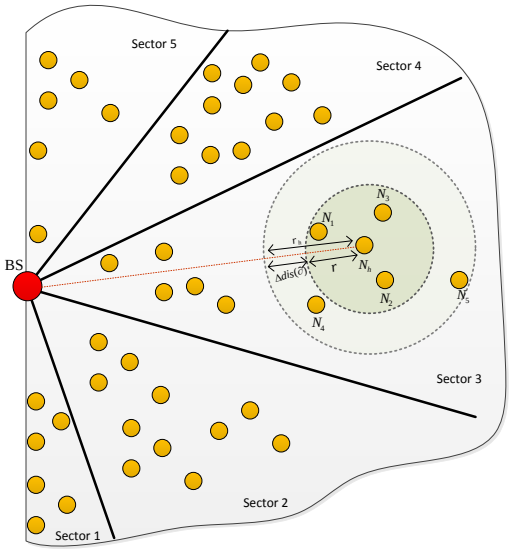

Figure 5.8 Cell Sectoring and a mobile node's communication distance.

It is assumed that the BS has knowledge of all SS profiles, such as their offered service sessions and device capabilities. In each of the created sectors, the BS provides a score for each SS in order to identify candidate monitors. For a set of N mobile nodes currently residing in a given sector, an initial node score, SC_h for $h = 1, \dots, |N|$, is calculated as follows:

$$SC_h = [w_{service} \times O_h + w_{power} \times E_h], \tag{5.1}$$

where $E_h \in [0,1]$ is the power-level stability score, which is a measure that reflects the node's residual capacity and needed transmission power for QoS reporting. $O_h \in [0,1]$ is the service session stability score and is calculated based on a node's service lifetime. Both E_h and O_h are scaled by the SP's assigned normalized weights $w_{service}$ and w_{power} respectively, to determine their significance.

Nodes with the highest score are selected as candidate monitors and are represented by the set $P \subset N$. Further details related to the calculations of these scores are provided in the following sections. The set P represents the candidate monitors that are capable of providing

the monitoring service, as they are the most stable nodes in terms of their energy and session connection time in the cell. Figure 5.9 illustrates an example of the division of responsibilities in a wireless environment. Details on the calculation of these scores are provided below.

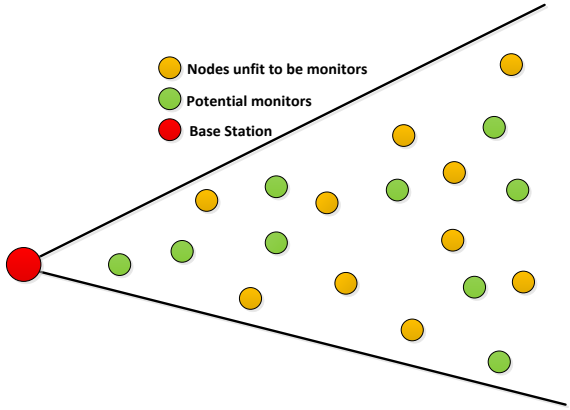

Figure 5.9 Potential monitors and unfit monitoring nodes.

5.4.1.2. Power-level Stability Scores

To calculate the score for each node with respect to the node's transmission and residual power capacity, it must be initially noted that the needed transmission power depends on the distance and the size of the data being sent [216]. As will be shown later, a multi-hop forwarding mechanism is adopted for the communication between the monitors and the BS. This is done in order to maintain fairness in terms of the power consumption between those monitors that are located closer to the BS and those that are at a farther distance. Also, a measure of the power consumption needed by a candidate monitor for transmitting data to the farthest node within its communication distance is considered. This will allow nodes that are located at a farther distance to the BS to compete with nodes that are closer to the BS.

To realize these goals, each node must select its served nodes by adjusting its communication radius r_h which is initially set to a given default value r. The BS then communicates a minimum allowable number of nodes, ∂, that h must serve and act as a monitor. If this number of SSs is not met using this initial value, then each node h attempts to increase its communication distance to r_h to serve more subscribers (Figure 5.8). The new length of node $h's$ communication radius, r_h, is set such that the number of nodes within the area of the communication distance, $num_{of_{nodes(\pi r_h^2)}}$, is at least ∂, unless the transmission

power required to send data to the furthest node within the communication distance is more than or equal to the transmission power required to send data to the BS. More precisely, h selects the maximum possible increase, $\Delta dis(\partial)$, in its communication range, where

$$r_h = r + \Delta dis(\partial) \tag{5.2}$$

s.t.

$$num_{of_{nodes\left(\pi r_h^2\right)}} \geq \partial \Leftrightarrow E_{transmit}^{h \to r_h}(t_c) < E_{transmit}^{h \to BS}(t_c) \tag{5.3}$$

where $E_{transmit}^{h \to r_h}(t_c)$ is the amount of energy required to send a packet from h to the farthest node within its communication area at the current time instant t_c, and $E_{transmit}^{h \to BS}(t_c)$ is the energy required to send a packet from h to BS at time t_c.

The power stability score for node h can then be derived according to the product of the residual and transmission power ratios, as follows:

$$E_h = \left(\frac{E_{residual}^h}{E_{residual}^{Max}}\right)\left(1 - \left(\frac{E_{transmit}^{h \to r_h}}{E_{transmit}^{\overline{N \to BS}}}\right)\right) \tag{5.4}$$

where $E_{residual}^h$ is the estimated current residual energy in node h, $E_{residual}^{Max}$ is a reference maximum energy that a mobile node can obtain with a fully charged battery, typically identical for all nodes in the case of this research. $E_{transmit}^{\overline{N \to BS}}$ is the maximum transmission power required by a node h to communicate with the BS. Algorithm 4 provides a detailed description of the steps involved in identifying node power-level stability.

Algorithm 4: Node power-level stability

Input: $r, E_{residual}^{Max}$
Output: Nodes energy level stability score

/* Calculate transmission energy consumption to BS */
for each node h **do**
    approximate distance D to BS using $RSSI$
    measure residual energy level $E_{residual}^h$
    send l bits of data to BS
    calculate transmission power $E_{transmit}^{h \to BS}$
end for
/* Determine node's communication distance */
for each node h **do**
    if $num_{of_{nodes\left(\pi r_h^2\right)}} < \partial$ **then**
        calculate $\Delta dis(\partial)$ to achieve $num_{of_{nodes\left(\pi r_h^2\right)}} = \partial$
        calculate $r_h = r + \Delta dis(\partial)$
    else
        set $r_h = r$

15: **end if**
16: calculate transmission power $E_{transmit}^{h \to r_h}$
17: **end for**
18: /∗ compute node′s power constraint score ∗/
19: calculate average transmission power $E_{transmit}^{\overline{N \to r}}$
20: **for** each node h **do**
21: calculate score E_h
22: **end for**

5.4.1.3. Service Request Stability Scores

Typically, nodes in a cellular network do not participate in all ongoing sessions (e.g., IPTV streaming sessions, online multiplayer game sessions, teleconferencing sessions). Also, a node may not participate in a session for its entire duration. For example, a node connected to the BS may not necessarily participate in all live video stream sessions used by its neighbors. Inspired by the work in [217], the work in this thesis follows the theory that the longer a node stays and participates in an environment in which a set of services are provided to the subscribers, the longer it will stay in the future. Therefore, in addition to a node's environmental presence, service usage frequency and duration represent a good indication of its potential as a monitor. These indications are captured through the calculation of a session stability score for a given service. To calculate this score, a weighted sum is obtained for the fraction of time a node has spent in the previous G sessions, i.e.,

$$O_h = \sum_{g=0}^{G} \left(\frac{T(h,g)}{T(g)}\right) \beta^g \tag{5.5}$$

where $T(h,g)$ is node h's participation duration in the session g and $T(g)$ is the session duration. β^g, $0 \leq \beta^g \leq 1$, is a forgetting weight factor that is used to give more significance to participating in recent sessions such that:

$$\sum_{g=0}^{G} \beta^g = 1 \tag{5.6}$$

$$\beta^0 > \beta^1 > \beta^2 > \cdots > \beta^G \tag{5.7}$$

where the most recent and the oldest services are indexed by $g = 0$ and $g = G$, respectively. In other words, the effect of participating in older service sessions will be gradually forgotten by the BS and will not have a strong impact on a node's future service session stability. On the contrary, newer sessions are given a higher weight.

5.4.2. Node and Monitor Matching (Phase II)

5.4.2.1. An Overview

In addition to node stability which was previously discussed, choosing monitors requires an identification of the similarity between a candidate monitor and the SSs. Therefore, this phase ensures a proper assignment of monitors to users based on their similarities with respect to their mobility, services and devices. Nodes are first grouped into clusters based on their semantic and movement similarities. More precisely, in this phase a set of monitors $M \subset P$ is selected from the previously obtained candidate set P. Each monitor $u \in M$ is assigned the responsibility of measuring and reporting service quality on behalf of a set of nearby similar nodes. The size of the set M is determined by the SP based on the number of SSs and the cell size as described in [206].

The monitor-to-user assignment process proceeds as follows: the BS first broadcasts to all the nodes the set P of candidate monitors. Each node, $h \in N - P$, i.e., each node in the set of nodes not considered as candidate monitors, requests a semantic similarity report between itself and all candidate monitors within its communication distance, in addition to submitting its current velocity and location to each of these nodes. Next, a new score, SIM_k, for each candidate monitor $k \in P$, is calculated as follows:

$$SIM_k = [w_{velocity} \times V_k + w_{similarity} \times S_k] \quad (5.8)$$

where V_k and S_k are the relative velocity and semantic similarity scores, respectively, for candidate node k, with respect to nearby nodes. Both V_k and S_k are given normalized weights $w_{velocity}$ and $w_{similarity}$, to determine their respective significance.

Once a score SIM_k is calculated for each potential monitor k, those with the highest score that are sufficient to serve all other SSs are selected as monitors. Each selected monitor measures and reports the service quality to the BS on behalf of all its assigned nodes within its cluster. As monitors leave the area and new users enter the sector, the BS periodically performs a search for new monitors to ensure all the SSs are covered by monitoring nodes. Details pertaining to the calculations of these 'similarity scores' are provided below.

5.4.2.2. Relative-Velocity

The fast-changing characteristics of mobile wireless environments make it difficult to discover relatively-stable monitors between moving nodes. Since the speed and direction of

nodes can be realized based on a node's recent movement history, it is practical to incorporate this information when choosing a QoS monitor.

The proposed algorithm chooses a potential monitor $k \in P$ that will most probably be within the communication distance of a monitored node $h \in N - P$, at the current and future time point. To do so, the relative velocity between node h and a potential monitor is used to determine if the two nodes are within communication distance at a future time point. The relative velocities of the previous time points are also considered. Given that two nodes are within communication distance of each other at previous time points, it is more probable that these two nodes will also be within the communication distance of each other in the future. A forgetting weight factor is used to give more significance to recent relative velocity measurements. A potential monitor k agrees to become a monitor for node h only if it is expected to be within the communication distance of node h at time t_{c+1}.

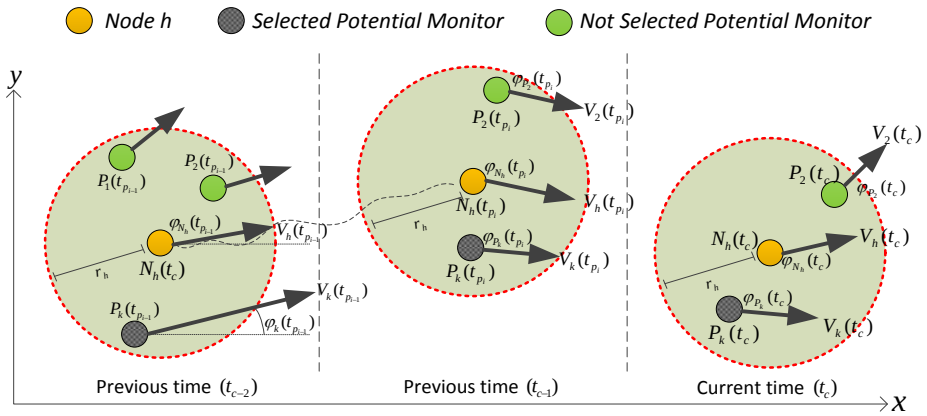

Figure 5.10 Relative velocity calculation.

To better illustrate this process, consider Figure 5.10, at time t_c, a potential monitor k is already within the communication distance r_h of a node h. The location information of h at the current time t_c and at the previous time points t_{c-1} and t_{c-2}, including its position, speed $\delta_h(t)$ and direction angle $\varphi_h(t)$, is sent to each potential monitor within its communication distance at time t_c. Each potential node k, then, calculates a score that will determine the node with the highest probability of being within node's h communication distance at time t_{c+1}, as follows:

$$\Delta v_{kh}^{(x)}(t_c) = \delta_k(t_c)\cos\varphi_k(t_c) - \delta_h(t_c)\cos\varphi_h(t_c) \tag{5.9}$$

$$\Delta v_{kh}^{(y)}(t_c) = \delta_k(t_c)\sin\varphi_k(t_c) - \delta_h(t_c)\sin\varphi_h(t_c) \tag{5.10}$$

$$\Delta V_{k-h}(t_c) = \frac{\Delta v_{kh}^{(x)}(t_c) + \Delta v_{kh}^{(y)}(t_c)}{2} \tag{5.11}$$

$$RV_{kh} = \sum_{i=0}^{C} \left(\frac{\Delta V_{k-h}(t_c - i)}{\overline{\Delta V}} \right) \delta^{c-i} \tag{5.12}$$

where $\Delta v_{kh}^{(x)}(t_c)$ and $\Delta v_{kh}^{(y)}(t_c)$ is the velocity difference between nodes k and h on the $x - axis$ and $y - axis$ respectively, $\Delta V_{k-h}(t_c)$ is the average velocity difference between nodes k and h, RV_{kh} is the total relative velocity difference for the considered time points, $\overline{\Delta V}$ is the maximum relative velocity difference found in the environment. δ^{c-i}, $0 \leq \delta^{c-i} \leq 1$, is a forgetting weight factor, the most recent and oldest relative velocity calculations are indexed by $i = 0$ and $i = C$, respectively.

When selecting a monitor for node h, the most optimal choice is to find k with a zero relative velocity such that $\Delta v_{kh}^{(x)}(t_c) = 0$ and $\Delta v_{kh}^{(y)}(t_c) = 0$. Ultimately, the potential monitor $k \in P$ at time instance t_c, with the smallest relative velocity value is given the highest relative velocity score, V_k.

$$V_k = 1 - \Delta V_{k-N} \tag{5.13}$$

$$\Delta V_{k-N} = \frac{\sum_{h=1}^{|N-P|} RV_{kh}}{\gamma} \tag{5.14}$$

where γ is the size of k's list (i.e. the number of nodes potential monitor k is responsible for monitoring), ΔV_{k-N} is the normalized velocity difference for potential monitor k. Algorithm 5 provides a detailed description of the steps involved in identifying node velocity similarity.

Algorithm 5: Node velocity similarity

Input: Potential Monitors $P \subset N$, Monitors $M \subset P$
Output: Velocity similarity score V_k
/* Calculate relative velocity similarity score */
1: **for** each node h **do**
2: **for** each potenital monitor k within node $h's$
3: communication distance at time t_c **do**
4: request current position $PS_h(t)$, speed $\delta_h(t)$, direction
5: angle $\varphi_h(t)$ for the cuurent and previous time points.
6: calculate relative velocity $\Delta v_{kh}^{(x)}(t)$ on the $x - axis$
7: calculate relative velocity $\Delta v_{kh}^{(y)}(t)$ on the $y - axis$
8: calculate $\Delta V_{k-h}(t)$
9: calculate RV_{kh}
10: store in $k's$ temporary list
11: **end for**
12: **for** each potential monitor k with $list.size(k) \neq 0$ **do**
13: calculate V_k

14: **end for**
15: choose potential monitor with the highest score V_k
16: **end for**

5.4.2.3. Semantic Similarity

In addition to relative velocities, the decision made to select monitors involves evaluating the semantic similarity between each potential monitor and the mobile node. The presence of an OWL-based ontology is considered in this dissertation. It has the ability to represent complex relationships between different contextual concepts. Through these ontologies, variations in the semantic descriptions of mobile nodes and the syntactic representations of their functionalities can be compared.

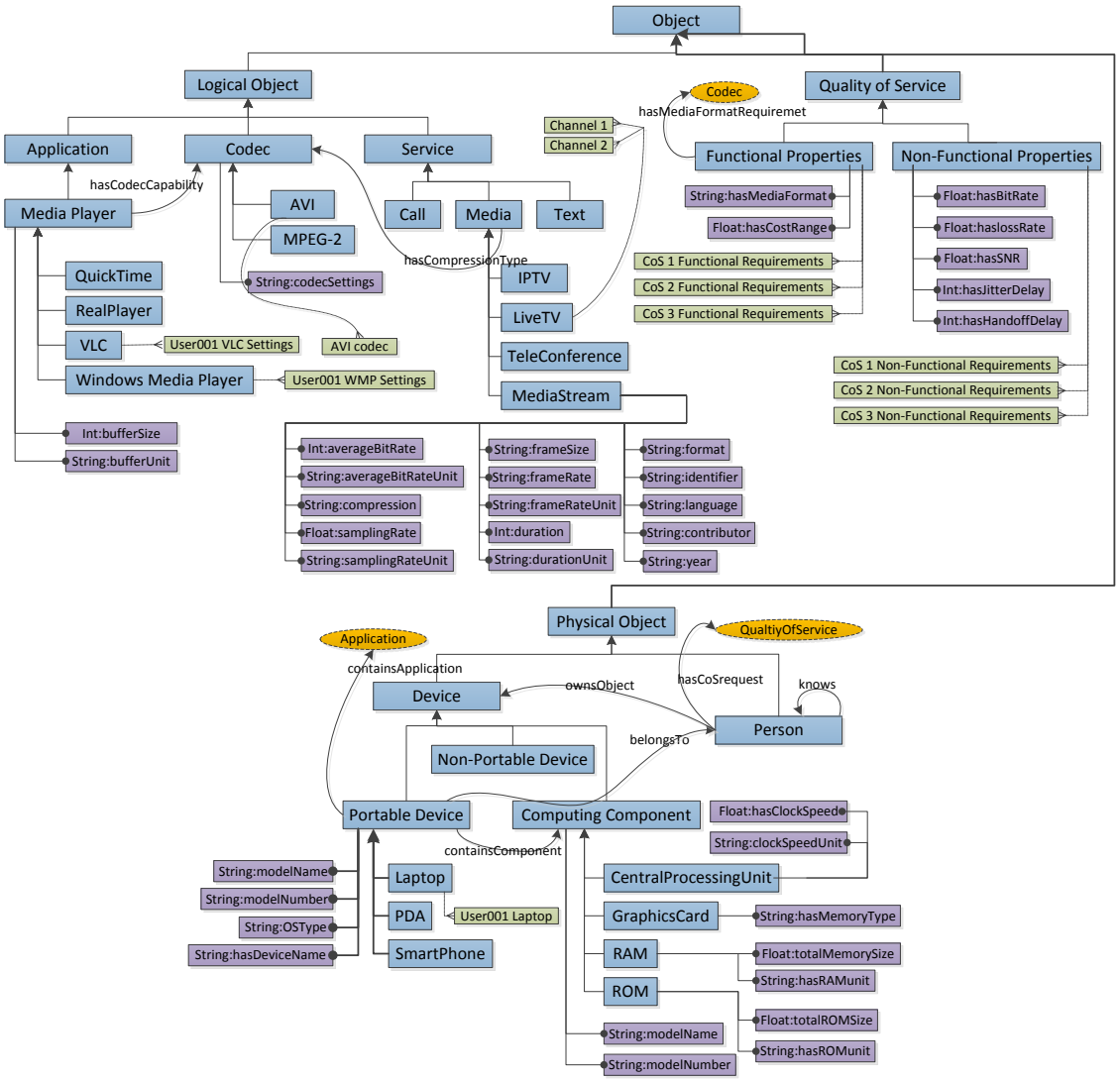

Figure 5.11 Simplified view of the ontology.

Context is any information of interest to a person within a system that can be sensed, inferred, or measured. Context information is thus infinite and should be simplified and fit to a system's domain requirements. The presented ontology covers an extended area of knowledge which contains the majority of contextual concepts that could possibly exist within a wireless environment. Since such environments provide services to users through their computing devices, the presence of the ontology that models users, devices, services and QoS is highly important. These characteristics are grouped under the Physical and Logical object entity classes described in Figure 5.11, in which physical characteristics of a device such as the CPU, RAM and ROM are part of the Physical Object class, while device application characteristics such as the media-player being used are grouped under the Logical Object class. Service and QoS entities are also modeled within the ontology in which service is contained within the logical object class, while QoS is at the same hierarchical level as the Physical and Logical object class entities.

Through these ontologies, variations in the semantic descriptions of mobile nodes and the syntactic representations of their functionalities can be compared, and this will largely aid in searching for the most similar nodes required in the clustering process. The decision made to select monitors involves evaluating semantic and syntactic property description similarity between a potential monitor PM_k and a mobile node h that was declared unfit to be a monitor. The similarity value is needed to rank and choose the most-fit nodes from the set of PMs. Extensive research has been conducted in regard to semantic comparison in web services, media services, and policy properties [218] [219] [220]. Node clustering in wireless environments requires some degree of support for semantic evaluation of mobile nodes and the requested services. Ontologies play a crucial role in the semantic-driven node clustering algorithm.

Semantic similarity between two nodes is identified by comparing the descriptive properties of ontology classes to determine the proportion of matching properties. In the case of a description mismatch, semantic distance is used to calculate the distance between concepts in the ontology. Distance refers to the number of class entities between a common ontology class or object property. This distance can determine whether concepts are related or disjoint [221] [222].

For each potential monitor k within the communication distance of node h, the BS

evaluates the semantic similarity between them. When comparing property descriptions of a potential monitor k with node h, the similarity function is illustrated with the following relationship [221]:

$$Sematic\_SIM[k,h] = \frac{|A \cap B|}{|A \cap B| + \omega(k,h)^{A \setminus B}|A \setminus B| + \omega(k,h)^{B \setminus A}|B \setminus A|} \quad (5.15)$$

where A and B correspond to the sets of ontology features describing nodes k and h, respectively. $A \cap B$ is the set of matching features in both nodes k and h, $A \setminus B$ is the set of features that are in k but not in h, $B \setminus A$ is the set of features that are in h but not in k, $\|$ is the cardinality of a set, and ω is a function that defines the relative importance of the non-common characteristics.

The similarity score defined in (5.15) has a range from 0 to 1, such that if an exact match is found in which every concept in k's description is also found in h's description, then a score of 1 is given. On the other extreme end, if there is no match, in other words, every concept found in k's description is not found in h's description, then the similarity score is 0 and the two nodes are considered disjoint. Most of the cases will fall under the third category such that the similarity score has the range $0 < Semantic\_SIM[k,h] < 1$. This indicates that an exact match is found in some descriptions while the other concepts do not have a match but are semantically near each other. This case is valid only if both descriptions fall within the same ontology branch.

The 'relative importance' values of the non-common descriptions $\omega(k,h)^{A \setminus B}$ and $\omega(k,h)^{B \setminus A}$ indicate how relatively important are descriptions in one mobile node that are not present in the other. The approach adopted to calculate such importance is to compare the degree of generalization between entity classes by determining the distance from the entity class found in one node but not found in the other to their immediate superclass that subsumes them. That is, their closest common ancestor hierarchically in the tree-structured ontology as seen in Fig.5.11. Hence, the function ω (5.16) - (5.17) can be expressed in terms of the depth and length of an entity class.

Semantic depth is defined as the number of concepts that are included in the shortest path from the concept being examined to the ontology's root conceptual property. On the contrary, semantic length is defined as the number of concepts that are included in the path from the concept being examined to the closest common ancestor property. This includes the ancestor property for the two mobile nodes being compared.

$$\omega(k,h)^{A\backslash B} = \sum_{a=1}^{|A|}\left[\frac{length(Desc_k^a, Desc_h^b)}{depth(Desc_k^a)}\right]/|A| \tag{5.16}$$

$$\omega(k,h)^{B\backslash A} = \sum_{b=1}^{|B|}\left[\frac{length(Desc_h^b, Desc_k^a)}{depth(Desc_h^b)}\right]/|B| \tag{5.17}$$

where $Desc_k^a$ is an ontology property for potential monitor k in which $a = 1 \ldots |A|$. $Desc_h^b$ is an ontology property for node h in which $b = 1 \ldots |B|$.

With this definition of ω, the non-common characteristics between mobile nodes' descriptions are considered less important than the common characteristics. This follows the research finding in [223] [224] that the closer the common ancestor is, the less important the different characteristics are, compared to the similar characteristics.

Finally, the potential monitor $k \in P$ with the highest similarity value is given the highest normalized semantic similarity score S_k.

$$S_k = \frac{\sum_{h=1}^{|N-P|} Sematic\_SIM[k,h]}{\gamma} \tag{5.18}$$

where γ is the size of k's list. The monitors $u \in M$, $u = 1 \ldots |M|$ are now chosen on the basis of the similarity score function (5.8) and the percentage of monitors required in the environment. The highest scoring nodes are chosen as QoS monitors. The monitor is now responsible for the nodes within its cluster to provide QoS feedback to the BS (Figure 5.12).

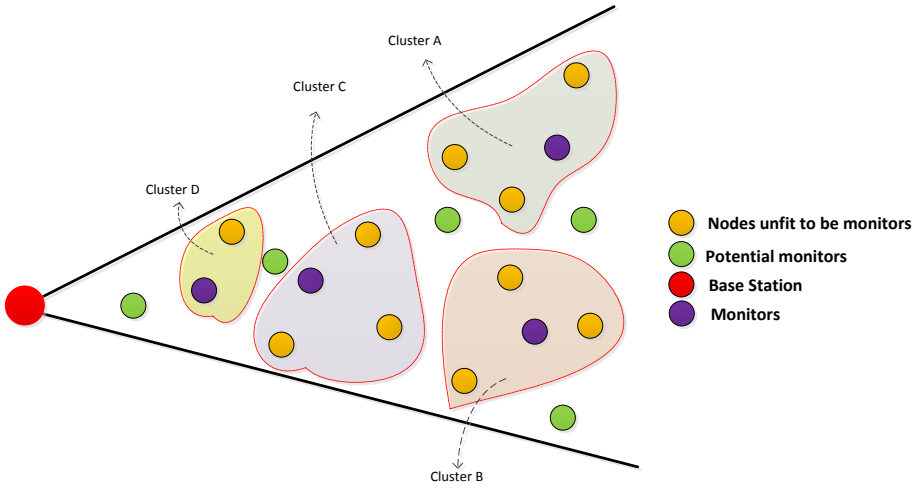

Figure 5.12 QoS monitoring clusters after Phase II.

5.4.3. Forwarding Node Identification (Phase III)

The set of monitors chosen in Phase II are distributed within each cell such that each monitor now serves a particular group of nodes. This distribution necessitates the presence of

a new set of nodes capable of forwarding the QoS feedback information for monitors that are distant to the BS. Hence, in Phase III, the set $D \subset P$ of candidate monitors which were not chosen as monitors are required to join an already established cluster. To join a cluster, each unassigned candidate monitor $d \in D$ must find the most similar monitor. The process of similarity identification is repeated such that the similarity score equation in (5.8) is used to find the most similar monitor within proximity for each remaining unassigned potential monitor (Figures 5.13 and 5.14).

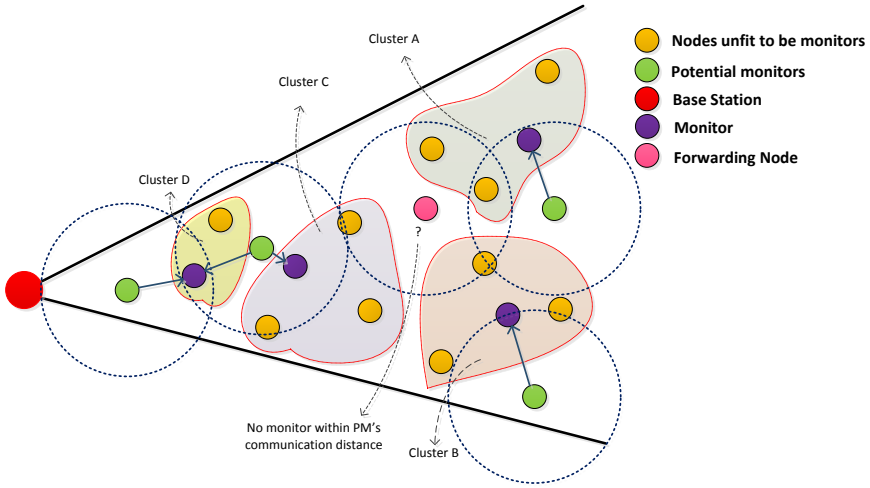

Figure 5.13 PMs joining neighboring cluster.

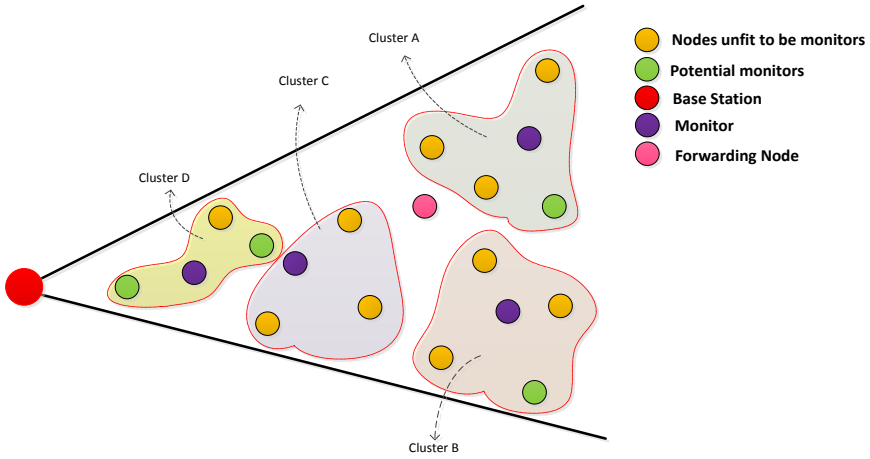

Figure 5.14 Final organization of QoS monitoring clusters after phase III.

Any remaining node d left that is not part of any cluster is used as a data forwarding node. The set of forwarding nodes $F \subset D$ are distinguished as nodes with relatively high

battery power availability and low transmission power requirement capabilities. A monitor will only consider forwarding nodes with relatively smaller distance to the BS. To perform such a task, the intermediate forwarding node will request the cluster-head ID from nodes within the intermediate forwarding node's communication distance (Figure 5.15). Intermediate node will then request from the BS to determine which monitors should forward QoS feedback data to the new intermediate node, if any. This is achieved by determining whether the estimated distance of an intermediate node to the BS is less than or greater than the distance of the monitor to the BS. If the first condition is true than the BS informs the monitor to forward its data through the new intermediate node until $E_{transmit}^{MN_u \to IN_v}(t_c) \geq E_{transmit}^{MN_u \to BS}(t_p)$, otherwise PM will forward QoS feedback data directly to the BS (Figure 5.16).

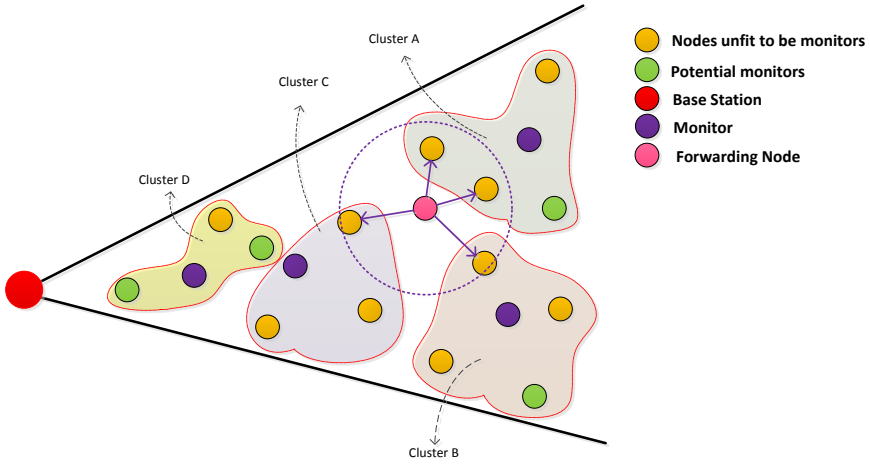

Figure 5.15 Identifying MNs within intermediate node's communication distance.

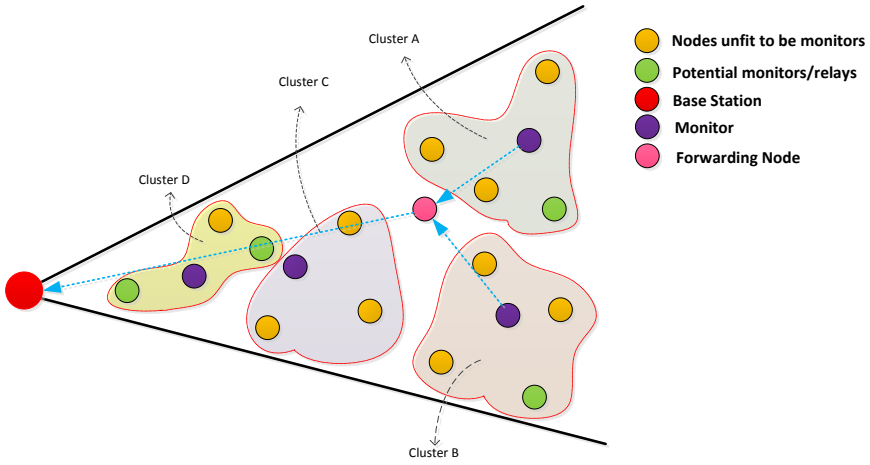

Figure 5.16 Forwarding QoS performance metrics to BS via Intermediate Nodes.

Therefore, some monitors will forward their QoS feedback to an intermediate node $f \in F$, while others which are not far away from the BS will basically forward its data directly to the BS. Algorithm 6 provides a detailed description of how forwarding nodes are assigned to monitors.

Algorithm 6: Identifying intermediate data forwarding nodes

Input: Candidate forwarding nodes $D \subset P$
Output: Intermediate forwarding nodes $F \subset D$

```
/* Determine which forwarding node is most suitable */
for each node d do
    for each node h within d's communication distance do
        request clusterhead of h
        send similarity comparison request to BS
        wait for forwarding node decision from BS
    end for
end for
/* Node forwarding decision of BS */
if E_transmit^{u→f}(t_c) < E_transmit^{u→BS}(t_c) then
    notify u to forward any data targeted to BS
    through f until E_transmit^{u→f}(t_c) > E_transmit^{u→BS}(t_p)
else
    reject u and f cooperation
```

5.5. EXPERIMENTAL EVALUATION

5.5.1. Overview

The work presented in this chapter focuses on the mechanism of choosing appropriate monitors in a cellular network to successfully report service performance without relying on individual QoS performance feedback. As such, in this section an evaluation is done on the performance of the research's solution by testing it against the individual performance feedback mechanism. The robustness of the presented monitor selection mechanism in terms of QoS monitor reporting accuracy is presented. Also, it will be shown that the presented reporting mechanism provides traffic overhead reduction and node power efficiency.

5.5.2. Simulation Setup

The experiments are conducted using NS-2. The reference topology is a 100 m × 100 m simulation area with a WiMAX BS placed at the center. Different scenarios have been conducted with up to 100 mobile nodes placed randomly in the cell. The minimum allowable

number of nodes within a cluster ∂ is set to 3, 5, and 7. The percentage of monitors is set at 20%. BonnMotion [225] is used to generate node movement scenarios according to the Gauss Markov mobility model. BonnMotion is a Java software that creates and exports mobility scenarios to network simulators such as NS-2, GloMoSim [226], and COOJA [227]. The node speed is considered to be in the range of 1-2 m/s. The packet size is set to 1240 bytes using constant bit rate (CBR) traffic with 0.1 second intervals originating at the BS. Node service quality feedback messages are sent periodically at 10-second intervals. The simulations run over a 3 GHz Core 2 Duo Intel processor, equipped with a Linux server and 4-Gbytes of memory.

Semantic similarity evaluations require formal ontology-based service descriptions. OntoCAT [228], an open source java-based library is used to parse, search through, and compare acquired service description files. OntoCAT provides a high level abstraction for interacting with ontology resources in the standard OWL format. It provides a seamless programming interface to query heterogeneous ontology resources where each resource is wrapped behind Java service commands. Furthermore, it provides a robust solution to perform the semantic similarity evaluation.

5.5.3. Measurement Accuracy

When a monitor is assigned to report performance measurements on behalf of the nodes within its cluster, the performance feedback should represent what the nodes are experiencing. In this experiment, several simulation runs have been conducted to test for measurement accuracy.

A simulation test with 100 nodes was performed to determine the delay jitter experienced by each node, in an attempt to compare the accuracy of the reported experienced service quality of the proposed mechanism with that of a pure client-side monitoring mechanism. When $\partial = 3$, the average reported jitter of the monitors over a 600 second simulation run is identical to the experienced jitter of the monitored nodes, as seen from Figure 5.17 (a). The average delay jitter for both monitors and monitored nodes is 0.3 ms and 0.295 ms, respectively, at the end of the simulation run. The maximum and minimum experienced jitter by the monitored node is 0.321 ms and 0.271 ms, respectively. On the contrary, monitors have maximum and minimum experienced jitter of 0.314 ms and 0.275 ms, respectively. It is

therefore clear that all monitors are good candidates for representing other nodes within their clusters and provide accurate performance quality feedback.

The same set of experiments were conducted repeatedly while increasing the minimum allowable number of nodes within a cluster to $\partial = 5$ and $\partial = 7$. Results show similar observations to simulation runs when $\partial = 3$. It can be noted from Figures 5.17 (b) and (c) that as the cluster size increases, the difference in the reported delay jitter becomes larger between monitors and monitored nodes. But this difference is almost insignificant, such that from Figure 5.17 (b), at the end of the simulation run, the average experienced delay jitter by the monitors and monitored nodes is 0.298 ms and 0.290 ms, respectively. As for the simulation conducted when $\partial = 7$, results at 600 s, seen in Figure 5.17 (c), show that the average experienced delay jitter by the monitors and the monitored nodes is 0.301 ms and 0.288 ms, respectively.

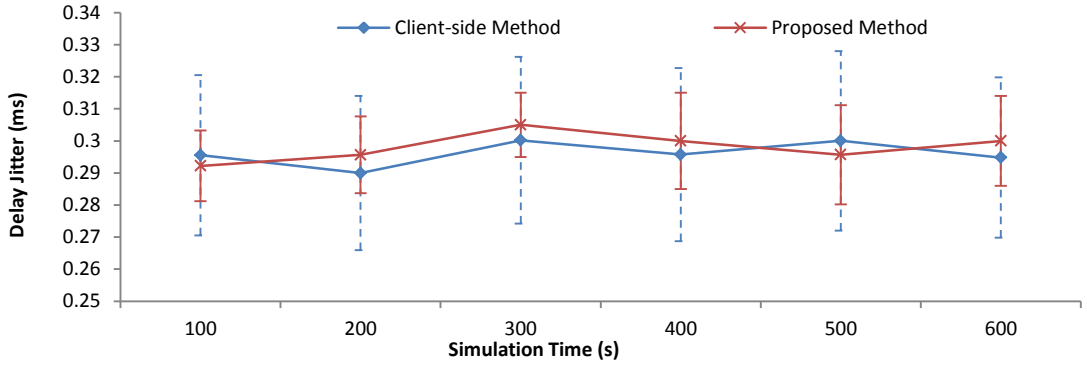

Fig. 5.17 (a). Evaluating the accuracy of reporting delay jitter when $\partial = 3$.

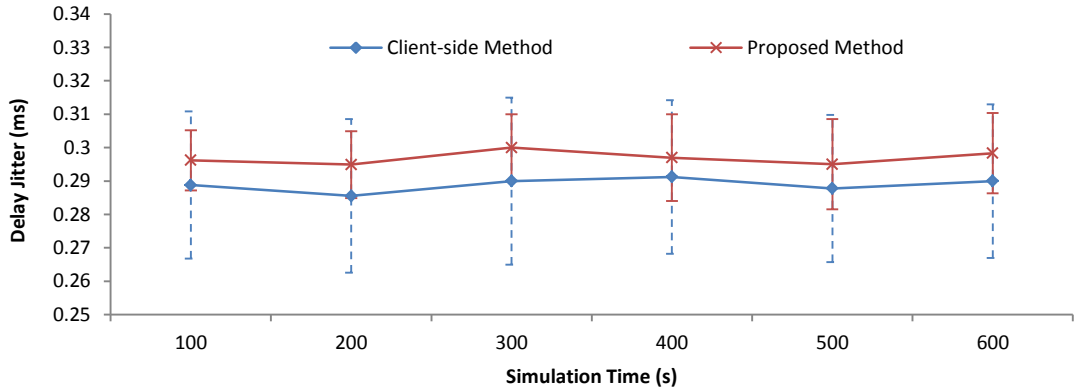

Fig. 5.17 (b). Evaluating the accuracy of reporting delay jitter when $\partial = 5$.

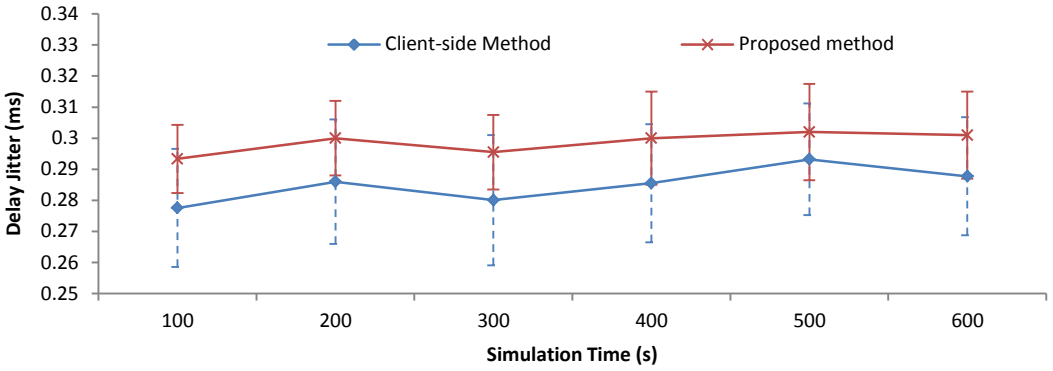

Fig. 5.17 (c). Evaluating the accuracy of reporting delay jitter when $\partial = 7$.

Figure 5.18 provides a summary of the experienced delay difference between the proposed mechanism and the client-side method while increasing the number of nodes within a cluster. It can be observed that the jitter difference is almost doubled each time the cluster size is increased by two nodes.

Although the results show a decrease in the accuracy of the monitor when reporting service quality as the number of nodes within a cluster increases, this difference is nearly negligible. It has been shown [229] that such small delay jitter and packet loss difference does not have an effect on the quality of video streaming in a wireless network. On the contrary, there are other trade-offs to choosing a smaller cluster size over a larger one in highly dense environments, such as the traffic overhead burden and the transmission energy consumption. Further details of these two aspects are discussed in the next section.

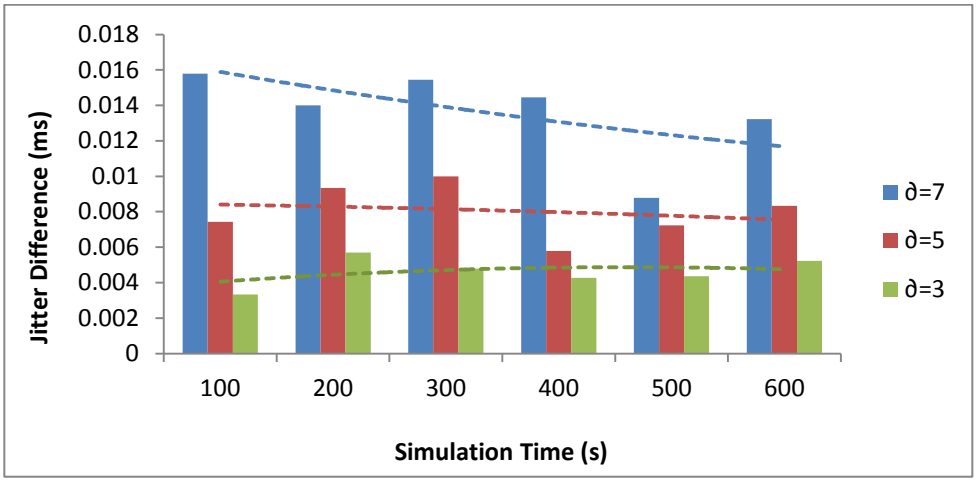

Figure 5.18 Difference between the experienced jitter of MNs to nodes within clusters with the total number of mobile nodes in the environment is 100 and ∂=3, 5, and 7.

Another observation is that a small cluster size provides almost similar results for a small and large number of nodes within the wireless environment as shown in Figure 5.19. The observed delay jitter difference between the monitors and monitored nodes is almost 0.0052 ms for an environment with 20 to 100 nodes. On the contrary, a larger cluster size provides enhanced results as the number of nodes increases in the environment. When $\partial = 7$, the jitter difference between the monitors and the monitored nodes with 20 and 100 mobile nodes in the environment is 0.0313 ms and 0.01323 ms, respectively.

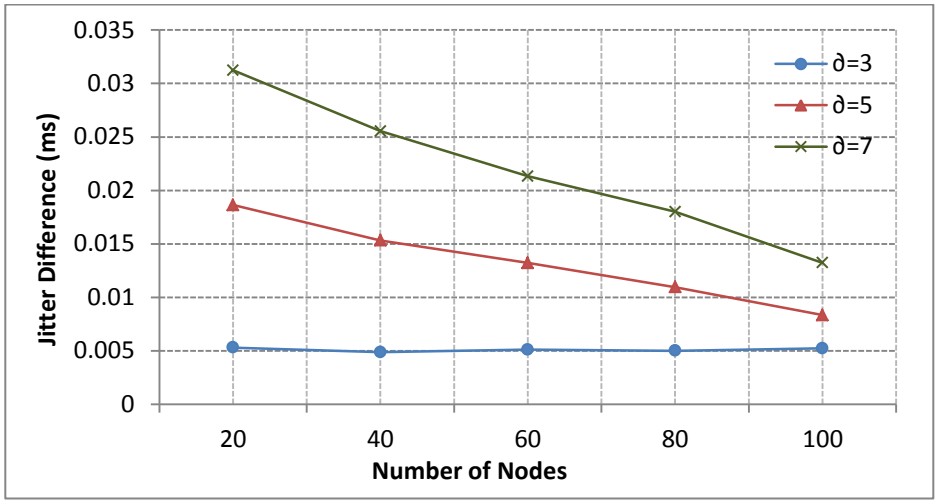

Figure 5.19 Experienced delay jitter differenced between MNs and nodes within clusters as the number of nodes within the environment increases.

The same set of simulation tests were used to determine the accuracy of monitors when reporting packet loss and throughput. Figure 5.20 provides an analysis of the average accumulated number of packets lost after each 100 s. When $\partial = 3$, at the end of the simulation run, the average reported accumulated number of packets lost after 100 s is identical to the experienced packet loss of the monitored nodes, as seen in Figure 5.20 (a). The average accumulated packet loss for both monitors and monitored nodes is 124 and 123 packets, respectively. The maximum and minimum number of lost packets for the monitored nodes is 128 and 90 packets, respectively. On the contrary, the maximum and minimum reported number of packets lost by the monitors is 120 and 92 respectively.

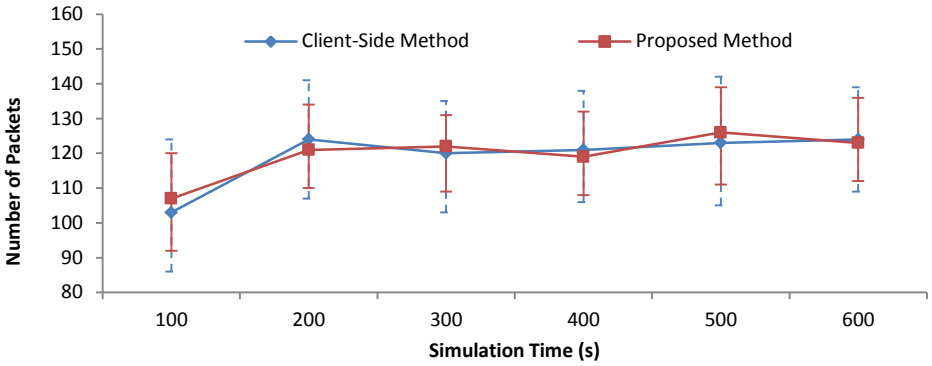

Figure 5.20 (a). Accuracy of reporting packet loss when $\partial = 3$.

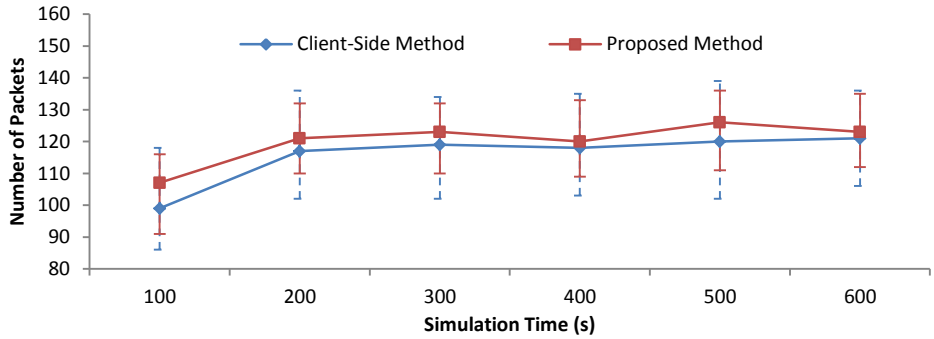

Figure 5.20 (b). Accuracy of reporting packet loss when $\partial = 5$.

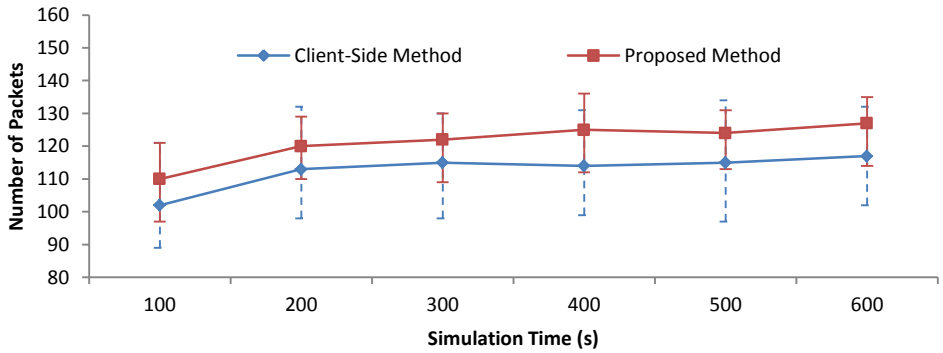

Figure 5.20 (c). Accuracy of reporting packet loss when $\partial = 7$.

Accuracy evaluations for packet loss reporting were also conducted while increasing the minimum allowable number of nodes within a cluster to $\partial = 5$ and $\partial = 7$ as shown in Figures 5.20 (b) and (c). Results show that all monitors are good candidates for representing other nodes within their clusters and provide accurate performance quality feedback.

A similar behavior is also reflected in Figure 5.21, which shows the throughput measurements reported by the monitors and the monitored nodes. The reported bit rate of the monitors over a 600 second simulation run is identical to the experienced throughput of the

monitored nodes for each CoS. Therefore, all selected monitors are good candidates for reporting accurate performance quality to the BS.

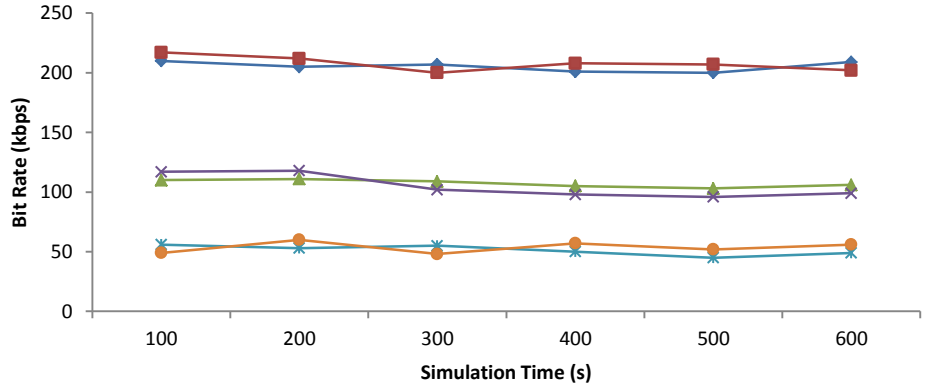

Figure 5.21 (a). Accuracy of reporting throughput when $\partial = 3$.

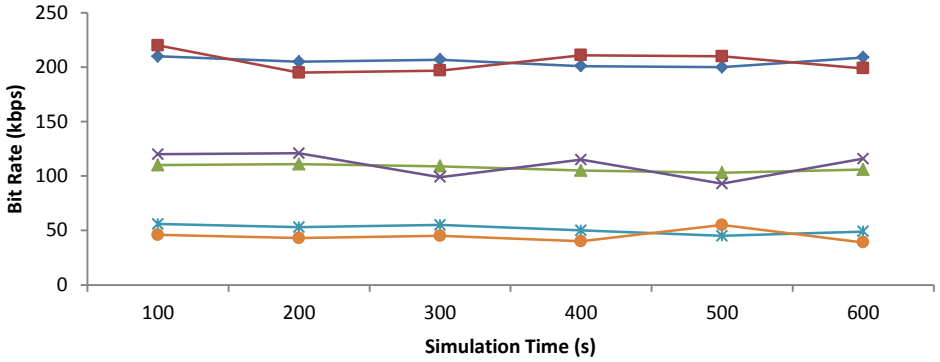

Figure 5.21 (b). Accuracy of reporting throughput when $\partial = 5$.

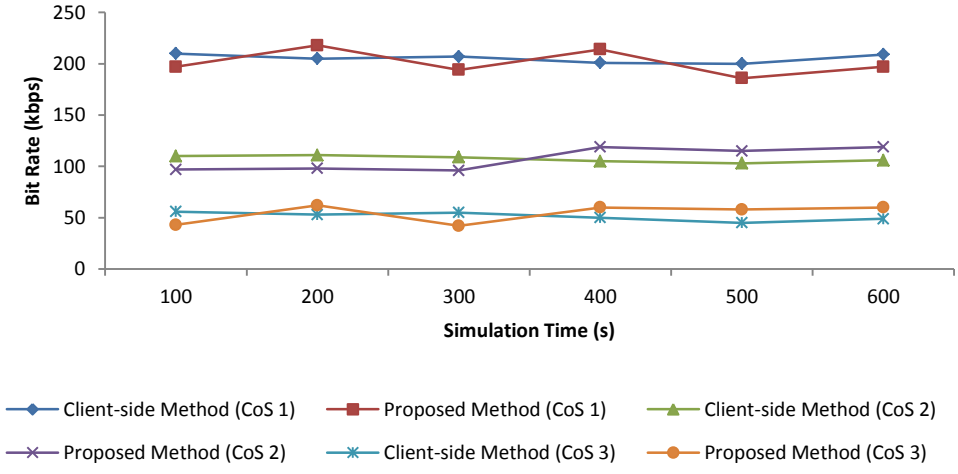

Figure 5.21 (c). Accuracy of reporting throughput when $\partial = 7$.

5.5.4. Traffic Overhead

It is apparent that with any clustering algorithm, communication traffic overhead is experienced when composing clusters. There is always a trade-off between a small increase in communication overhead and system efficiency. For example, an increase in traffic overhead at the beginning of cluster establishment for QoS monitoring would result in less traffic later on. It is therefore necessary to consider such overhead when evaluating the performance of the proposed mechanism. The BS requests service quality feedback messages to be sent from nodes periodically at 10-second intervals.

As a result, from Figure 5.22, with an environment of 100 nodes, it is observed that when clustering is not involved, the same number of packets are sent to the BS periodically every 10 seconds until the simulation is terminated, accumulating to a total of 6000 packets (see Figure 5.23). Moreover, a 60 s clustering delay is incurred on the system with a total of 1090, 1120, and 1145 number of packets when the minimum enforced cluster size is 3, 5, and 7, respectively. On the contrary, traffic overhead is considerably reduced for quality reporting to the BS after establishing clusters, since only monitors are sending messages to the BS either directly or through intermediate data forwarding nodes. From Figure 5.22, it is evident that the larger the size of the cluster, the lower the number of messages communicated periodically to the BS. When $\partial = 7$, 28 messages are communicated to the BS. While when $\partial = 3$ and $\partial = 5$, 60 and 42 messages are communicated back, respectively. The total number of accumulated packets sent for clustering and quality feedback purposes are 4103, 3346, and 2629 packets for $\partial = 3$, $\partial = 5$, and $\partial = 7$, respectively.

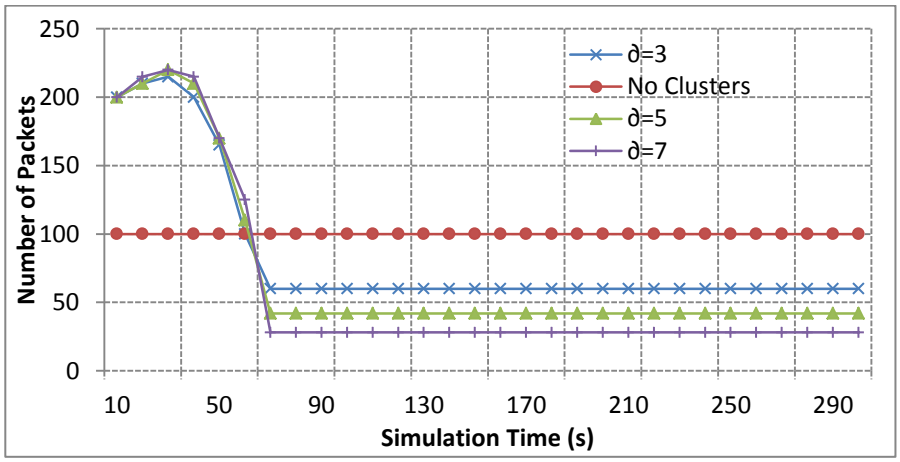

Figure 5.22 QoS feedback messages sent to BS when 100 nodes are present in the environment.

The proposed monitoring mechanism performs well when more nodes are present in the environment as shown in Figure 5.23. This can be explained by the fact that with a small number of nodes within the environment, less quality feedback messages are required to be reported. So, in an environment with 20 nodes, a total of 1200 packets are sent back to the BS after a simulation run of 600 s and no clustering communication overhead is incurred, such that each mobile node is responsible for reporting its quality performance to the BS directly. While for the clustered solution, a total of 854, 670, and 526 packets are incurred for both constructing the clusters and quality feedback reporting when $\partial = 3$, $\partial = 5$, and $\partial = 7$, respectively.

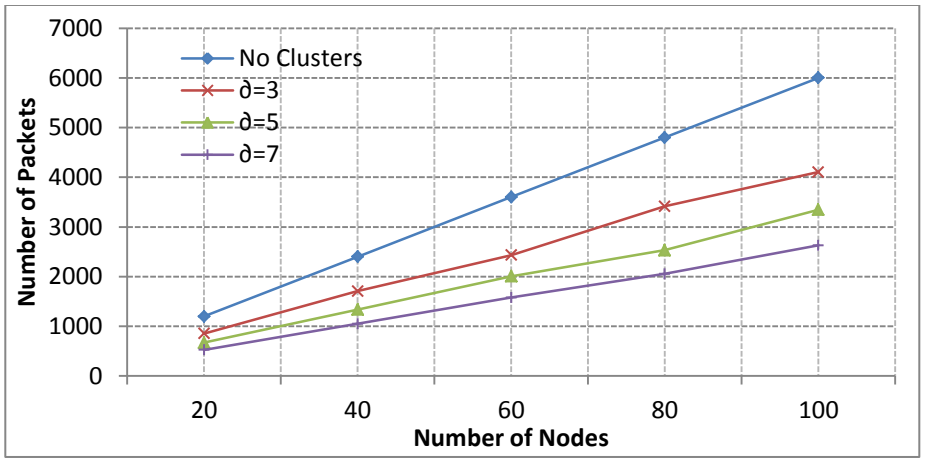

Figure 5.23 Total accumulated number of messages sent to the BS at the end of a simulation run against the number of nodes present in the environment.

5.5.5. Transmission Power Consumption

One of the main objectives of the presented QoS monitoring mechanism is to decrease the overall network transmission power consumption used for service quality reporting in order to prolong the lifetime of mobile nodes within the environment.

For the individual client monitoring mechanism, a total average of 3.39 mJ is used each 10 s to report to the BS service quality performance of 100 mobile nodes, as seen in Figure 5.24. Only direct transmission from the node to the BS is used in this case. Such direct transmission technique results in more power consumption as the distance increases between the mobile node and the BS. It is therefore insufficient to have each node report its service performance directly to the BS.

On the contrary, with the proposed monitoring mechanism, an average of around 1 mJ is used to report performance to the BS. This decrease in power consumption is accomplished through a reduced number of nodes that report to the BS, in which only monitors are responsible for reporting. QoS feedback is sometimes sent to intermediate data forwarding nodes to be forwarded to the BS. This decreases the monitor's energy consumption and prolongs the lifetime of monitors by reducing the energy required to report QoS performance.

Cluster size has an effect on node transmission energy consumption as well. It can be observed from Figure 5.24, that having a cluster size of $\partial = 7$ consumes the least transmission energy over the other two cluster sizes. Conversely, it is evident that more energy is consumed when $\partial = 7$ during the clustering cycle when compared to $\partial = 3$ and $\partial = 5$. This outcome is a result of an increased number of communication messages required to form clusters and derive intermediate forwarding nodes. A total of 61.8 mJ, 60 mJ, and 58.7 mJ are consumed for QoS performance reporting purposes when $\partial = 3$, $\partial = 5$, and $\partial = 7$, respectively, after a 600 s simulation run when 100 nodes are present in the environment. On the contrary, the client-side reporting mechanism consumes a total of 201.8 mJ at the end of a simulation run, which is almost triple what the adopted algorithm achieves.

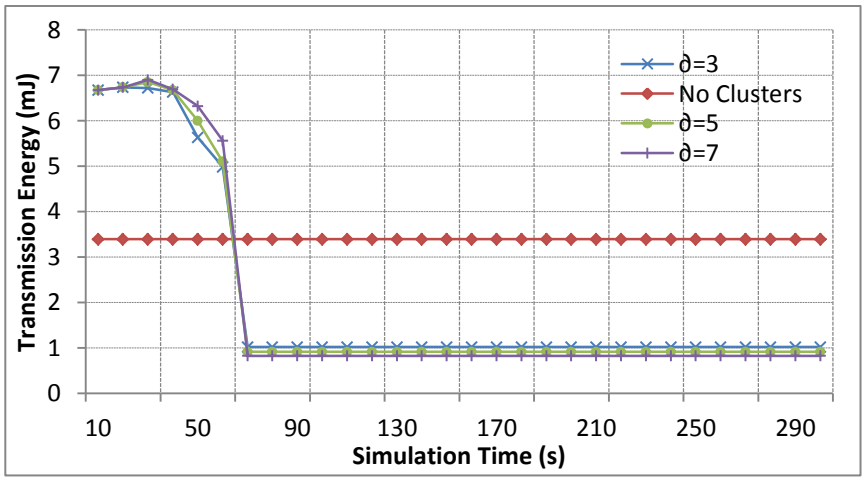

Figure 5.24 Average transmission power consumption for reporting node performance quality to the BS every 10 s for a simulation scenario with 100 nodes.

It is clear from Figure 5.25 that the proposed mechanism performs well as the number of nodes increase within the environment. This can be explained by the fact that less quality

feedback messages are required to be reported when a small number of nodes are present in the environment. Thus, more energy is consumed in clustering as opposed to reporting service quality feedback. From the figure, it can be seen that in an environment of size 20 nodes, a total of 40.7 mJ is consumed to report performance quality to the BS after a simulation run of 600 s and no clustering communication overhead is incurred. Each mobile node is responsible of reporting its service performance to the BS directly. While for the clustered solution, a total of 25.7 mJ, 24.9 mJ, and 24.3 mJ of energy are consumed for both constructing the clusters and quality feedback reporting. This demonstrates that the adopted method performs well in both small and large scale environments.

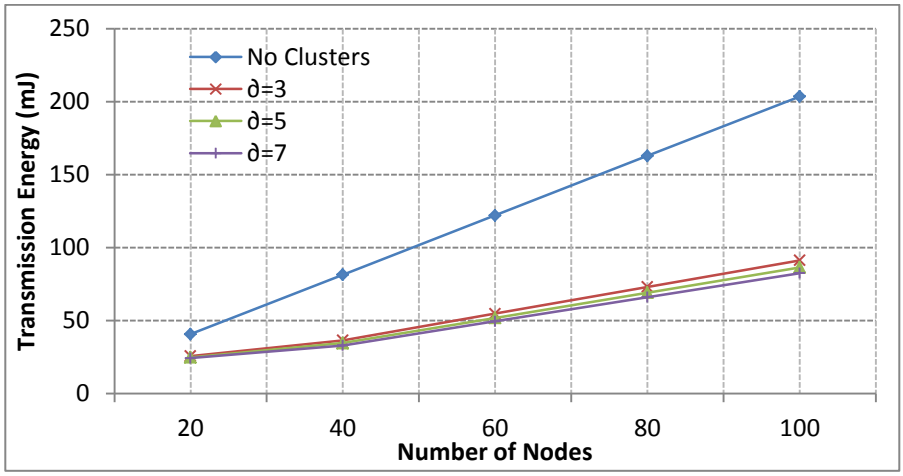

Figure 5.25 Total transmission power consumption for reporting node performance quality to the BS after 600 s of a simulation run against the number of nodes present in the environment.

In summary it is clear that the proposed QoS monitoring method decreases the overall network traffic overhead and also reduces the overall network transmission power consumption. The proposed monitoring technique adapts well when the number of users increases. The solution did not introduce critical performance overhead into the system such that it takes only around 60 s to construct such clusters.

5.6. SUMMARY

This chapter presented a QoS monitor selection mechanism for cellular data networks. The mechanism involves three phases. The first phase identifies stable nodes within the environment based on residual and transmission power capabilities, in addition to a service request score. The second phase identifies node similarity in terms of relative-velocity and

semantic nearness. The presented mechanism provides the benefit of performing client-side quality measurement retrieval while reducing the overall network traffic overhead and transmission power consumption. The algorithm performs well in dense and mobile environments with limited power availability of service subscriber mobile nodes and the absence of performance monitoring devices. Results show that the developed solution is characterized as highly stable with accurate service quality measurements. Additionally, only small overall traffic overhead and transmission power consumption is observed.

CHAPTER 6

CONCLUSION AND FUTURE RESEARCH DIRECTION

This chapter outlines the contributed research work and discusses plans and directions for future work. The chapter is organized as follows: Section 6.1 summarizes the conducted research and the main contributions in the area of policy-based network management in wireless heterogeneous networks. Section 6.2 illustrates the limitations of this current research and sheds a light on the planned future research direction. Finally, Section 6.3 provides some concluding remarks.

6.1 CONDUCTED RESEARCH WORK

The focus of the conducted research thus far, has been the development of an automated policy-based management system that dynamically configures network, handover, and SLA policies. The first step towards achieving that goal included a literature study of the management problem from a service provider's point of view and the difficulties associated with mobility. Two questions were addressed: What are the requirements of a SP profit- and user-satisfying aware service management system? and Why do the current management approaches not satisfy these requirements? Based on the identified limitations of current research, a novel framework for an automated policy configuration system has been designed. The architecture has been presented as a multi-component model that automatically adapts to the dynamicity of the network conditions. In the following, a summary of the main contributions of the current research work is given:

- *A simulator-assisted multi-component network and service management system architecture*. A simulator-assisted framework was presented that determines the optimal decisions for the cooperating providers. This work introduced a complete breakdown of

the architecture that incorporates the advantages of real-time simulation tools to continuously achieve optimal network decisions in the form of adaptive policy configurations. Since in policy-based management systems, the events that occur, the conditions, and the actions that must be performed are denoted into policy terms, in this research the simulator-assisted architecture estimates the impact of different policy adaptation decisions and guides the decision making process for adapting the behavior of network components and different management components such as handoff and service level agreements. It also provides an automated self-managed system that is capable of reacting efficiently to contextual changes in heterogeneous wireless mobile environments. The proposed topology's simulation results illustrated the effectiveness of the proposed scheme in terms of QoS performance, access point load balancing, and service provider profit increase.

- *A novel tabu-search-enhanced variable configuration selection strategy mechanism* has been presented. To perform the task of choosing the optimal policy configurations requires a method to continuously explore the space of all possible configurations. Thus, a fast local search procedure was developed based on the tabu-search heuristic approach. The proposed solution was applied to a discrete variable selection problem to acquire the optimal value of each variable. The new modified tabu-based search algorithm was called "Iterated Local and Global-Tabu Search (IGL-TS)". This method decouples the configuration learning process from that of the actual configuration adaptation step. Evaluation results demonstrated that using local search techniques is considered an effective method to find near-optimal solutions to service provider profit maximization problems. The modifications applied to the search algorithm provide great influence on performance, speed of convergence and running time.

- *A novel QoS monitor selection mechanism in cellular data networks*. A client-side performance monitoring scheme was presented that relies on measurements collected by a carefully chosen subset of mobile service subscribers. The scheme involved three phases. The first phase identifies stable nodes within the environment based on residual and transmission power capabilities, in addition to a service request score. The second

phase identifies node similarity in terms of relative-velocity and semantic nearness. The third phase identifies intermediate packet forwarding nodes to decrease the transmission distance between the monitors and the BS. Due to the reduced number of reporting monitors, the proposed scheme efficiently reduces the overall network traffic overhead caused by QoS performance feedback generated by all the service subscribers in traditional client-side approaches. The proposed work also minimizes the average per-user transmission power needed to report these measurements and, hence, prolongs the lifetime of the mobile subscribers' device batteries within the environment.

6.2 LIMITATIONS AND FUTURE RESEARCH WORK

There are still many possible directions that future work in this area may take. The main focus of possible future research work can be divided into five directions as follows: scalable and efficient service provider-side network simulation, profit management in cases of SP joining and leaving the cooperation, QoS monitor management in cases of node joining and leaving the environment, incorporating the VHO component of the architecture within other systems, and system deployment for in-network testing.

6.2.1. Scalable and Efficient SP-side Network Simulation

The focus of the conducted research has been the development of a policy-based handoff and SLA management system in heterogeneous wireless networks. The research has adhered to the importance of meeting user quality requirements while increasing SP profit through cooperation. In this work a solution was presented that relied on a network simulator to evaluate the effects of applying different handover and SLA policies on the user-perceived service quality and SP profit. However, the main limitation of the current adapted simulator is that it should know the information for the entire network. This may be difficult to apply in a real system because knowing the information of the whole network can be challenging.

To overcome this limitation, the simulator must be designed for parallel network simulations such that the network is partitioned into domains and the simulation time into intervals. Each domain should be simulated independent of and concurrent with other domains over the same simulation time interval. At the end of each interval, only statistical traffic data, including packet drop rates, average packet delays and similar data are

exchanged between domain simulators. The large memory size required by the simulation software can be eliminated with the support of distributed simulators. Each participating simulator will only possess data related to the part of the network it simulates. This approach allows support of parallel simulation of large-scale networks with infrequent synchronization and achieves significant simulation speedups. The plan is to extend this research work to investigate the feasibility of applying parallel simulation techniques without affecting the current network management system model.

6.2.2. Profit Management in Cases of SP Joining and Leaving the Cooperation

It is assumed throughout this dissertation that all the SPs cooperate to increase their profit. However, it is also possible that some of the SPs do not cooperate for purposes such as monopoly. Another issue that will be looked at is the situation of determining how to continue providing a seamless service in cases of a new SP joining the cooperation or one of the providers leaving the cooperation. To overcome these limitations it is necessary to look at another level of management, namely, the service provider cooperation level management. This level of management would consider enforcement of policies for cooperation agreements between SPs. Such policies, for example, would provide a relief time duration to handoff all service subscribers in an efficient and timely manner without loss of service quality, before leaving cooperation or temporarily shutting down the service.

6.2.3. Management of QoS Monitors in Cases of a Node Joining and Leaving the Environment

The high level of dynamicity associated with mobile environments requires proactive adaptations to dynamically replace QoS monitors in cases of node failure. The current approach for network monitors selection attempts to adapt monitors based on the current node availability and predicted node duration and movement speed and direction within the environment. It is highly probable that some nodes might leave the environment or do not stay within the predicted movement direction. Thus, it is important to consider such cases in the monitor selection scheme. The plan is to extend this research work to investigate the feasibility of applying proactive learning during service runtime. This will provide a better

monitor selection solution that is considered more stable. In addition, the plan is to extend the mechanism to include the selection of 'QoS monitor bystanders'. These 'by-standing' nodes are responsible for measuring and sending service quality measurements to the BS in case of a monitor failure.

6.2.4. Incorporating the Simulator-Assisted VHO Scheme within Service Specific Overlay Networks

The work presented in this dissertation is part of an ongoing project [230] [231] developed in the 'Intelligence for Mobile Autonomic and Cognitive Networks Laboratory'. The project focuses on the dynamic construction of Service Specific Overlay Networks (SSONs) for the underlying physical network to deliver customized media to end-users. The developed service composition architecture considers node mobility. The presented management architecture will be incorporated within the developed SSON architecture to solve the issue of node migration between neighbouring cells. An initial possible solution was provided in [191]. The work discussed some solutions pertaining to predictive SSON establishment in neighboring networks prior to any vertical handover of the media client's connection. However, further research should be conducted regarding the costs, both financial and in terms of delay, associated with such an approach.

6.2.5. System Deployment for In-Network Testing

The performance of the developed system was tested using simulation tools. Results showed the effectiveness of the proposed scheme in terms of QoS improvement, access point load balancing, service provider profit increase, reduction in the overall network traffic and overall node power consumption. Despite the accuracy of such simulation tools, it is more accurate and industry-acceptable to test developed systems and schemes on real networks. Thus, based on this research, the plan is to implement and adapt the the developed management system on a small scale real network. This will give further support and evidence towards the system's robustness.

6.3 CONCLUDING REMARKS

Network management in heterogeneous wireless networks is a complex problem

comprising a large number of challenging issues. The next generation wireless mobile communications will be based on the cooperation of the providers of the heterogeneous underlying infrastructure integrating different wireless access technologies and service providers. In addition, service subscribers are no longer passive; they are involved in maintaining and managing network and service provider resources. In addition, future mobile users need to enjoy seamless mobility and ubiquitous access to services. This dissertation presented comprehensive and effective improvements on different facets of network management. In addition, this research will facilitate the evolution of wireless system management for the next generation network and provide enhanced satisfaction for both service providers and subscribers.

LIST OF PUBLICATIONS

Journal Publications

- I. Al Ridhawi, N. Samaan, and A. Karmouch, "A Tabu-Search-Enhanced Online-Simulator for Adaptive Wireless Networks," [to be submitted] to IEEE Transactions on Dependable and Secure Computing, 2013.
- I. Al Ridhawi, N. Samaan, and A. Karmouch, "A QoS Monitor Selection Mechanism for Cellular Data Networks," [under review] to IEEE Transactions on Wireless Communications, 2013.

Conference Publications

- I. Al Ridhawi, N. Samaan, and A. Karmouch, "Simulator-Assisted Joint Service-Level-Agreement and Vertical-Handover Adaptation for Profit Maximization," in Proc. 12th IEEE/IPSJ International Symposium on Applications and the Internet, pp.74,82, 16-20 July 2012.
- Y. Al Ridhawi, I. Al Ridhawi, N. Samaan, and A. Karmouch, "An Adaptive Vertical Handover in Service Specific Overlay Networks," in Proc. 3rd Workshop on Enablers for Ubiquitous Computing and Smart Services, 16-20 July 2012.
- I. Al Ridhawi, N. Samaan, and A. Karmouch, "A Policy-Based Simulator for Assisted Adaptive Vertical Handover," in Proc. IEEE International Symposium on Policies for Distributed Systems and Networks, pp.41-48, 6-8 June 2011.
- Y. Al Ridhawi, I. Al Ridhawi, and A. Karmouch, "A Context-aware and Location Prediction Framework for Dynamic Environments," in Proc. 7th IEEE International Conference on Wireless and Mobile Computing, 10-12 October 2011.
- Y. Al Ridhawi, I. Al Ridhawi and A. Karmouch, "Policy-Based Personalized Context dissemination for Location-Aware Services," in Proc. 7th International ICST Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services, 6-9 December 2010.

- I. Al Ridhawi, M. Aloqaily, A. Karmouch, and N. Agoulmine, "A location-aware user tracking and prediction system," in Proc. IEEE conference on Global Information Infrastructure Symposium, pp.1-8, 23-26 June 2009.

BIBLIOGRAPHY

[1] C. Singh, S. Sarkar, A. Aram, and A. Kumar, "Cooperative Profit Sharing in Coalition-Based Resource Allocation in Wireless Networks," in IEEE/ACM Transactions on Networking, vol.20, no.1, pp.69-83, Feb. 2012.

[2] S. Fernandes and A. Karmouch, "Vertical Mobility Management Architectures in Wireless Networks: A Comprehensive Survey and Future Directions," in IEEE Communications Surveys & Tutorials, vol.14, no.1, pp.45-63, First Quarter 2012.

[3] P.A. Gutierrez, I. Miloucheva, D. Wagner, C. Niephaus, A. Flizikowski, N.V. Wambeke, F. Armando, C. Chassot, and S.P. Romano, "NETQOS Policy Management Architecture for Flexible QoS Provisioning in Future Internet," in Proc. The Second International Conference on Next Generation Mobile Applications, Services and Technologies, 2008, pp.53-58, 16-19 Sept. 2008.

[4] C. Fan, A.M. Schlager, A. Udugama, and V. Pangboonyanon, "Managing Heterogeneous Access Networks Coordinated policy based decision engines for mobility management," in Proc. 32nd IEEE Conference on Local Computer Networks, 2007, pp.651-660, 15-18 Oct. 2007.

[5] H. Lu and I. Faynberg, "An Architectural Framework for Support of Quality of Service in Packet Networks," in IEEE Communications Magazine, vol. 41, pp. 98–105, Jun. 2003.

[6] R. Braden, D. Clark, and S. Shenker, "Integrated Services in the Internet Architecture: an Overview," Network Working Group, Jun. 1994.

[7] S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang, and W. Weiss, "An Architecture for Differentiated Services," Network Working Group, Dec. 1998.

[8] A.F. Nadeau and C. Srinivasan, "Multiprotocol Label Switching (MPLS) Management Overview," Network Working Group, Sept. 2003.

[9] M. Lee, Y. Xiaohui, D. Marconett, S. Johnson, R. Vemuri, and S. Yoo, "Autonomous Network Management Using Cooperative Learning for Network-Wide Load Balancing in Heterogeneous Networks," in Proc. IEEE GLOBECOM 2008, Global Telecommunications Conference, 2008, pp.1-5, Nov. 30 2008-Dec. 4 2008.

[10] E. Exposito and J. Montalvo, "An Ontology-Based Framework for Autonomous QoS

Management in Home Networks," in Proc. 6th International Conference on Networking and Services (ICNS), pp.117-122, 7-13 March 2010.

[11] S. Chung, C. Srisathapornphat, and C. Jaikaeo, "Adaptive autonomous management of ad hoc networks," in Proc. IEEE/IFTP on Network Operations and Management Symposium, pp. 891- 893, 2002.

[12] N. Samaan and A. Karmouch, "An automated policy-based management framework for differentiated communication systems," in IEEE Journal on Selected Areas in Communications, vol.23, no.12, pp. 2236-2247, Dec. 2005.

[13] F. Khan and E. Huh, "An adaptive resource management for mobile terminals on vertical handoff," in Springer annals of telecommunications, vol.63, no.7-8, pp.435-447, Aug. 2008.

[14] H. Honglin, J. Zhang, X. Zheng, Y. Yang, and W. Ping, "Self-configuration and self-optimization for LTE networks," in IEEE Communications Magazine, vol.48, no.2, pp.94-100, February 2010.

[15] C. Frenzel, H. Sanneck, and B. Bauer, "Automated rational recovery selection for self-healing in mobile networks," in Proc. International Symposium on Wireless Communication Systems, pp.41-45, 28-31 Aug. 2012.

[16] J.M. Chang, W.T. Hsiao, J.L. Chen, and H.C. Chao, "Mobile Relay Stations Navigation-Based Self-Optimization Handover Mechanism in WiMAX Networks," in Proc. 4th International Conference on Ubiquitous Information Technologies & Applications, pp.1-5, 20-22 Dec. 2009.

[17] S. Hariri, Q. Guangzhi, R. Modukuri, C. Huoping, and M. Yousif, "Quality-of-protection (QoP)-an online monitoring and self-protection mechanism," in IEEE Journal on Selected Areas in Communications, vol.23, no.10, pp.1983- 1993, Oct. 2005.

[18] R.S. Sutton, A.G. Barto, "Reinforcement Learning: An Introduction," Cambridge, MA: MIT Press, 1998.

[19] W. Zhuang, Y.S. Gan, K.J. Loh, and K.C. Chua, "Policy-based QoS-management architecture in an integrated UMTS and WLAN environment," in IEEE Communications Magazine, vol.41, no.11, pp.118- 125, Nov. 2003.

[20] L. Huang, S. Kumar, and C.C Kuo, "Adaptive resource allocation for multimedia

QoS management in wireless networks," in IEEE Transactions on Vehicular Technology, vol.53, no.2, pp. 547- 558, March 2004.

[21] X. Li, H.M. Chan, T. Hung, K.H. Tong, and S.J. Turner, "Design of an SLA-Driven QoS Management Platform for Provisioning Multimedia Personalized Services," in 22nd International Conference on Advanced Information Networking and Applications, pp.1405-1409, 25-28 March 2008.

[22] G. Wang, A. Chen, C. Wang, C. Fung, and S. Uczekaj, "Integrated quality of service (QoS) management in service-oriented enterprise architectures," in Proc. 8th IEEE International Enterprise Distributed Object Computing Conference, pp. 21- 32, 20-24 Sept. 2004.

[23] Q. Liang, H.C. Lau, and X. Wu, "Robust Application-Level QoS Management in Service-Oriented Systems," in Proc. IEEE International Conference on e-Business Engineering, pp.239-246, 22-24 Oct. 2008.

[24] R. Zoubairi, Z. Jarir, and M. Erradi, "Dynamic QoS management in mobile services framework," in Proc. 2011 International Conference on Multimedia Computing and Systems, pp.1-6, 7-9 April 2011.

[25] Y. Bai, Y. Chen, C. Li, S. Liu, and D. Qian, "An Architecture of Policy-Based Application-aware Network QoS Management for Large-scale Heterogeneous Networks," in Proc, Conference on Future Generation Communication and Networking, vol.1, pp.22-26, 6-8 Dec. 2007.

[26] S.K. Ray, K. Pawlikowski, and H. Sirisena, "Handover in Mobile WiMAX Networks: The State of Art and Research Issues," in IEEE Communications Surveys & Tutorials, vol.12, no.3, pp.376-399, Third Quarter 2010.

[27] S. Mohanty, "A new Architecture for 3G and WLAN integration and Inter-System Handover Management," in Springer Wireless Networks, vol. 12, No. 6, pp.733-745, April 2006.

[28] P. Neves, J. Soares, S. Sargento, H. Pires, and F. Fontes, "Context-aware media independent information server for optimized seamless handover procedures," in Elsevier Computer Networks, vol.55, no.7, pp.1498–1519, May, 2011.

[29] SK. Lee, K. Sriram, K. Kyungsoo, H.K. Yoon, and N. Golmie, "Vertical Handoff Decision Algorithms for Providing Optimized Performance in Heterogeneous

Wireless Networks," in IEEE Transactions on Vehicular Technology, vol.58, no.2, pp.865-881, Feb. 2009.

[30] A. Ahmed, L.M. Boulahia, and D. Gaïti, "Cooperative Agent Based Vertical Handover Scheme for Heterogeneous Networks," in Proc. 6th Advanced International Conference on Telecommunications, pp.410-415, 9-15 May 2010.

[31] J.M. Kang, J. Strassner, S. Seo, and J.W. Hong, "Autonomic personalized handover decisions for mobile services in heterogeneous wireless networks," in Elsevier Computer Networks, vol. 55, no. 7, pp. 1520-1532, May 2011.

[32] W. Song, W.J. Chung. D. Lee, C. Lim, S. Choi, and T. Yeoum, "Improvements to seamless vertical handover between mobile WiMAX and 3GPP UTRAN through the evolved packet core," in IEEE Communications Magazine, vol.47, no.4, pp.66-73, April 2009.

[33] H.H. Choi, O. Song, Y.K. Park, and J.R. Lee, "Performance Evaluation of Opportunistic Vertical Handover Considering On–Off Characteristics of VoIP Traffic," in IEEE Transactions on Vehicular Technology, vol.59, no.6, pp.3115-3121, July 2010.

[34] E.S. Navarro and V.W. Wong, "Comparison between Vertical Handoff Decision Algorithms for Heterogeneous Wireless Networks," in Proc. 63rd IEEE Vehicular Technology Conference, vol.2, pp.947-951, 7-10 May 2006.

[35] R.A.V. Ramirez and R. Ramos, "A vertical handoff decision algorithm which considers the uncertainty during the decision making process," in Proc. IFIP International Conference on Wireless and Optical Communications Networks, pp.1-6, 28-30 April 2009.

[36] I. Lassoued, J.M. Bonnin, Z. Hamouda, and A. Belghith, "A Methodology for Evaluating Vertical Handoff Decision Mechanisms," in Proc. 7th International Conference on Networking, 2008, pp.377-384, 13-18 April 2008.

[37] S. Balasubramaniam, and J. Indulska, “Vertical handover supporting pervasive computing in future wireless networks,” in Elsevier Computer Communications,” vol. 27, no.8, pp. 708–719, 2004.

[38] J. Hou and D.C. O’Brien, “Vertical handover-decision-making algorithm using fuzzy logic for the integrated Radio-and-OW system”, in IEEE Transactions on Wireless

Communications, vol. 5, no. 1, pp. 176 – 185, Jan. 2006.

[39] K. Herrmann, G. Muhl, and K. Geihs, "Self-management: the solution to complexity or just another problem?," in IEEE Distributed Systems Online, vol.6, no.1, Jan. 2005.

[40] H. Ludwig, A. Keller, A. Dan, R. King, and R. Franck, "A Service Level Agreement Language for Dynamic Electronic Services," in Proc. 4th International Workshop on Electronic Commerce Research, vol. 3, pp. 43–59, 2003.

[41] ISO-QoS, Quality of Service Basic Framework, ISO/IEC JTC1/SC21/WG1 N1145, International Standards Organisation, UK, 1994.

[42] Y. R., R. Guerin, and D. Pendarakis, "A Framework for Policy-based Admission Control," in IETF RFC 2753, Informational, Jan. 2000.

[43] I. Aib and R. Boutaba, "On Leveraging Policy-Based Management for Maximizing Business Profit," in IEEE Transactions on Network and Service Management, vol.4, no.3, pp.25-39, Dec. 2007.

[44] C. Blum and A. Roli, "Metaheuristics in combinatorial optimization: Overview and conceptual," in ACM Computing Surveys, vol. 35, no. 3, pp. 268-308, 2003.

[45] W. Kasch, J. Ward, and J. Andrusenko, "Wireless network modeling and simulation tools for designers and developers," in IEEE Communications Magazine, vol.47, no.3, pp.120-127, March 2009.

[46] F. Glover, E. Taillard, and D. Werra, "A user's guide to tabu search," in Springer Journal of Annals of Operations Research, vol.41, no.1, pp.1-28, March 1993.

[47] J.Y. Yu and P.H. Chong, "A survey of clustering schemes for mobile ad hoc networks," in IEEE Communications Surveys & Tutorials, vol.7, no.1, pp.32-48, First Qtr. 2005.

[48] M. Krebs, A. Stein, and M.A. Lora, "Topology Stability-Based Clustering for Wireless Mesh Networks," in Proc. IEEE Conference of Global Telecommunications Conference, pp.1-5, 6-10 Dec. 2010.

[49] C. Zhikui, Y. Song, L. Liang, and X. Zhijiang, "A clustering approximation mechanism based on data spatial correlation in wireless sensor networks," in Proc. IEEE Conference of Wireless Telecommunications Symposium (WTS), pp.1-7, 21-23 April 2010.

[50] T. Wang and Z. Yang, "A Location-Aware-Based Data Clustering algorithm in Wireless Sensor Networks," in Proc. 11th IEEE Singapore International Conference on Communication Systems, 2008, pp.1-5, 19-21 Nov. 2008.

[51] S.M. Jung, Y.J. Han, and T.M Chung, "The Concentric Clustering Scheme for Efficient Energy Consumption in the PEGASIS," in Proc. 9th International Conference on Advanced Communication Technology, vol., no.1, pp.260-265, 12-14 Feb. 2007.

[52] G. Koloniari and E. Pitoura, "A Game-Theoretic Approach to the Formation of Clustered Overlay Networks," in IEEE Transactions on Parallel and Distributed Systems, vol.23, no.4, pp.589-597, April 2012.

[53] C. Smith and D. Collins, "First Generation (1G)," in 3G Wireless netowrks, 1st ed. New York, 2002, Ch. 1.

[54] A. Damnjanovic, J. Montojo, W. Yongbin, J. Tingfang, T. Luo, M. Vajapeyam,T. Yoo, O. Song, and D. Malladi, "A survey on 3GPP heterogeneous networks," in IEEE Wireless Communications, vol.18, no.3, pp.10-21, June 2011.

[55] Bluetooth CIG, "Specification of the Bluetooth system, Version 1.1," February 22 2001, available from www.bluetooth.com.

[56] Intel, "Ultra-Wideband (UWB Technology): Enabling high-speed wireless personal area networks," 2005, available from http://www.intel.com/technology/comms/uwb/download/ultra-wideband.pdf.

[57] I. Al Ridhawi, M. Aloqaily, A. Karmouch, and N. Agoulmine, "A location-aware user tracking and prediction system," in Proc. IEEE Conference on Global Information Infrastructure Symposium, pp.1-8, 23-26 June 2009.

[58] Y. Al Ridhawi, I. Al Ridhawi, and A. Karmouch, "A context-aware and location prediction framework for dynamic environments," in Proc. 7th IEEE International Conference on Wireless and Mobile Computing, pp.172-179, 10-12 Oct. 2011.

[59] Y. Al Ridhawi, I. Al Ridhawi, and A. Karmouch, "Policy-Based Personalized Context dissemination for Location-Aware Services," in Proc. 7th International ICST Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services, December 6-9 2010, Sydney, Australia.

[60] P. Bahl and V. Padmanabhan, “RADAR: An In-Building RF-Based User Location and Tracking System,” in IEEE International Conference on Computer Communications, IEEE CS Press, Los Alamitos, Calif., pp. 775-784, 2000.

[61] W. Griswold, P. Shanahan, S. Brown, R. Boyer, M. Ratto, R. Shapiro, and T. Truong, "ActiveCampus: experiments in community-oriented ubiquitous computing," in IEEE Computer, vol.37, no.10, pp. 73-81, Oct. 2004.

[62] M. Ratto, R. Shapiro, T. Truong, and W. Griswold, “The ActiveClass Project: experiments in ecouraging classroom participation,” in International Journal of Computer Support for Collaborative Learning, 2003.

[63] K. Kyamakya and K. Jobmann, "Location management in cellular networks: classification of the most important paradigms, realistic Simulation framework, and relative performance analysis," in IEEE Transactions on Vehicular Technology, vol.54, no.2, pp. 687- 708, March 2005.

[64] N. Samaan and A. Karmouch, "A Mobility Prediction Architecture Based on Contextual Knowledge and Spatial Conceptual Maps," in IEEE Transactions on Mobile Computing, vol. 4, no. 6, pp. 537-551, Nov.-Dec. 2005.

[65] N. Samaan and A. Karmouch, “An evidence-based mobility prediction agent architectures,” in Proc. 5th IEEE/IFIP Workshop on Mobile Agents for Telecommunication Applications, October 8–10, Marrakech, Morocco, 2003.

[66] J. Orwant, “Doppelganger Goes to School: Machine Learning for User Modeling,” M.S. thesis, Dept. of Media Art and Sciences, Massachusetts Inst. of Technology, Sept. 1993.

[67] D. Ashbrook and T. Starner, “Learning Significant Locations and Predicting User Movement with GPS,” in Proc. 6th International Symposium on Wearable Computers, pp. 101-108, Oct. 2002.

[68] S. Pack and Y. Choi, "Fast handoff scheme based on mobility prediction in public wireless LAN systems," in IEE Proceedings-Communications, vol.151, no.5, pp. 489-495, 24 Oct. 2004.

[69] Y. Zhang, J. Hu, J. Dong, Y. Yuan, J. Zhou, and J. Shi, "Location Prediction Model Based on Bayesian Network Theory," in Proc. IEEE Conference on Global Telecommunications, 2009, pp.1-6, Nov. 30 2009-Dec. 4 2009.

[70] Y. Zhang, W. Zhuang, and A. Saleh, "Vertical Handoff between 802.11 and 802.16 Wireless Access Networks," in Proc. IEEE Conference on Global Telecommunications, pp.1-6, Nov. 30 2008-Dec. 4 2008.

[71] J. Kang, D. Kum, Y. Li, and Y. Cho, "Seamless Handover Scheme for Proxy Mobile IPv6," in Proc. IEEE International Conference on Networking and Communications in Wireless and Mobile Computing, pp.410-414, 12-14 Oct. 2008.

[72] A.R. Prasad, A. Zugenmaier, and P. Schoo, "Next generation communications and secure seamless handover," in Proc. 1st International Conference on Security and Privacy for Emerging Areas in Communication Networks, pp. 267- 274, 5-9 Sept. 2005.

[73] P. Vidales, J. Baliosian, J. Serrat, G. Mapp, and F. Stajano, A. Hopper, "Autonomic system for mobility support in 4G networks," in IEEE Journal on Selected Areas in Communications, vol.23, no.12, pp. 2288- 2304, Dec. 2005.

[74] M. Rawashdeh and A. Karmouch, "Seamless video handoff in session mobility over the IMS network," in Proc. IEEE International Symposium on a World of Wireless, Mobile and Multimedia Networks & Workshops, pp.1-6, 15-19 June 2009.

[75] L. Zhaojun, "An autonomic hierarchical mobility management framework for 3GPP heterogeneous networks," in Proc. Future Network and Mobile Summit, pp.1-8, 16-18 June 2010.

[76] I. Al Ridhawi, N. Samaan, and A. Karmouch, "A Policy-Based Simulator for Assisted Adaptive Vertical Handover," in Proc. IEEE International Symposium on Policies for Distributed Systems and Networks, pp.41-48, 6-8 June 2011.

[77] S. Dhar, A. Ray, and R. Bera, "Design and Simulation of Vertical Handover Algorithm for Vehicular Communication", in Proc. International Journal of Engineering Science and Technology, vol 2,no 10, pp. 5509-5525, October,2010.

[78] Q. Song and A. Jamalipour, "A network selection mechanism for next generation networks," in Proc. IEEE International Conference on Communications, vol.2, pp. 1418- 1422, pp. 16-20 May 2005.

[79] M. Rehan, M. Yousaf, A. Qayyum, and S. Malik, “A Cross-Layer user centric vertical handover decision approach based on MIH local triggers,” in IFIP Advances

in Information and Communication Technology: Wireless and Mobile Networking, vol. 308, ch. 32, pages 359–369, 2009.

[80] H.Y. Huang, C.Y. Wang, and R.H. Hwang, “Context-Awareness Handoff Planning in Heterogeneous Wireless Networks,” Lecture Notes in Computer Science, Vol. 6406 pp. 430-444, 2010.

[81] M. Kassar, B. Kervella, and G. Pujolle, "An overview of vertical handover decision strategies in heterogeneous wireless networks," in Elsevier Computer Communications, vol. 31, pp. 2607-2620, June 2008.

[82] M. Atiquzzaman and A.S. Reaz, "Survey and classification of transport layer mobility management schemes," in Proc. 16th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications, vol.4, pp.2109-2115, 11-14 Sept. 2005.

[83] A. Taha, H.S. Hassanein, and H.T. Mouftah, "Extensions for Internet QoS paradigms to mobile IP: a survey," in IEEE Communications Magazine, vol.43, no.5, pp. 132-139, May 2005.

[84] C. Perkins, “IP mobility support for IPv4, IETF RFC 334,” 2002.

[85] A. Tanenbaum, “The Network layer,” in Computer Netowrks, 4th ed., 2003, ch. 5, sec. 5.2.9, pp. 283-284.

[86] D. Johnson, C. Perkins, and J. Arkko, “Mobility support in IPv6, IETF RFC 3775,” 2004.

[87] N. Montavont and T. Noel, “Handover management for mobile nodes in IPv6 networks,” in IEEE Communications Magazine, vol. 40, no. 8, pp. 38–43, 2002.

[88] S. Dixit, “Wireless IP and Its Challenges for the Heterogeneous Environment,” in International Journal of Wireless Personal Communications, vol. 22, no. 2, pp. 261–273, 2002.

[89] R. Moskowitz and P. Nikander, “Host Identity Protocol (HIP) Architecture, IETF RFC 4423,” May 2006.

[90] P. Eronen, “IKEv2 Mobility and Multihoming Protocol (MOBIKE), IETF RFC 4555,” June 2006.

[91] T. Kivinen and H. Tschofenig, “Design of the IKEv2 Mobility and Multihoming (MOBIKE) Protocol, IETF RFC 4621,” August 2006.

[92] E. Fogelstroem, A. Jonsson, C. Perkins, “Mobile IPv4 Regional Registration, IETF RFC 4857,” 2007.

[93] H. Soliman, C. Castelluccia, K. E. Malki, and L. Bellier, “Hierarchical mobile IPv6 mobility management (HMIPv6), IETF RFC 4140,” 2005.

[94] R. Koodli, “Fast handovers for Mobile IPv6, IETF RFC 0468,” 2005.

[95] A. Misra, S. Das, A. Dutta, A. McAuley, and S.K. Das, “IDMP-based fast handoffs and paging in IP-based 4G mobile networks,” in IEEE Communication Magazine, vol. 40, pp. 138-145, 2002.

[96] R. Ramjee, K. Varadhan, L. Salgarelli, S.R. Thuel, S.Y. Wang, and T. Porta, "HAWAII: a domain-based approach for supporting mobility in wide-area wireless networks," in IEEE/ACM Transactions on Networking, vol.10, no.3, pp.396-410, Jun 2002.

[97] S. Gundavelli, K. Leung, V. Devarapalli, K. Chowdhury, and B. Patil, “Proxy Mobile IPv6, Internet Draft, draft-ietf-netlmm-proxymip6-10, NETLMMWG,” Feb. 2008.

[98] IETF, “SIP: Session Initiation Protocol,” Tech. Rep., June 2002.

[99] R. Hsieh, Z. Zhou, and A. Seneviratne, “S-MIP: a seamless handoff architecture for mobile IP,” in Proc. 22nd Annual Joint Conference of the IEEE Computer and Communications Societies, vol. 3, 2003.

[100] C. Politis, K. Chew, and R. Tafazolli, “Multilayer mobility management for all-IP networks: pure SIP vs. hybrid SIP/mobile IP,” in Proc. 57th IEEE Semi-annual Vehicular Technology Conference, vol. 4, 2003.

[101] Q. Wang, M. Abu-Rgheff, and A. Akram, “Design and evaluation of an integrated mobile IP and SIP framework for advanced handoff management,” in Proc. IEEE International Conference on Communications, vol. 7, 2004.

[102] Q. Wang and M. Abu-Rgheff, “A multi-layer mobility management architecture using crosslayer signalling interactions,” in Proc. 5th European Personal Mobile Communications Conference, pp. 237–241, 2003.

[103] G.N. Stone, B. Lundy, and G.G. Xie, “Network policy languages: A survey and a new approach,” in IEEE Network, vol. 15, no. 1, pp. 10–21, Feb. 2001.

[104] H. Derbel, N. Agoulmine, and M. Salaün, "ANEMA: Autonomic network management architecture to support self-configuration and self-optimization in IP networks," in Elsevier Computer Networks, vol. 53, no. 3, pp. 418-430, 2009.

[105] K. Murray and D. Pesch, "Call Admission and Handover in Heterogeneous Wireless Networks," in IEEE Internet Computing, vol.11, no.2, pp.44-52, March-April 2007.

[106] P. Vidales, R. Chakravorty, and C. Policroniades, "PROTON: a policy-based solution for future 4G devices," in Proc. 5th IEEE International Workshop on Policies for Distributed Systems and Networks, pp. 219- 222, 7-9 June 2004.

[107] G. Nyberg, C. Ahlund, and T. Rojmyr, "SEMO: A Policy-based system for Handovers in Heterogeneous Networks," in Proc. International Conference on Wireless and Mobile Communications, pp.62, 29-31 July 2006.

[108] D.D. Clark, "Policy routing in internet protocols. IETF Network Working Group, RFC 1102," May 1989.

[109] M. Nossik, F. Welfeld, and M. Richardson, "PAX PDL: A Non-Procedural Packet Description Language," IETF Network Working Group, Tech. Rep., Sept. 30, 1998.

[110] N. Brownlee, "SRL: A language for describing traffic flows and specifying actions for flow groups," IETF Network Working Group, Tech. Rep. RFC 2723, Aug 1999.

[111] J. Lobo, R. Bhatia, and S. Naqvi, "A policy description language," in Proc. 6th National conference on Artificial intelligence and the eleventh Innovative Applications of Artificial Intelligence conference, pp. 291–298, Menlo Park, 1999.

[112] R Wies, "Policies in network and systems management-Formal definition and architecture," in Journal of Network System Management vol. 2, no. 1, pp. 63–83, 1994.

[113] M.Z. Hasan, "An active temporal model for network management databases," in Proc. 4th International Symposium on Integrated network management, pp. 524–535, London, 1995.

[114] S. Jajodia, P. Samarati, and V.S. Subrahmanian, "A logical language for expressing authorisations" in Proc. IEEE Symposium on Security and Privacy, 1997.

[115] R. Ortalo, "A flexible method for information system security policy specification," in Proc. 5th European Symposium on Research in Computer Security, pp. 67–84, London, 1998.

[116] H. James, R. Pandey, and K. Levitt, "Security policy specification using a graphical approach," Technical Report CSE-98-3, University of California, Davis Department of Computer Science, 1998.

[117] T. Koch, C. Krell, and B. Kramer, "Policy definition language for automated management of distributed systems," in Proc. IEEE Second International Workshop on Systems Management, pp. 55–64, June 19–21, 1996.

[118] S. Godik and T. Moses, "eXtensible Access Control Markup Language (XACML) version 1.0," OASIS, XACML Technical Committee, 18 February 2003.

[119] K. Twidle, N. Dulay, E. Lupu, and M. Sloman, "Ponder2: A Policy System for Autonomous Pervasive Environments," in Proc. 5th International Conference on Autonomic and Autonomous Systems, pp. 330–335, April 2009.

[120] G. Waters, J. Wheeler, A. Westerinen, L. Rafalow, and R. Moore, "Policy framework architecture," IETF Network Working Group, Feb 1999.

[121] "CIM Policy Model," DMTF White paper, February 2001.

[122] V. Koutsonikola and A. Vakali, "LDAP: framework, practices, and trends," in IEEE Internet Computing, vol.8, no.5, pp. 66- 72, Sept.-Oct. 2004.

[123] C. Basile, A. Cappadonia, and A. Lioy, "Network-Level Access Control Policy Analysis and Transformation," in IEEE/ACM Transactions on Networking, vol.20, no.4, pp.985-998, Aug. 2012.

[124] O. Dohndorf, J. Kruger, H. Krumm, C. Fiehe, A. Litvina, I. Luck, and F. Stewing, "Tool-Supported Refinement of High-Level Requirements and Constraints Into Low-Level Policies," in Proc. IEEE International Symposium on Policies for Distributed Systems and Networks, pp.97-104, 6-8 June 2011.

[125] J. Rubio-Loyola, J. Serrat, M. Charalambides, P. Flegkas, G. Pavlou, and A. Lluch-Lafuente, "Using linear temporal model checking for goal-oriented policy refinement frameworks," in Proc. IEEE International Workshop on Policies for Distributed Systems and Networks, pp. 181–190, 2005.

[126] J. Rubio-Loyola, J. Serrat, M. Charalambides, P. Flegkas, and G. Pavlou, "A functional solution for goal-oriented policy refinement," in Proc. 7th IEEE International Workshop on Policies for Distributed Systems and Networks, pp. 133–144, 2006.

[127] T. Braun, M. Diaz, J. Gabeiras, and T. Staub, “Motivations and Basics,” in End-to-End Quality of Service Over Heterogeneous Networks, 2008, ch. 1, pp. 1.

[128] Y. Al Ridhawi, I. Abdeljaouad, G. Kandavanam, and A. Karmouch, "An architecture for autonomic management of overlay networks," in Proc. IEEE International Global Communication Conference, pp.77-82, 5-9 Dec. 2011.

[129] R. Goyette and A. Karmouch, "Using AHP/TOPSIS with cost and robustness criteria for virtual network node assignment," in Proc. IEEE International Conference on Communications, pp.5885-5889, 10-15 June 2012.

[130] D. Niyato and E. Hossain, "A Cooperative Game Framework for Bandwidth Allocation in 4G Heterogeneous Wireless Networks," in Proc. IEEE International Conference on Communications, vol.9, pp.4357-4362, June 2006.

[131] S. SungHoon, L. SeungChan, and S. JooSeok, "Policy Based Intelligent Vertical Handover Algorithm in Heterogeneous Wireless Networks," in Proc. International Conference on Convergence Information Technology, pp.1900-1905, 21-23 Nov. 2007.

[132] F. Hou, P. Ho, X. Shen, and A. Chen, "A Novel QoS Scheduling Scheme in IEEE 802.16 Networks," in Proc. IEEE Wireless Communications and Networking Conference, pp.2457-2462, 11-15 March 2007.

[133] H. Qiu, Y. Li, J. Wu, and X. Gu, "Compensation Buffer Sizing for Providing User-Level QoS Guarantee of Media Flows," in Proc. IEEE International Conference on Communications, pp.90-94, 19-23 May 2008.

[134] A. Kind, X. Dimitropoulos, S. Denazis, and B. Claise, "Advanced network monitoring brings life to the awareness plane," in IEEE Communications Magazine, vol.46, no.10, pp.140-146, October 2008.

[135] R. Mehrotra, A. Dubey, S. Abdelwahed, and W. Monceaux, "Large Scale Monitoring and Online Analysis in a Distributed Virtualized Environment,“ in Proc. 8th IEEE International Conference and Workshops on Engineering of Autonomic and Autonomous Systems, pp.1-9, 27-29 April 2011.

[136] A. Kamel and A. Al-Fuqaha, "Client-side architecture for mobile service QoS monitoring using Generalized Extreme Value theorem," in Proc. IEEE International Global Communication Conference, pp.690-694, 5-9 Dec. 2011.

[137] G. Gardikis, L. Boula, G. Xilouris, A. Kourtis, E. Pallis, M. Sidibe, and D. Negru, "Cross-layer monitoring in IPTV networks," in IEEE Communications Magazine, vol.50, no.7, pp.76-84, July 2012.

[138] A. Bulut, N. Koudas, A. Meka, and A. Singh, D. Srivastava, "Optimization Techniques for Reactive Network Monitoring," in IEEE Transactions on Knowledge and Data Engineering, vol.21, no.9, pp.1343-1357, Sept. 2009.

[139] M. Matthew, B. Chun, and E. David, "The ganglia distributed monitoring system: Design, implementation, and experience," in Elsevier Parallel Computing, vol. 30, no. 7, pp.817-840, 2004.

[140] L. Huang and M.J. Neely, "The Optimality of Two Prices: Maximizing Revenue in a Stochastic Communication System," in IEEE/ACM Transactions on Networking, vol.18, no.2, pp.406-419, April 2010.

[141] H. Kameda and E. Altman, "Inefficient Noncooperation in Networking Games of Common-Pool Resources," in IEEE Journal on Selected Areas in Communications, vol.26, no.7, pp.1260-1268, September 2008.

[142] W. Chai, K. Ho, M. Charalambides, and G. Pavlou, "A Policy-Driven Network Management System for the Dynamic Configuration of Military Networks", in Lecture Notes in Computer Science: Scalability of Networks and Services, 2009.

[143] Q. Jia "Efficient Computing Budget Allocation for Simulation-Based Policy Improvement," in IEEE Transactions on Automation Science and Engineering, vol.9, no.2, pp.342-352, April 2012.

[144] W.B. Powell, "Approximate Dynamic Programming: Solving the Curse of Dimensionality," in Wiley-Interscience, 2nd ed., New York, 2007.

[145] X.R. Cao, "Stochastic Learning and Optimization: A Sensitivity-Based Approach. In Springer, New York, 2007.

[146] R. Ribeiro, "Fuzzy multiple attribute decision making: A review and new preference elicitation techniques," in Journal of Fuzzy Sets and Systems, vol. 78, n. 2, pp. 155 – 181, 1996.

[147] W. Zhang, "Handover decision using fuzzy MADM in heterogeneous networks", in Proc. IEEE Wireless Communications and Networking Conference, vol. 2, pp. 653– 658, 2004.

[148] H.J. Wang, R.H. Katz, and J. Giese, “Policy-enabled handoffs across heterogeneous wireless networks”, in Proc. IEEE Workshop on Mobile Computing Systems and Applications, pp. 51–60, 1999.

[149] X. Yan, Y. Ahmet, and S. Narayanan, "A survey of vertical handover decision algorithms in Fourth Generation heterogeneous wireless networks," in Elsevier Computer Networks, vol. 54, no. 11, pp. 1848-1863, Aug. 2010.

[150] X. Fafoutis and V. Siris, "Handover Incentives for Self-Interested WLANs with Overlapping Coverage," in IEEE Transactions on Mobile Computing, 2011.

[151] W. Lee and D. Cho, "Enhanced Group Handover Scheme in Multiaccess Networks," in IEEE Transactions on Vehicular Technology, vol.60, no.5, pp.2389-2395, Jun 2011.

[152] E. Navarro, Y. Lin, and V.W. Wong, "An MDP-Based Vertical Handoff Decision Algorithm for Heterogeneous Wireless Networks," in IEEE Transactions on Vehicular Technology, vol.57, no.2, pp.1243-1254, March 2008.

[153] M. Liu, Z. Li, X. Guo, and E. Dutkiewicz, "Performance Analysis and Optimization of Handoff Algorithms in Heterogeneous Wireless Networks," in IEEE Transactions on Mobile Computing, vol.7, no.7, pp.846-857, July 2008.

[154] M. Abramson and H. Wechsler, "Tabu search exploration for on-policy reinforcement learning," in Proc. International Joint Conference on Neural Networks, vol.4, pp. 2910- 2915, 20-24 July 2003.

[155] S.P. Singh, T. Jaakkola, M.L. Littman, and C. Szepesvari. "Convergence results for single-step on-policy reinforcement-learning algorithms," in Journal of Machine Learning, vol. 38, no. 3, pp. 287–308, 2000.

[156] J. Pacheco, S. Casado, and L. Núñez, "A variable selection method based on Tabu search for logistic regression models," in European Journal of Operational Research, vol. 199, no. 2, pp. 506-511, Dec. 2009.

[157] Y. Vorobeychik, “Probabilistic analysis of simulation-based games”, in ACM Transactions on Modeling and Computer Simulation, vol. 20, no. 3, pp. 1–25, 2010.

[158] J. April, M. Better, F. Glover, and J. Kelly, “New advances and applications for marrying simulation and optimization”, in Proc. 36th conference on Winter simulation, pp. 80–86, 2004.

[159] Y. Liu, B.K. Szymanski, and A. Saifee, “Genesis: a scalable distributed system for large-scale parallel network simulation”, in Elsevier Computer Networks, vol. 50, no. 12, pp. 2028–2053, 2006.

[160] I. Aib and R. Boutaba, “PS: A Policy Simulator,” in IEEE Communications Magazine, vol. 45, no. 4, pp. 130 –136, April 2007.

[161] G. Gardikis, L. Boula, G. Xilouris, A. Kourtis, E. Pallis, M. Sidibe, and D. Negru, "Cross-layer monitoring in IPTV networks," in IEEE Communications Magazine, vol.50, no.7, pp.76,84, July 2012.

[162] J. Tang and X. Zhang, "Cross-Layer-Model Based Adaptive Resource Allocation for Statistical QoS Guarantees in Mobile Wireless Networks," in IEEE Transactions on Wireless Communications, vol.7, no.6, pp.2318-2328, June 2008.

[163] M. Sidibe and A. Mehaoua, "QoS monitoring framework for end-to-end service management in wired and wireless networks," in IEEE/ACS International Conference on Computer Systems and Applications, pp.964,968, March 2008.

[164] R. Clegg, S. Clayman, G. Pavlou, L. Mamatas, and A. Galis, "On the Selection of Management/Monitoring Nodes in Highly Dynamic Networks," in IEEE Transactions on Computers, vol. 62, no. 6, pp. 1207-1220, June 2013.

[165] J. Laiho, K. Raivio, P. Lehtimaki, K. Hatonen, and O. Simula, "Advanced analysis methods for 3G cellular networks," in IEEE Transactions on Wireless Communications, vol.4, no.3, pp.930,942, May 2005.

[166] A. Kamel and A. Al-Fuqaha, "Client-side architecture for mobile service QoS monitoring using Generalized Extreme Value theorem," in Proc. IEEE GLOBECOM Workshops, pp.690-694, 5-9 Dec. 2011.

[167] A. Kamel, A. Al-Fuqaha, and D. Benhaddou, "Client-based QoS data selection and modeling using generalized extreme value theorem and linear opinion pool," in Proc. IEEE International Conference on Communications, pp.7045,7049, 10-15 June 2012.

[168] R. Jurca, B. Faltings, and W. Binder, “Reliable QoS monitoring based on client feedback.,” in Proc. ACM 16th International Conference on World Wide Web, New York, USA, pp. 1003-1012.

[169] M. Serhani, R. Dssouli, A. Hafid, and H. Sahraoui, "A QoS broker based architecture for efficient Web services selection," in Proc. IEEE International Conference on Web Services, vol. 1, pp.113-120, July 2005.

[170] A. Michlmayr, F. Rosenberg, P. Leitner, and S. Dustdar, "Comprehensive QoS monitoring of Web services and event-based SLA violation detection," in Proc. 4th ACM International Workshop on Middleware for Service Oriented Computing, New York, USA, pp. 1-6, 2009.

[171] R. Clegg, S. Clayman, G. Pavlou, L. Mamatas, and A. Galis, "On the Selection of Management/monitoring Nodes in Highly Dynamic Networks," in IEEE Transactions on Computers, 2012.

[172] Chao Zhang, Huarui Yin, Weidong Wang, and Guo Wei, "Selective cooperation based on overall power minimization in wireless networks," in Proc. 3rd International Conference on Communications and Networking, pp.498-503, 25-27 Aug. 2008.

[173] D. Wei, Y. Jin, S. Vural, K. Moessner, and R. Tafazolli, "An Energy-Efficient Clustering Solution for Wireless Sensor Networks," in IEEE Transactions on Wireless Communications, vol.10, no.11, pp.3973,3983, November 2011.

[174] A. Quiroz, N. Gnanasambandam, M. Parashar, and N. Sharma, "Robust clustering analysis for the management of self-monitoring distributed systems," in Journal of Cluster Computing, vol.12, no.1, pp. 73-85, 2008.

[175] H. Luu and X. Tang, "On the Construction of Rings Overlay for Robust Data Collection in Wireless Sensor Networks," in Proc. IEEE Conference on Wireless Communications and Networking, pp.1-6, 18-21 April 2010.

[176] C. Sreng, H. Yanikomeroglu, and D. Falconer, "Relayer selection strategies in cellular networks with peer-to-peer relaying," in Proc. 58th IEEE conference on Vehicular Technology, pp. 1949- 1953 Vol.3, 6-9 Oct. 2003.

[177] M. Peng and W. Wang, "Investigation of Cooperative Relay Node Selection in Heterogeneous Wireless Communication Systems," in Proc. IEEE International Conference Workshops on Communications, pp.174-178, 19-23 May 2008.

[178] M. Baidas and A. MacKenzie, "An Auction Mechanism for Power Allocation in Multi-Source Multi-Relay Cooperative Wireless Networks," in IEEE Transactions on Wireless Communications, vol.11, no.9, pp.3250-3260, September 2012.

[179] I. Chaabane, S. Hamouda, and S. Tabbane, "A novel relay selection scheme for LTE-advanced system under delay and load constraints," in Proc. IEEE Conference Workshops on Wireless Communications and Networking, pp.263-267, 1-1 April 2012.

[180] K. Vardhe, D. Reynolds, and B. Woerner, "Joint power allocation and relay selection for multiuser cooperative communication," in IEEE Transactions on Wireless Communications, vol.9, no.4, pp.1255-1260, April 2010.

[181] J. Wieselthier, G. Nguyen, and A. Ephremides, "On the construction of energy-efficient broadcast and multicast trees in wireless networks," in Proc. 19th IEEE Conference on Computer and Communications Societies, pp.585-594 vol.2, 2000.

[182] W. Wang, V. Srinivasan, and K. Chua, "Extending the Lifetime of Wireless Sensor Networks Through Mobile Relays," in IEEE/ACM Transactions on Networking, vol.16, no.5, pp.1108-1120, Oct. 2008.

[183] N. Aslam, W. Robertson, W. Phillips, and S. Sivakumar, "Relay Node Selection in Randomly Deployed Homogeneous Clustered Wireless Sensor Networks," in Proc. IEEE Conference on Sensor Technologies and Applications, pp.418-423, 14-20 Oct. 2007.

[184] M. Nahas, A. Haj-Ali, T. Chakerian, and M. Rihani, "Asynchronous relay selection protocol for distributed cooperative networks," in Proc. International Symposium on Wireless Communication Systems, pp.456-460, 28-31 Aug. 2012.

[185] Y. Chen, G. Yu, P. Qiu, and Z. Zhang, "Power-Aware Cooperative Relay Selection Strategies in Wireless Ad Hoc Networks," in Proc. 17th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications, pp.1-5, 11-14 Sept. 2006.

[186] T. Ng and W. Yu, "Joint optimization of relay strategies and resource allocations in cooperative cellular networks," in IEEE Journal on Selected Areas in Communications, vol.25, no.2, pp.328-339, February 2007.

[187] "NS-2 Network Simulator," [Online]. Available: http://www.isi.edu/nsnam/ns/".

[188] J. Shin, J.W. Kim, and C.J. Kuo, "Quality-of-service mapping mechanism for packet video in differentiated services network," in IEEE Transactions in Multimedia, vol. 3, no. 2, pp. 217–230, Jun. 2001.

[189] G. Ghinea, J.P. Thomas, and R.S. Fish, “Mapping quality of perception to quality of service: The case for a dynamically reconfigurable communication system,” in Journal of Intellectual Systems, vol. 10, no. 5/6, pp. 607–632, 2000.

[190] W. Kim, B. Lee, J. Song, Y. Shin, and Y. Kim, “Ping-Pong Avoidance Algorithm for Vertical Handover in Wireless Overlay Networks”, in Proc. 66th IEEE Vehicular Technology Conference. pp. 1509 –1512, 2007.

[191] Y. Al Ridhawi, I. Al Ridhawi N. Samaan, and A. Karmouch, "An Adaptive Vertical Handover in Service Specific Overlay Networks," in Proc. 3rd Workshop on Enablers for Ubiquitous Computing and Smart Services, 16-20 July 2012.

[192] H. Stiller and B. Stiller, “SLO Auditing Task Analysis, Decomposition, and Specification,” in IEEE Transactions on Network and Service Management, vol. 8, no. 1, pp. 15 –25, march 2011.

[193] A. Ouda, H. Lutfiyya, and M. Bauer, “Automatic Policy Mapping to Management System Configurations”, in Proc. IEEE International Symposium on Policies for Distributed Systems and Networks, pp. 87 –94, July 2010.

[194] T.L. Saty “How to make a decision: The analytic hierarchy process,” in The INFORMS Journal on the Practice of Operation Research, Vol. 24, No. 6., pp. 19-43, 1994.

[195] “NIST IT Laboratory,” [Online] Available: http://www.nist.gov/itl/antd/emntg/ssm_seamlessandsecure.cfm.

[196] P. Pawar, B.J. van Beijnum, M. van Sinderen, A. Aggarwal, P. Maret, and F. De Clercq, “Performance evaluation of the context-aware handover mechanism for the nomadic mobile services in remote patient monitoring”, in Elsevier Computer Communications, vol. 31, no. 16, pp. 3831–3842, 2008.

[197] Q. Song and A. Jamalipour, “A network selection mechanism for next generation networks”, in IEEE International Conference on Communications, vol. 2, pp. 1418–1422, 2005.

[198] S. Zanakis, J. Evans, and A. Vazacopoulos, “Heuristic methods and applications: A categorized survey,” in European Journal of Operational Research, vol. 43, no. 1, pp. 88-110, Nov. 1989.

[199] I.H. Osman and G. Laporte, “Metaheuristics: A bibliography,” in Annals of Operations Research, vol. 63, pp. 513–623, 1996.

[200] C. Blum and A. Roli, “Metaheuristics in combinatorial optimization: Overview and conceptual comparison,” in ACM Computer Survey, vol. 35, no. 3, pp. 268-308, September 2003.

[201] F. Glover, “Future paths for integer programming and links to artificial intelligence,” in Journal of Computer Operations and Research, vol. 13, pp. 533–549, 1986.

[202] F. Glover, “Heuristics for integer programming using surrogate constraints,” in Journal of The Decision Science Institute, vol. 8, pp. 156-166, 1977.

[203] R. Durret, “Markov Chains,” in Probability: Theory and Examples, 4th ed. Cambridge, U.K., Cambridge University Press, 2010, Ch. 6.

[204] R.A. Aziz, M. Ayob, and Z. Othman, "The effect of learning mechanism in Variables Neighborhood Search," in Proc. 4th Conference on Data Mining and Optimization, pp.109-113, 2-4 Sept. 2012.

[205] “OpenTS – Java Tabu Search,” [Online] Available: http://www.coin-or.org/Ots/index.html.

[206] F. Ricciato, "Traffic monitoring and analysis for the optimization of a 3G network," in IEEE Wireless Communications, vol.13, no.6, pp.42-49, Dec. 2006.

[207] C. Molina-Jimenez, S. Shrivastava, J. Crowcroft, and P. Gevros, "On the monitoring of contractual service level agreements," in Proc. 1st IEEE International Workshop on Electronic Contracting, pp.1,8, 6 July 2004.

[208] C. Molina-Jimenez, S. Shrivastava, J. Crowcroft, and P. Gevros, "QoS Monitoring of Service Level Agreements," Information Society Technologies, Newcastle, U.K., TAPAS D10 Rep., May 2004.

[209] K. Anagnostakis, S. Ioannidis, S. Miltchev, M. Greenwald, J. Smith, and J. Ioannidis, "Efficient packet monitoring for network management," in Proc. IEEE/IFIP Symposium on Network Operations and Management, pp. 423- 436, 2002.

[210] D. Schuehler, J. Lockwood, “TCP-Splitter: A TCP/IP flow monitor in reconfigurable hardware,” in Proc. 10th IEEE Symposium on High Performance Interconnects, pp.127,131, 2002.

[211] Y. Ariba, F. Gouaisbaut, S. Rahme, and Y. Labit, "Robust control tools for traffic monitoring in TCP networks," in Proc. IEEE conference on Control Applications & Intelligent Control, pp.525-530, 8-10 July 2009.

[212] A. Quiroz, M. Parashar, N, Gnanasambandam, and N. Sharma, "Design and evaluation of decentralized online clustering" in ACM Transactions on Autonomous and Adaptive Systems, vol.7, no.3, article 34, October 2012.

[213] A. Quiroz, M. Parashar, and I. Rodero, "Autonomic management of distributed systems using online clustering," in Proc. IEEE International Symposium on Parallel & Distributed Processing, Workshops and Phd Forum, pp.1-4, 19-23 April 2010.

[214] Z. Li and H. Shen, "Game-Theoretic Analysis of Cooperation Incentive Strategies in Mobile Ad Hoc Networks," in IEEE Transactions on Mobile Computing, vol.11, no.8, pp.1287-1303, Aug. 2012.

[215] C. Saraydar and A. Yener, "Adaptive cell sectorization for CDMA systems," in IEEE Journal on Selected Areas in Communications, vol.19, no.6, pp.1041,1051, Jun 2001.

[216] X. Chen and N. Rowe, "Saving Energy by Adjusting Transmission Power in Wireless Sensor Networks," in Proc. IEEE Global Telecommunications Conference, pp.1-5, 5-9 Dec. 2011.

[217] F. Wang J. Liu, and Y. Xiong, "On Node Stability and Organization in Peer-to-Peer Video Streaming Systems," in IEEE Systems Journal, vol.5, no.4, pp.440-450, Dec. 2011.

[218] G. Ting, W. Haiyang, Z. Naihui, and L. Fei, "An Improved Way to Facilitate Composition-Oriented Semantic Service Discovery," in Proc. IEEE International Conference on Computer Engineering and Technology, pp. 156-160, 2009.

[219] K. Iqbal, M. L. Sbodio, V. Peristeras, and G. Giuliani, "Semantic Service Discovery using SAWSDL and SPARQL," in Proc. 4th IEEE International Conference on Semantics, Knowledge and Grid, pp. 205-212, 2008.

[220] E. Sirin, B. Parsia, and J. Hendler, "Filtering and Selecting Semantic Web Services with Interactive Composition Techniques," in IEEE Intelligent Systems Magazine, vol. 19, no. 4, pp. 42-49, 2004.

[221] M. Rodriguez and M. Egenhofer, "Determining semantic similarity among entity classes from different ontologies," in IEEE Transactions on Knowledge and Data Engineering, vol.15, no.2, pp. 442- 456, March-April 2003.

[222] S. Xia, Z. Hu, and Q. Niu, "An Approach of Semantic Similarity Measure between Ontology Concepts Based on Multi Expression Programming," in Proc. 6th Web Information Systems and Applications Conference, pp.184-188, 18-20 Sept. 2009.

[223] A. Tversky, "Features of Similarity," Psychological Rev., vol. 84, pp. 327-352, 1977.

[224] C. Krumhansl, "Concerning the Applicability of Geometric Models to Similarity Data: The Interrelationship between Similarity and Spatial Density," Psychological Rev., vol. 85, pp. 445-463, 1978.

[225] "BonnMotion: A mobility scenario generation and analysis tool," [Online]. http://net.cs.uni-bonn.de/wg/cs/applications/bonnmotion/.

[226] "GloMoSim: Global Mobile Informarion Systems Simulation Library," [online]. Available: http://pcl.cs.ucla.edu/projects/glomosim/.

[227] "Cooja," [Online]. Available: http://www.contiki-os.org/.

[228] T. Adamusiak, T. Burdett, N. Kurbatova, K. Velde, N. Abeygunawardena, D. Antonakaki, M. Kapushesky, H. Parkinson, and M. Swertz, "OntoCAT – Simple Ontology Search and Integration in Java, R and REST/JavaScript," BMC Bioinformatics,2011.

[229] Y. Ennaji, M. Boulmalf, and C. Alaoui, "Experimental analysis of video performance over wireless local area networks," in Proc. IEEE International Conference on Multimedia Computing and Systems, pp.488,494, 2-4 April 2009.

[230] Y. Al Ridhawi, "Dynamic Composition of Service Specific Overlay Networks," Ph.D. Thesis, University of Ottawa, Canada, 2013

[231] I. Abdeljaouad, "Self-Configuration and Monitoring of Service Specific Overlay Networks," Ph.D Thesis, University of Ottawa, Canada, 2013.