The Pennsylvania State University

The Graduate School

College of Engineering

RESOURCE ALLOCATION IN WIRELESS NETWORKS

A Thesis in

Computer Science and Engineering

by

JaeSheung Shin

Submitted in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

May 2007

The thesis of JaeSheung Shin has been reviewed and approved* by the following:

Thomas F. La Porta
Distinguished Professor of Computer Science and Engineering
Thesis Adviser
Chair of Committee

Guohong Cao
Associate Professor of Computer Science and Engineering

John Metzner
Professor of Computer Science and Engineering / Electrical Engineering

Aylin Yener
Associate Professor of Electrical Engineering

Sencun Zhu
Assistant Professor of Computer Science and Engineering

Raj Acharya
Professor of Computer Science and Engineering
Head of the Department of Computer Science and Engineering

* Signatures are on file in the Graduate School

Abstract

The critical design issues for wireless networks include provisioning of seamless communication with Quality-of-Service (QoS) guarantees, improved service accessibility, reliable data transfer, and high communication performance. However, limited bandwidth and processing power of the network elements and scarce, and time/location-dependent radio resources make the design space much more complex and challenging.

The overall objective of this research is to design and analyze the resource allocation schemes for wireless networks that can provide high communication performance.

Driven by two main evolution issues in wireless cellular networks, core network evolution to all-IP wireless networks and radio access network evolution to high-capacity wireless networks based on multi-frequency, multi-hop transmission, in this research we investigate two kinds of resource allocation in wireless networks: network resource allocation and radio resource allocation. The resource allocation algorithms proposed in this research focus on perfect load balancing across each network element and dynamic multi-hop relay network formation in wireless cellular and wireless mesh network environment. The relay network formation includes path discovery to a GW or a relay mesh router, path selection, frequency allocation, and local tuning to reduce interference in resulting relay networks further.

The advantages of these algorithms are demonstrated through extensive simulation. The results show that the algorithms achieve significantly improved throughput gain.

Table of Contents

List of Tables

List of Figures

Acknowledgements

First of all, I would like to thank God who has been always with me and my family, blessed us, and kept us safe.

This thesis owes very much to my Ph.D. advisor, Dr. Thomas F. La Porta, and I would like to thank him from the bottom of my heart for his priceless guidance, advice, and encouragement during entire graduate school period. He not only trained me to be a professional but also took care of me like a family member. Thanks to his admission to the Department of Computer Science and Engineering, I could start a new life at the Penn State University. He always led me to be open-minded and flexible for the way of thinking. I learn from him about how to become an excellent researcher as well as a thoughtful advisor. Everything that he showed me must be a model of my behavior for my entire life.

I am also very grateful to my Ph.D. committee members, Dr. Guohong Cao, Dr. John Metzner, Dr. Aylin Yener, Dr. Sencun Zhu, for their valuable advices and comments.

It is always great pleasure to remind the colleagues in Networking and Security Research Center (NSRC), especially HeeSook Choi, Patrick Traynor, Raju Kumar, Yan Sun, Kameswari Kotapati, Hui Song, Jing Zhao. I cannot forget their great advices and helps.

I have a great memory to study and spend time with Korean colleagues in the Department of Computer Science and Engineering, Jongman Kim, Ingyu Lee, Dongkook Park, Byungwon On, Younghyun Oh, Sungmin Bae, Seungwoo Son, Yoonchan Jhi, Jieun Jeong, Hyunsuk Noh, Jungsup Kim, Youngjae Kim, Boram Lee, Jaehyun Lim, Gunwoo

Nam, Hyunyoung Kil, Hungsik Kim, Yoojin Hong, Sunam Ryu, and so on. I will never forget the Korean colleagues in the Department of Electrical Engineering who started the study at Penn State Univ. with me, Byounghee Kim, Kyounghwan Lee, Heayoung Yoon.

I thank Prof. Keumsuk Lee, Dr. Sunbae Lim, Prof. Hyuckjae Lee, Prof. Youngnam Han very much for helping me prepare studying abroad.

Above all, I deeply grateful to Rev. Sangki Eun, Rev. Juyong Lee, and Rev. David Alas leading State College Korean Church and all brothers and sisters, especially young adult group and choir members, in the church for their sincere advice and prayers for me and my family.

I dedicate this thesis to my father, Booil Shin, my mother, Keumrye Choi, father-in-law, Dongwon Lee, and mother-in-law, Ilsoon Jang, who always trust and support us. I cannot give enough thanks to my wife's sister and her husband for their encouragement and support. I am also grateful to all the other family members, who are not listed here, for their silence trust and help.

Last but not the least, I would like to express my most sincere gratitude to my lovely wife, Soyeon Lee, for her endless support, prayers, and love, and my adorable daughter, Grace Yedam Shin. They are the reason why I live!

To my lovely wife and daughter

Chapter 1. Introduction

The efficient allocation of resources is a fundamental problem for any modern telecommunication network. In wireless networks, increasing demand for accommodating mobiles with diverse service requirements in an environment with scarce, and time- and location-dependent radio resources makes the efficient utilization of resources more difficult. Hence, it is necessary to design resource allocation schemes considering both the service requirements of mobiles and the state of radio environment.

Driven by the need for increased network capacity, global roaming, and the growing use of wireless data over the Internet, two main evolution issues in wireless cellular networks have been considered. The first is the *core network evolution* to all-IP wireless networks for full support of session-based IP applications. The second is the *radio access network evolution* to high-capacity wireless networks based on multi-frequency, multi-hop transmission. Moreover, recently the integration of various wireless network technologies has been successfully completed with the realization of wireless mesh networks to provide ubiquitous wireless services.

In this thesis, we investigate two kinds of resource allocation in wireless networks: (1) network resource allocation to support core network evolution and (2) radio resource allocation to support the evolution of radio access networks and wireless mesh networks. If an integrated scheme that addresses these two kinds of resource allocation is considered, it will achieve a significant performance gain.

1.1. Network Resource Allocation

Session-based IP applications, such as Internet telephony, are an important component of the emerging mobile Internet. The ubiquitous availability of these services is critical to the success of the mobile Internet. Since all-IP networks will be deployed in phases and current mobile telecommunication systems will be in operation for decades to come, the interworking between current cellular services and all-IP services is a key problem.

One method to support interworking is to use multi-mode terminals. Another method is to allow subscribers to use multiple devices that are accessible via a single user address. These methods introduce a user mobility infrastructure into the network.

Based on these methods, three interworking approaches to support seamless roaming in a wireless mobile environment are considered [51]. The first approach called *master-slave mobility management* requires denoting a master protocol (network) and slave protocols (networks). In this approach, the master mobility manager is fundamentally the same as the mobility manager native to the master protocol. On the other hand, the slave mobility manager is responsible for providing the interworking and interoperability function.

The second approach called *federated mobility management* utilizes the native mobility manager in each network. In order to realize seamless roaming across heterogeneous networks, this approach exploits a new entity called user mobility component that contacts the mobility managers of the participating networks through a dedicated database interface instead of the communication protocols for mobility management or call control.

The third approach called *unified mobility management* integrates mobility management functions of multiple network protocols in one logical entity, and adds user mobility management capabilities on top of this multi-protocol management system. This approach is based on a specific instance of a unified mobility management system proposed in [17][18][27].

Our contribution is to compare the performance of these approaches and to find a way to achieve better performance for the seamless roaming to IP-based services.

We perform a comparative performance analysis. We model each component in the network as an M/M/1 queuing system. Based on the processing time determined through a combination of experimental measurement of an extensive system prototype and inference, we measure the average call delivery delay of the Session Initiation Protocol (SIP) terminated calls of each approach and compare them.

Due to the network hierarchy of cellular networks, several network elements have relatively high processing load compared to the other elements, and hence are easily overloaded as the call or mobility rate increases. In order to achieve perfect load balancing across each network element and get better performance for the seamless roaming for IP-based services, we propose an efficient assignment scheme of processing resources to each network element in this thesis. That is, processing resources of each network element are adjusted so that their processing rates keep constant and thus the load of each element is balanced perfectly.

The analytical results show that we can reduce the average call delivery delay of the network as a result of load balancing. They also show that the unified mobility management approach is most efficient if a great deal of interworking is required, and as more users invoke IP-based services.

1.2. Radio Resource Allocation

Traditional wireless networks typically rely on a centralized architecture in which all mobile nodes communicate only with a Base Station (BS) via a single hop wireless link. The protocols and the radio spectrum used on this link are predetermined and fixed by a Wireless Service Provider (WSP) according to regulatory policies. This network model has resulted in several limitations: (1) even if a large number of BSs are deployed, there are still places where the communication platform may fail to provide successful communication; (2) in a densely populated area, the area that a single BS covers reduces in size to avoid congestion, and hence the number of BSs deployed increases. This increases the cost of deployment; and (3) wireless network capacity is limited by the scarcity of available radio spectrum, therefore new WSPs face a high barrier to entry because of the need to own spectrum.

As one of the solutions to overcome these limitations, many researchers have looked into building multi-hop wireless networks. In these networks, mobile nodes forward data on behalf of other mobile nodes to a BS or a destination node. This network architecture provides several benefits. The first is the low deployment cost because the total number of BSs deployed can be reduced. Second, network fault tolerance is increased, because if a BS fails, the relay network can forward traffic to another functioning BS. This results in improved network coverage. Third, network capacity is increased because nodes may transmit at lower power and traffic may be aggregated resulting in more efficient channel access and transmission scheduling.

However, in multi-hop wireless networks comprised of single frequency and single radio nodes, the broadcast nature of wireless transmission coupled with multi-hop relaying limits the usable capacity of the network due to increased interference among relaying nodes. Several research efforts have addressed this problem with the consideration of using time-division multiplexing or multiple radios at a node.

Even though the time-division multiplexing does not require additional radios, it incurs a large overhead to synchronize the nodes involved in communication and increased transmission delay. Alternatively, including multiple radios in each node so that they can be assigned different frequencies may ensure that network links can utilize diversity and avoid interference with adjacent links. However, this approach may increase the node's cost.

With the advent of agile radios, mobile nodes with a single radio may shift frequencies very quickly so that they can operate dynamically in spectrum that is most suitable for their communication based on availability, interference level, or business arrangement. Therefore, the combination of relay networks with agile radios provides for even more powerful enhancements. However, the agile radios are still not commercially viable.

Taking this into account, we consider two types of Multi-Frequency, Multi-Hop (MF-MH) wireless networks that consist of nodes equipped with an agile radio or multiple radios, respectively: (1) *MF-MH wireless cellular networks* comprised of mobile nodes with an agile radio and (2) *MF-MH wireless mesh networks* comprised of mobile nodes with multiple radios.

In MF-MH wireless cellular networks, there is a centralized BS in each cell. All data traffic goes through the BS. We consider an environment in which when the BS schedules a large number of nodes that have poor signal quality and hence low bandwidth, it may choose to make new spectrum available on which a relay network is formed. This information is broadcast over the cellular control channels so that all nodes within the cell receive it simultaneously. Mobile nodes may choose to form a relay network operating on the introduced spectrum. When the relay network is created, each group of nodes within the relay network select a gateway (GW) that acts as an access point to the BS. Then the nodes establish a path to the GW and forward data to it.

At a later time, if the quality of service from the BS improves or the spectrum in which the relay network is operating is no longer available to the WSP, the relay network is dissolved and nodes return to using the cellular interface with the BS.

The MF-MH wireless mesh network consists of mesh clients and mesh routers forming an infrastructure. As with the GWs in MF-MH wireless cellular networks, relay mesh routers serve as access points for mesh clients and form a multi-hop mesh backbone connected to the GW mesh routers. The GW mesh routers act as bridges between the mesh network and the Internet. They correspond to the BS in MF-MH wireless cellular networks.

In these networks mentioned above, the relay networks are dynamically formed and dissolved if needed. In order to avoid contention and interference, each relay network operates on different radio spectrum. Within a relay network, the multiple orthogonal frequencies[1] are used to establish non-interfering links so that multiple nodes within

---

[1] *The frequency is also called "channel".*

range of each other may transmit simultaneously without relying on a MAC protocol or distributed scheduling algorithm to resolve contention.

Our contribution is to propose a set of distributed algorithms for relay network formation and to evaluate the performance of proposed algorithms in terms of the throughput of the resulting relay network.

The relay network formation algorithms include path discovery to a GW or a relay mesh router, path selection, resource allocation, and local tuning to improve the performance of resulting relay networks. Since each client may have multiple radios, the interface assignment in a mesh client is considered in mesh network environment.

We measure the throughput of the resulting relay network in three scenarios: (1) relay network using a single frequency; (2) relay network using multiple frequencies assigned by our distributed frequency allocation; and (3) relay network using multiple frequencies with optimal assignment. The results show that by exploiting multiple frequencies, we can achieve higher overall network throughput and average node throughput. Our proposed algorithms can achieve close to optimal performance.

Furthermore, the availability of multiple frequencies allows the mobile nodes to form diversity paths between each other. The broadcasting nature of the wireless medium enables an intermediate relay node to overhear the data broadcasted by the source, which then relays the information to the destination on a new frequency. The path diversity helps to reduce the bit error rate, and to increase the network throughput. We also show that network formation algorithms that exploit such a diversity provided by intelligent resource allocation can achieve better end-to-end performance.

1.3. Outline

The rest of the thesis is organized as follows. In chapter 2, we review previous work related to our research. In chapter 3, we discuss the network resource allocation in wireless cellular networks. In chapters 4 and 5, we discuss the relay network formation and reliability-aware radio resource allocation for MF-MH wireless cellular networks. In chapter 6, we discuss the relay network formation and radio resource allocation for MF-MH wireless mesh networks. Finally, chapter 7 summarizes the results and concludes the thesis.

Chapter 2. Related Work

In the following subsections we briefly review previous work related to our research.

2.1. Load Balancing

There has been a great deal of work on the distribution of processing and communication tasks evenly across a computer network so that no single device is overloaded. Distributed communication networks usually have a modular architecture. An arbitrary number of processing elements can be executed on various machines in the networks. In these networks, load balancing is especially important because it is difficult to predict the number of requests that will be issued to a network device.

In a distributed environment, some research efforts have focused on homogeneous distributed systems [43][73] and some on heterogeneous distributed systems [19][69][71]. In [71], algorithms to determine the optimal load in a heterogeneous distributed system based on a queuing network model are proposed. In these algorithms, a job may be processed at the host or it may be transferred to another host. In [69], an algorithm that finds the optimum assignment of jobs to sites in a distributed system is presented. An improvement in system performance through file placement and process assignment is presented in [19].

While the algorithms mentioned above are based on static load balancing policy, a dynamic load balancing policy in a distributed system is introduced in [20][21]. In [20], the simulation model allows process migration to different sites depending on the loads on the hosts.

In this thesis, based on static policy using a queuing network model, we achieve the load balancing across the wireless cellular core network. We adjust processing resources of network elements depending on loads on the elements so that all elements have the same utilization at expected operating conditions.

2.2. Interworking between Heterogeneous Wireless Cellular Networks

In order to support session-based IP applications in wireless cellular networks, the interworking and migration between current network services and all-IP services are required. There are several approaches for providing the interworking capabilities based on the mobility management model in a heterogeneous mobile network environment. The details are described in [51].

In this thesis, we compare the trade-offs of the three interworking approaches using a comparative performance analysis and investigate which approach is suitable for supporting the migration to all-IP mobile networks.

2.2.1. Master-Slave Mobility Management Approach

This model is a generalization of the GSM/ANSI-136 Interoperability Team (GAIT) approach [81]. In this approach, one of the protocols acts as a master protocol. A master mobility manager is introduced in the master network. This master mobility manager is fundamentally the same as the mobility manager native to the master protocol, potentially with minor extensions to support user mobility. It stores user's service and location information, as well as user mobility information such as mappings between user addresses and terminal addresses. For all other networks, which we call slave networks, a slave mobility manager relative to a master protocol is introduced. Although the slave mobility manger will internally process some protocol specific requests, common mobility manager functions (e.g. registration and call delivery) are achieved by interactions with the master mobility manager. A slave mobility manager is required to store users' service profile information specific to their native network type. Furthermore, it maintains a dynamic data mapping for multi-protocol operations. Hence, the slave mobility manager is viewed as a network element in master network from the point of view of the master network.

While this approach may be simple to deploy if the master protocol is based on an existing system and works efficiently if call requests are predominantly on a single network type, its performance degrades as interworking is required.

2.2.2. Federated Mobility Management Approach

This approach is one of the key rationales behind the current definition of standard open interfaces for presence and availability information, such as the PAM (presence and awareness management) interface in Parlay [82]. It utilizes the native mobility manager in each network. Each mobility manager acts independently, and mobility management operations between mobility managers and call control points (CCP) are handled separately in each network. In order to realize seamless roaming across heterogeneous networks, this approach exploits a new entity called a user mobility component (UMC). The UMC contacts the mobility managers of the participating networks (e.g. Universal Mobile Telecommunication Systems (UMTS) or Session Initiation Protocol (SIP)) through a database interface instead of the communication protocols for mobility management or call control. In other words, the mobility manager databases act as federated data servers [66]. The UMC also belongs to one of the participating networks, which we call a main network. The UMC participates in communication activities through the protocol of the main network.

This approach allows a fair amount of isolation between different networks, but may require multiple data accesses over a wide area that might adversely affect performance.

2.2.3. Unified Mobility Management Approach

This approach integrates mobility management functions of multiple network protocols in one logical entity, and adds user mobility management capabilities on top of this multi-protocol management system. This approach is based on a specific instance of a Unified Mobility Manager (UMM) proposed in [17][18][27]. The UMM provides traditional cellular networks with standard Home Location Register (HLR) functionality, maintaining the registration and location information of the mobile terminal and the value added service settings such as call forwarding activation, barring of incoming calls, and so on. For SIP networks, it works as a SIP location server, maintaining the user's contact addresses together with related service subscriptions. The protocol interworking capability of the UMM facilitates efficient realization of user mobility in a multi-protocol mobile network infrastructure. The UMM has a novel architecture, called Common

OPerationS (COPS), that abstracts mobility management functions so they can be accessed by multiple network interfaces.

This facilitates efficient interworking, but incurs moderate overhead if only a single protocol is being used in a session.

2.3. Multi-hop Wireless Cellular Networks

There has been a great deal of work on multi-hop wireless networks to improve cellular network performance [1][24][40][46][75]. Some studies [45][72] showed that an equivalent service can be provided to the end-user with a reduced BS density if wireless relays are used for multi-hopping, thus reducing the infrastructure cost. Moreover, the spectral efficiency of currently deployed single hop systems can also be significantly improved. Thus, the practicality of multi-hop traffic relaying is justified according to these cost models.

In [1][24], the authors have addressed relaying in GSM networks and in UMTS networks. In this thesis, we mainly investigate a third generation (3G) wireless environment in which the sharing of communication resources is done via a combination of regulating power and time division multiplexing.

This is similar to the Unified Cellular and Ad hoc Network architecture (UCAN) system [46]. For example, in the 1xEV-DO system, the BS schedules only a single node for downlink transmission at any instant, and transmits at full power. The bit rate achieved during each time interval depends on its signal quality to the mobile node, which is roughly a function of distance. If the BS can schedule nodes with better signal quality more often, then a higher average bit rate for the network can be achieved. To maximize throughput of 3G networks, the UCAN system proposes that the BS transmit all downlink data to mobile nodes with high signal quality, and that these nodes then forward data to other nodes in the network through a high speed relay network operating in a different spectrum than the 3G interface, specifically using an 802.11 network. In this way, the downlink from the BS can always run at its maximum rate and all users achieve higher throughput. It is reported that UCAN can achieve improvements of the average and maximum throughput of up to 37% and 82%, respectively.

Our work is most similar to the UCAN system. However, UCAN is different from ours in that the relay networks in UCAN operate on a *single* frequency and a *persistent* 802.11 network exists for use as the relay network.

2.4. Path Discovery

A vast amount of research has been conducted in ad hoc routing protocols. Ad hoc routing protocols are typically subdivided into two main categories: *proactive* routing protocols and *reactive* routing protocols.

Reactive routing protocols are characterized by on-demand path discovery mechanisms initiated when a source needs to communicate with a destination that it does not know how to reach. The path discovery is usually in the form of query flood. The reactive ad hoc routing protocols such as AODV [56] and DSR [29] have some optimization features such that (1) a node may passively learn a route to a destination, for example if it is part of a longer path to that destination, and (2) a node that has previously learned a route may immediately return this route in response to a request without a further search. These features can greatly reduce the number of messages required to find routes. Moreover, the traditional reactive routing protocols are still considered better for small scale networks with a path length of few hops like a single cell environment [53]. Comparing AODV with DSR, some studies [11][50] show that AODV outperforms DSR in the situations with heavy routing load and higher mobility. AODV is also better that DSR in terms of scalability and CPU and memory usage.

Motivated by these observations, we use a reactive routing protocol, specifically the modified version of AODV, for relay network formation.

2.5. Path Diversity

An intense research effort is currently being directed to understanding the performance limits of wireless relay networks [25][33][52]. To increase the transmission range and capacity of wireless ad-hoc networks, data packets can be delivered to the destination not only by a direct link, but by one or more diversity paths with the help of intermediate relay nodes. Such relay assisted communication systems are shown to achieve the benefits of spatial diversity without requiring physical antenna arrays [37][64][65].

In wireless relay networks, the relay node overhears the data from the source, performs appropriate signal processing of the received data, and forwards the data to the destination. The destination combines the received signals from the source and the relay to achieve lower bit error rate, hence a higher network goodput. Realistic scenarios require the orthogonality of frequency between the transmission from the source and from the relay for simultaneous transmission and reception at the relay. This orthogonality can be realized by frequency division [14][38][41], time division [25][37][52], or code division [64][65] multiplexing.

Methods of forwarding the data at the relay include the so called *Amplify and Forward* (AF), and *Decode and Forward* (DF) schemes [22][36]. In AF, the relay simply amplifies what it receives from the source and forwards it to the destination. In DF, the relay decodes the signal from the source and re-encodes and forwards the newly encoded data to the destination. As long as there is no decoding error at the relay, DF performs better than AF because the destination can receive the two identical signals from the diversity paths [36].

The possibility of a decoding error is low for reasonably large received Signal-to-Noise plus Interference Ratio (SNIR) values at the relay. If the SNIR is not large, the signal may be decoded in error, and DF can perform worse than AF [36][38]. In order to prevent this error propagation, the relay nodes typically employ a threshold rule [22][36] by which they decide to perform DF only when the received SNIRs are larger than a threshold.

In this thesis, we consider the diversity case where all relay nodes employ DF. However, unlike most of these previous studies that mainly focused on the bit error probability within single two-hop diversity, we consider the relay path which consists of multiple two-hop diversified paths and define the end-to-end error probability of the whole relay path.

2.6. Frequency Allocation

Much of the previous work on resource allocation in wireless networks have focused on the frequency assignment that applies to link-based transmission within a given frequency band. The frequency allocation algorithms are classified into two categories: centralized and distributed.

Most of the centralized algorithms define the frequency allocation as a graph coloring problem which has been shown to be NP-complete or NP-hard and solve it using some heuristics. The large majority use greedy approaches in which nodes with either a high priority or the highest degree of connectivity in a network choose colors first [7][13][44]. Algorithms that favor the nodes with the smallest degree have also been explored [60]. In [47], the nodes with the maximum colors in a local neighborhood are picked first.

In most of the distributed algorithms, based on the knowledge of one or two-hop connectivity, neighbors' allocated frequency information, and their priority information, each node determines the best frequencies for its own transmission [7][13][44]. Some approaches do not require negotiation between nodes [48][49].

In this thesis, we propose a link-based frequency allocation algorithm in a distributed manner. In our algorithm, each node obtains the used frequency information of neighboring nodes one or two-hops away to choose non-interfering frequencies. We also propose a local tuning algorithm to improve the throughput of resulting relay networks.

2.7. Multi-Radio Wireless Networks

Recently, there has been a great deal of effort on advanced wireless networks in which nodes can simultaneously communicate with their neighbors using multiple radio interfaces over multiple orthogonal frequencies [4][32][35]. In [35], it is shown that the network throughput can be significantly improved when mobile nodes are equipped with multiple interfaces and enabled to utilize multiple frequencies. In order to form multiple orthogonal links efficiently, several frequency allocation schemes are proposed in [4][32].

Motivated by these observations, we consider multi-hop wireless networks that consist of nodes equipped with multiple radios operating on multiple orthogonal frequencies. In order to improve the interface utilization of a node and hence increase network throughput, we propose an interface assignment algorithm.

2.8. Wireless Mesh Networks

A great deal of work has recently focused on Wireless Mesh Networks (WMNs) that aim to integrate various wireless networks and thus provide ubiquitous wireless services [3][4][8][10][12]. In [3], the authors classify the architecture of WMNs into three categories: client WMN, infrastructure WMN, and hybrid WMN. A client WMN provides peer-to-peer communications among mesh clients without any infrastructure. An infrastructure WMN includes mesh routers forming a multi-hop backbone network for mesh clients. This backbone can be built using various types of radio technologies including IEEE 802.11, 802.15, and 802.16. A hybrid WMN is the combination of infrastructure WMN and client WMN.

Unlike mesh clients, mesh routers are generally supposed to be stationary. Thus, the wireless multi-hop backbone of infrastructure WMNs is static. Most previous research efforts have mainly focus on the efficient construction of the multi-hop backbone.

In this thesis we focus on the dynamic access to the wireless mesh backbone. We envision an infrastructure WMN in which mesh clients dynamically form a multi-hop relay network to connect to a backbone. The mobility of the mesh clients and the dynamic nature of these relay networks motivate the need for distributed algorithms to form such relay networks.

Chapter 3. Resource Allocation for Core Network Elements in Single Hop Wireless Cellular Networks

3.1. Introduction

The wireless communications industry is headed towards all-IP mobile networks [55][58], both through the advance of standards efforts such as the 3rd Generation Partnership Project (3GPP) [83], and with the proliferation of competing technologies such as WiFi. Session-based services, such as voice-over-IP and multimedia communication, are critical for the success of these networks. The fact that it will take several years to finalize standards for all-IP networks, and many more for these networks to be deployed with great density, dictates that there will be a long migration from current circuit-based networks and services to the all-IP environment. In addition, access networks and core networks may evolve at a different pace, requiring further interworking. Therefore, the interworking and migration between current network services and all-IP services is a key problem. Figure 3-1 shows an overview of a likely scenario in which multiple network types must interwork to provide seamless session-based services.

From a cellular telecommunications standpoint, the IP Multimedia Subsystem (IMS) [83] in all-IP mobile networks is the target network for future IP-based wireless services. The IMS can be considered a special instantiation of the SIP framework for Internet telephony service, in that the roles of various call control and service execution servers are clearly defined, and some interfaces that are unspecified in SIP have been specified for the

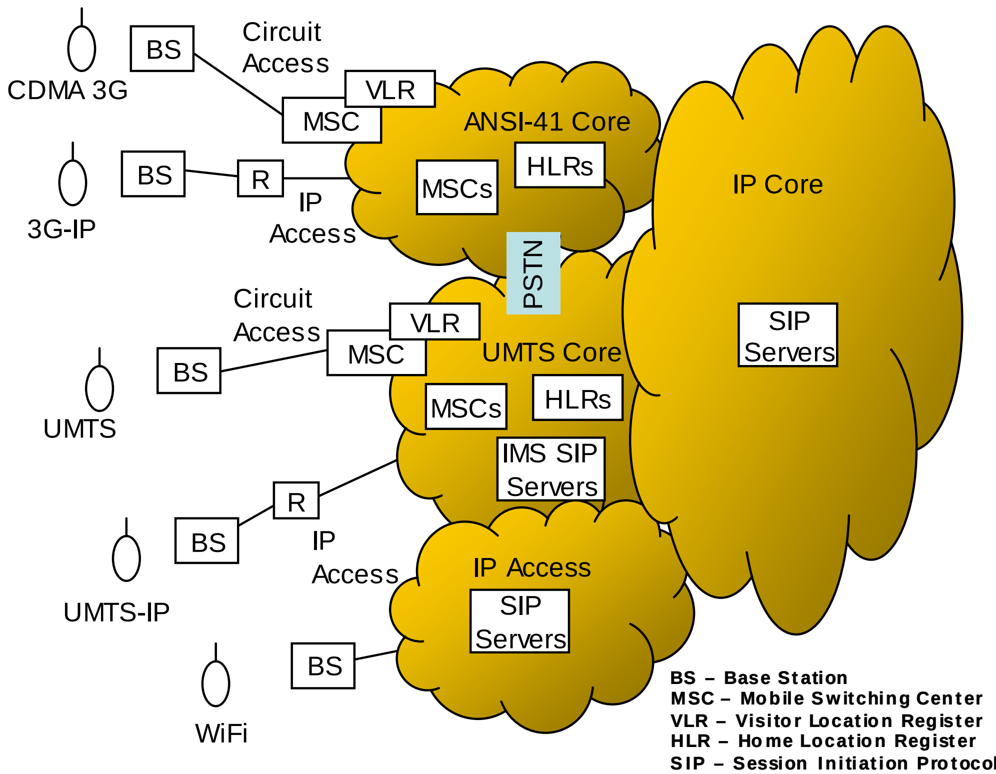


Figure 3-1. Heterogeneous network environment in the future

IMS architecture.

One approach to supporting seamless roaming between a mobile telephony network and emerging all-IP mobile networks is to use multi-mode terminals. Multi-mode terminals must satisfy two requirements. First, they must operate over multiple physical levels and low-layer protocols in order to access multiple systems. Second, they must support the application level signaling capabilities to request services from these networks. This is not always practical and may not be possible when considering a mix of mobile and fixed terminals. Another approach is to allow subscribers to use multiple devices that are accessible via a single user address. This may be accomplished by introducing a user mobility infrastructure into the network. The user mobility infrastructure enables users to switch between end devices and still get the same, personal services. The user mobility concept requires changing the traditional m-to-1 address-to-terminal relationship in cellular networks to a more general m-to-n relationship between addresses and terminals.

Some existing mobile network protocols support user mobility to a limited degree. The ANSI-41 flexible alerting and multiple access hunting features [84] allow 1-to-*n* address-to-terminal relationships; however, the destination terminals are confined to ANSI-41 or PSTN terminals. SIP allows forking to multiple destinations, but requires all signaling messages to be converted to SIP even if only a circuit switched network needs to be involved in the call set-up. In order to efficiently support user mobility in a heterogeneous network environment, a user mobility infrastructure is needed that can accept calls from any packet or circuit core network and allow users to roam into the entire range of mobile access networks.

In this thesis, we examine three approaches for supporting seamless roaming in a wireless, mobile environment; (1) master-slave mobility management approach, (2) federated mobility management approach, and (3) unified mobility management (UMM) approach. Emphasizing support for SIP terminals as our focus is on enabling the mobile Internet, we evaluate the performance of these three approaches through a comparative analysis using a technique similar to that in [18], [74] and [39].

In [18], the authors compared the performance of the UMM approach with a type of gateway approach for delivering SIP-originated calls to UMTS devices. In [39], the authors analyzed the signaling overhead for some interworking scenarios. However, in this thesis we consider the much more general case of multiple combinations of call originating and terminating devices. In addition to the call processing, we consider the signaling load for user's mobility. Moreover, we propose the load balancing scheme across the network elements so that each network element keeps constant utilization and thus the load of each element is balanced perfectly. As a result, we reduce the call delivery delay and hence get better performance.

3.2. Comparative Performance Analysis

3.2.1. Overview

In order to compare the performance of the three approaches: master-slave, federated, and unified mobility management, we choose the call setup signaling latency because this is the most important factor in the telecommunications signaling systems.

We first describe the methodology used to compare the performance of the different systems, and then discuss the analytical results. We refer to each approach as a *configuration*, and each mix of terminal types, e.g., call originated from an ANSI terminal to be delivered to a SIP device, as a *scenario*. Because the focus of this chapter is on providing mobile Internet services, we concentrate on the performance of calls delivered to SIP devices. We varied several parameters, shown in Table 3-1, including the offered call arrival rate (λ_{call}), offered mobility rate (λ_{mm}), call mix (*o_call_mix* and *d_call_mix*), and addressing type mix (user address or terminal address – *P(um)*).

For this evaluation we assume that an ANSI-41 network serves as the master network in the master-slave configuration, and as the main network in the federated configuration. We make this choice because cellular telecommunication networks are the dominant mobile networking technology today and will be the starting point for any interworking solutions. Likewise, we assume that only single-mode phones are used.

3.2.2. Example of Signaling Flows

Figures 3-2, 3-3, and 3-4 show the representative flows of SIP-terminated calls for the UMM, master-slave, and federated mobility management configurations, respectively. In this example, a user owns two types of terminals, one for ANSI-41 and another for SIP, and the latter is assumed to be active at the time when a call set-up request arises. It is further assumed that the telephone number of the user, namely a user address, belongs to the ANSI-41 network.

There are two main challenges to the UMM approach. First, the system must support multiple protocols and be easily extendable to accommodate new protocols. Second, the UMM must efficiently implement a protocol interworking mechanism with

Variable	Description
λ_{call}	Offered call arrival rate to system including calls of all types
λ_{mm}	Offered location update arrival rate to system including terminals of all types
o_call_mix	Percentage of calls originated from an ANSI-41 network
d_call_mix	Percentage of calls to ANSI-41 destinations
P(um)	Percentage of calls delivered via a user address (as opposed to terminal address)

Table 3-1. Analysis Parameters

any combination of the protocols. The *Common OPerationS* (*COPS*) architecture [27] addresses these issues. The heart of this architecture is the *COPS interface* that embodies the generalized mobility management methods that are common to many mobile networks. The COPS interface is defined between two key components in the COPS architecture, a *protocol-dependent logic server* (*PDLS*) and a *core logic server* (*CLS*). The architecture can support multiple types of networks with different protocols. For each network type, a specific PDLS is defined which terminates the respective protocol interface and implements protocol-specific service logic. When a PDLS receives a request for a service that does not require interworking, it provides the service directly without contacting the CLS. If the received request may *potentially* require interworking with other networks, the PDLS invokes the COPS interface to pass the control to the CLS. The CLS provides protocol-independent services and determines if interworking is necessary. If required, it communicates with the appropriate PDLS of a target protocol through the COPS interface to accomplish protocol interworking. In the event that there is no interworking, CLS uses the same COPS message to achieve the service, but sends the message to a PDLS of the same protocol type as the originating network.

In Figure 3-2, *ISUP IAM* is first delivered to the home MSC of the ANSI-41 network to establish the call to a mobile user. The home MSC then interrogates the ANSI-41 HLR (mobility manager) for the user location with an *ANSI-41 Location Request* message. Since the UMM is acting as the *ANSI-41* HLR, it receives this message. Internally, the location request message is received at the *ANSI-41* PDLS, translated to a *COPS Request Location* (RL) message, and sent to the CLS. The CLS then contacts the UMM database to determine the user associated with the called address, and selects one, or several of the

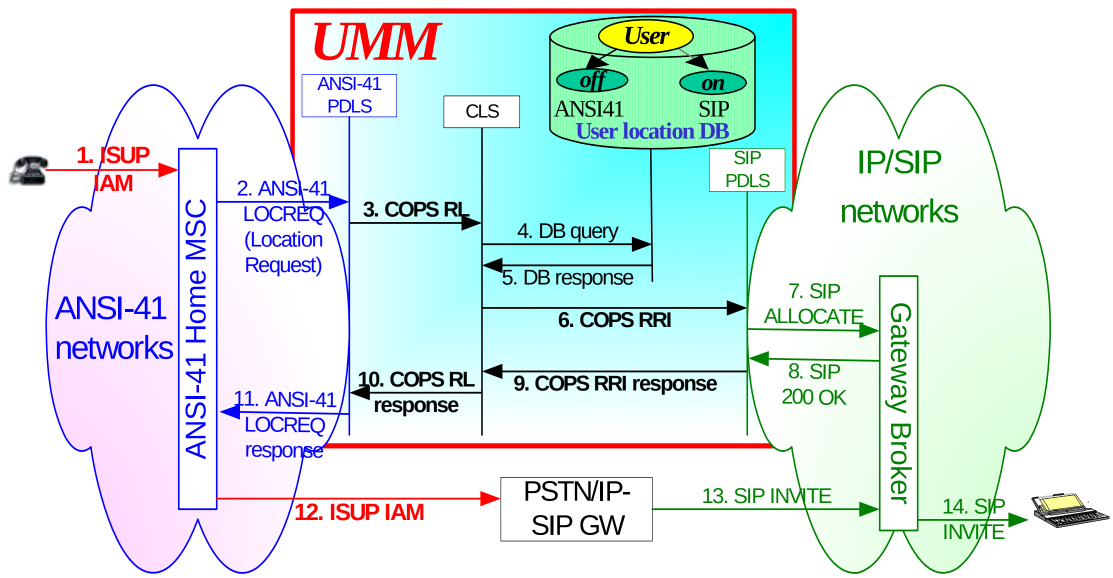


Figure 3-2. Call flow example of UMM approach: ANSI-41 to SIP

user's devices to terminate the call. In this example, the user's SIP terminal is selected and the current SIP contact universal resource identifier (URI) is retrieved from the UMM database. Refer to [18] for more details on UMM database support of user mobility infrastructure.

Because the originating network is a cellular network, which does not understand a SIP contact address, we must emulate the temporary number allocation process at the destination network. The temporary number allocation process allocates a short-lived routable telephone number to extend a call leg towards the destination network, e.g. serving MSC, in traditional circuit-switched cellular networks. The destination network must also maintain a mapping from the temporary number to the identity of the final destination, namely a SIP contact address in this cellular-to-IP interworking scenario. The *SIP ALLOCATE* method is proposed as an extension to SIP to accomplish this goal.

The CLS first sends a *COPS Request Route Information* (RRI) message to the PDLS of the destination network, namely the SIP PDLS. Then, the SIP PDLS selects an appropriate PSTN/IP-SIP gateway and sends a *SIP ALLOCATE* message to a *gateway broker* to retrieve a temporary routing number. A mapping from the routing number to the SIP contact address is also stored. The routing number is sent back to the SIP PDLS in a 200 OK response and is eventually delivered back to ANSI-41 home MSC via *COPS RRI*, *RL* responses and *ANSI-41 LOCREQ* response. The home MSC then sends an *ISUP IAM* using the temporary number as its destination to the PSTN/IP-SIP gateway. The gateway initiates a *SIP INVITE* message to the corresponding gateway broker, which completes the call delivery by sending a *SIP INVITE* message with an appropriate SIP contact address obtained from the mapping table.

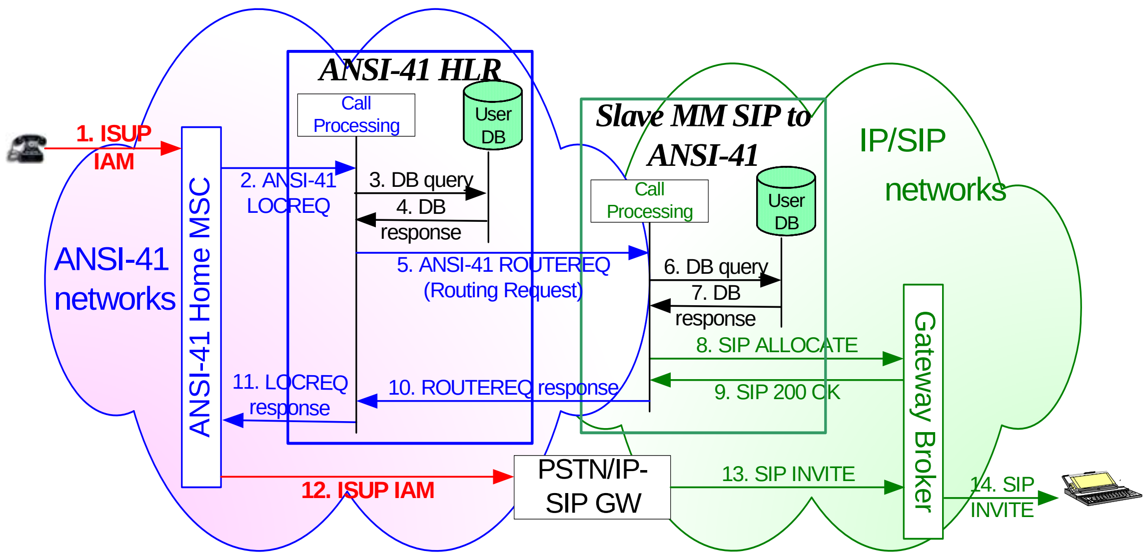


Figure 3-3. Call flow example of master-slave approach: ANSI-41 to SIP

In Figure 3-3, a call is placed from a PSTN device through an ANSI-41 core network to a mobile SIP device using the master-slave network configuration. The ANSI-41 home MSC acts as the originating MSC and terminates the standard ANSI-41 and ISUP protocols for location management and call control, respectively. The ANSI-41 HLR is the master mobility manager. The slave mobility manager terminates ANSI-41 from the master mobility manager, and translates the mobility management messages into SIP for the slave network. The PSTN/IP SIP gateway performs the dual function for the call control messages. The gateway broker and SIP terminal use the standard SIP protocols.

In Figure 3-4, the same call scenario is shown using a federated mobility management configuration. The User Mobility Component (UMC) accesses mobility management and service profile information directly from the ANSI-41 and SIP location server databases. The ANSI-41 home MSC uses standard ANSI-41 and ISUP protocols; likewise, the SIP location server, SIP Proxy and SIP terminal use standard SIP.

3.2.3. Network Modeling

We model each component in the network as an M/M/1 queuing system. Each message that arrives at a component, e.g., a MSC, will experience a certain queuing delay and service time. Combined, this is the sojourn time of a message through the network element. Summing the sojourn times of the messages required to complete a task at all elements they pass through yields the time taken to carry out the task. This method is based on that described in [74] with some simplifying assumptions. While we make simplifying assumptions, none affect the main goal of our analysis – comparing the performance of the various configurations under like conditions.

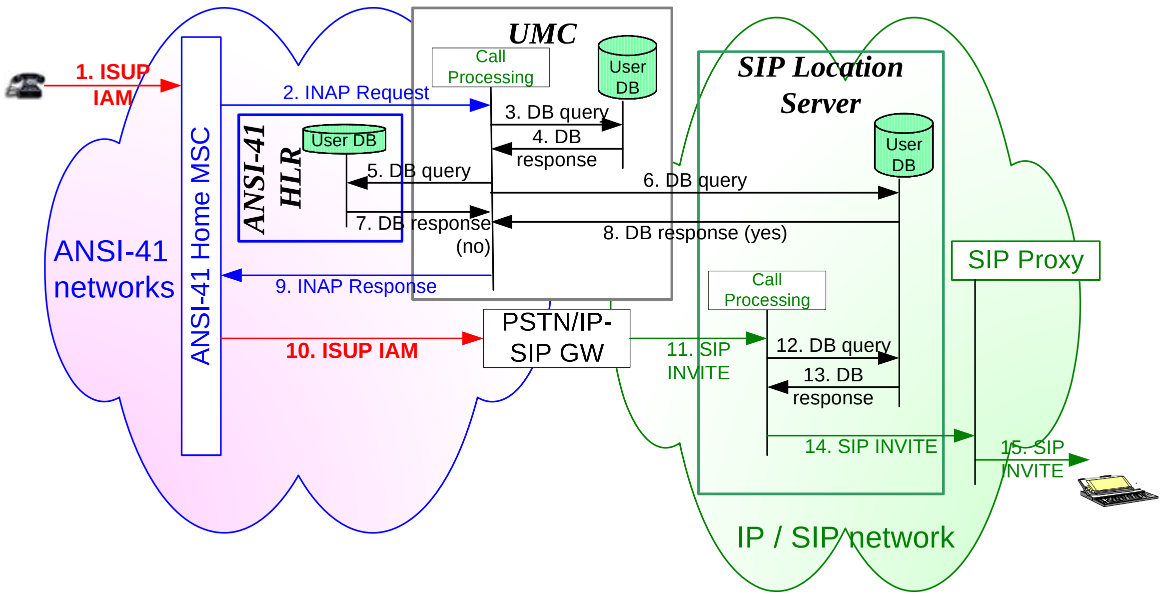


Figure 3-4. Call flow example of federated approach: ANSI-41 to SIP

By relaxing the assumptions stated below, the absolute values of the figures may change, but the relative performance of the configurations, and hence main conclusions, will not.

To determine the sojourn time of each message we first calculate the load on each element of the system. We assume that service requests (call establishment/release requests) arrive with a Poisson distribution. Because there is significant aggregation in the network between elements, we can assume that the arrival process at each element is also Poisson [74]. For simplicity we assume processing times are exponential. To determine the load on each element we developed extensive message flows for all possible call scenarios and configurations. These include procedures for call establishment for network originated calls, call release, and location updates. Note that when determining our results, we considered different mixes of call scenarios.

Consider network element *e* (e.g. MSC, ANSI-PDLS, CLS, etc.) and network configuration *c* (i.e. master-slave, federated, or unified). Let $M^{(e,c,m)}$ denote the set of messages processed at element *e* in configuration *c* for a specific mix of scenarios *m*. From [30], the utilization, $\rho^{(e,c,m)}$, of network element *e* in configuration *c* with scenario mix *m* is given by

$$\rho^{(e,c,m)} = \sum_{i \in M^{(e,c,m)}} \frac{\lambda_i}{\mu_i^{(e)}} \tag{3-1}$$

where λ_i is the mean arrival rate of messages of type *i* and $\mu_i^{(e)}$ is the mean service rate of message *i* for element *e*. The messages considered include those related to call establishment, release, and location update procedures. The rates for the call procedures are equal as we assume all calls are completed and hence eventually released. The rates for the location update procedures are set independently of the call procedures.

To determine the average call setup latency for each scenario, we need to know the average sojourn time of message *i* through element *e*, $E_e(T^i)$, which is given by

$$E_e(T^i) = \frac{1}{\mu_i^{(e)}(1-\rho^{(e,c,m)})} \quad \text{, with } \ i \in M^{(e,c,m)} \qquad \text{(3-2)}$$

For example, consider the UMM system shown in Figure 3-2. The flow for an ANSI originated call to a SIP terminal is shown. The time taken to deliver a call originated from the MSC to the SIP proxy is the time taken for the following messages to be processed: *ISUP IAM* (MSC), *ANSI-41 LOCREQ* (ANSI-41 PDLS), *COPS RI* (CLS), *DB Query* (User Location DB), *DB Response* (CLS), *COPS RRI* (SIP PDLS), *SIP ALLOCATE* (Gateway Broker), *SIP 200 OK* (SIP PDLS), *COPS RRI Response* (CLS), *ANSI-41 LOCREQ Response* (MSC), *ISUP IAM* (Gateway), *SIP INVITE* (Gateway Broker, a.k.a. a SIP Proxy). Therefore, we sum the sojourn times of these messages to determine the call delivery time. The service rates for each message and network element were determined through a combination of experimental measurement of an extensive system prototype and inference.

In order to eliminate the impact of specific processing platforms on our results, we use relative processing times in this analysis. To do this, we normalize the total processing times for all messages required in an MSC to originate and terminate an ANSI-

ANSI call in the unified mobility management configuration to be equal to one, i.e., $L^{(MSC,UMM,ANSI\_ANSI)} = 1$. We chose this element, configuration and scenario as our baseline because it is the current standard. We then normalize all other processing times to these values.

3.2.4. Load Balancing and Normalization

In a typical network, there will be a different number of each type of network elements. For example, a UMM-based network may contain *X* MSCs, *Y* ANSI-PDLSs, *Z* SIP-PDLSs, etc. In a properly engineered network, the ratio between the different network elements is chosen to have all elements equally loaded. For each configuration and scenario under consideration, we calculate the optimal ratio between all involved network elements in this respect. To do this we consider the overall processing load in each network configuration, and the processing load of each element in the network.

To balance the load across each element in configuration *c* we apply more processors to the elements that experience higher load. This has the effect of adjusting the message arrival rates at the individual elements. To determine the effective call arrival and mobility rates at each element *e* for configuration *c* under scenario *m*, while holding total processing resources constant across configurations, we first determine the average utilization of each element if perfect load balancing was achieved, $\rho_{ave}^{(c,m)}$, using

$$\rho_{ave}^{(c,m)} = \frac{\sum_{e \in EL^c} \sum_{i \in M^{(e,c,m)}} \frac{\lambda_i}{\mu_i}}{|EL_c|} \qquad (3\text{-}3)$$

We define $\lambda_{eff\_mm}^{(e,c,m)}$ and $\lambda_{eff\_call}^{(e,c,m)}$ to be the effective rate of mobility procedure arrivals and call arrivals at an element, respectively, when the load on the system is perfectly balanced. Therefore, we can also express $\rho_{ave}^{(c,m)}$ as

$$\rho_{ave}^{(c,m)} = \lambda_{eff\_call}^{(e,c,m)} \sum_{i \in M_{call}^{(e,c,m)}} \frac{1}{\mu_i} + \lambda_{eff\_mm}^{(e,c,m)} \sum_{j \in M_{mm}^{(e,c,m)}} \frac{1}{\mu_j} \tag{3-4}$$

where $M_{call}^{(e,c,m)}$ is the set of call related messages processed at element *e* in configuration *c* for scenario *m*, and $M_{mm}^{(e,c,m)}$ is the set of mobility related messages processed at element *e* in configuration *c* for scenario *m*. Note that the ratio of call arrivals to mobility procedure arrivals is maintained when the system is load balanced. This can be expressed as

$$\frac{\lambda_{eff\_call}^{(e,c,m)}}{\lambda_{eff\_mm}^{(e,c,m)}} = \frac{\lambda_{call}}{\lambda_{mm}} \tag{3-5}$$

We use equations (3-3) ~ (3-5) to determine the effective call arrival rates at each element. In order to do a fair comparison between the different network configurations, we have fixed the total processing cost (resources) for all scenarios. For this purpose, we first determine the processing load on network element *e* for configuration *c* with a scenario call mix m, $L^{(e,c,m)}$ by

$$L^{(e,c,m)} = \sum_{i \in M^{(e,c,m)}} 1/\mu_i^{(e)} \tag{3-6}$$

To get the total processing load for configuration c, $L^{(c,m)}$, we use

$$L^{c,m} = \sum_{e \in EL^{c}} L^{(e,c,m)} \tag{3-7}$$

where EL^{c} is the set of network elements employed in configuration c. We use the processing load for the UMM configuration, $L^{UMM,m}$, as our baseline, and hence normalize all other configurations to this load in order to compare networks of the same cost. To do this normalization, we calculate the processing factor, $PF^{(c,m)}$, for each configuration c using

$$PF^{(c,m)} = \frac{L^{UMM,m}}{L^{c,m}} \tag{3-8}$$

Finally, we apply the normalization factor to $\lambda_{eff\_mm}^{(e,c,m)}$ and $\lambda_{eff\_call}^{(e,c,m)}$ to obtain the normalized effective rate of mobility procedure arrivals and call arrivals, $\overline{\lambda_{eff\_mm}^{(e,c,m)}}$ and $\overline{\lambda_{eff\_call}^{(e,c,m)}}$, respectively, as

$$\overline{\lambda_{eff\_mm}^{(e,c,m)}} = \frac{\lambda_{eff\_mm}^{(e,c,m)}}{PF^{(c,m)}} \tag{3-9}$$

and

$$\overline{\lambda_{eff\_call}^{(e,c,m)}} = \frac{\lambda_{eff\_call}^{(e,c,m)}}{PF^{(c,m)}} \tag{3-10}$$

These rates are used to determine the message arrival rates in equation (3-1) by considering the number of messages per call and mobility procedure arrival. Using equation (3-2), and the derived message flows, we determine the mean sojourn time for each element in the system. We then determine the time taken to deliver a call to a mobile user by summing the sojourn times of each message required to carry out this task for each element.

3.2.5. Analysis Results

We first examine the impact of load balancing across network elements. Due to the network hierarchy of cellular networks, several network elements have relatively high processing load compared to the other elements and hence are easily overloaded as the call or mobility rate increases. For example, Figure 3-5 (a) shows the utilization of each network element when using the UMM approach with incoming and outgoing call rate equal to 0.5 and fixed mobility equal to 0.31. We see that the utilization of the ANSI-MSC is relatively higher than the other elements before load balancing. Therefore, the average sojourn time of signalling message through ANSI-MSC must increase rapidly as the offered call rate increases, and hence has severe impact on the average call setup latency of whole network.

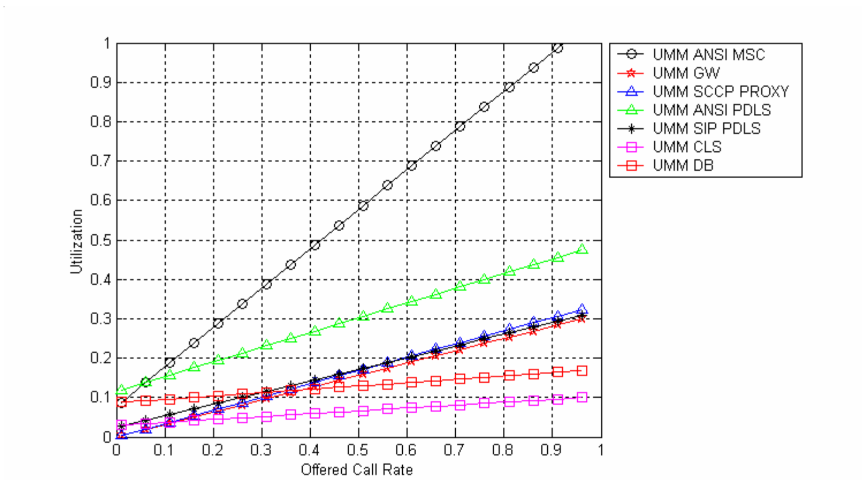
In order to achieve perfect load balancing across each network element, we determine the average utilization of each element and determine the effective call arrival rates at each element in equation (3) – (5) so that all elements have the same average utilization as shown in Figure 3-5 (b) and hence the load of each element is balanced perfectly.

Figure 3-6 shows the impact of the load balancing. Each approach is based on different network architecture and each network element processes a different number of signaling messages according to the architecture. Thus, some approaches have highly loaded network elements and some approaches don't. In Figure 3-6, we see that the perfect load balancing is useful for effectively reducing the average call setup latency of the approaches which have highly loaded network elements.

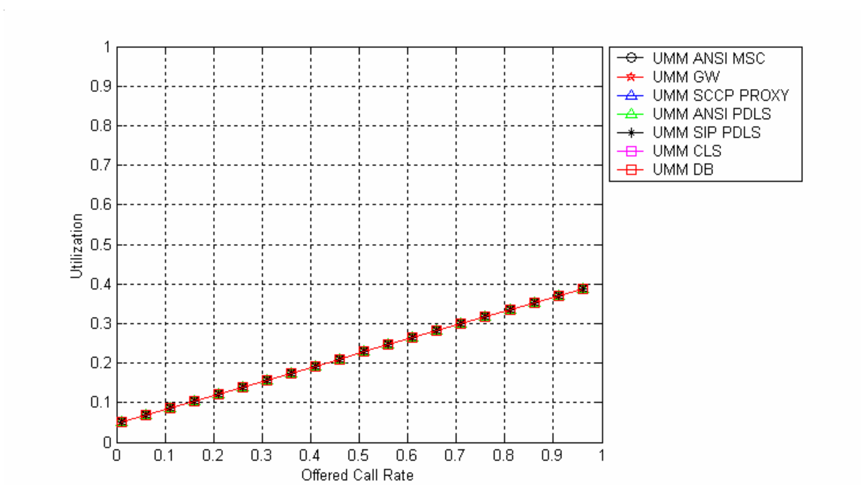
Based on a perfectly load-balanced network, we measure the call setup latency of the SIP terminated calls in each approach varying the parameters in Table 3-1. Interworking occurs whenever a call is terminated on a different network type than on which it was originated. Informally, the percentage of calls requiring interworking is simply the percentage of ANSI-41 originated calls terminating on SIP devices plus the percentage of SIP originated calls terminating on ANSI-41 devices. More formally, the percentage of interworking, *I*, is:

$$I = o\_call\_mix(1 - d\_call\_mix) + (1 - o\_call\_mix)(d\_call\_mix) \qquad (3\text{-}11)$$

Figure 3-7 shows the impact of mobility on the different configurations for the case of devices being equally divided between SIP and ANSI-41. There is little impact on the performance of all the scenarios, with the performance of all the systems improving as the mobility rate decreases, i.e., more calls are made per move.



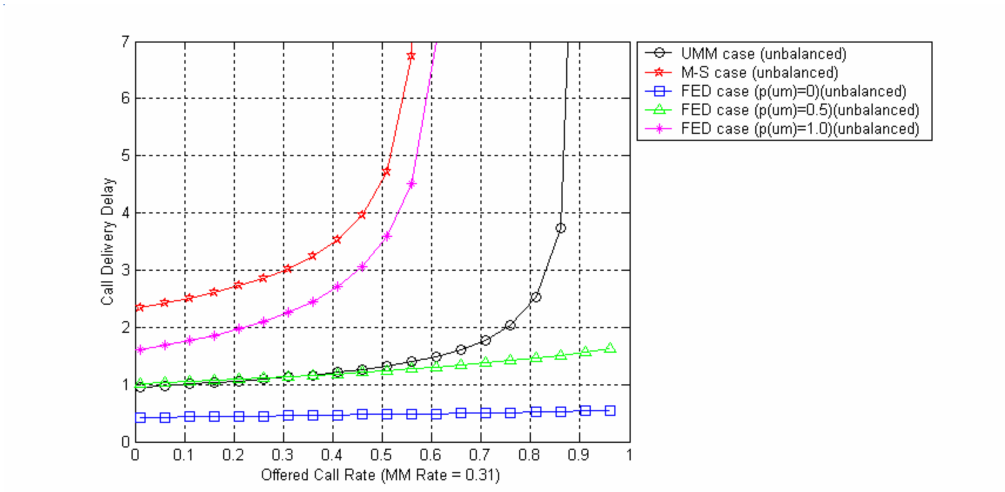
(a) Before load balancing



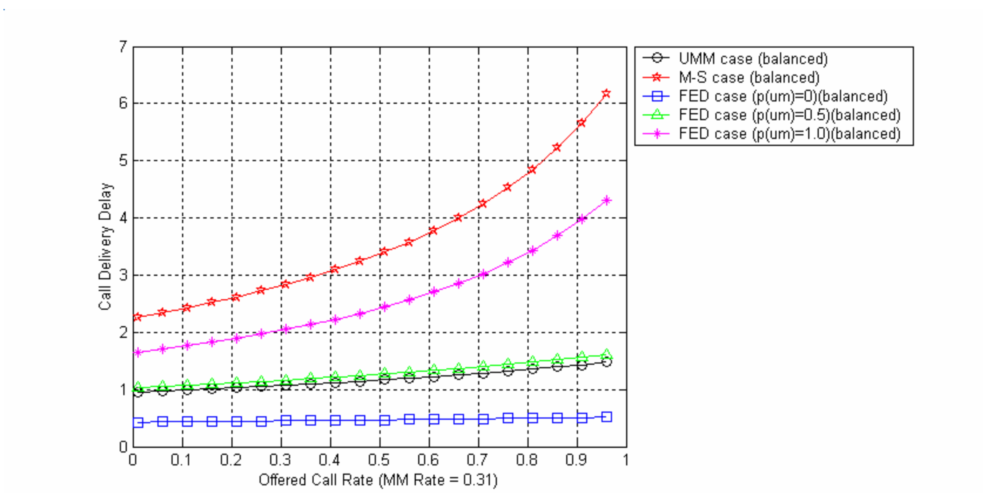
(b) After load balancing

Figure 3-5. Utilization of network elements in UMM approach

(o_callmix = d_call_mix = 0.5, mobility_rate = 0.31)



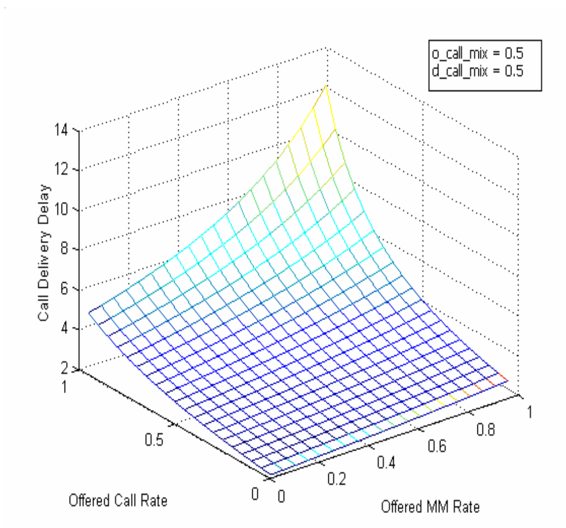
(a) Before load balancing



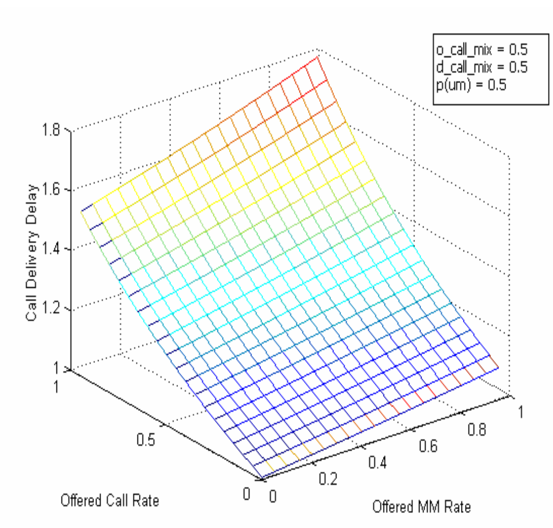
(b) After load balancing

Figure 3-6. Average call setup latency of each approach

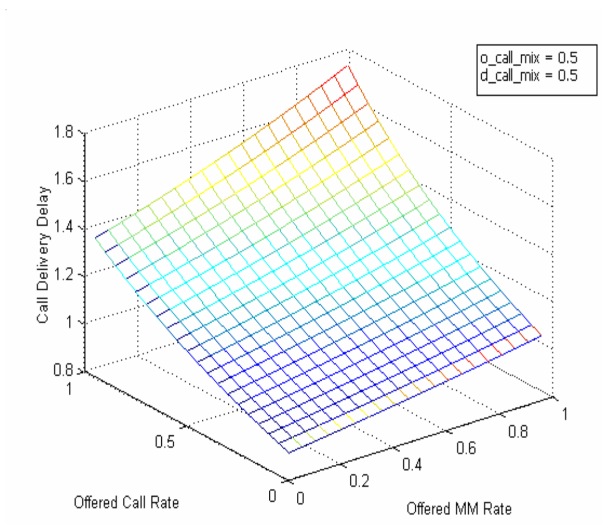
(o_callmix = d_call_mix = 0.5, mobility_rate = 0.31)



(a) Master-Slave



(b) Federated



(c) UMM

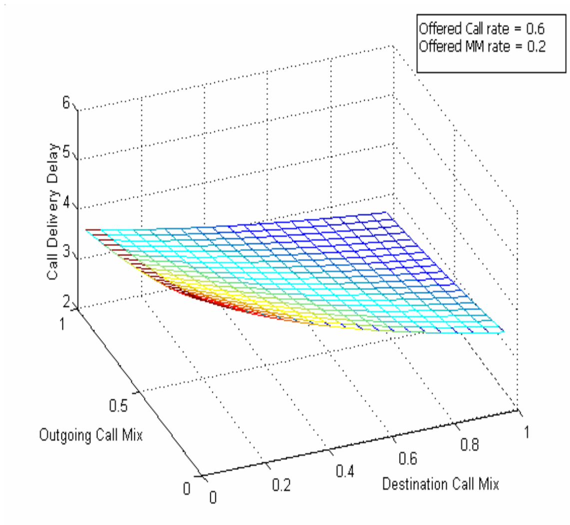
Figure 3-7. Call delay vs. call mobility ratio

Figure 3-8 shows the impact of the call mix on the different configurations when none of the systems are in an overloaded state. Recall from Table 3-1 that *destination call mix* refers to the percentage of ANSI-41 terminated calls, and *originating call mix* refers to the percentage of ANSI-41 originating calls.

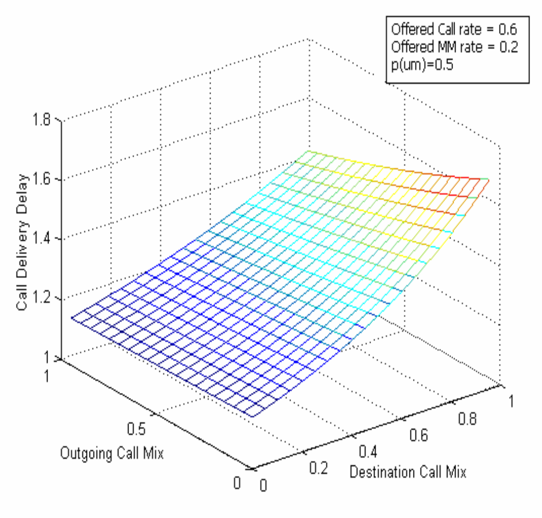
Figure 3-8 (a) shows the performance of the master-slave configuration. We see that this configuration performs poorly when most calls are delivered to SIP terminals (low values of destination call mix). That is because interworking is required with the core ANSI-41 network which results in significant messaging overhead. The master-slave approach is more efficient for ANSI-41 terminated calls (high destination call mix) because the call is being delivered via the master protocol and hence less messaging is required.

Figure 3-8 (b) shows the performance of the federated configuration. In this configuration, ANSI-41 terminated calls require incur the most delay because of the extra processing associated with ANSI-41. The most overhead is incurred when all calls are originated by SIP users and delivered to ANSI-41 terminals because of the interworking required and name translation. Figure 3-8 (c) shows the performance of the UMM configuration. The UMM is least efficient when the system is pure ANSI-41 because of the high processing cost of this protocol. It is more efficient for SIP originated calls because it is effective at performing address translations.

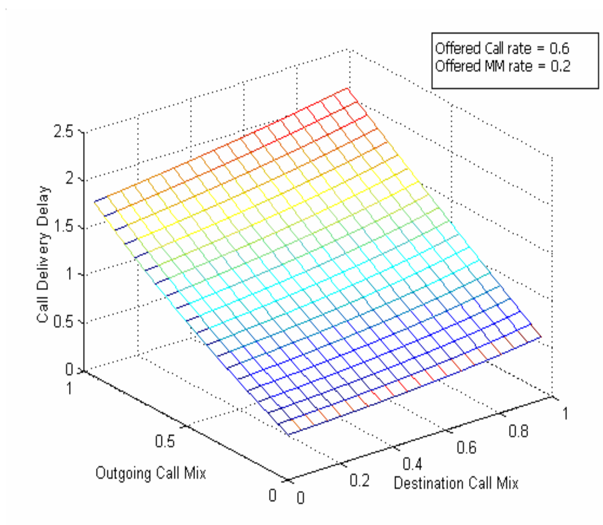
Figure 3-9 shows the performance difference between the UMM and federated configurations. The surface represents the values of call delay of the UMM configuration minus the call delay of the federated configuration. Therefore, when the value on the vertical axis is below 0, the UMM configuration is more efficient. As shown in the figure, even if the performance of the federate configuration improves as more SIP terminals



(a) Master-Slave



(b) Federated



(c) UMM

Figure 3-8. Call delay vs. call mix

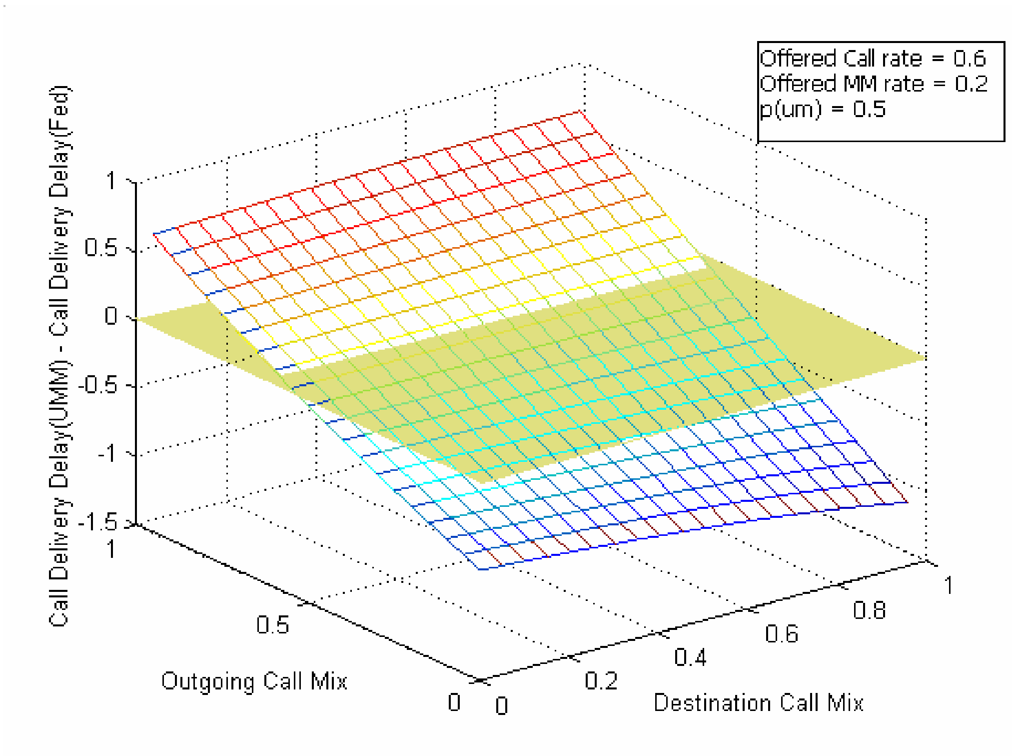


Figure 3-9. Comparison of UMM vs. federated mobility management

are involved (Figure 3-8(b)), the UMM outperforms the federated approach as more SIP terminals are involved in calls. This is because the federated system requires significant messaging via the UMC to coordinate the networks (see Figure 3-4) which reduces the benefit of the simplified SIP processing.

3.3. Concluding Remarks

Given the current state of wireless networks, and the investment already made in current telecommunication networks, the migration to all-IP networks will occur over many years. Therefore, the systems must be evaluated on their ability to be easily introduced into the current networks, provide efficient interworking during the migration period, and eventually efficiently support services in a predominantly IP environment.

We consider three interworking approaches to support session-based IP applications. The master-slave approach is easy to introduce into existing systems because, while it requires new equipment to be deployed in new networks, most existing equipment is untouched. But, its performance is not as efficient as the other approaches. The federated approach is very efficient when there is a single dominant network protocol and little data access is required to deliver a call. However, this advantage dissipates as more data access is required, for example to resolve a user address or to access data to provide interworking functions, such as determining which of multiple possible terminals or networks to route a call. In addition, due to the nature of a federated database system, user data tends to be spread over multiple databases which require synchronization and perhaps multiple provisioning systems. The unified mobility management approach is very efficient for cases in which interworking and user mobility support (e.g., address resolution) is required. But, the cost of this approach is that, if the services are not required, there is some extra overhead. The UMM has other benefits for deployment as well. Because the UMM supports native interfaces to several networks, and handles all interworking functions, it is the only new network element that needs to be introduced into a network to enable roaming; no other elements require modification.

We compare the performance of these three approaches through a comparative analysis using a technique. In order to get better performance, the perfect load balancing across network elements is also considered.

In summary, the UMM approach is highly suitable for supporting the migration to all-IP mobile networks because it implements the basic functions to support interworking in an integrated fashion. This approach may have limitations as more advanced services are introduced and multiple providers provide services on a single call, because the management of data in a single database may become impractical. In these cases, the combination of using the UMM for basic session services and interworking, and a federate approach for providing access to other service-specific data, may be an attractive solution.

Chapter 4. Network Formation and Resource Allocation for Multi-Frequency, Multi-Hop Wireless Cellular Networks

4.1. Introduction

Mobile nodes in traditional wireless cellular networks communicate through centralized base stations (BS) in a pre-defined spectrum. To improve the performance of such cellular networks, several studies on multi-hop wireless cellular networks have been undertaken [1][24][40][46][75]. We investigate relaying in a third generation (3G) wireless system, CDMA2000 1xEVDO [6][28]. In the 1xEVDO system, the bit rate achieved during each time interval depends on its signal quality to the mobile node. If the BS can schedule nodes with better signal quality more often, then a higher average bit rate for the network can be achieved.

In this chapter, we envision a multi-hop wireless cellular network that uses 1xEVDO scheduling [67]. Multiple relay networks are dynamically formed when performance on the radio access network is degraded. A disjoint frequency band for each relay network is allocated dynamically by the BS. In this way, multiple non-interfering relay networks may operate in parallel.

The dynamic nature of these relay networks motivates the need for an explicit procedure for mobile nodes to form a relay network. Moreover, it requires that every mobile node be able to communicate over a wide range of frequency bands. We assume that each mobile node is equipped with an agile radio [16] in addition to a cellular interface to meet this requirement. The cellular interface is leveraged so that the BS may broadcast information during relay network formation.

We also allow the band used by a single relay network to be divided into multiple orthogonal frequencies to construct a relay network comprised of non-interfering links. This allows multiple nodes within communication range of each other to transmit simultaneously without relying on a MAC protocol or distributed scheduling algorithm to resolve contention and prevent collisions.

In this chapter, we present three relay network formation algorithms. Each algorithm first determines gateway (GW) nodes which are best suited for acting as a bridge between the relay network and the BS. The algorithms then discover a path from each node through the relay network to the GW node. Each algorithm also provides a simple and distributed frequency allocation scheme to build relay networks with non-interfering links to improve network throughput.

We also propose two enhancements to increase network throughput of resulting relay networks. In the first enhancement, each node detects interference from remote nodes based on its received Signal-to-Noise-plus-Interference Ratio (SNIR), and may dynamically switch channels to reduce interference. Second, we support GW reselection so that a new GW with a higher achievable data rate may be used.

We compare these algorithms in terms of the overhead of the relay network formation. We also measure the throughput of the resulting relay network in three scenarios: (1) relay network using a single frequency; (2) relay network using multiple frequencies assigned by our distributed frequency allocation; and (3) relay network using multiple frequencies with optimal assignment. We also quantify the benefits of the enhancements.

The results lead us to conclude that having nodes outmost from the BS initiate route discovery first is the best approach for reducing the relay network formation overhead. The results also show that using simple and distributed frequency allocation we can achieve high throughput gains over using networks that uses only a single frequency. This simple frequency assignment algorithm achieves 80-85% of the optimal average throughput. Moreover, they show that the enhancements further improve average throughput.

The rest of this chapter is organized as follows. In section 4.2, we briefly discuss related work in this area. In section 4.3, we present the network model for dynamic multi-frequency, multi-hop wireless cellular networks. Basic operations used for relay network formation and frequency allocation are explained in section 4.4. In section 4.5, we describe two enhancements to increase network throughput of resulting relay networks. In section 4.6, we present our simulation environment and results. In section 4.7, we discuss the related issues of this chapter. We conclude this chapter in section 4.8.

4.2. Network Model

In this chapter, we focus on a single cell environment in which there is a BS and several mobile nodes. In Figure 4-1, seven mobile nodes communicate with the BS initially. The BS advertises a new frequency band, B_{r1}, available on which a relay network is formed. Then, the nodes form a relay network r_1 operating on B_{r1}. At some time later, the BS advertises frequency band B_{r2} on which a new relay network r_2 is formed.

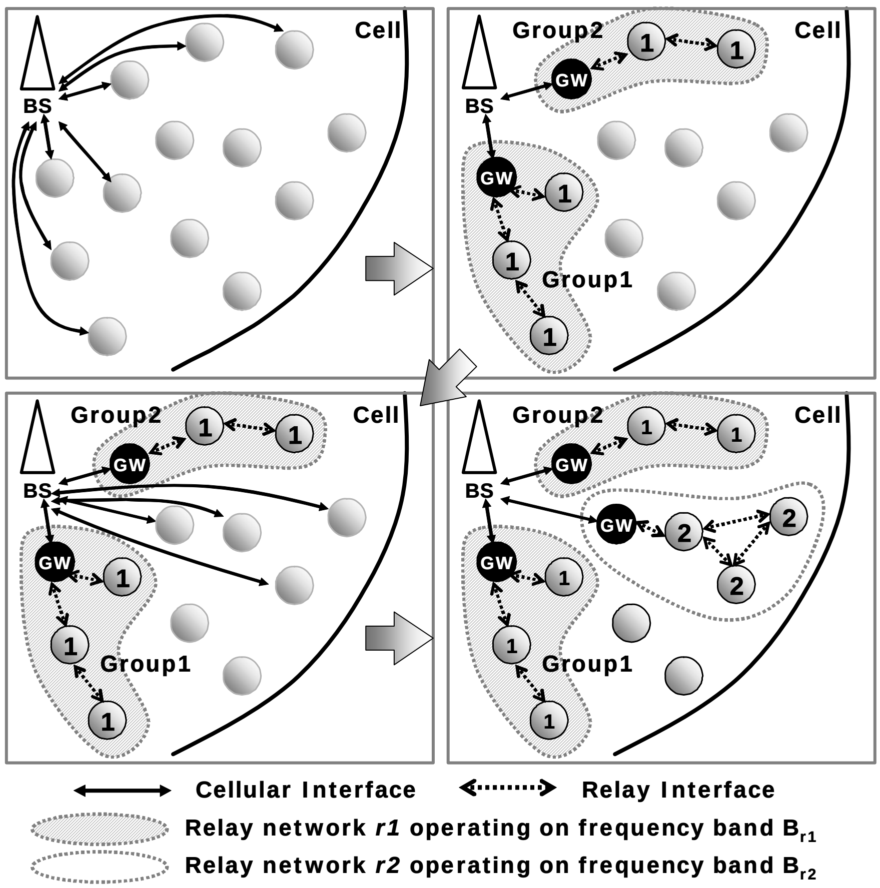


Figure 4-1. Example of relay network formation

4.2.1. Frequency Band Allocation

A frequency band is allocated by the BS whenever a relay network is formed. Let S be the set of all available radio frequency bands, $S = \{B_1, B_2, \ldots, B_N\}$, and R be the set of relay networks currently operating, $R = \{r_1, r_2, \ldots, r_R\}$. If B_{ri} and B_{rj} are frequency bands on which relay network r_i and $r_j \in R$ operate, respectively, they should satisfy the requirements: (1) $B_{ri} \in S$ and $B_{rj} \in S$, (2) $B_{ri} \neq B_{rj}$.

4.2.2. Frequency Allocation

Each frequency band may be divided into multiple orthogonal frequencies. For example, a band B_{ri} consists of the set of orthogonal frequencies, $F_{ri} = \{f_1, \ldots, f_K\}$, where K is the maximum number of orthogonal frequencies in B_{ri}. These frequencies may be used to construct a relay network comprised of non-interfering links so that multiple nodes within range of each other may transmit simultaneously without relying on a MAC protocol or distributed scheduling algorithm to resolve contention and prevent collisions.

4.2.3. Dynamic Formation and Dissolution of the Relay Network

The formation and the dissolution of each relay network are performed through the following steps:

- If the BS schedules a large number of nodes that have poor signal quality, it may advertise a frequency band on which a relay network may be formed. The frequency advertisement is broadcast over the cellular control channel. This information includes all available orthogonal frequencies in the band and a control channel used for the relay network formation.

- Some mobile nodes form a relay network operating on the introduced band.
- Once the relay network is created, the BS schedules only the GW nodes for transmission. Nodes in the relay network communicate with the BS via the GWs.
- At a later time, the relay network may be dissolved and nodes return to using the cellular interface with the BS.

In this chapter, we present the relay network formation algorithms used in step (2) and compare them in terms of the overhead of the relay network formation. We also measure the network performance of the resulting relay network in step (3).

4.3. Network Formation and Frequency Allocation Algorithms

Each relay network is formed in two phases. In Phase I, GW nodes are chosen for each group. A group consists of a set of mobile nodes which are reachable to each other. The transmission radius of a node in the relay network (e.g. 250m in 802.11) is very small compared to the cellular coverage (e.g. 10~20 km). Thus a relay network generally consists of several isolated groups of mobile nodes. In Figure 4-1, the relay network *r1* consists of two isolated groups of mobile nodes. Each group needs at least one GW node to act as a bridge between the BS and the group.

Phase II of relay network formation consists of two steps. In the first step, the nodes join the relay network by establishing a path to one of the GWs. In the second step, while returning a route reply (RREP) message to the source node of a route request (RREQ), the GW and intermediate nodes on the reverse path assign orthogonal frequencies to the links on the path.

Since the relay networks are dynamically formed and dissolved in our network, we use a modified version of AODV [56] as the ad-hoc routing protocol to find the path from the mobile nodes to the GW. In this modified AODV, the RREQ contains path information like DSR [29]. Each intermediate node appends its identification to the RREQ before forwarding it. Thus, upon receiving the RREQ, the GW node can learn the members of specific groups within the relay network.

We leverage two optimization features of reactive ad-hoc routing protocols. First, a node may passively learn a route to a destination if it is part of a longer path to the destination. In this case, it will not launch its own RREQ. Second, a node that has previously learned a route may immediately return this route in response to a request without a further search.

These two features can greatly reduce the number of messages flooded to find routes to the GWs. In order to make the utmost use of the passive route learning, intuitively the furthest node from the BS is the best choice to launch a route request first. This will greatly reduce the load at the GW node. To fully leverage the immediate response to a RREQ, scheduling the nodes nearest the BS to launch a RREQ first is the best choice.

Motivated by these observations, we propose the three relay network formation algorithms which use node's location information - two centralized algorithms: Furthest First (FF), Nearest First (NF); and a distributed algorithm: Locally Outmost First (LOF). The algorithms dictate the scheduling by which nodes send out their own RREQ.

4.3.1. Phase I – GW Discovery

In our environment, the BS broadcasts the frequency band information on which a relay network is formed over the cellular control channels so that all nodes within the cell receive it simultaneously. A pre-agreed upon channel within the relay network frequency band is defined as the signaling channel used to establish and maintain the relay network.

To select GW nodes, every node forming a new relay network periodically broadcasts a neighbor advertisement (NADV) message over the control channel as soon as it receives new frequency band information. The NADV contains the identification of the sending node, and a metric indicative of the received signal quality from the BS, and the distance of the node from the BS.

```
GW_Discovery( )
---------------------------------------------------------
1: D_i = Distance from the BS of node i ;
2: Receive NADV from all neighbors;
3: D_k = Distance from the BS of neighbor k;
4: If ( D_i == min (D_i , D_k) ) for all neighbors k,
5:     then node i acts as a GW ;
```

Figure 4-2. GW discovery algorithm

Whenever a node receives a NADV message from its neighbors, it compares its signal strength to the BS with their own. If the node has better signal quality to the BS compared to all one-hop neighbors, it acts as a GW node as shown in Figure 4-2. If several one-hop neighbors have the same signal quality, one may be selected according to additional metrics such as processing power, battery lifetime, etc.

4.3.2. Phase II – Joining the Relay Network

In this section we discuss the procedure by which nodes join a relay network. This consists of two steps: initiating route discovery and assigning frequencies on each link in the relay network.

4.3.2.1. Step 1: Initiating a Route Discovery

The difference among our various network formation algorithms is the schedule by which nodes initiate route discovery.

Baseline(BL) **–** As a baseline, we consider the case in which every node joining the relay network initiates a route discovery just after GW discovery phase without any specific scheduling for initiating route discovery. This is very simple, but runs the risk of severe congestion on the relay network during the formation phase because every node sends out *RREQ* almost simultaneously. It may also overload the GW node with many *RREQ*s during a short time period.

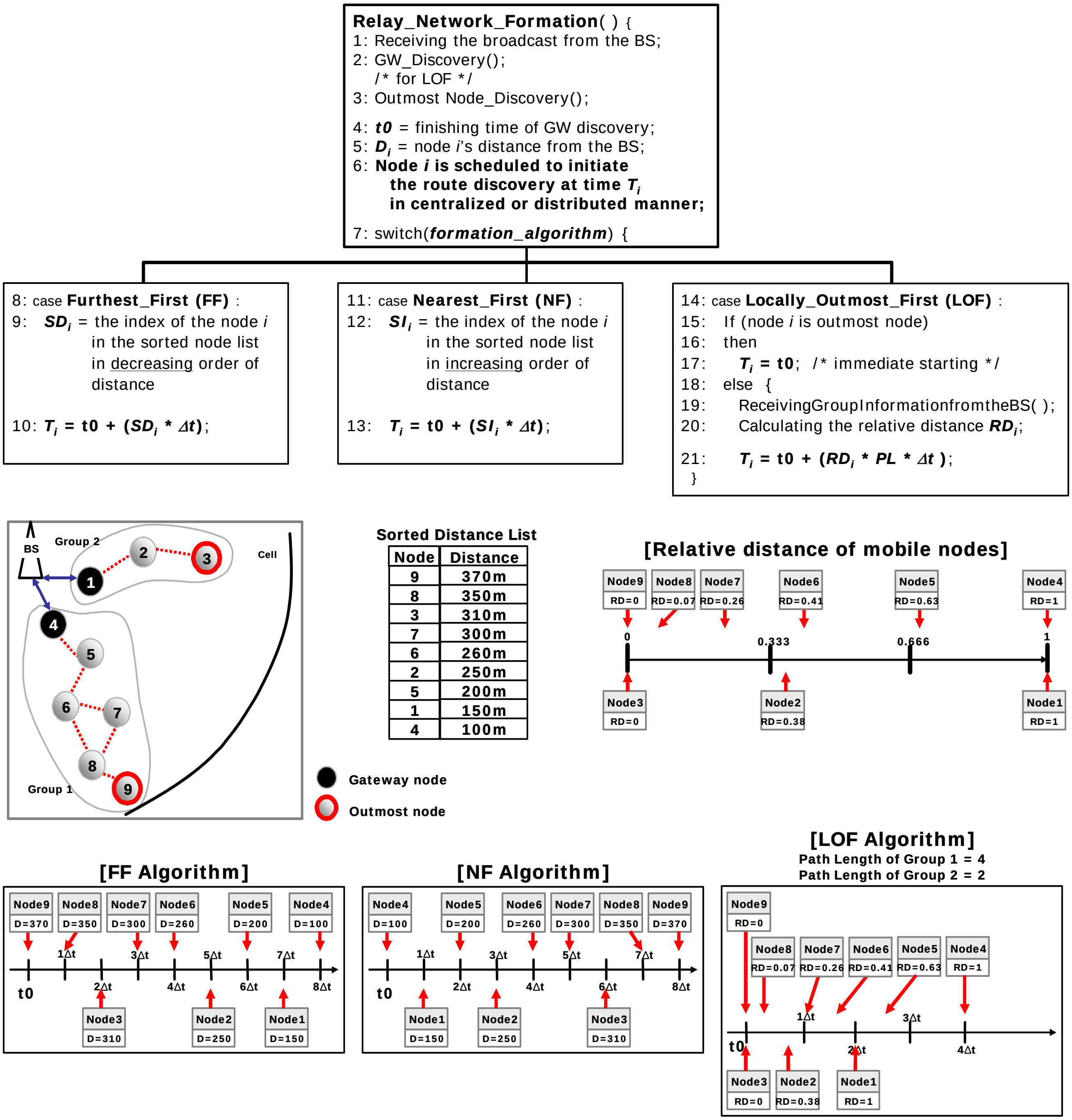


Figure 4-3. Formation algorithms and example of scheduled initiation of route discovery

<u>Furthest First(FF)</u> **–** To solve the problems of BL, we exploit the ability of nodes to passively learn routes. This optimization reduces both the congestion caused by the potential *RREQ* storm in the relay network and the load on the GW. From this observation, the best schedule is to have the nodes furthest from the BS launch *RREQ*s first, so that as many nodes will passively learn a route to the BS as possible.

Note that the formation algorithm schedules the *initiation* of each node's route discovery. It is impossible to anticipate which node will be located at the end of the longest path from the GW nodes until all nodes complete the route discovery. However, the nodes furthest from the BS generally have the greatest possibility to have the longest path from the GW nodes. This algorithm thus forces the furthest node to launch *RREQ* first.

There are two drawbacks with this algorithm. First, it is impractical for a node to learn the exact location of all other nodes. We therefore assume that the BS may act as a central controller and track each node's location in the cell. The BS forces the node furthest from it to launch a *RREQ* first by sending a pre-defined signal to the node over the cellular interface.

Second, if a node is not on the path between the furthest node and the BS, it will not passively learn a route. However, it is impossible to anticipate which nodes may or may not passively learn a route to the BS in advance. To ensure that all nodes learn a route, the BS schedules each node to launch its own *RREQ* in decreasing order of distance from the BS at every certain time interval, Δt. Thus, if a node passively learns a route before its scheduled time, the time slot assigned to the node is wasted. This introduces unnecessary latency into the network formation process.

As shown in Figure 4-3, node 9 (furthest from the BS) is scheduled to send out its *RREQ* at time *t0;* node 8 (the next furthest node) is scheduled at $t0+\Delta t$, and so on. After node 9 launches its *RREQ* at time *t0*, nodes 5, 6, and 8 may passively learn a route to the BS. Thus, if they receive the *RREP* before $t0+6\Delta t$, $t0+4\Delta t$, and $t0+\Delta t$, respectively, they will not send out their own *RREQ*s and the time slots assigned to the nodes are wasted.

Nearest First(NF) **–** When a node already having a path to the GW receives a *RREQ*, it can immediately return a *RREP* to the source node. In so doing, the majority of the *RREQ* traffic is replied to without being forwarded to the GWs. Therefore, having the node nearest to the BS initiate a *RREQ* first will also lower load on the network. As with FF, this scheme requires the BS to act as a central controller.

For example, in Figure 4-3, if node 6 has a route to the BS before receiving the *RREQ* generated by node 9 at time $t0+8\Delta t$, it returns a *RREP* to node 9 immediately.

In addition to requiring a centralized controller, this algorithm has the additional drawback that almost all nodes must launch a *RREQ*, at least to reach the node preceding them to the GW.

Locally Outmost First (LOF) **–** FF and NF reduce the number of signaling messages flooded in the network. However, mobile nodes may experience long latency due to the strict, sequential scheduling. They are, moreover, not practical because the centralized scheduler must know the distance to all nodes. We present a *distributed* algorithm, called *Locally Outmost First*, to achieve the benefits of FF and NF, and the potential low latency of the baseline case. Without any centralized controller, this algorithm allows each node to make a schedule for its own route discovery based on its distance and the group information broadcast by the BS.

This algorithm is composed of two steps - outmost node discovery and scheduling. A relay network consists of several isolated groups of mobile nodes as shown in Figure 4-1; each group can form the relay network independently. In this algorithm, all outmost nodes in *each group* initiate a route discovery first just after outmost node discovery. Thus, several paths will be discovered simultaneously. The *outmost node* has the greatest distance from the BS compared to all neighbors within its transmission range. Each group has at least one outmost node. For example, nodes 9 and 3 act as the outmost node in groups 1 and 2, respectively, in Figure 4-3.

(1) Outmost Node Discovery

This procedure can be overlapped with GW discovery. During the GW discovery, each node compares its distance from the BS with all neighbors' distances. As shown in Figure 4-4, if a node has a greater distance compared to all neighbors, it becomes an outmost node.

```
Outmost_Node_Discovery( )
1: D_i = Distance from the BS of node i ;
2: D_k = Distance from the BS of neighbor k;
3: if ( D_i == max (D_i , D_k) ) for all neighbors k,
4:     then node i acts as a outmost node;
```

Figure 4-4. Outmost node discovery algorithm

(2) Scheduling

The GW node initially delivers the group information to the BS when it receives *RREQs* sent from the outmost nodes which passed through a new path. The BS then broadcasts the group information so that mobile nodes in the cell exploit it. The group information contains the ID of the GW, the distance of the GW (D_G), the ID of the outmost node that sent the *RREQ*, the distance of the outmost node (D_O), and the path length (*PL*).

With the characteristic of flooding, all nodes in the group receive at least one *RREQ* sent from the outmost nodes. If a node receives a *RREQ* from an outmost node and the BS broadcasts group information including the outmost node and a GW, the node belongs to the same group as the outmost node and the GW.

Scheduling ( )

1: D_i = Distance from the BS of node *i*;
2: D_G = Distance from the BS of GW;
3: D_O = Distance from the BS of the outmost node;
4: PL = Path length = hop count between the outmost node and the GW;
5: Relative distance of node *i*

$$RD_i = 1 - \left(\frac{D_i - D_G}{D_O - D_G} \right) \tag{4-1}$$

6: T_i = Scheduled time of node *i*'s route discovery
$= t0 + (RD_i \times PL \times \Delta t)$

Figure 4-5. Scheduling of route discovery algorithm

Moreover, the node may be on one of the paths between the outmost node and the GW. Thus, based on the broadcast group information and the information in received *RREQ*, each node can calculate its *relative distance* between the GW node and the outmost node. It then makes a schedule for its own route discovery with the relative distance as shown in Figure 4-5. The relative distance has a value in the range [0..1]. A smaller value indicates that the node is closer to the outmost node.

As shown in Figure 4-3, since nodes 3 and 9 are outmost nodes, they send out *RREQs* simultaneously at time *t0*. Then, the intermediate nodes can calculate their relative distances based on the group information. For example, node 6 belongs to group 1 in which the GW node and the outmost node are nodes 4 and 9, respectively. Node 4's distance from the BS is 100m and node 9's distance is 370m. Thus, the relative distance of node 6 is 0.4. The path length is 4. Thus, node 6 makes a schedule for its route discovery at time $t0+1.6\Delta t$. If it does not passively learn a route before $t0+1.6\Delta t$, it sends out its own *RREQ* at the scheduled time. Since this scheduled time is inversely proportional to its relative distance, this algorithm makes use of the passive route learning.

4.3.2.2. Step 2: Assigning Frequency

Each node establishes a single path to a GW node. The GW node returns a *RREP* to the source node on behalf of the BS. While returning the *RREP*, the GW and intermediate nodes on the reverse path are responsible for assigning a non-interfering frequency to links on the path. In this section, we present a simple distributed frequency allocation algorithm.

We define the Used Frequency Information (*UFI*) of a node as the set of frequencies used on its all incident links. To make a local frequency allocation when a path is being established, a node requires the *UFI* of all neighbors. This information is received in two ways: first, the *NADV* messages periodically broadcast include the *UFI* of a node. Second, the *UFI* is included in the *RREP* generated by a node. Figure 4-6 shows the algorithm and an example of the frequency allocation on a path.

Each node i in the relay network r_i maintains a set of available frequencies in the relay network, Af_i. Initially $Af_i = F_{ri}$, all frequencies in the relay network. When node i receives UFI_j from the *NADV* generated by node j, it recalculates $Af_i = Af_i \cap \overline{UFI}_j$. When node i receives a *RREP* from node *i-1*, it further recalculates $Af_i = Af_i \cap \overline{UFI}_{(i-1)}$. It then assigns a frequency to its next-hop-link by choosing from the resultant Af_i as shown in Figure 4-6.

If the node receiving the *RREP* has no available non-interfering frequencies, it may select a frequency that is already chosen by the previous nodes. In this case, the MAC protocol will resolve contention between the competing links, thus lowering network performance. In order to alleviate the degradation of network performance, the node picks up a frequency from $UFI_{(i-1)}$ except f_{i-1}.

Frequency_assignment( )

Af_i = Set of available frequencies of node i ;
UFI_i = Set of frequencies used by node i ;
UFI_j = Set of frequencies used by neighboring node j ;
$UFI_{(i-1)}$ = Set of frequencies used by the previous node on the reverse path ;
f_{i-1} = Frequency for upstream link of node i assigned by the previous node ;

/ Initial Condition */*
$Af_i = F_{ri}$;

/ Upon receiving NADV from a neighboring node j */*
/ recalculate Af_i */*
$Af_i = Af_i \cap \overline{UFI_j}$; return;

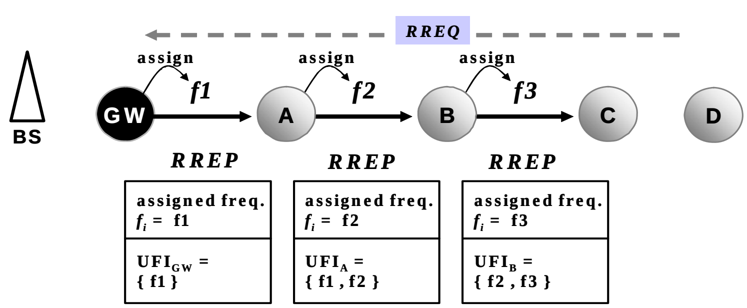
/ Upon receiving RREP from the previous node on the reverse path */*
/ recalculate Af_i and UFI_i */*
$Af_i = Af_i \cap \overline{UFI_{(i-1)}}$;
$UFI_i = UFI_i \cup \{ f_{i-1} \}$;

/ Assign a frequency */*
if (Af_i != Ø)
  / There is an available frequency */*
  then { choose a frequency f_i from Af_i and assign it to its next-hop-link ;
      $Af_i = Af_i \cap \overline{\{ f_i \}}$;
      $UFI_i = UFI_i \cup \{ f_i \}$;
    }

  / There is no available frequency */*
  else { pick up a frequency f_i from $UFI_{(i-1)}$ *except* f_{i-1} *and*
      assign it to its next-hop-link ;
      $UFI_i = UFI_i \cup \{ f_i \}$;
    }

insert UFI_i and f_i into *RREP* message and return to the next node ;

(a) Frequency allocation algorithm



(b) Example of frequency allocation

Figure 4-6. Frequency allocation

4.3.3. Transient Behavior

In order to transfer downlink data to the destination node, BS needs to keep a list of nodes on the relay network. Thus, the GW node delivers the group membership information to the BS when it receives a *RREQ* that passed through a new path. Since some intermediate nodes which have a route to the GW may immediately return a *RREP* to the source node, some *RREQ*s do not reach the GW. Thus, group membership information maintained by the BS may be incomplete even if the source node already joined the relay network. In this case, the BS will continue to use its cellular interface to communicate with the source node in the downlink until it receives data from the node via the relay network on the uplink. At this time, the BS will store a record of the node being in the relay network group served by a GW and transfer the next downlink data through the relay network.

Due to mobility, links in the relay networks may break and new nodes may join a relay network. In these cases, the effected nodes issue *RREQs* as if they were joining a relay network for the first time. Any node receiving the *RREQ* responds with an immediate *RREP* as during relay network formation. Since the nodes already maintain the used frequency information, frequency allocation on the new link is made as during network formation.

4.3.4. Multiple GWs in the Same Isolated Group

In order to reduce the relay network formation delay, our GW discovery algorithm utilizes the information carried by *NADV* messages that traverse only one hop. Therefore, depending on the topology (e.g. concave hull), there may be multiple GWs in the same isolated group.

In this case, the group is divided into multiple subgroups logically. Each subgroup has only one GW and consists of nodes which establish a path to this GW. A node in this isolated group may receive multiple *RREP*s from the GWs; such a node selects a GW that has better signal quality to the BS than itself and is the fewest hops away. The frequencies assigned to links on the paths unselected have a soft state. Thus, they are released unless a data packet is transmitted before the timer expires.

4.4. Enhancements

In addition to the basic algorithms described in section 4.3, we propose two enhancements to improve the throughput of resulting network: *local tuning* and *GW reselection*.

4.4.1. Local Tuning

When multiple frequencies are available for a relay network, all nodes make frequency allocations according to the algorithm described in section 4.3.

In this algorithm, each node assigns a frequency to the next-hop-link based on the Af_i calculated. This may not guarantee non-interfering frequency allocations for the following reasons. First, when choosing a frequency, each node cannot consider the subsequent frequency allocation of other nodes on a different branch of the relay network. As shown in Figure 4-7, while establishing a path, the *UFI* maintained in nodes D and B will not include frequency *f3*. They thus may each assign *f3* to the link to node E and C, respectively, which will result in contention on these links. In this case, the MAC protocol will arbitrate the transmission of the nodes and they will achieve lower throughput.

Second, the proposed frequency allocation algorithm ignores the fact that a node's interference range may be greater than its transmission range. As shown in Figure 4-7, node A and node X are not within transmission range and belong to different groups. They thus cannot exchange their *UFI*. Moreover, the RTS-CTS exchange will not occur between these nodes. As a result, these two nodes may assign the same frequency, f_2, to the next-hop-link. This leads the nodes to interfere with each other resulting in higher bit error rates and lower throughput at node B.

4.4.1.1. Local Tuning Algorithm

In general, interference from remote nodes causes a drop in received SNIR at the interfered node so that it experiences a higher bit error rate. This local tuning algorithm consists of two phases in which each node measures received SNIR of all its incident links periodically and reassigns a new frequency to the interfered link if SNIR of the link has fallen off markedly.

(1) Phase 1: Interference Detection

Based on the statistical information of the radio interface module, the MAC layer periodically calculates a moving average of the received SNIR. If the MAC layer detects a decrease of the received SNIR below a threshold, it notifies the routing (AODV) layer.

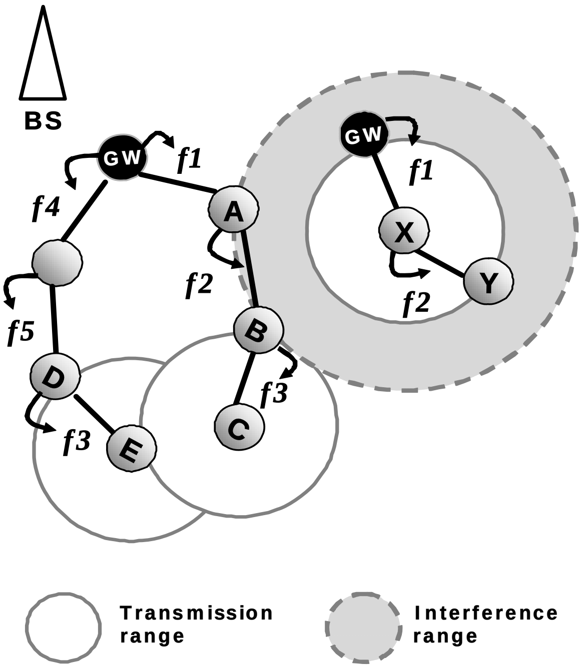


Figure 4-7. Potential interference

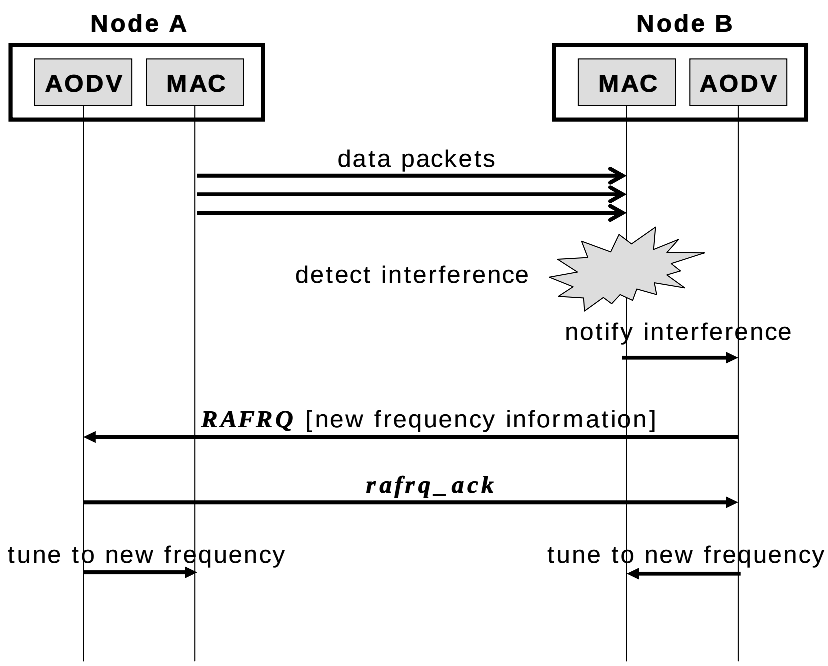


Figure 4-8. Signaling flows for local tuning

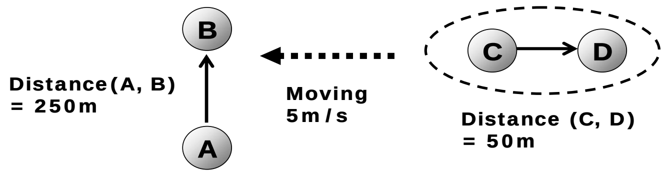
(2) Phase 2: Frequency Reassignment

Based on the information received from MAC layer, the routing layer recognizes the interfered link and determines the address of the node which sends data on this link. It then selects a new frequency from its stored Af_i, inserts it into a reassign frequency request (*RAFRQ*) message, and sends it to the node as shown in Figure 4-8. After sending or receiving the acknowledgement of the *RAFRQ* message, the routing layer reassigns the new frequency to the interfered link.

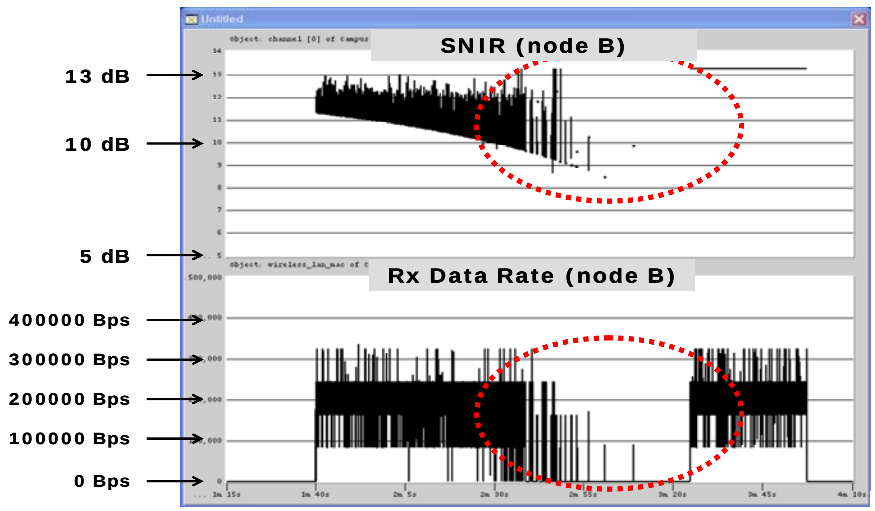
4.4.1.2. Effect of Local Tuning

We measure the effect of local tuning with a simple scenario. In this scenario, there are two pairs of nodes, as shown in Figure 4-9-(a). Nodes A and C are transmitting a fixed amount of data, 6 Mbits, to nodes B and D, respectively, on the same frequency. The distance between nodes A and B is 250m and the distance between nodes C and D is 50m. Thus, node D has a higher received SNIR than node B. The pair of nodes C and D are moving toward the pair of nodes A and B at a speed of 5m/s.

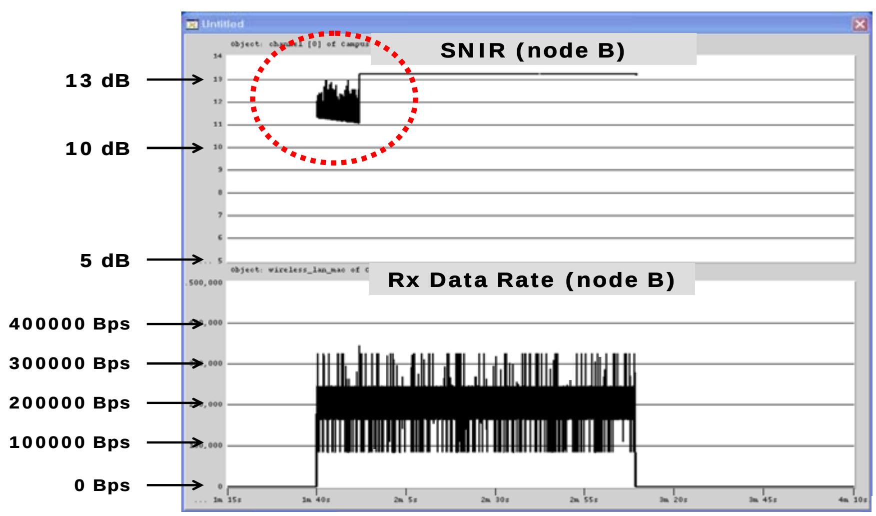
Without local tuning, the received SNIR of node B degrades rapidly as the two pairs of nodes get closer. This causes a higher bit error rate at node B and the received data rate of node B drops down to 0 bps as shown in Figure 4-9-(b). After node C completes the transmission of data, or moves away, node B will receive rest of the data from node A. With local tuning, node B is able to detect the interference from nodes C and D and switch to a new frequency. As a result, two pairs of nodes can transmit data in parallel without any interference so that they can achieve higher throughput as shown in Figure 4-9-(c).



(a) Scenario



(b) Without local optimization



(c) With local optimization

Figure 4-9. Effect of local tuning

The frequency switching of a pair of nodes can trigger a cascade of frequency switching in the network. Our local tuning algorithm is inspired by the work in [31] which claims that if every node selects its channel within a finite number of frequency switches, frequency allocation reaches a stable state where nodes cease changing frequencies.

4.4.2. GW Reselection

The GW acts as a bridge between the BS and an isolated group of nodes in the relay network. Thus, the data rate of the GW node has a great impact on the average throughput of this group of nodes.

In the 1xEVDO system, based on the received SNIR, mobile nodes transmit a 4-bit data rate control (*DRC*) sequence to the BS to request a specific data rate [6][28]. A GW with a higher DRC can achieve a higher data rate. In this chapter, we consider a scenario in which the average throughput of the group of nodes improves because of GW reselection as shown in Figure 4-10.

4.4.2.1. GW Reselection and Re-registration

If a non-GW node in the group moves toward the BS it may be able to achieve a higher data rate with the BS than the current GW. In this case, this node should be selected as a new GW replacing the current GW. For example, Figure 4-10 shows the scenario in which node 4 is selected as a new GW with higher data rate. As shown in Figure 4-11, GW reselection and re-registration are performed through the following procedure:

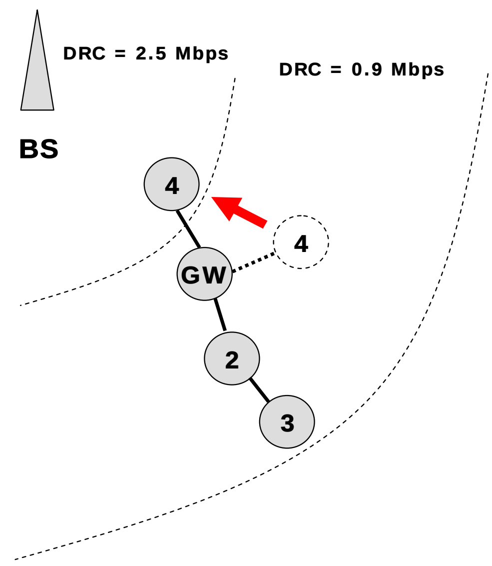


Figure 4-10. Example of GW reselection scenario

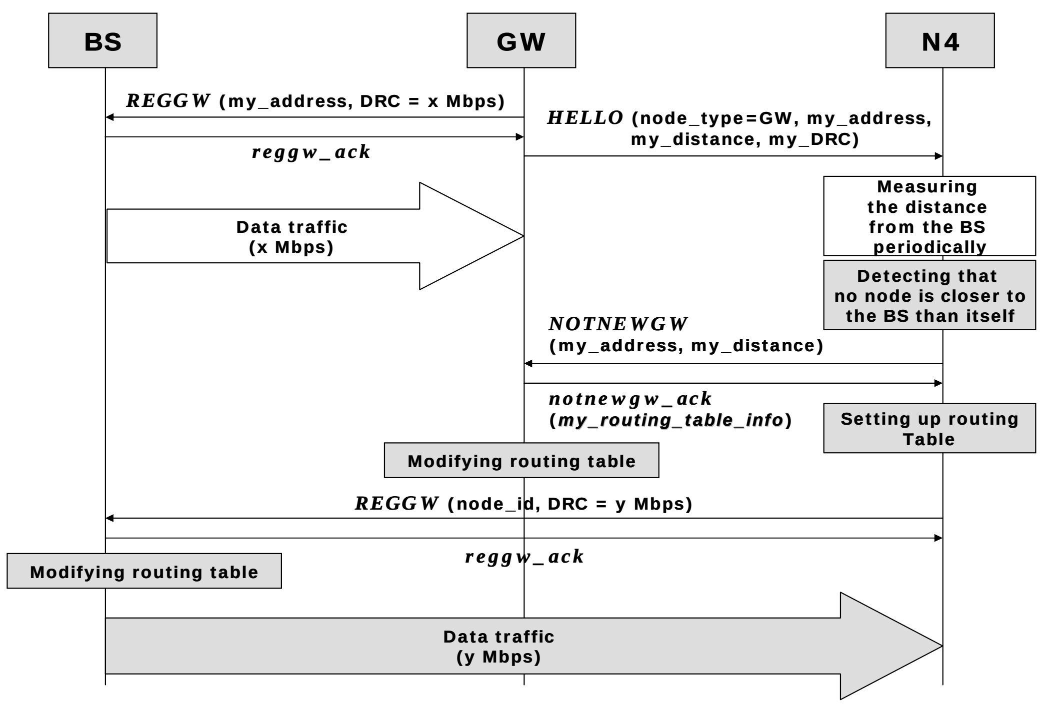


Figure 4-11. Signaling flows for GW reselection

When a node is selected as a GW, it sends a register GW (*REGGW*) message to the BS. This message contains ID of the node and its expected DRC. If the expected DRC is affordable, the BS acknowledges *REGGW* message and then transmits data to the GW with the data rate corresponding to the DRC.

We assume that the *HELLO* message generated by a GW contains the status information of the GW, for example, node ID, distance from the BS, and current DRC. Based on periodically received *HELLO* messages, the neighboring nodes of a GW know the current status of the GW. If a node gets closer to the BS so that it may achieve a higher DRC than current GW, it sends a notify new GW (*NOTNEWGW*) message to current GW. Upon receiving a *NOTNEWGW* message, the current GW returns the acknowledgement of *NOTNEWGW* message containing all its routing information to the new GW. The new GW also sends a *REGGW* message to the BS which contains its expected DRC. Then, the BS acknowledges *REGGW* message, updates its routing information, and then transmits data to the new GW with the corresponding data rate.

At this time, the former GW establishes a link to the new GW over which it forwards all the data it receives from the relay network.

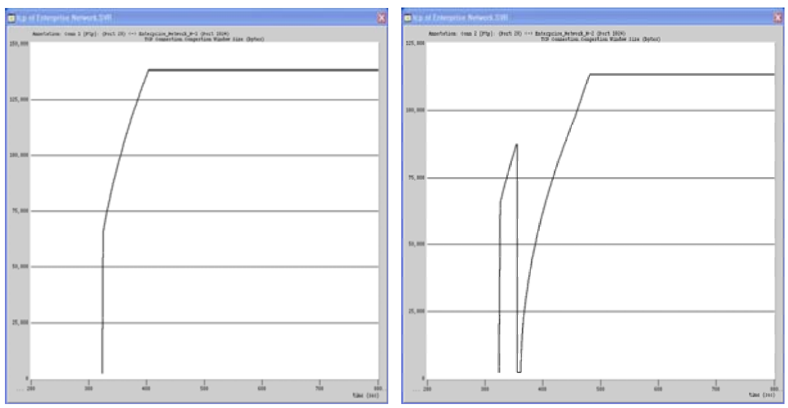
4.4.2.2. Effect of GW Reselection

Based on the scenario in Figure 4-10, we measure the average throughput of the nodes. Figure 4-12-(a) shows that we can achieve higher average throughput of the nodes if a node which has a higher DRC than the current GW is selected as a new GW.

However, GW reselection requires that at least the BS, former GW, and new GW update their routing information while they are transmitting or relaying data. Some data packets may be dropped during this period. Moreover, the BS stops data transmission to its former GW and resumes transmission to the new GW. As a result, the TCP congestion window size of all nodes in the group may be reduced. In fact, our trace, shown in Figure 4-12-(b), shows that the congestion window of the nodes drops to 0.

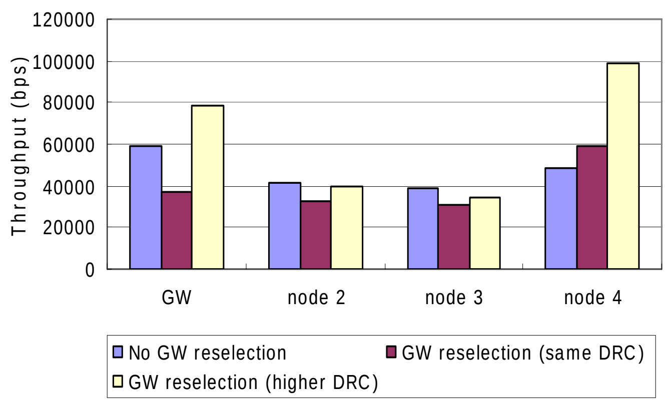
As shown in Figure 4-12-(c), however, we still achieve improved average throughput if the new GW has a higher data rate to the BS than the former GW. We also show that if the new GW does not achieve higher throughput than the former GW, overall throughput will decrease due to the drop in the congestion window.

(a) Comparison of average throughput of the nodes



No GW reselection **GW reselection**

(b) Congestion window size of former GW



(c) Throughput of the nodes

Figure 4-12. Effect of GW reselection

4.5. Performance Evaluation

We simulated the relay network formation algorithms in NS-2 v.2.1b9a to compare them in terms of the overhead during network formation. OPNET v.11.5 was used to measure the throughput gains achieved by the single and multi-frequency relay networks.

4.5.1. Simulation Environment

The air interface to all nodes when no relay network is in operation is based on a 1xEVDO. This interface is also used between the GW nodes and the BS when a relay network is in operation. We use a simplified approximation of a 1xEVDO system in which the BS schedules the mobile nodes on the forward cellular link with three classes of data rates: 2.5Mbps, 921Kbps, and 153Kbps.

The relay network uses IEEE 802.11a which provides up to 12 orthogonal frequencies. But 4 frequencies out of 12 are dedicated to point-to-point transmission. Thus, we assume that the relay network uses only 8 frequencies in these simulations.

The experiments are based on a 4000x4000m^2 6-sector cell with up to 100 mobile nodes. In this scenario, each node in the relay network downloads 6Mbits data from an FTP server. Table 4-1 summarizes the simulation parameters.

The value of Δt used in the formation algorithms discussed in section 4.4 is initially broadcast by the BS. In a 1xEvDO system, each mobile node can occupy maximum 16 timeslots (each of which is 1.67ms) during its turn. Thus, in order to synchronize with this system, Δt is set to 26.72ms in this simulation.

Parameter	Value
Cell size (BS at center)	4000m X 4000m
N (# of nodes / cell)	1 → 100
# of sectors / cell	6 sectors
DRC_i (Classes of HDR links)	DRC_3 : 2457 Kbps DRC_2 : 921 Kbps DRC_1 : 153 Kbps
Downloaded file size	6 Mbits / user
Application	FTP/TCP
Advertised frequency band	Frequency band used for 802.11a
# of orthogonal frequencies	8
Air interface range	250 m

Table 4-1. Simulation parameters

4.5.2. Comparison of the Formation Algorithms

We use three metrics to compare the performance of our relay network formation algorithms: *signaling traffic*, *formation latency*, and *GW load* as shown in table 2.

These metrics indicate the *overhead* of the relay network formation. Signaling traffic generated indicates the degree of network congestion during network formation. Since relay networks are typically formed when the network is experiencing poor performance, formation latency is critical. The processing load at GWs is proportional to the traffic intensity of the cellular interface between the BS and GWs during the formation process.

For the comparison, 100 different network topologies are generated in each case of 10 to 100 nodes. Each data point is the average over the runs.

The results of the simulation are shown in Figures 4-13 ~ 16. It is clear that all algorithms incur trade-offs. In general, the algorithms with strict scheduling like NF and FF have the highest latency (Figure 4-13) due to their sequential nature, while those with more parallelism have lower latency at the expense of higher signaling traffic (Figure 4-14) and GW load (Figure 4-15).

NF has the highest latency, but the fewest messages to hit the GW nodes. Even though BL has low latency, the signaling traffic and the load at GW node during network formation is about three times of that when using FF and NF. LOF has good performance in terms of signaling traffic and load at the GW node, but high latency at high node density.

Therefore, we define the *weighted overall overhead* of each algorithm as shown in Table 4-2. It is the summation of relative value of the three metrics which are given different weights according to their importance.

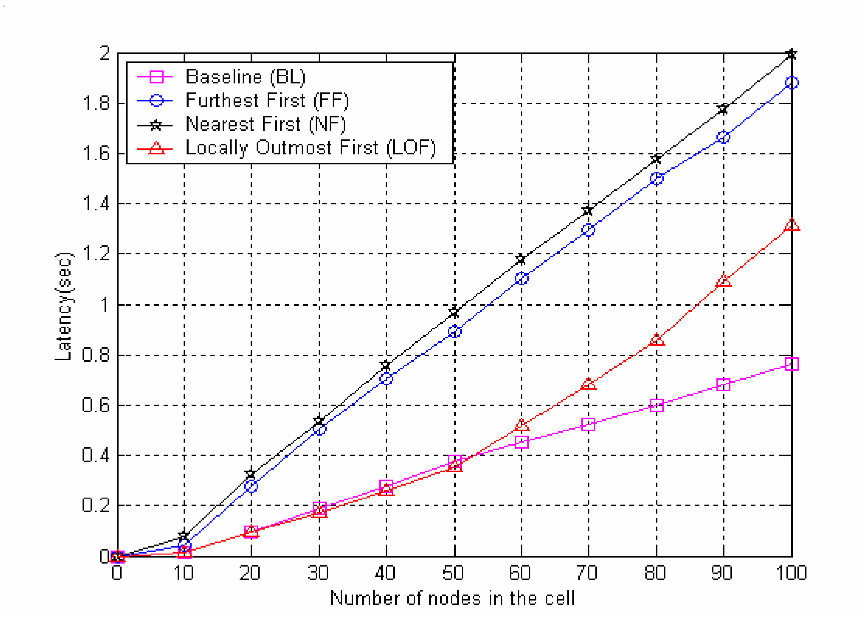


Figure 4-13. Formation latency

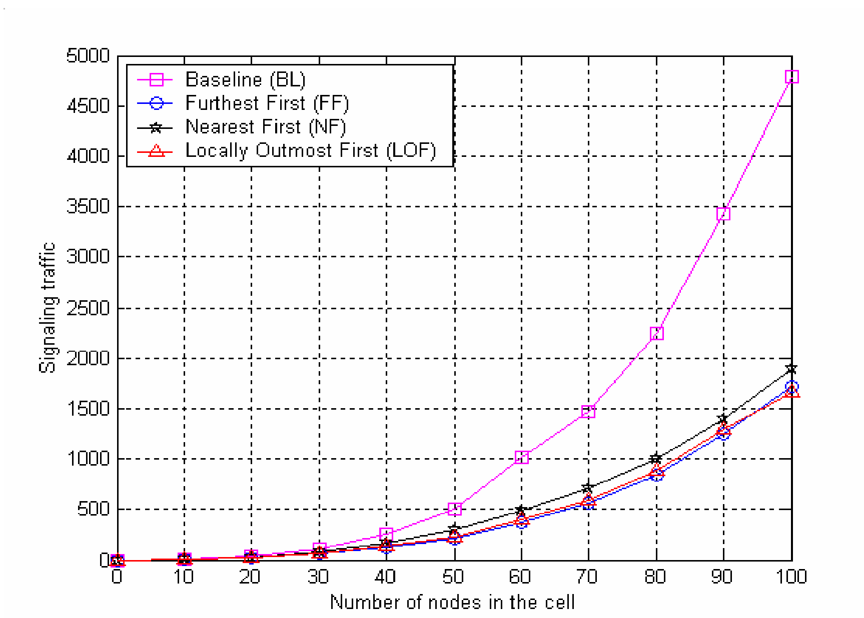


Figure 4-14. Signaling traffic

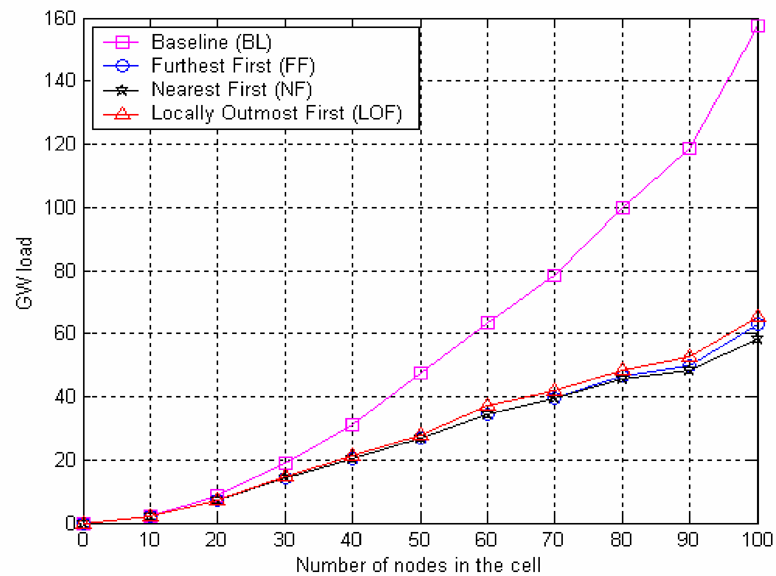


Figure 4-15. Load at GWs

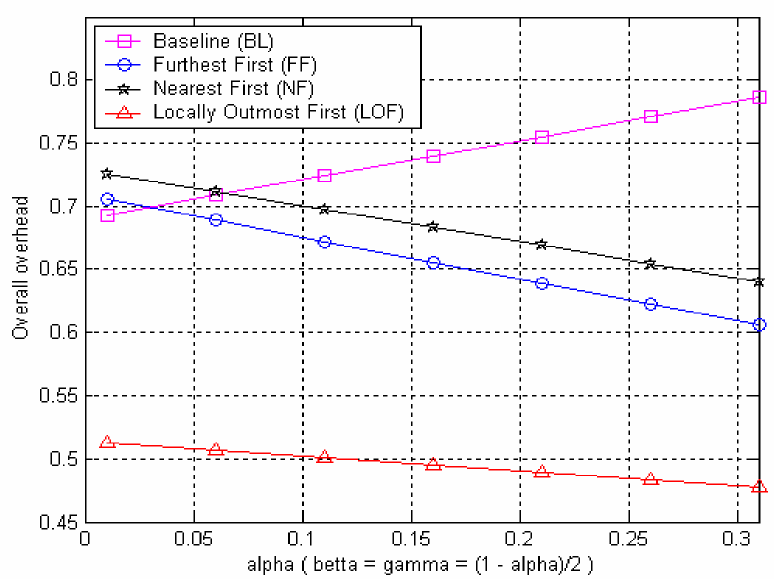


Figure 4-16. Overall overhead ($\beta = \gamma = (1-\alpha)/2$)

In Figure 4-16, β and γ are set to be equal, meaning that low latency and low load on the GW are of equal importance. We evaluate the overhead over the range when the signaling traffic is considered highly unimportant ($\alpha = 0.01$), through the case in which all three metrics are of equal importance ($\alpha = \beta = \gamma = 0.33$). As shown, LOF is the best performing algorithm in terms of overall formation overhead.

4.5.3. Throughput of the Relay Network

Since LOF is the best overall performing algorithm, we measure the throughput of the relay networks resulting from LOF using OPNET. We consider both single and multi-frequency relay networks formed by LOF. We also consider the multi-frequency relay networks with optimal frequency allocation as discussed below.

We use two metrics to compare the throughput of the relay networks: *overall network throughput* and *average node throughput* as shown in table 2.

The overall network throughput indicates the total throughput of the relay network as measured at the BS until all nodes complete their FTP transaction. Thus, it generally depends on the throughput of the node with the lowest data rate. The average node throughput is calculated based on the average FTP response time of the nodes in the relay network.

1) Single Frequency Relay Networks

In this section, we examine the performance gain of the single frequency relay network over the pure 1xEVDO system. Our simplified system supports data rates on the forward link of 2.5Mbps, 921Kbps, and 153Kbps.

Performance of the formation algorithms (*S* = type of formation algorithm (e.g. BL, FF, NF, LOF))		
Name	Description	Formula
Signaling Traffic	Total number of routing messages received by all mobile nodes forming the relay network	**M(S) =** $\sum_{i=1}^{N} msg_i$ **, where** msg_i **= # of received messages by node i**
Formation Latency	The time elapsed between the first *RREQ* and all nodes having a route to the BS	**L(S) =** ***t_final – t_init*** **, where** ***t_final*** **= time when all nodes have a route to the BS** ***t_init*** **= time when the first** ***RREQ*** **is sent out**
GW Load	Total number of routing messages received by all GW nodes	**G(S) =** $\sum_{g=1}^{G} g\_msg_g$ **, where** g_msg_g **= # of received messages by GW node g**
Weighted Overall Overload	Summation of relative overhead compared to the other algorithms each of which has different weight according to the importance of the metric	**RM(S)=relative signaling traffic of algorithm S=** $\frac{M(S)}{\max(M(BL), M(FF), M(NF), M(LOF))}$ **RL(S)=relative formation latency of algorithm S =** $\frac{L(S)}{\max(L(BL), L(FF), L(NF), L(LOF))}$ **RG(S)=relative GW load of algorithm S =** $\frac{G(S)}{\max(G(BL), G(FF), G(NF), G(LOF))}$ **WO(S)=weighted overall overhead of alg. S =** $(\alpha \times RM(s)) + (\beta \times RL(s)) + (\gamma \times RG(s))$ $\alpha + \beta + \gamma = 1$

Throughput of the resulting relay network		
Name	Description	Formula
Overall Network Throughput	The total throughput of the relay network as measured at the BS until all nodes complete their FTP transaction	Rt_i **= FTP response time of node** ***i*** ***FS*** **= Download file size of each node** **Longest Response time = max (** Rt_i **), 1 <= *i* <= N** **Overall network throughput** $= \frac{FS \times N}{\max(Rt_i)}$
Average Node Throughput	Average throughput of each nodes in the relay network	**Average node throughput =** $\frac{FS}{\left(\frac{\sum_{i=1}^{N} Rt_i}{N}\right)}$

Table 4-2. Performance metrics

Considering a time period in which the DRC of all nodes is constant, if all nodes in the 1xEVDO network are backlogged, each node is scheduled for the same amount of time at the rate supported by their DRC. The network throughput under these conditions is simply the weighted average of the throughput of all nodes in the 1xEVDO network.

Thus when there are *m* active nodes in the sector, the maximum achievable data rate of node *i*, R_hdr_i, and the maximum average data rate of the nodes, $\overline{R\_hdr}$, are given by

$$R\_hdr_i = DRC_i \times \frac{1}{m} \text{ , where } DRC_i = \text{data rate of node } i \quad (4\text{-}2)$$

$$\overline{R\_hdr} = \frac{\sum_{i=1}^{m} R\_hdr_i}{m} \quad (4\text{-}3)$$

When using a relay network, the BS schedules the GWs instead of the members of the relay network. In this way, all nodes on the relay network share the data rate sustained on the link between the GW and the BS. We assume the BS schedules GWs proportionately with the number of nodes on the relay network that they support. For example, a GW terminating a relay network with 4 nodes will be scheduled twice as often as a GW node terminating a relay network with 2 nodes. A single node is treated as a GW supporting a relay network of 1 node.

Therefore a node can never achieve higher throughput acting as a single node than when a member of a relay network. In fact, if a node joins a relay network in which the GW node has a higher DRC than itself, its throughput will be increased. Thus, the maximum achievable data rate of node *i*, R_relay_i, occurs when a node is a member of a relay network. Likewise, the maximum average relay network throughput, $\overline{R\_relay}$ is achieved when all nodes join a relay network. These values are given by:

$$R\_relay_i = DRC_{GW} \times \frac{1}{m} \text{ , where } DRC_{GW} \text{ = data rate of GW} \quad (4\text{-}4)$$

$$\overline{R\_relay} = \frac{\sum_{i=1}^{m} R\_relay_i}{m} \quad (4\text{-}5)$$

We compare the simulated throughput achieved with a single frequency relay network with the theoretical maximum achievable throughput of a 1xEVDO system in which no relay network is used. The throughput was obtained using OPNET according to the parameters in Table 4-1. Figure 4-17 shows the results obtained from three random topologies of a cell with 30 nodes. It can be clearly seen from Figure 4-17 that a 1xEVDO system operating with relay networks has better node throughput than a pure 1xEVDO system. Note that these results are extremely conservative: the 1xEVDO system is ideal in that it includes no protocol overheads or impact of errors.

We also note that the performance gain achieved by the relay network varies considerably from topology to topology. In Figure 4-17, the performance gain of the relay network over the pure 1xEVDO system is lowest in topology 2 and highest in topology 3.

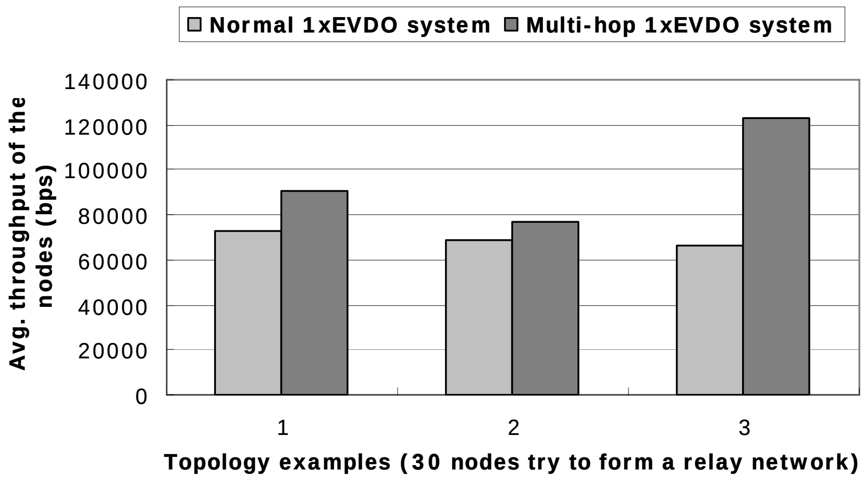


Figure 4-17. Examples of throughput gain

		# of nodes with DRC=3 (2457 Kbps)	# of nodes with DRC=2 (921 Kbps)	# of nodes with DRC=1 (153 Kbps)
Topo-logy 1	Normal	8 [7]	16 [2]	6
	Relaying	15	11	4
Topo-logy 2	Normal	6 [4]	17	7
	Relaying	10	13	7
Topo-logy 3	Normal	5 [14]	16 [9]	9
	Relaying	19	11	0

Table 4-3. Examples of nodes' data rate

This can be explained as follows: A relay network improves the performance of a cellular network by exploiting the fact that GW nodes have throughput equal to or higher than the nodes they serve. From Table 4-3 it can be seen that, in topology 2, there are seven nodes with DRC = 1(153Kbps). When the relay network is formed in topology 2, all the nodes that have DRC = 1 originally keep the same data rate because the GW connecting the BS with these nodes also has DRC = 1. Thus these nodes achieve no performance improvement. The slight overall improvement in performance is because the 4 nodes that have DRC = 2(921Kbps) connect to a GW node with DRC = 3(2.5Mbps), thus increasing their throughput. In the case of topology 3, all the nodes that have DRC = 1 are connected through GW nodes that have DRC = 2. Thus a considerable improvement in performance is achieved in this topology.

2) Multi-Frequency Relay Networks

In this section we quantify the performance gains achieved by using a multi-frequency relay network. We consider the relay network formed using the LOF algorithm.

We measure the throughput of each resulting relay network in the following three scenarios.

- **Scenario 1**: The relay network uses a single frequency on all links.
- **Scenario 2**: The relay network may use up to all 8 available frequencies. All nodes in the relay network make frequency allocations according to the frequency allocation scheme in section 4-4.
- **Scenario 3**: As with scenario 2, the relay network may use up to 8 frequencies.

Unlike scenario 2, the frequencies are assigned by a centralized assignment scheme based on the interference constraint in [4]. In order to maximize network throughput, this algorithm assigns a frequency to the link based on an interference constraint that considers a node's interference range as well as transmission range. This scheme makes optimal frequency allocations resulting in no contention and minimum interference.

Besides these basic scenarios, we also quantify the benefits of the local tuning in the case that all nodes can detect interference and reassign a new frequency if needed according to the local tuning algorithm, in addition to scenario 2.

Figures 4-18 and 4-19 show that if we exploit multiple frequencies, we can achieve higher overall network throughput and average node throughput. Figure 4-20 shows the percentage of the throughput achieved by our frequency allocation scheme of section 4 when the optimal throughput which the relay network can achieve is set to 100%. It shows that our frequency allocation scheme achieves at least 84% of the average throughput of the optimal channel assignment algorithm. It also clearly shows that using local tuning results in higher average throughput.

We note that as the number of nodes increases in the network, the improvement in performance of the optimal frequency allocation increases (Figure 4-19). This is because it is more likely that the distributed frequency allocation algorithm will result in two nodes in different relay networks that are within interference range being assigned the same link frequency (see discussion of Figure 4-7).

Figure 4-21 illustrates this effect. This figure shows the nodes' received SNIR in each scenario with the same topology. It can be clearly seen that there is considerable improvement in SNIR when multiple frequencies are used, and that in some cases, the SNIR when using optimal frequency allocation is higher than when using distributed frequency allocation. However, Figure 4-21 shows that the local tuning algorithm improves the SNIR to a value significantly close to that obtained with optimal channel assignment.

Since Figures 4-19 and 4-20 plot the average throughput of all nodes in the relay network, and not only for nodes that experience an SNIR improvement due to the local tuning, the improvement in throughout is somewhat obscured.

4.6. Discussion

Research [34][63] has shown that in multi-hop wireless cellular networks, the transmissions from a mobile terminal to the BS are broken into multiple wireless hops, and hence require less total transmission power than single hop cellular networks.

In this chapter, we assume the use of 802.11 interfaces for relay networks. In [26], it is shown that the 802.11 ad-hoc mode reduces the transmission power significantly compared to the cellular mode and achieves higher network throughput.

However, multi-hop relaying may lead to higher energy consumption for gateway nodes. In this chapter, the GW is chosen depending only its signal quality from the BS. We can consider an improved GW selection metric that includes the residual battery power of the node in addition to the signal quality from the BS. Based on this metric, the GW can be reselected periodically for balancing energy consumption.

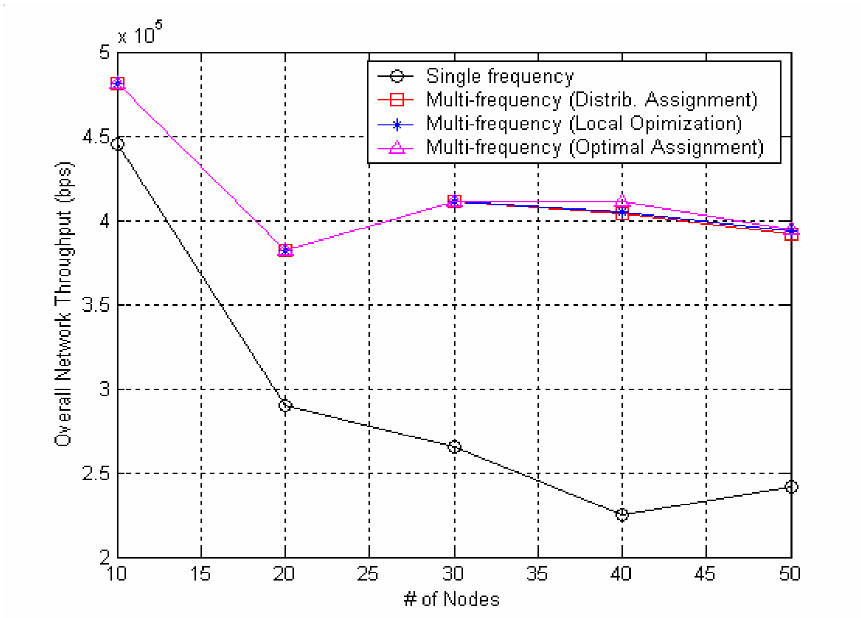


Figure 4-18. Overall network throughput

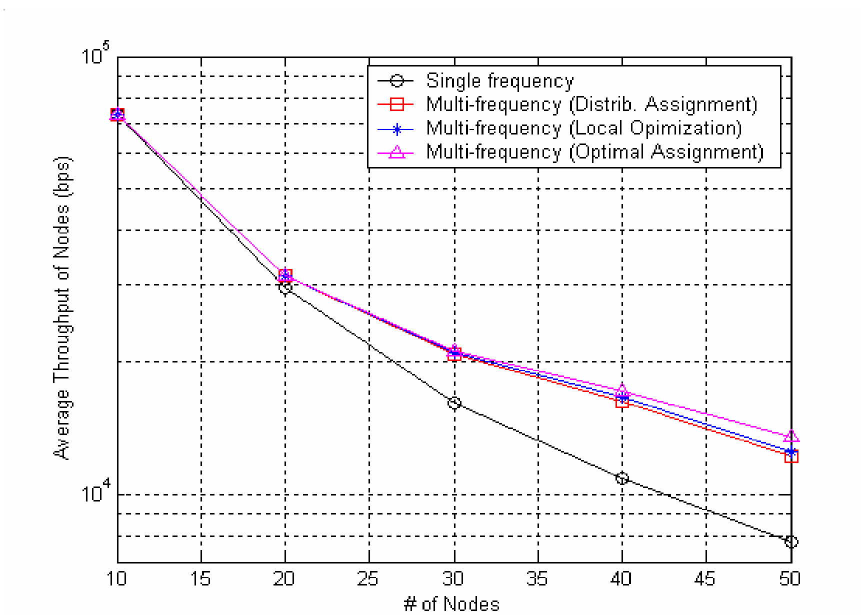


Figure 4-19. Average throughput of the nodes in all scenarios

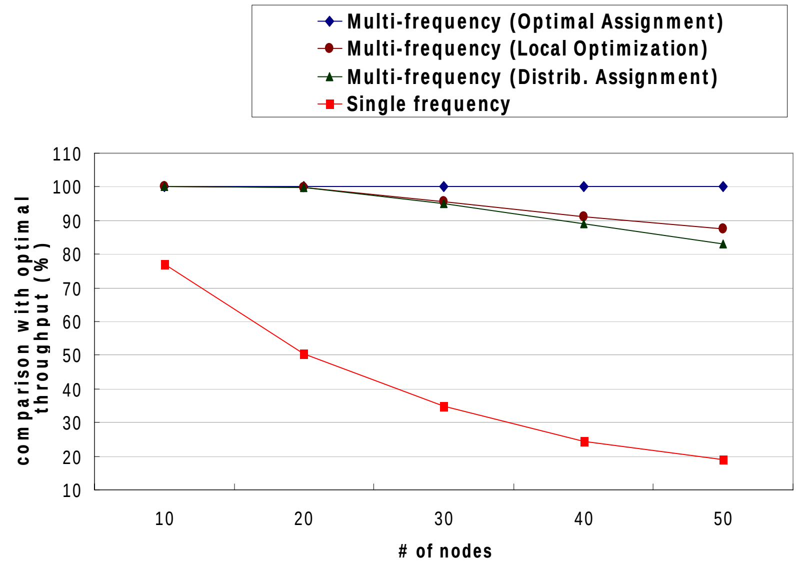


Figure 4-20. Throughput Comparison

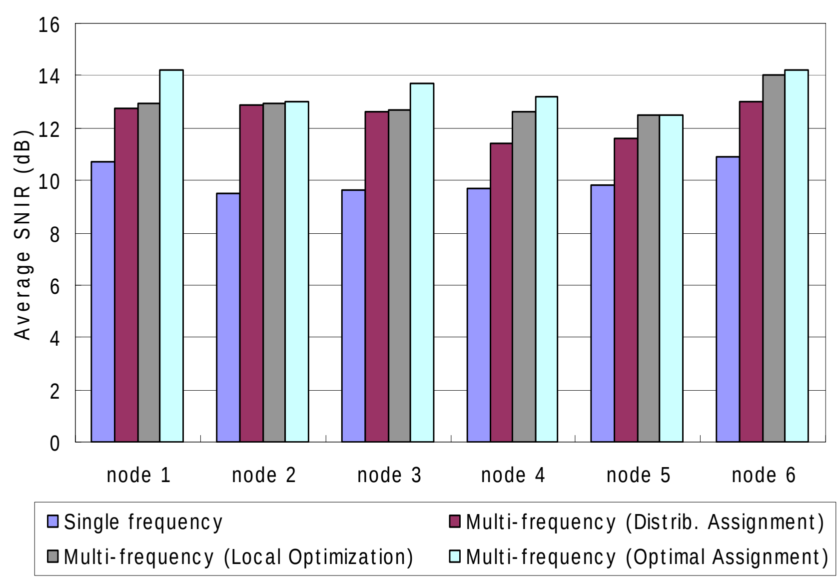


Figure 4-21. Example of node's SNIR (Example of 50 node topology)

4.7. Concluding Remarks

In this chapter we analyzed the formation of relay networks for dynamic multi-frequency, multi-hop wireless cellular networks. We propose two centralized algorithms and one distributed algorithm for network formation. While establishing paths to the GW nodes, mobile nodes can make non-interfering frequency allocations to the relay links based on limited hop information. As a result, the number of interfering links can be reduced and hence we can achieve improved network throughput. For further improvement of network throughput, we also propose two enhancements to support local tuning and GW reselection.

Our results show that schemes scheduling nodes furthest from the BS to initiate route discovery first make good use of passive route discovery and hence reduce the relay network formation overhead. Moreover, they can build efficient relay networks which attain high throughput. The distributed LOF algorithm achieves a good trade-off between relay network formation overhead and latency.

We measured the throughput of the relay network resulting from the LOF algorithm. The results show that by exploiting multiple frequencies, we can achieve higher overall network throughput and average node throughput. Our simple distributed frequency allocation scheme achieves 80-85% of the optimal average node throughput. Moreover, the enhancements further improve average throughput.

Chapter 5. Reliability-Aware Resource Allocation for Multi-Frequency, Multi-Hop Wireless Cellular Networks

5.1. Introduction

The availability of multiple frequencies allows mobile nodes to form diversity paths between each other. The broadcasting nature of the wireless medium enables an intermediate relay node to overhear the data broadcasted by the source, and then relay the information to the destination on a new frequency. The destination can detect and recover errors in the packet delivered from the source quickly by comparing it with the packet received from the intermediate relay node. Therefore, the path diversity helps reduce the bit error rate, and increase throughput. Such relay-assisted schemes have been attracting much interest recently due to the performance improvement they provide [25][33][37][52][64][65].

In this chapter, we posit that relay network formation algorithms that exploit diversity provided by intelligent resource allocation lead to better end-to-end performance. We extend the relay network formation algorithms so that mobile nodes not only communicate over multiple frequencies but exploit two-hop diversity paths when possible. We compare the performance of the formation algorithms based on the quality of the resulting relay networks considering end-to-end error rate, percentage of non-interfering hops on each path, and percentage of hops with diversity on each path. Our results show that network formation algorithms which enforce some order in assigning frequencies are more effective than those that allow nodes to randomly join the relay network.

In this chapter, we make the following contributions:

- We investigate relay network formation algorithms that provide frequency allocation to establish non-interfering links between mobile nodes. Moreover, they allow the mobile nodes to establish two-hop diversity paths between adjacent nodes to reduce bit error rate and increase throughput.
- We compare the performance of the formation algorithms based on the metrics indicating the reliability of resulting relay networks (e.g. end-to-end error rate, percentage of non-interfering hops on each path, and percentage of hops with diversity on each path.)

The rest of this chapter is organized as follows. In section 5.2 we extend the basic relay network formation algorithms to accommodate diversity and multiple frequencies per relay network. In section 5.3, we evaluate the performance of the network schemes by providing simulation results. In section 5.4, we discuss the related issues of this chapter. In section 5.5, we describe enhanced diversity setup and frequency allocation algorithms to provide performance improvement. We conclude this chapter in section 5.6.

5.2. Network Formation with Two-hop Diversity

The network formation algorithms presented in Section 4.3.2 were designed to support relay networks operating in multiple frequencies. These algorithms include the frequency allocation that makes each link in the relay networks orthogonal, if possible. This allows nodes within transmission range to transmit simultaneously without requiring a MAC protocol to resolve contention, thus increasing network throughput and providing isolation of paths.

In this section we extend these algorithms to add two-hop diversity, using an intermediate relay node, to each hop of the path whenever possible. This spatial and frequency diversity reduce the bit error rate on a path and hence increase network throughput further.

Our network model is shown in Figure 5-1. Mobile nodes cooperate to form a relay path to a BS through a GW node. Nodes may play a combination of three roles as follows:

- ***Source nodes*** generate traffic.
- ***Intermediate nodes*** on the path forward traffic from the source to the GW.
- ***Relay nodes*** provide a two-hop diversity path between two nodes with a direct link.

Nodes may play all of these roles simultaneously. In Figure 5-1, node 1 is a source. Nodes 2, 3, and 4 are intermediate nodes; node 3 is a source node as well. Node 5 is a relay node which adds two-hop diversity to the link between nodes 2 and 3; it could be also a source node. Node 6 is the GW for these nodes to the BS.

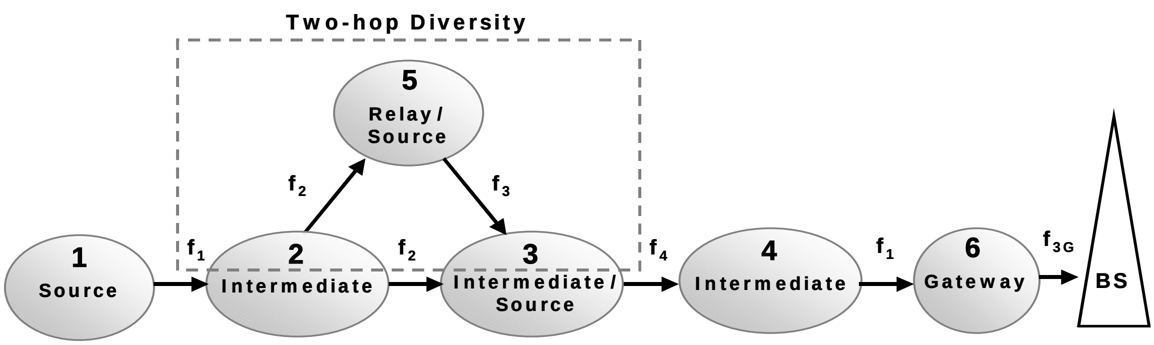


Figure 5-1. Network Model

All of the network formation algorithms discussed have the following properties:

- The BS advertises the frequencies available in a relay network.
- Mobile nodes are equipped with an agile radio so that they are able to operate on all advertised frequencies.
- Even if multiple frequencies are available within each relay network, every mobile node uses a common control frequency to exchange all control messages (e.g. *RREQ*, *RREP*, *NADV*) for the network formation.
- Even if each relay network can operate on multiple frequencies, a mobile node uses only one frequency to transmit data at a time, i.e., nodes do not transmit data over multiple frequencies simultaneously.

In order to reduce the end-to-end bit error probability, and ultimately improve the throughput of the network, each node can use two-hop diverse paths comprised of a single relay node. In Figure 5-1, node 2 transmits data to node 3 (its next hop) on frequency f_2. This transmission is also received by node 5. Node 5 retransmits the data to node 3 on frequency f_3. Thus, node 3 receives the data twice on two different frequencies. In order to build a two-hop diversity path without any interfering links, each node should have maximum 6 available frequencies for uplink and downlink transmission of a source node.

The relay network formation schemes consist of three main phases. Phases I and II are similar in purpose to those described in chapter 4 with extensions to support SNIR measurement and diversity set-up. To determine if adding diversity to a direct link will reduce an end-to-end error rate, SNIR measurements are made during the phase I. Frequency allocations for both direct hops and diversity paths are made during the phase II. Finally, for certain nodes that have learned routes passively, diversity is added if possible, during phase III of the network formation after initial paths are established.

5.2.1. Phase I - GW Discovery and SNIR Measurement

Before choosing a frequency or deciding if diversity should be added to a hop, each node calculates the local error probability given its local measured SNIRs from its neighbors. Every node joining the relay network initially broadcasts a *NADV* message as part of the GW discovery. The *NADV* messages are broadcast over all available frequencies in the relay network; receiving nodes measure and store the SNIR value of all received signals in their neighbor table. For each neighbor, the table contains the ID of the node, its received signal quality from the BS, and its received SNIR from the neighbor on each frequency. Based on the received signal quality from the BS for each node, the GW node is chosen.

5.2.2. Phase II - Joining the Relay Network

In the path setup phase, the nodes join the relay network by forming a path through one of the GWs to the BS. The *RREQ* sent from a node is forwarded by several intermediate nodes to the GW. When received, the GW replies with a *RREP* back to the source. Note, the GW only replies to the first instance of the *RREQ* message from the same source. The *RREP* is delivered to the source node in reverse direction. The path established by this exchange is called the *direct path*. If the *RREQ* is received by a node that already has a path to the GW, this node may generate an immediate *RREP* response back to the source. If there are multiple paths from the source node to the GW or replying node, some intermediate nodes on the direct path may receive several duplicates of the *RREQ* before receiving the *RREP*. These nodes are called *joint nodes*.

Figure 5-2 shows an example of a direct path and a joint node. In this example, the group contains seven nodes and one GW. The *RREQ* generated by node 1 is forwarded to the GW by nodes 3, 6, and 7. The *RREP* is returned by these nodes to the source. Thus, in this case, the direct path for source node 1 consists of node 1, 3, 6, and 7. Since there are multiple paths from node 1 to 6 and from node 3 and 6, node 6 may receive multiple duplicates of the *RREQ* before receiving the *RREP* from node 7. As a result, node 6 detects that it is a joint node.

5.2.2.1. Direct Hop Frequency Allocation

While establishing the path, a frequency must be assigned to each hop. This includes both direct hops and diversity hops. This frequency allocation is performed by intermediate nodes on the direct path. When each intermediate node receives a *RREP*, it is responsible for assigning suitable up and downlink frequencies for the link between itself and the next node on the reverse path before forwarding *RREP* to the node.

In our modified AODV protocol, the *RREQ* contains the SNIR information of the uplink node, and the *RREP* contains the frequencies used by the downlink node on the reverse path. Thus, a node can determine the frequencies available for the next hop. Based on the SNIR information of the next node, the node selects frequencies which have the highest SNIR value over the available frequencies.

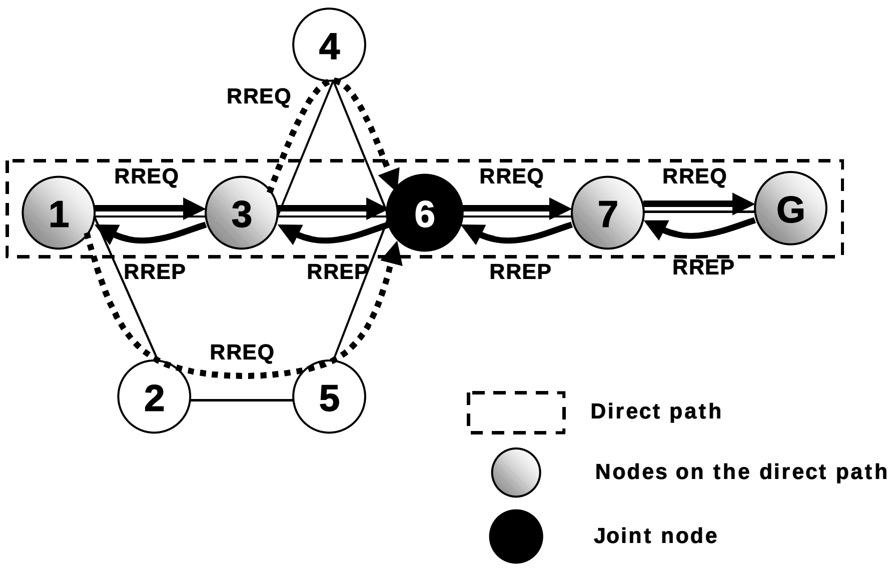


Figure 5-2. Example of the direct path and joint node

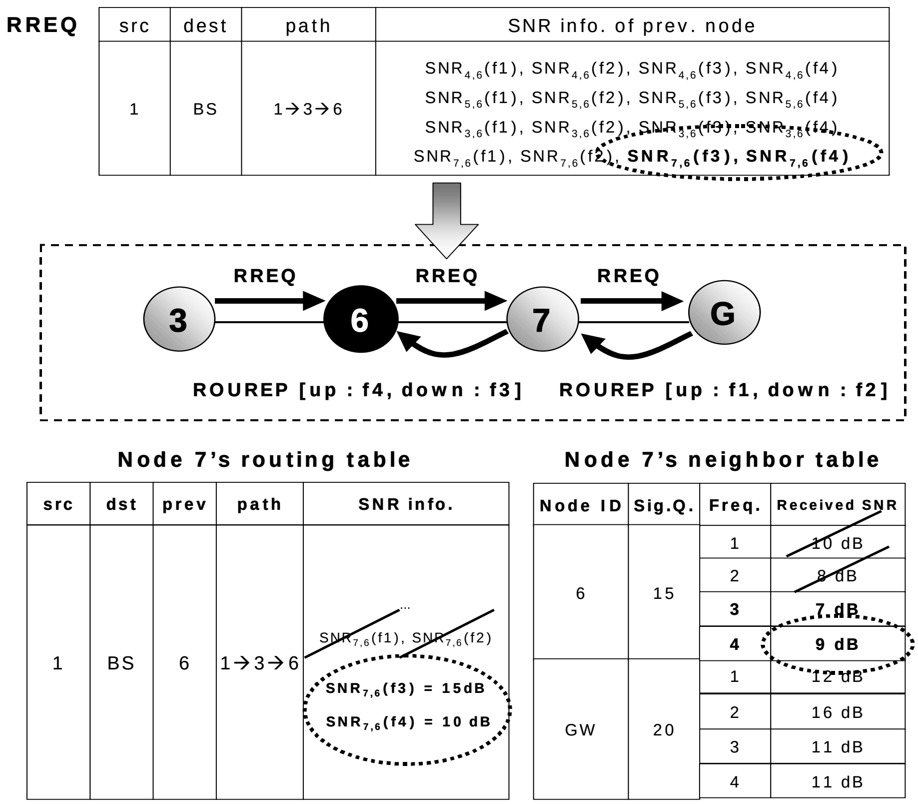


Figure 5-3. Frequency allocation

Figure 5-3 shows the frequency allocation algorithm and an example of the frequency allocation at node 7 in Figure 5-2. In this example, we have 4 frequencies for the relay network. During the SNIR measurement procedure, node 7 keeps the SNIR value of all received signals and path information sent from node 6 and the GW in its neighbor table. The *RREP* sent from GW node indicates that, for the source node 1, the GW has assigned the frequency f_1 for the uplink and f_2 for downlink to the link between node 7 and GW node. Node 7 consults its routing table and neighbor table to select two frequencies for its hop to node 6. It excludes f_1 and f_2 and selects the remaining frequencies with the highest SNIR value.

If the node receiving the *RREP* has no available non-interfering frequencies to choose from, it may select a frequency that is already assigned to other link. In this case, a MAC layer protocol must be used to resolve contention between the competing links, thus lowering network performance. After selecting the frequencies, the node inserts the assigned frequency information into *RREP* and forwards it to the next node on the reverse path.

5.2.2.2. Detecting Possible Diversity Paths

In order to reduce the end-to-end error probability each node will add diversity to each hop of the path if possible. Here we consider two-hop diversity paths which consist of a direct link and one relay node. In Figure 5-2, the only alternative path between two adjacent nodes with a single intermediate hop is between nodes 3 and 6 via node 4.

Based on the path information in the received *RREQs*, each joint node checks if the previous node of a *RREQ* is equal to the previous node of another *RREQ* in two-hop distance. If there is a *RREQ* satisfying the condition, there is a two-hop diverse path between the previous node on the forward path and itself. For example, in Figure 5-2, node 6 receives 3 *RREQ*s. From the path information of the *RREQ*, it detects that it receives a *RREQ* from node 3 directly, and a duplicate of the *RREQ* from node 3 through node 4. Thus, it realizes that there is a two-hop diverse path between node 3 and itself.

5.2.2.3. Diversity Setup

If the joint node detects a two-hop diverse path, it attempts to add a diversity path. Setting up the diversity path includes checking if there are available frequencies to be assigned to the diverse path with a link SNIR above the threshold, reselecting the frequencies for the direct hop, and selecting the most suitable frequencies for the diverse path. If there is no available frequency for the diverse path, then diversity is not established.

During the initial direct hop frequency allocation procedure, the up and downlink frequencies which have the highest SNIR value of the direct hop over the available frequencies are selected. However, in order to maximize the benefits of diversity, the joint node reselects the uplink frequency, f_u, which has the highest SNIR value of the hop from the next node on the reverse path to the relay node instead of the direct hop. In the same way, it reselects downlink frequency, f_d, which has the highest SNIR value of the hop from itself to the relay node.

Then, the joint node selects the uplink frequency for the relay node, f_{u_relay}, which has the highest SNIR value of the hop from the relay node to itself, and the downlink frequency for the relay node, f_{d_relay}, which has the highest SNIR value of the hop from the relay node to the next node on the reverse path. If diversity is established, the joint node inserts the assigned frequency information into the *RREP* and forwards it the next node. It also sends *RREP* to the relay node.

5.2.3. Phase III - Adding Diversity to the Relay Node

Relay nodes may act as a source node as well as shown in Figure 5-1. When receiving the *RREP* from the joint node at the end of diversity setup procedure, the relay node passively learns a route to the GW. Therefore, the relay node may not participate in the procedures to establish a diversity hop from itself to the joint node. In order to get improved network performance, the joint node can add diversity for the relay node in this phase.

In order to do this, the joint node assigns up and downlink frequencies and establishes diversity in the same way described in the previous section. While this phase is logically separate from the path setup phase, the joint node can perform these procedures simultaneously with setting up a diverse path in the previous phase. Thus, the joint node can insert all assigned frequency information for two diverse paths into the *RREP* and forward it the next node and the relay node.

While this procedure complicates path setup, it provides performance improvements discussed in section 5.4.

5.3. Performance Evaluation

In the following subsections, we present out performance evaluation results.

5.3.1. Simulation Environment

In order to evaluate the relay network formation algorithms, we simulate our protocols using ns-2. Table 5-1 summarizes the simulation parameters. We use IEEE 802.11b with a 115-meter communication range as the common control frequency for delivering routing messages. The CMU scenario generation tool was used to create a network consisting of up to 80 mobile nodes within a square cell of $886 \times 886\,\mathrm{m}^2$. The BS is located in the center of the cell. We vary the number of frequencies available within a relay network from 8 to 12. Each simulation is run 30 seconds and each data point in the result graphs is the average of 100 runs with different topologies.

5.3.2. Performance Metrics

We evaluate each scheme against three metrics: the average end-to-end error probability of each node, the percentage of hops with diversity, and the percentage of hops with interfering links. All mobile nodes in the relay network except the GW nodes can be source nodes for a path; the GW nodes are always the destination for a path.

The end-to-end error probability is given by

$$P_{e2e} = 1 - \prod_{i=1}^{N} \left(1 - P_{e,i}\right) \tag{5-1}$$

, where N is the total number of hops between the source and the destination.

Variable	Value
Duration	30 sec.
Runs	100 runs with different topologies
Air interface range	115m
MAC for common control signal	802.11
Cell Size (BS at center)	886m x 886m
Routing protocol	Modified AODV
Packet size	1500 bytes
Nodes per cell	1 ➔ 80 nodes
Frequencies available within a relay network	8 ➔ 12 frequencies
Δt	20 msec

Table 5-1. Simulation Parameters

$P_{e,i}$ is the error probability at the *i*th hop given by

$$P_{e,i} = \begin{cases} P_{err\_direct}(\gamma_{sd}) = Q\left(\sqrt{2\gamma_{sd}}\right) & \text{if no diversity or } \gamma_{sr} < \gamma_{th} \\ P_{err\_diversity}(\gamma_{sd}, \gamma_{sr}, \gamma_{rd}) = 1 - P^{d}_{c|\gamma_{sr} \geq \gamma_{th}} & \text{if } \gamma_{sr} \geq \gamma_{th} \end{cases} \tag{5-2}$$

, where γ_{sd} is the received SNIR from the source to the destination, γ_{sr} is from the source to the relay, γ_{sd} is from the relay to the destination, γ_{th} is the threshold SNIR, and Q( ) is the Gaussian tail function. $P^{d}_{c|\gamma_{sr} \geq \gamma_{th}}$ is the correct decision probability at the destination given that γ_{sr} is greater than or equal to γ_{th}. The details of error probability analysis are described in [38][68].

The average end-to-end error probability, $P_{ava.e2e}$, is the average error probability of all source and destination pairs in the network and is given by

$$P_{avg.e2e} = \frac{\sum_{i=1}^{S} P_{e2e,i}}{S} \tag{5-3}$$

, where S is the number of the source nodes in the network and $P_{e2e,i}$ denotes the end-to-end error probability of the *i*th source and the corresponding GW node pair.

During the frequency allocation procedure, if there is no available non-interfering frequency for a node to select for a link, it will choose an interfering frequency. In this case a MAC protocol is required on the link to resolve contention in real-time. This reduces end-to-end throughput, potentially severely, as shown in [15][23]. The *percentage of hops with interfering links* is defined by the average number of hops assigned an interfering frequency to the total number of hops of the path from each source node to a corresponding GW node.

Percentage of hops with interfering links =

$$\frac{\sum_{i=1}^{S} \frac{\#\ of\ hops\ with\ interfering\ links\ on\ the\ path\ from\ node\ i\ to\ a\ GW\ node}{total\ \#\ of\ hops\ on\ the\ path\ from\ node\ i\ to\ a\ GW\ node}}{S} \tag{5-4}$$

If a joint node detects a two-hop diverse path, it will select a frequency to provide diversity if one is available that meets the target SNIR threshold. The *relative diversity percentage* is defined as the average ratio of the number of hops having diversity to the number of hops with possible diversity on the path from each source node to a corresponding GW.

Relative diversity percentage =

$$\frac{\sum_{i=1}^{S} \frac{\#\ of\ hops\ with\ actual\ diversity\ on\ the\ path\ from\ node\ i\ to\ a\ GW\ node}{\#\ of\ hops\ with\ possible\ diversity\ on\ the\ path\ from\ node\ i\ to\ a\ GW\ node}}{S} \tag{5-5}$$

A hop is considered to have possible diversity if a two-hop diverse path to it exists, i.e., the only reason it will not have diversity is if no frequency above the SNIR threshold is available.

The *absolute diversity percentage* is defined as the ratio of number of hops with diversity to the total number of hops in a path.

Absolute diversity percentage =

$$\frac{\sum_{i=1}^{S} \frac{\#\ of\ hops\ with\ actual\ diversity\ on\ the\ path\ from\ node\ i\ to\ a\ GW\ node}{total\ \#\ of\ hops\ on\ the\ path\ from\ node\ i\ to\ a\ GW\ node}}{S} \tag{5-6}$$

We obtained results from two different scenarios described as follows.

- Scenario I – equal SNIR on each link: Every link is assumed to have the same received SNIR value of 10 dB. The number of frequencies in each relay network is varied from 8-12.
- Scenario II – different SNIR on each link: Each link has a random SNIR value in the range between 0dB and 20dB. The number of frequencies in each relay network is 12.

5.3.3. Evaluation

In the following results, we compare the performance of the relay network formation algorithms. As a baseline we include a case in which no diversity is added to a path. This allows us to compare the performance gains of the formation algorithms based on their ability to add diversity. These results are summarized in Table 5-2. We found, as expected, that given a fixed SNIR value, the end-to-end error probability without diversity is proportional to the number of hops on the path from each node to the GW. Figure 5-4 shows the average number of hops on the path from each node to the GW for each algorithm.

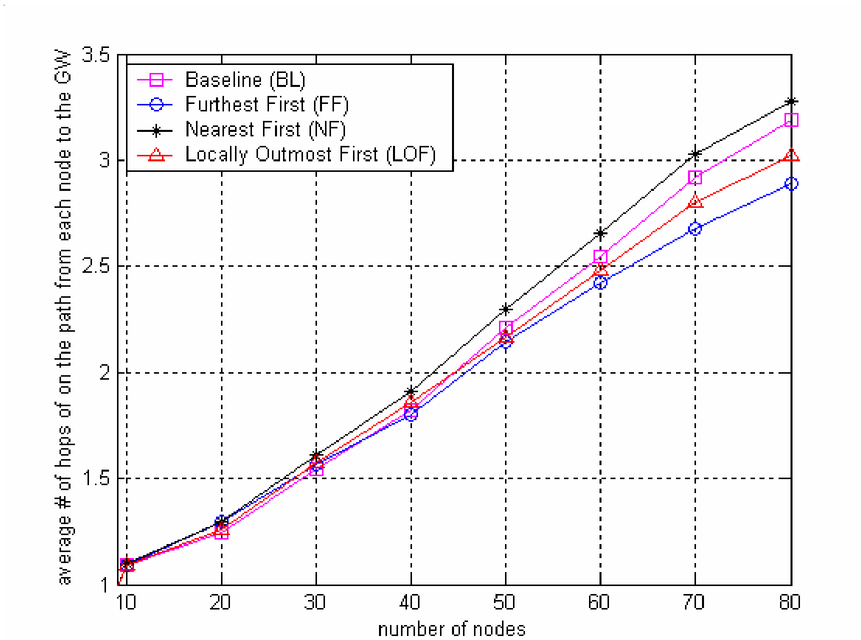


Figure 5-4. Average number of hops of each path

Algorithm	Nodes/network			
	10	**30**	**50**	**80**
BL (Baseline)	0.000008360	0.000011844	0.000016841	0.000024629
NF (Nearest First)	0.000008461	0.000012391	0.000017660	0.000025147
FF (Furthest First)	0.000008403	0.000012035	0.000016479	0.000022186
LOF (Locally Outmost First)	0.000008389	0.000012100	0.000016564	0.000023175

Table 5-2. Average end-to-end error probability of each node without diversity

5.3.3.1. Scenario I

Figure 5-5 shows the average end-to-end error probability for all five formation algorithms versus the number of nodes in the network when diversity is included. Figure 5-6 shows how varying the number of frequencies available in the relay network affects the average end-to-end error probability for each node. In NF and FF, even if we reduce the number of available frequencies in a relay network to 8, when using these algorithms, joint nodes still have enough frequencies to achieve almost 100% diversity. This illustrates the efficiency of frequency allocation when using strict ordering in scheduling join requests. BL is affected most by the reduction in frequencies because it does not allocate the frequencies efficiently and achieves low diversity.

In Figure 5-6, we observe that the algorithms that enforce an ordered path discovery starting with nodes furthest from the BS (FF and LOF) have the best performance. These results can be explained by examining Figures 5-7 and 5-8.

As shown in Figures 5-7 and 5-8, BL results in a low percentage of diversity hops both in a relative and absolute sense. With a small number of nodes, BL has the best performance because each node sends out its own *RREQ* which results in the shortest paths to the GW. However, as the number of nodes increases, this simultaneous broadcast of *RREQ* causes each intermediate node to receive many *RREP*s. This triggers a large number of frequency allocations resulting in few frequencies being available for diversity.

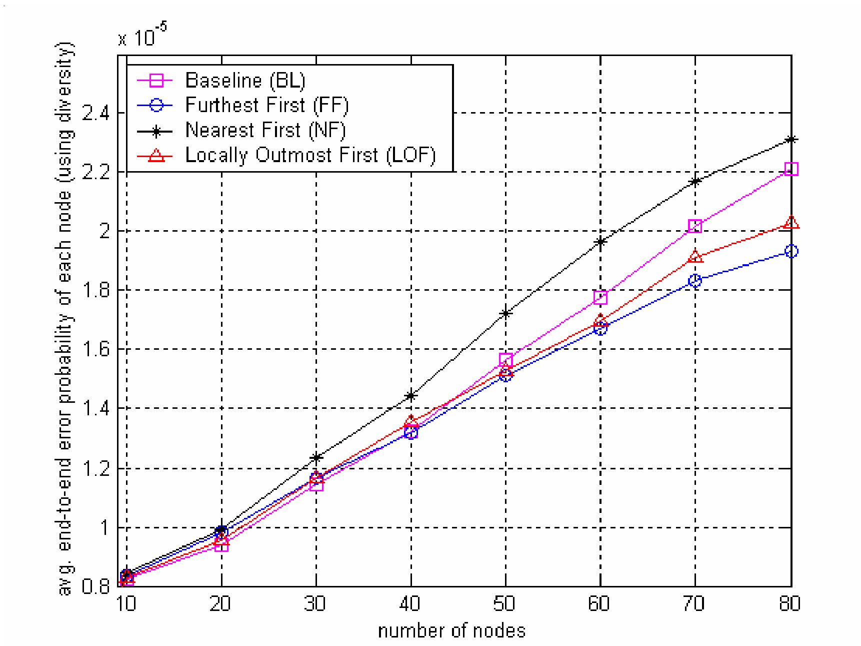
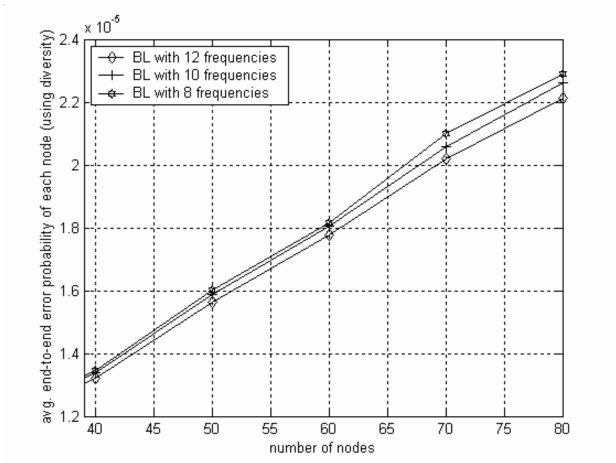
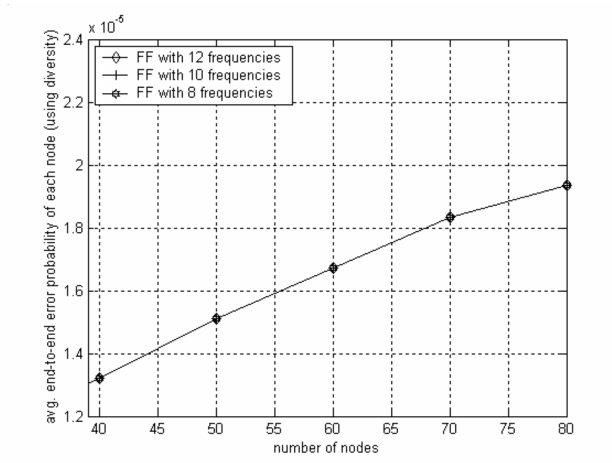


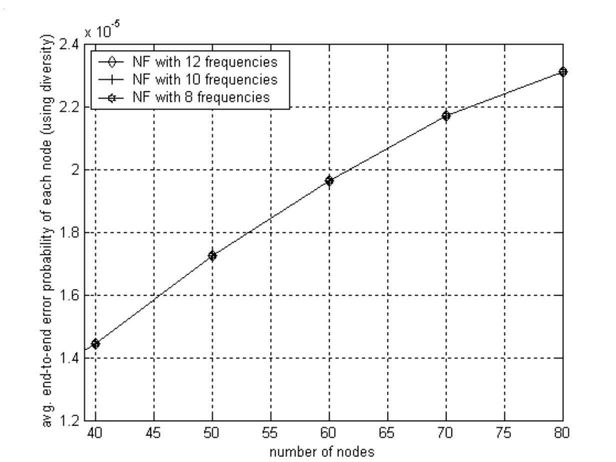
Figure 5-5. Average end-to-end error probability of each node with diversity



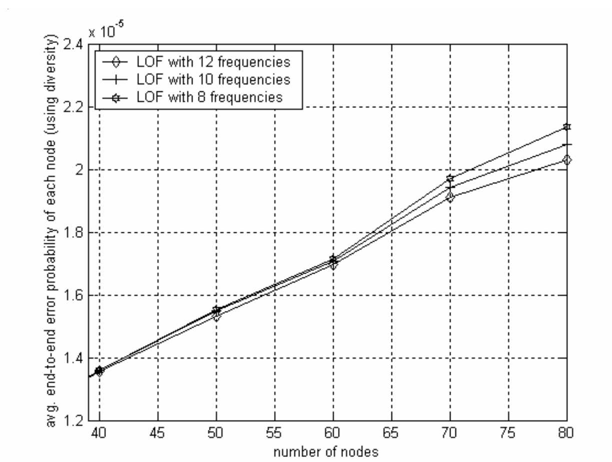
(a) Baseline (BL) Scheme



(b) Furthest First (FF) Scheme



(c) Nearest First (NF) Scheme



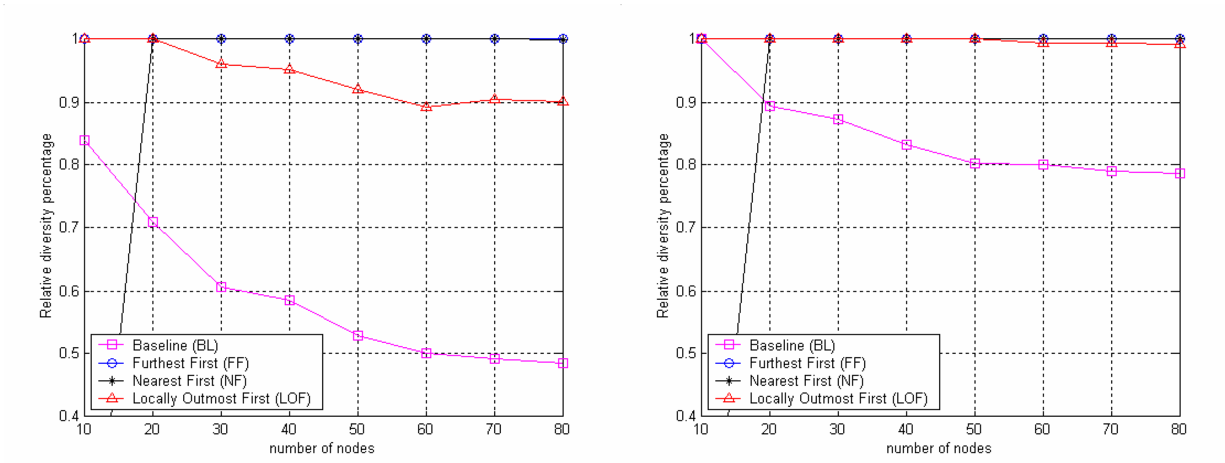
(d) Locally Outmost First (LOF) Scheme

Figure 5-6. Average end-to-end error probability using diversity with various frequencies

NF also performs poorly with respect to error rate. There are two reasons for this performance. First, NF tends to result in long average paths as shown in Figure 5-4. Second, NF also results in relay networks with few joint nodes, and therefore does not have an opportunity to add much diversity to its paths. As shown in Figures 5-7 and 5-8, NF achieves a high percentage of relative diversity, meaning that it assigns frequencies efficiently, but has a low percentage of absolute diversity. This latter fact is a result of having very few joint nodes and is the main reason for the high error rates.

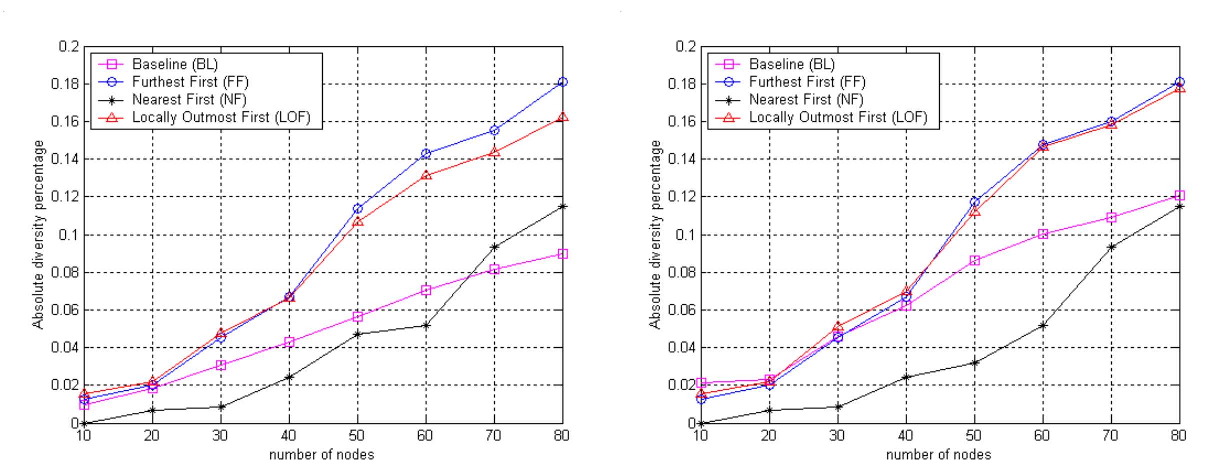
On the other hand, with FF the node furthest from the BS establishes a direct path with diversity first, allowing intermediate nodes on this path to share the frequency allocations. Many nodes adjacent to the path also passively learn the route and this share the same path. Therefore, intermediate nodes on the direct path do not have to assign new frequencies when they act as source nodes, or when they are part of the path for a different source node. As a result, fewer frequencies are used on the direct path making more frequencies available for adding diversity. As shown Figures 5-7 and 5-8, FF results in a relay network with both high relative and absolute diversity. As a result, FF has the lowest end-to-end error probability.

In Figure 5-7, we find that the percentage of diversity is inversely proportional to the degree of parallelism of the formation algorithms. With a high degree of parallelism, many nodes send out their own *RREQ* resulting in few nodes sharing a path to the GW. The existence of multiple direct paths uses the available frequencies thus reducing diversity and increasing the number of interfering links.



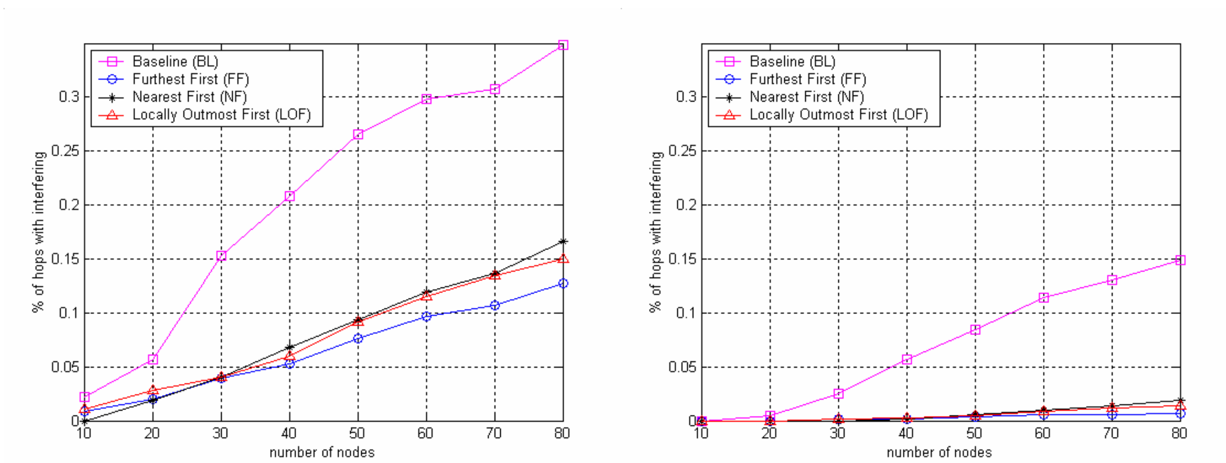
(a) # of available frequencies = 8 **(b) # of available frequencies = 12**

Figure 5-7. Relative diversity percentage on the path



(a) # of available frequencies = 8 **(b) # of available frequencies = 12**

Figure 5-8. Absolute diversity percentage on the path



(a) # of available frequencies = 8 **(b) # of available frequencies = 12**

Figure 5-9. The percentage of hops with interfering links

Figure 5-9 shows the percentage of hops with interfering links on the path from each node to the GW node. This result shows that the ordered algorithms that perform frequency allocation in order (either nearest or furthest first) are most efficient and therefore result in fewer interfering links.

5.3.3.2. Scenario II

In this case, every node has a random SNIR value in the range of [0 .. 20] dB. Therefore, diversity may not be established in some cases because no link above the SNIR threshold exists over which to setup the diversity hop. However, the successful establishment of diversity will improve performance greatly because some direct hop links may have poor SNIR values. Figure 5-10 shows the average end-to-end error probability of each node when diversity is included in the path setup.

The average end-to-end error probability depends on the SNIR values of the nodes. This graph shows that we obtain similar results as those shown in Figure 5-5. NF generally has the highest end-to-end error probability and the FF-based algorithms (FF and LOF) have the lowest end-to-end error probability as the number of nodes increases.

5.4. Discussion

In the following two subsections we discuss the impact of adding diversity to relay nodes acting as sources, and optimal frequency allocation.

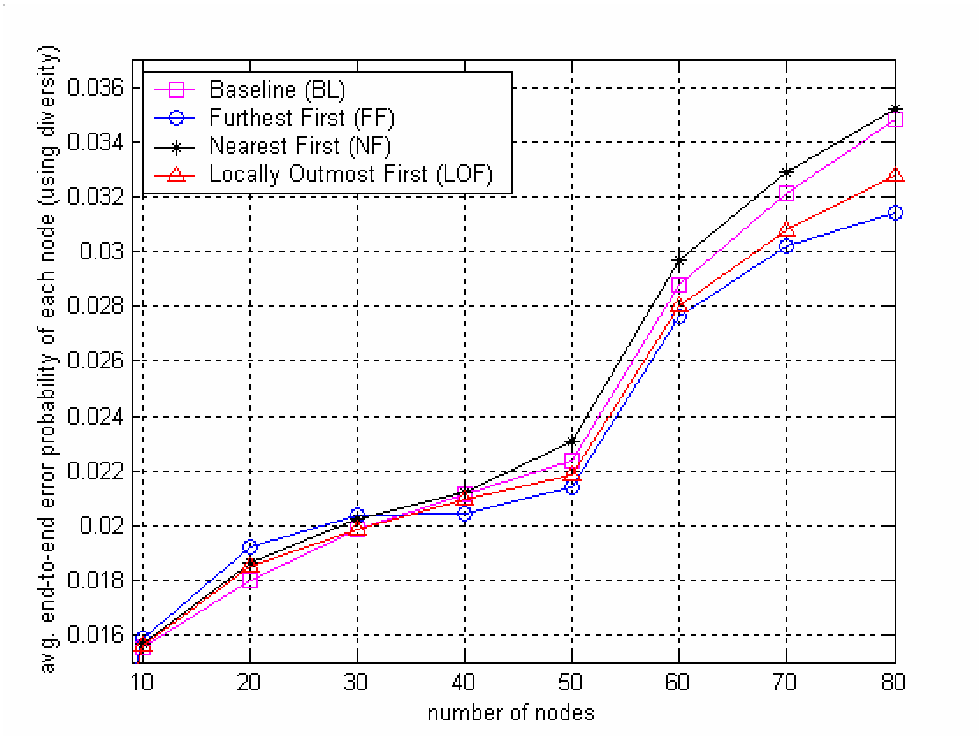


Figure 5-10. Example of avg. probability of error of each node with random SNIR (with diversity)

5.4.1. Adding Diversity to the Relay Node

As discussed in chapter 4, if a node passively learns a route, it will not participate in the procedures to establish a diversity hop from itself to its next hop node. If this node is acting as a relay node on a diversity hop for another path, there is the possibility that a diverse path can be established for this relay node when it is acting as a source as well. In order to add diversity to the relay node, the joint node sets up a diverse path for the relay node simultaneously when setting up a direct path for a source node. It assigns additional frequencies for uplink and downlink transmission of the relay node. Then it inserts all assigned frequency information into the RREP and sends it to the relay node as well as the next node on the reverse path.

Table 5-3 shows the percentage of nodes that are successful when attempting to add diversity using this method. Nodes that fail do so because there are not enough frequencies available for them to add the diversity hop. Each node that is successful will see an error rate improvement on this hop.

Because BL does not assign frequencies efficiently as mentioned in Section 5.4.3.1, it does not benefit from this phase except in cases of very small relay networks. As the number of nodes increase, the improvement of LOF decreases because they cannot add diversity to some relay nodes due to the lack of available frequencies. However, Table 5-3 shows that FF and NF algorithm do not suffer from this because they efficiently assign frequencies. Note that NF cannot add diversity when the networks size is small. In addition, even if NF can add diversity to almost all relay nodes, it has the lowest absolute diversity percentage on the path as shown in Figure 5-8. In contrast, FF-based algorithms (FF and LOF) can take full advantage of this phase.

Algorithm	Nodes/network			
	10	**30**	**50**	**80**
BL	100 %	75 %	70 %	67 %
NF	—	100 %	100 %	100 %
FF	100 %	100 %	100 %	100 %
LOF	100 %	100 %	94 %	92 %

Table 5-3. Percentage of nodes attempting to add diversity in Phase III that are successful

5.4.2. Optimality of the Formation Schemes

In this chapter, our formation schemes are based on single-path AODV routing protocol. Thus, every node first establishes a single direct path to the GW node, and then performs frequency allocation and adds diversity where possible. In general, the direct path is the shortest path from a source node to the GW node. But, because the end-to-end error probability of the network is affected by the SNIR value of each node as well as the hop count of the path, the shortest path may not be the path that yields the best performance. One possible solution is to extend path selection to multiple rounds so that several end-to-end paths may be evaluated before one is selected and committed.

Such a multi-round algorithm may also lead to more efficient frequency allocation. In the current algorithms, while returning the *RREP*, intermediate nodes assign the most suitable frequencies to the link for uplink and downlink transmission to the next hop with limited information as to the entire path. Thus, even if each node picks up the most suitable frequency for its own link, it may not be the globally optimal assignment. A multi-round algorithm would allow nodes to obtain more global information when performing frequency selection.

5.5. Concluding Remarks

In this chapter we studied the formation of relay networks for dynamic multi-radio, multi-hop wireless cellular networks. Based on the three network formation algorithms described in chapter 4, our algorithms allocate frequencies and create diversity hops for the paths between a source node and the GW node to the BS. The frequency allocations are made to reduce the number of interfering links in the network. We compare the algorithms in terms of the average end-to-end error probability, percentage of diversity of each path, and percentage of interfering links on each path in the resulting relay networks.

We found that algorithms that order the path discovery starting with nodes furthest from the BS perform best. This is because these algorithms afford the highest amount of path sharing and therefore result in the most efficient frequency assignments.

Chapter 6. Network Formation and Resource Allocation for Multi-Frequency, Multi-Hop Wireless Mesh Networks

6.1. Introduction

Wireless mesh networks (WMNs) are actively being considered as a solution to provide last-mile Internet access. WMNs have several advantages including rapid and low-cost deployment and the provision of ubiquitous network connectivity using various radio technologies. There are many academic testbeds and commercial deployments of WMNs using IEEE 802.11 a/b/g [8][61] and 802.15 [86]. IEEE 802.16 WiMax mesh networks [87] aim at providing broadband wireless access in metropolitan area networks.

In typical deployments, the mesh clients are directly connected to a relay mesh router. The relay mesh routers form a multi-hop mesh backbone to provide connectivity to the Internet. Most research on WMNs has targeted the construction of this backbone network to achieve optimal performance.

We envision an environment in which mesh clients dynamically form a multi-hop relay network to connect to a mesh backbone. Multi-hop communication will result in increased network coverage and reduced total transmission power. There has been a great deal of work on multi-hop extensions in wireless LANs [88][89] and in GSM and 3G wireless cellular networks [24][46].

In [89], the authors show that a multi-hop extension can result in a throughput improvement for wireless LANs of over 60% using a single channel and 70% when two channels are used. In [46], it is reported that the multi-hop relaying can achieve improvements of the average and maximum throughput of up to 37% and 82%, respectively, in 3G networks.

Conversely, it has been shown that multi-hop wireless networks may suffer from serious performance degradation as path lengths increase because adjacent links on a flow compete with each other. Multi-hop communication utilizing multiple channels and multiple interfaces has been proposed to overcome this limitation.

We apply the multi-hop extension to provide access to a backbone WMN. We assume every mesh client in our network is equipped with an agile radio. Even with a single interface, mesh clients may dynamically construct a relay network of non-interfering links to connect to the mesh backbone. The mobility of the mesh clients and the dynamic nature of these relay networks motivate the need for distributed algorithms to form such relay networks.

In this chapter, we propose a set of distributed algorithms to form the relay network. The algorithms include path discovery, resource allocation, path selection and local tuning. Our contributions include:

- a distributed algorithm for dynamically building relay networks comprised of non-interfering links;
- a path selection algorithm that allows mesh clients to choose from a set of paths to achieve high throughput;
- a local tuning algorithm to overcome performance degradation caused by the interference from remote clients.

We evaluate the performance of our proposed algorithms through simulations in OPNET.

The rest of this chapter is organized as follows. In Section 6.2, we present our network model. Algorithms used for relay network formation are explained in Section 6.3. In Section 6.4, we describe the local tuning algorithm. We present our simulation environment and results in Section 6.5. In section 6.6, we discuss the related issues of this chapter. We conclude this chapter in section 6.7.

6.2. Network Model

Our network consists of mesh clients and mesh routers. The mesh routers are classified into two types: relay mesh routers and GW mesh routers. Relay mesh routers serve as access points for mesh clients and form a multi-hop mesh backbone connected to GW mesh routers. The GW mesh routers act as bridges between the mesh backbone and the Internet. Every mesh client is equipped with an agile radio. They form a *multi-hop access relay network* in order to connect to a relay mesh router.

As shown in Figure 6-1, the network has two logical layers. The first layer is the multi-hop mesh backbone comprised of relay mesh routers. Relay mesh routers form this backbone using a variety of technologies including 802.11, 802.15, and 802.16 [90]. This backbone provides access to the Internet via the GW mesh router.

Since mesh routers are stationary, the topology of mesh backbone does not change frequently. The mesh backbone is formed during the network deployment and initialization phase. Therefore, the formation overhead of the mesh backbone is not critical; rather, the performance of the resulting backbone network is more important. There has been a great deal of work on constructing this backbone network to achieve optimal performance [4][10][70].

The second layer consists of wireless multi-hop access relay networks of mesh clients to reach one of the relay mesh routers. Mesh clients are mobile and when a group of clients move into a new router's serving area, they may form a new relay network on-demand, or join one of the existing relay networks already in service. The overhead for this network formation (e.g. formation delay and signaling traffic) has a significant impact on the mesh clients' performance. Due to the mobility of mesh clients and the dynamic nature of these relay networks, centralized algorithms are rendered impractical. In this chapter we focus on the formation of these relay networks.

The relay mesh routers periodically broadcast a list of available frequency bands they support and the internal frequency structure of the band. Let S be the set of all available radio frequency bands in the mesh network, $S = \{B_1, B_2, \ldots, B_N\}$, and R be the set of relay networks currently operating, $R = \{r_1, r_2, \ldots, r_R\}$. If B_{ri} and B_{rj} are frequency bands on which relay network r_i and r_j, operate, respectively (r_i, $r_j, \in R$), they satisfy the requirements: (1) $B_{ri} \in S$ and $B_{rj} \in S$, (2) $B_{ri} \neq B_{rj}$

Each frequency band may be divided into multiple non-overlapping frequencies. For example, a band B_{ri} consists of the set of non-overlapping frequencies, $F_{ri} = \{f_1, \ldots, f_K\}$, where K is the maximum number of non-overlapping frequencies in B_{ri}. These frequencies are used to construct a relay network comprised of non-interfering links so that multiple nodes within range of each other may transmit simultaneously without relying on a MAC protocol or distributed scheduling algorithm to resolve contention and prevent collisions.

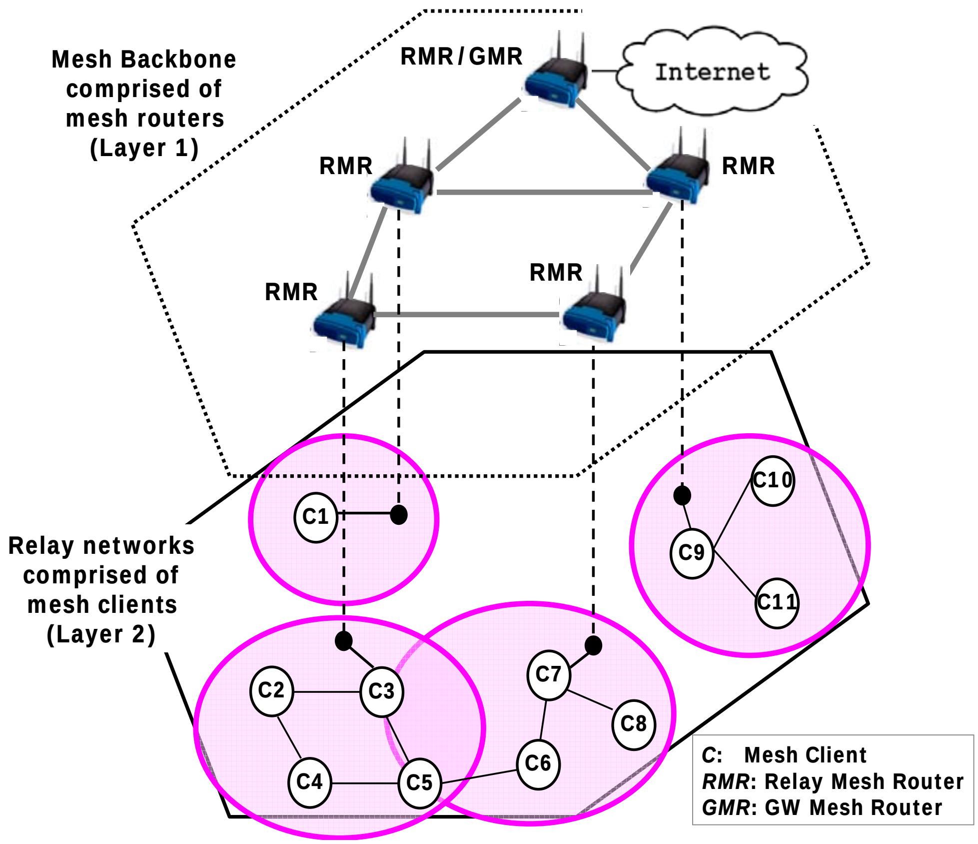


Figure 6-1. Network model

6.3. Relay Network Formation

Several steps are required to form a relay network that achieves high data rates. First, mesh clients *discover* paths to one or more mesh routers. As these paths are being discovered, frequencies are *allocated* to links in an attempt to build paths that have no contention on them. At the conclusion of this phase, a mesh client has a choice of paths over which it may attach to the mesh backbone. The mesh client *selects* the path over which it expects to achieve the highest rate. This selection is made based on several metrics as described below. Once the path is selected, the resources along that path are reserved, and the resources along the paths that were not selected are released.

In the following subsections we describe the operations for discovering paths, including allocating frequencies to links, and the algorithm for path selection.

6.3.1. Basic Operations for Path Discovery

To discover paths from a mesh client to a relay mesh router, we use a modified version of AODV. As with AODV, in order to get one-hop neighborhood information and local link availability, mesh clients broadcast *HELLO* messages periodically. Mesh clients broadcast a route request (*RREQ*) message to find paths to a relay mesh router. Intermediate clients set up the reverse path to the source and then forward the *RREQ* to the relay mesh router. When receiving a *RREQ*, the relay mesh router returns a route reply (*RREP*) message to the source of the *RREQ*.

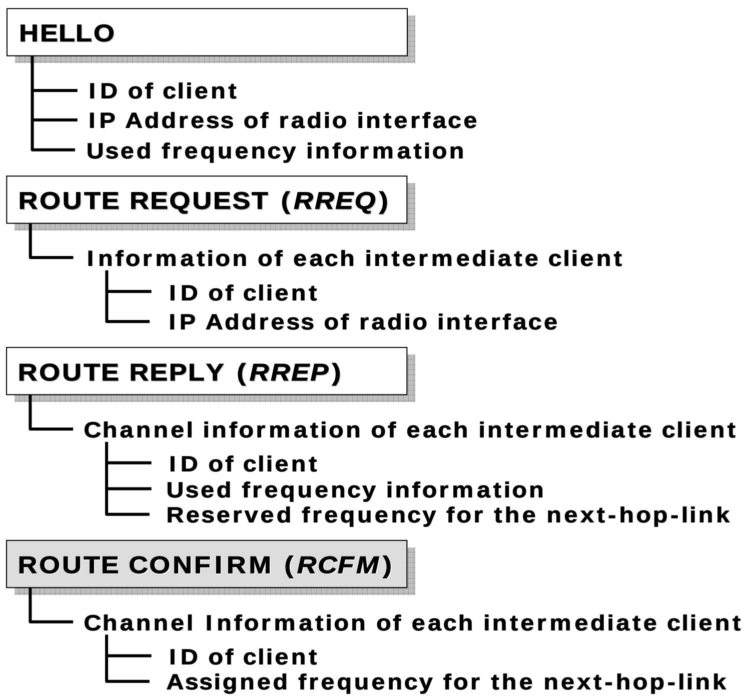


Figure 6-2. Message and information added to original AODV

Figure 6-2 shows the messages and additional information required for modified AODV. The *RREQ* carries path information as in DSR [29]. Each intermediate client appends its ID and the IP address of its radio interface to the *RREQ* before forwarding it. Upon receiving the *RREQ*, the mesh router learns the group of clients on the path.

Upon receiving a *RREQ*, the relay mesh router returns a *RREP* to the source. While returning the *RREP*, the mesh router and intermediate clients on the path reserve a non-interfering frequency for the next-hop-link, and append their used-frequency information and the reserved frequency for the next-hop-link to the *RREP*. The details of frequency allocation are explained in Section 6.3.2.

The source client may have multiple paths to the same mesh router, or paths to multiple mesh routers as shown in Figure 6-3. The client selects a primary path according to a path selection metric and returns a route confirmation (*RCFM*) message to the corresponding mesh router along the primary path. After forwarding the *RCFM*, each intermediate client actually assigns the reserved frequency to the link. The details of path selection are explained in Section 6.3.3.

6.3.2. Frequency Allocation

While returning the *RREP*, the mesh router and intermediate clients on the path are responsible for reserving a non-interfering frequency for the next-hop-link on the path. Upon receiving a *RCFM* message from the source client, they assign the reserved frequency to the link.

The algorithm used for frequency allocation is the same as the algorithm described in Section 4.3.2.2.

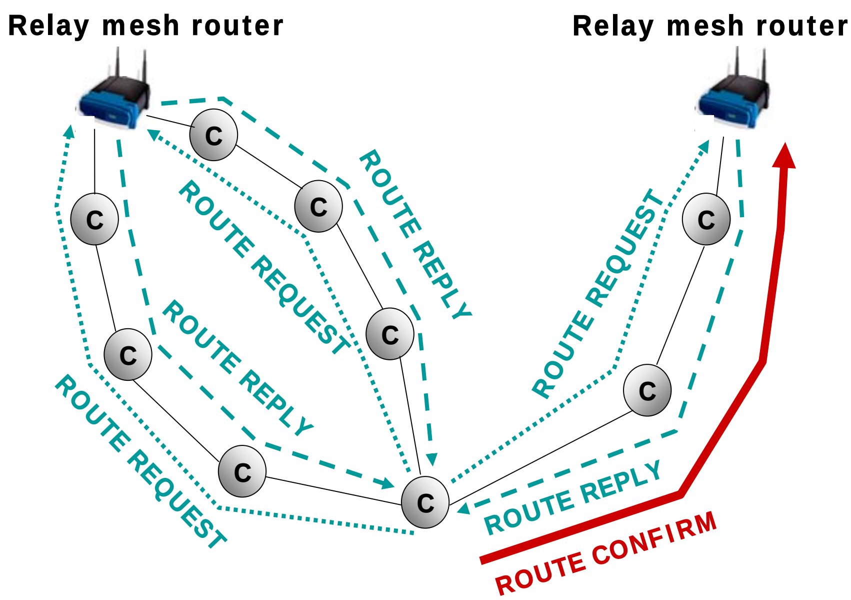


Figure 6-3. Path selection

6.3.3. Path Selection

Depending on network topology, mesh clients may have multiple paths to the same mesh router, or paths to multiple mesh routers, e.g., mesh clients 2-8 in Figure 6-1. In [9][12], the authors propose a high throughput path selection metric based on the expected number of data transmissions (ETX) of a packet and the expected transmission time of a packet on the link (ETT). They assume that frequencies are pre-allocated, or allocation is determined by some external agency. Below we define a set of metrics and a local algorithm that is run in a mesh client to choose a path to a mesh router.

6.3.3.1. Path Selection Metric

In our proposed algorithms, frequency allocation is integrated with the routing process. It is only after the source mesh client selects a primary path that frequencies are assigned. Therefore, it is impossible for each client to measure the actual ETX or ETT of all links on the paths when selecting a primary path. A naïve solution is for clients to estimate the ETX or ETT of all adjacent links on all available frequencies by probing them in advance. However, this is very costly because the clients have to potentially test $O(n^4)$ pairs even if they check only pairwise interference of two links that share an end point [54]. This is impractical for the dynamic relay networks that we consider here.

We propose a path selection algorithm considering that in a multi-hop wireless network, throughput is mainly affected by three factors: *path length*, *link rate*, and *shared rate of a link*. While under certain conditions, one factor may be more important than the others, under most normal operating conditions all three will impact performance to some degree. We define measures of these factors and weight them to provide a basis for path selection.

Path Length. The network throughput is inversely proportional to the path length, *L*:

$$T^* = \frac{1}{L} \tag{6-1}$$

Link Rate. If two adjacent links operate on the same frequency, only one link can be active at a time because of the contention between two links. We define the ratio of throughput of the links when they are active simultaneously to their throughput when they are active individually. The ratio for two contending links, assuming a perfect access protocol, is 0.5 [54]. Likewise, if three contending links operate on the same frequency, the ratio drops to 0.3 because only one link may be active among them. We use this simplifying assumption to derive our link rate path metric.

In reality, the link rate also depends on the signal quality of the link. However, when selecting a primary path, it is difficult for clients to know the actual signal quality of the links. Thus, they simply assume that all links on the path have the same rate, *R*, without any contention.

As shown in Figure 6-2, the *RREP* carries the UFI and the reserved frequency for the next-hop-link of all intermediate clients on the path. Using this information, the source client determines the number of links contending with a link *k,* denoted by C_k. The estimated rate of link *k* is given by $R/(C_k+1)$. Thus, the average estimated link rate of a path is defined as

$$\frac{\left(\sum_{k=1}^{L} \frac{R}{(C_k+1)}\right)}{L} \tag{6-2}$$

The normalized average estimated link rate of a path is defined as

$$R^* = \frac{\left(\frac{\sum_{k=1}^{L} \frac{R}{(C_k+1)}}{L} \right)}{R} = \frac{\left(\sum_{k=1}^{L} \frac{1}{(C_k+1)} \right)}{L} \tag{6-3}$$

Shared Rate of a Link. Since multiple mesh clients establish a multi-hop path connected to a relay mesh router, a link on the path is shared by subsequent clients on the path. Let S_k denote the number of clients that share a link k. Using the path information carried by a *RREP*, the client calculates the estimated shared rate of a link k defined as

$$\frac{\left(\frac{R}{(C_k+1)} \right)}{S_k} \tag{6-4}$$

Figure 6-4 shows an example of the estimated shared rate of a link. In example 1, links (1,2) and (2,3) contend with each other, thereby reducing their estimated rate to 6 Mbps. Also link (1,2) is shared by five nodes – nodes 2, 3, 4, 5, and 6. Therefore the shared rate of this link is 6 / 5 Mbps.

From these figures we observe the bottleneck link of a path is a function of the rate of the link (how much contention it has) and how many flows traverse it. This second factor is directly related to its position on the path: The closer the link is to the relay mesh router, the more mesh client flows it will be serving. To simplify our metric, we assume that the bottleneck link is the first link with contention (i.e., first unchannelized link) on the relay path closest to the relay mesh router.

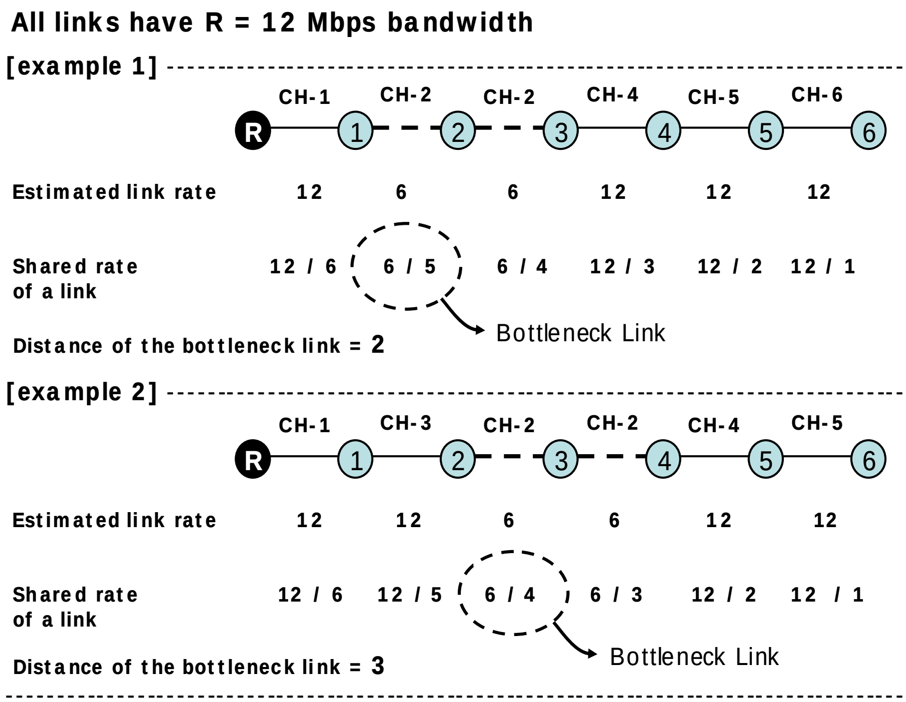


Figure 6-4. Estimation of the shared bandwidth

Based on the UFI and the reserved frequencies of all intermediate clients, the source client determines the distance of the bottleneck link from the mesh router, ***D***. The normalized distance of the bottleneck link from the relay mesh router is defined as

$$D^* = \frac{D}{L} \tag{6-5}$$

$D = L$ if there is no contending link on the path.

We define the path metric as the weighted average of these three normalized factors as follows:

$$(\ \alpha \times T^*\ )\ +\ (\ \beta \times R^*\ )\ +\ (\ \gamma \times D^*\ ) \tag{6}$$

, where $\alpha + \beta + \gamma = 1$.

Each normalized factor in equation (6-6) has a value in the range [0..1]. The source client selects the path with the maximum value of the metric.

The weights of the three terms in the path selection metric have a large impact on the performance of the relay network. In order to determine the weights α, β, and γ, we analyze the impact of each factor on the end-to-end path rate.

From (6-1) and (6-3), the ratio of the impact of link rate to path length is given by $\sum_{k=1}^{L} \frac{1}{(C_k + 1)}$. When a link k has no link contending with it, C_k is equal to 0. Therefore, if there are x contending links on a path of length L, the ratio is

$$\frac{\alpha}{\beta} = (L - x) + \sum_{k=1}^{x} \frac{1}{(C_k + 1)} \tag{6-7}$$

From (6-1) and (6-5), the ratio of the impact of shared rate of a link to path length is given by

$$\frac{\alpha}{\gamma} = D \qquad (6\text{-}8)$$

The expected value of *D* is given by

$$E(D) = 1 + \sum_{k=1}^{L-1}\left(k \times \left(\prod_{x=1}^{k-1}(1-p)\right) \times p\right) \qquad (6\text{-}9)$$

, where p = probability that a link experiences contention.

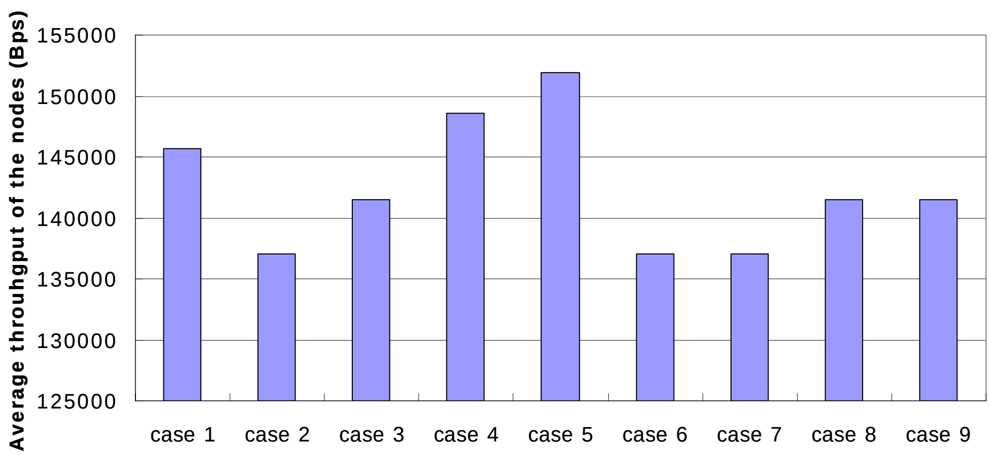
We simulated a network in which 50 mesh clients and 4 relay mesh routers are randomly deployed with eight non-interfering frequencies available. From these experiments, we observe that our proposed algorithms form a relay network in which each path has about 15% (p = 0.15, x = 0.15*L) of its links experiencing contention on the average, and each link experiencing contention does so with an average of 2.2 links (C_k = 2.2). The average path length of the clients is 6 (L = 6).

Based on these numbers, and from (6-7), the ratio of α to β should be 5.4:1. Likewise, the expected location of the first contending link on a path is D = 2.9. Thus, from (6-8), the ratio of α to γ should be 2.9:1. With the constraint that $\alpha + \beta + \gamma = 1$, we estimate, that in this scenario good settings are $\alpha \approx 0.6$, $\beta \approx 0.1$, and $\gamma \approx 0.3$.

To confirm these results, we measured the resulting throughput of the relay networks formed in this scenario for several different settings of α, β, and γ. The results, summarized in Figure 6-5, show that in fact, these settings achieve the highest average throughput. Note that we compared these settings against those that include only one of each of the three factors (cases 1-3).

case	α	β	γ	case	α	β	γ
1	1	0	0	6	0.3	0.6	0.1
2	0	1	0	7	0.1	0.6	0.3
3	0	0	1	8	0.3	0.1	0.6
4	0.6	0.3	0.1	9	0.1	0.3	0.6
5	0.6	0.1	0.3				

(a) Weight of path selection metric



(b) Average throughput of the clients

Figure 6-5. Weight of metric vs. average throughput

6.3.3.2. Dynamic Weight Selection

In a WMN, mesh clients join and leave a relay over time. As a result, the paths established to the relay mesh router and the assigned frequencies of the links on the paths will also change. This in turn may lead to a change of the link rate on the path and may result in a new bottleneck link on a path. Thus, we need a weight selection scheme dynamically adaptable to the network status.

In order to calculate the weights, α, β, and γ, we need current network status information represented by the average value of L, p, and C_k (denoted by $\overline{L}$, $\overline{p}$, and $\overline{C_k}$, respectively) of existing nodes in the relay network.

The values of L, p, and C_k of each client in the relay network are gathered by the relay mesh router periodically. Based on the gathered information, the relay mesh router calculates $\overline{L}$, $\overline{p}$, and $\overline{C_k}$ and finally calculates α, β, and γ as described in Section 6.3.3.1. It then broadcasts the updated weights in the relay network.

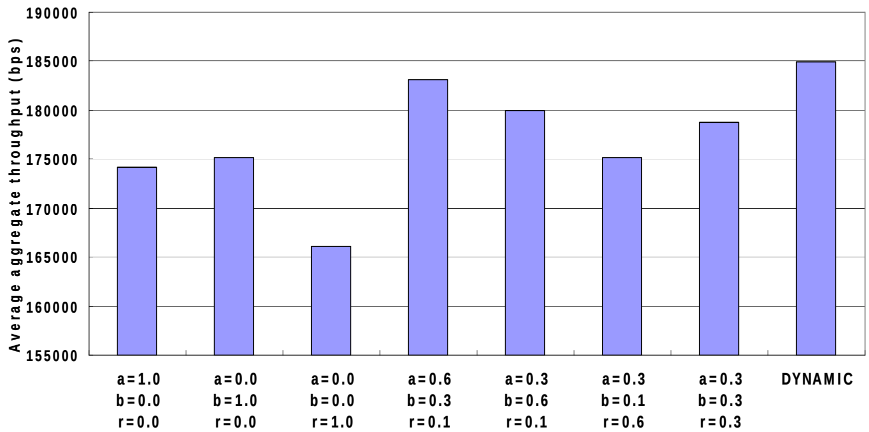
The new weights are delivered to clients intending to join the relay network by the relay mesh router when the new client establishes a path. Based on these new weights, the client calculates the path selection metric and chooses a primary path. In order to minimize the overhead of information exchange for the weight updates between a relay mesh router and the mesh clients, all information is appended to data packets exchanged between them.

In effect, by weighting the factors of path length, link rate and shared rate of a link, we are balancing their impact. In certain circumstances this balance may not lead to an optimal choice of a path. The circumstances under which this occurs are typically those of extreme conditions, for example when there are no contending links in a network or during extreme congestion. We found that under these circumstances, the selection of paths does not have a large impact on the overall network performance because either all paths provide very good performance, or none do. To evaluate the effectiveness of our dynamic algorithm under circumstances when path selection will impact network performance, we ran simulations under different conditions and determined the average rate achieved by the nodes.

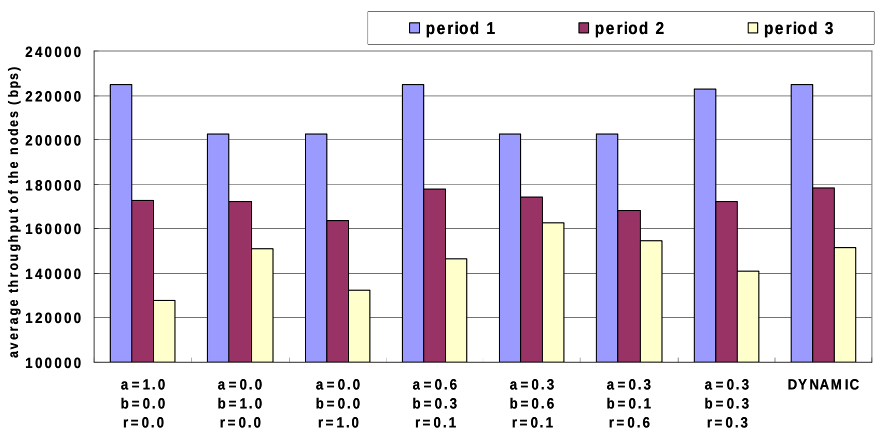
In all periods of the simulation, the average path length, $\overline{L}$ = 3. During the first period (period 1), few users join the mesh network. At this time, the probability of a link experiencing contention, $\overline{p}$, and the number of contending links, $\overline{C_k}$, are very low, 0.01 and 0.1, respectively. As the number of clients joining the relay network increases, the amount of contention and existence of bottleneck links also increase. During this period (period 2), $\overline{p}$ = 0.3 and $\overline{C_k}$ = 1.3. During the last period (period 3), the network is heavily congested. The probability for contention and number of contending links increase to $\overline{p}$ = 0.6 and $\overline{C_k}$ = 1.7.

In Figure 6-6-(a), we show the average rates achieved for the various settings over the duration of the simulation. As can be seen, the use of a single parameter to select the path (α, β, or γ has a value equal to 1.0 as shown in the first three bars) results in lower data rates.

For all periods, the dynamic algorithm that attempts to include and balance all three metrics outperforms a selection based on a single metric. In some cases preset weights do outperform the dynamic weights. This is because the dynamic algorithm attempts to balance the impact of the metrics whereas the preset weights are biased towards some



(a) Average aggregate throughput



(b) Snapshot of the throughput measured during only each time period.

Figure 6-6. Performance of the dynamic weight selection

factors. However, over the long term when a network experiences a range of operating conditions, the dynamic scheme performs best.

This is shown in Figure 6-6-(b). We see that the average rate achieved by the dynamic algorithm outperforms all of the other selected weights during at least one period, while it underperforms some of the pre-selected weights during a specific period. However, even for these cases the performance achieved by the dynamic weights was within 93% or the best performing set of weights for the period in question. Furthermore, the dynamic weights achieved higher overall average throughput as shown in Figure 6-6-(a).

6.4. Local Tuning of the Resulting Relay Network

The throughput of multi-hop wireless networks can be degraded by two following factors:

- *Contention on a path*: This includes contention with adjacent nodes and nodes operating on the same frequency on the same path.
- *Interference from remote nodes*: This includes interference from remote nodes using the same frequency on the same or a different path.

Since AODV is a shortest path routing protocol, a mesh client that is a large distance from a client, but still within the client's transmission range is likely selected as the next hop. Our path discovery and frequency allocation algorithm ensures that clients within two hops do not use the same frequency if they have available non-overlapping frequencies. Thus, our algorithm avoids the contention on the path. If the paths contain some contending links, our path selection algorithm estimates the throughput of the path considering the contending links and selects a path with the best expected throughput.

However, in general, the interference range of a wireless device is larger than its transmission range. In this case, the mesh clients may suffer from interference with clients on the same path more than two hops away. They may also experience interference from clients on different paths within interference range, or when a client operating on the same frequency moves into its interference range.

In order to interference from remote nodes, we use the local tuning algorithm described in Section 4.4.1.

The frequency switching of a pair of clients can trigger a cascade of frequency switching in the network. However, if every client selects its frequency within a finite number of frequency switches, the frequency allocation reaches a stable state where nodes cease changing frequencies [31].

6.5. Performance Evaluation

We simulated the set of proposed algorithms in OPNET and measured the throughput gains achieved by the single and multi-frequency relay networks and compared it with the optimal throughput which the relay network can achieve.

In this simulation, the mesh clients use IEEE 802.11a which provides up to 12 non-interfering frequencies. The clients use 8 frequencies from 5.15 to 5.35 GHz band for the access relay networks. The experiments are based on a 4000x4000m^2 network with 4 relay mesh routers and up to 50 mesh clients. Each client downloads 10 Mbytes from a FTP server in the Internet. Table 6-1 summarizes the simulation parameters.

Parameter	Value
Cell size	4000 m X 4000 m
N (# of mesh clients / cell)	10 → 50
# of relay mesh routers / cell	Up to 4 routers
Download file size	10Mbytes / client
Application	FTP / TCP
Frequency band & # of orthogonal frequencies	Frequency band used for 802.11a, 8
Air interface range	250m

Table 6-1. Simulation parameters

In this simulation, in order to quantify the throughput gain of the relay network comprised of mesh clients, we set up the backbone network such that all relay mesh routers are connected to the GW mesh router directly with 1000BASE-T high speed Ethernet link. Thus, the transmission delay between the GW mesh router and the relay mesh router can be ignored. Because OPNET does not support mobile nodes with agile radios, we emulate the agile radio environment by assigning multiple radios to the mobile nodes.

We measure the performance of the relay network in terms of the average throughput of the mesh clients forming the relay network. We quantify the performance gains achieved by using a multi-frequency relay network. Based on the results described in Section 6.3.3., the weights α, β, and γ are set dynamically.

We measure the average throughput in the following three scenarios and quantify the benefits of the local tuning as well:

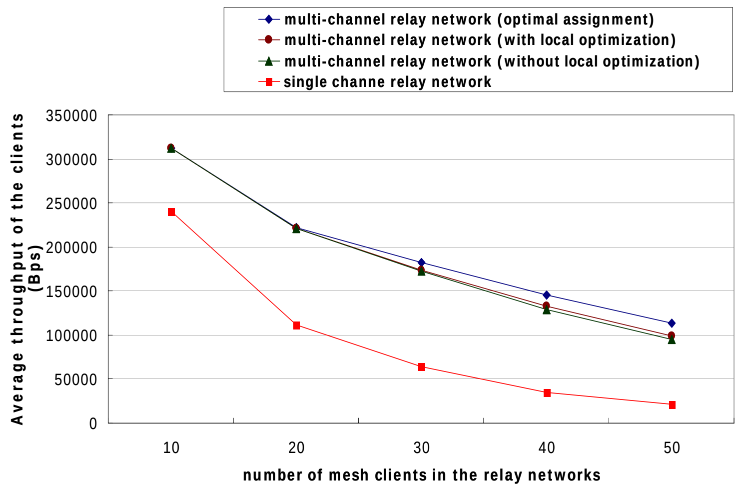
- **Scenario 1**: In [42], the authors show that for multi-hop configuration some commercial products (e.g. Netgear cards) may provide only one orthogonal frequency. Thus, in this scenario, even if mesh clients have multiple radio interfaces, they tune all interfaces to the same frequency.
- **Scenario 2**: The relay network uses up to 8 frequencies. All clients in the relay network make frequency allocations according to the frequency allocation scheme in Section 3.
- **Scenario 3**: As with scenario 2, the relay network uses up to 8 frequencies. Unlike scenario 2, the frequencies are assigned by a centralized allocation scheme based on the interference constraint in [4]. This scheme assigns a

frequency to a link based on an interference constraint that considers a node's interference range and transmission range. It makes optimal frequency allocations resulting in no contention and minimum interference.

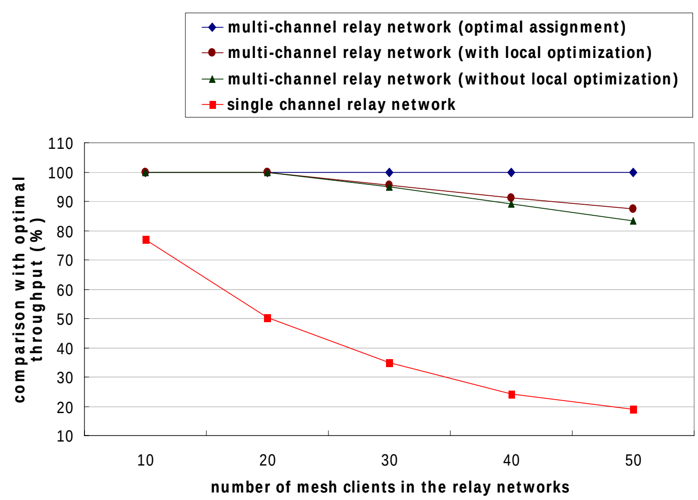
Figure 6-7-(a) shows that if we leverage multiple frequencies, higher average network throughput can be achieved. Figure 6-7-(b) shows the percentage of the throughput achieved by our proposed formation algorithms when the optimal throughput which the relay network can achieve is set to be 100%. It shows that our proposed formation algorithm achieves at least 83% of the average throughput as the optimal frequency allocation algorithm. It also shows that using local tuning improves average throughput up to 6%.

In Figure 6-7, we observe that as the number of clients increases in the relay network, the improvement in performance of the optimal frequency allocation in scenario 3 increases. This is because it is more likely that the frequency allocation algorithm will result in several nodes that are within interference range being assigned the same frequency.

Figure 6-8 illustrates this effect. This figure shows an example of the received SNIR of individual nodes in each scenario with the same topology. It is clearly seen that there is a considerable improvement in SNIR when multiple frequencies are used. Figure 6-8 also shows the effect of local tuning. With local tuning the SNIR is close to when using optimal frequency allocation. The performance improvement of local tuning is not apparent in Figure 6-6 because this figures shows the average rate of all nodes whereas the local tuning algorithm benefits the lower performing nodes only, thus resulting in more balanced performance.



(a) Average throughput of the mesh clients



(b) Throughput Comparison

Figure 6-7. Throughput of the relay networks

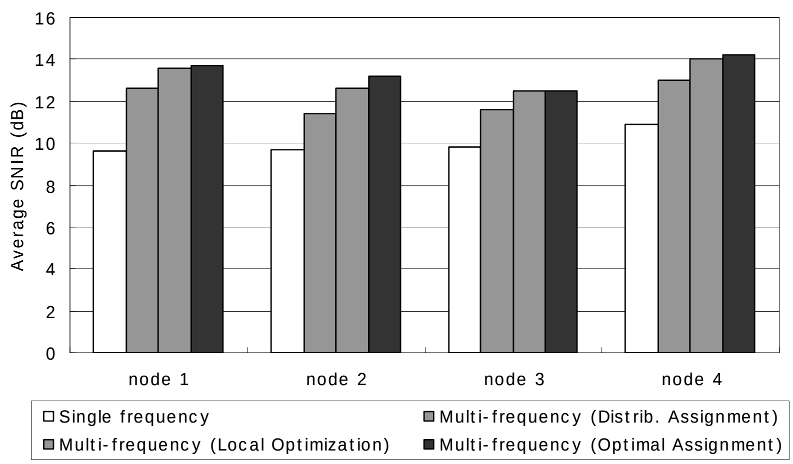


Figure 6-8. Example of node's SNIR (number of nodes = 40)

6.6. Discussion

6.6.1. Path Selection Metric Considering Link Life Time

Link failures and path recovery caused by mesh client mobility have a serious impact on the network performance especially when the mesh clients use TCP-based applications. Therefore, in addition to the terms in the path selection metric described in Section 6.3.3, we can consider the expected lifetime of a link. Su et al. [91] define the link expiration time based on the mobile nodes' position, velocity, and direction. The extension of our path selection algorithm to incorporate this new metric is left for future work.

6.6.2. Interface Assignment

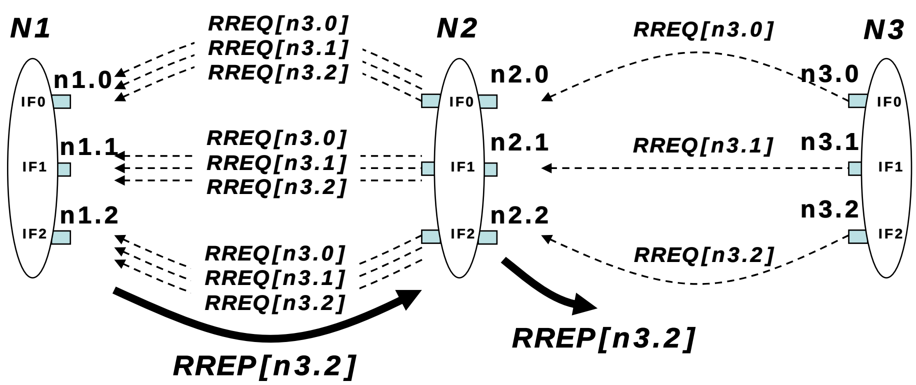
A node equipped with multiple radio interfaces is capable of distributing data over all interfaces simultaneously. However, this requires the destination node to reorder received packets. It also degrades the network throughput when the node has connections to multiple destinations with different data rates.

Our modified version of AODV allows the mesh clients to assign an interface to the path established for each connection. Since AODV is originally designed for single-interface nodes, it does not include any interface assignment and has the following problems in a multi-radio environment:

- In original AODV, nodes discard the duplicates of a *RREQ* by checking its source address. When a node broadcasts a *RREQ* in a multi-radio environment, the message is generated from all interfaces of the node. Each *RREQ* has a different source address corresponding to the source node's outgoing interface. The intermediate nodes cannot distinguish which *RREQ* is generated from the same source by only checking the source address of *RREQ*. They thus forward all *RREQs*. As a result, a total of $O(i^n)$ messages are flooded in the network, where i is the maximum number of interfaces of the nodes and n is the path length. Figure 6-4 shows an example of original AODV operating in a multi-radio environment. In this example, all nodes have 3 interfaces. When node N3 sends out a *RREQ*, three *RREQ*s are generated. Node N2 rebroadcasts each *RREQ* on all interfaces. Thus, nine messages are generated by node N2.

- In AODV, upon receiving a *RREQ*, the node stores the source address of the *RREQ* and the address of the outgoing interface of the previous-hop-node in the routing table. When receiving a *RREP*, the node retrieves the previous-hop-node's address using the destination address of the *RREP,* and then forwards the *RREP* through the corresponding interface. As shown in Figure 6-9, sometimes the same interface is used for both the reverse and forward paths. In this case, the node *cannot assign any new frequency to the next-hop-link on the reverse path* because the interface should be used for the link to the previous-hop-node on the path. This decreases the frequency utilization of the network. Moreover, since these two adjacent links contend with each other, this results in throughput degradation.

Pirzada, et al. [57] propose multi-radio AODV and solve this problem by having the intermediate nodes broadcast a *RREQ* on all interfaces except the incoming interface of the *RREQ*. This is based on the assumption that all nodes have the same number of interfaces with identical frequency allocation.



<Routing Table of N2>

	src	dst	next	interface
Reverse path	N1.0	N3.2	N3.2	IF2
Forward path	N3.2	N1.2	N1.2	IF2

Figure 6-9. Original AODV operating in a multi-radio environment

If a node has only one interface or different frequency allocation from its neighbors so that only the incoming interface has a common frequency with the neighbors, this protocol does not work. In Figure 6-9, if nodes N1, N2, and N3 have only one common frequency on IF 2, node N2 cannot forward the *RREQ* because IF 2 is the incoming interface on which the *RREQ* is received from node N3.

We propose an enhanced multi-radio AODV protocol. Unlike the solution of [57], our protocol allows each intermediate node to broadcast the *RREQ* on all interfaces and hence works even when several nodes have a single interface. Our enhancements include:

- ***Identification of the Source.*** As shown in Figure 6-2, the source and intermediate clients append their ID and the addresses of all their interfaces to the *RREQ*. Therefore, each intermediate client can distinguish and discard the duplicates of a *RREQ* based on the ID of the source client and forward only the first *RREQ* received. Moreover, the receiving client can obtain the address information of all interfaces of the preceding nodes along the path with just the first *RREQ*. This enhancement can significantly reduce the number of messages flooded in the network.
- ***Interface Selection.*** The mesh clients assign an interface to the path as follows:
  - Upon receiving a *RREP*, if the same interface is used for both reverse and forward paths, the intermediate mesh client finds a new available interface which has a common frequency with next hop.
  - If there is no interface available, it assigns the same interface that was used for receiving the *RREP*.

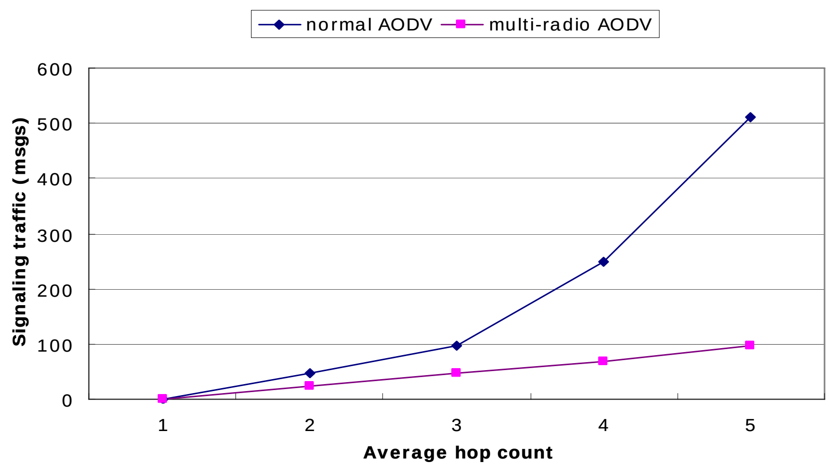
- It updates its routing table according to the assigned interface and forwards the *RREP* to the next hop.

As shown in Figure 6-2, the *HELLO* messages broadcast periodically carry the Used Frequency Information (UFI) of each interface of one-hop neighbors. The client therefore can find an interface that has a common frequency with the next-hop-node. This enhancement can increase the frequency utilization of the network and increase network throughput.

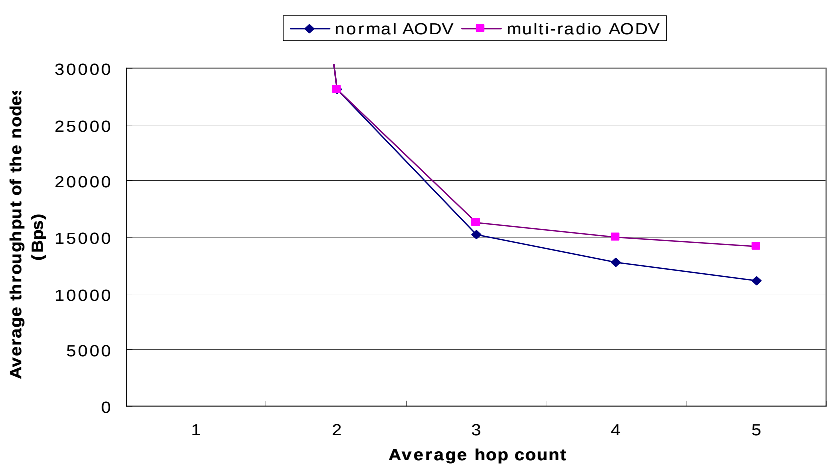
To isolate the impact of our multi-radio AODV protocol we performed an experiment with a simple topology in which every client has three radios and has a path to a mesh router with a length of up to 5. Figure 6-10 shows that when using original AODV the total number of signaling messages flooded in the network increases exponentially and the average throughput of mesh clients decreases exponentially as the average path length of each client increases. Our interface assignment algorithm reduces the amount of signaling traffic flooded in the network significantly and improves the network throughput.

6.7. Concluding Remarks

In this chapter, we propose relay-based access to wireless mesh networks. While there have been many studies focusing on centralized routing and frequency allocation algorithms for multi-hop backbone networks of mesh routers, in this paper we study the formation of dynamic relay networks that consist of mesh clients. We propose a set of



(a) Signaling traffic



(b) Average throughput of the nodes

Figure 6-10. Effect of interface assignment

algorithms including path discovery, resource allocation, and path selection. We also propose a local tuning algorithm to improve network throughput.

The results show that by leveraging multiple frequencies, we can achieve higher average node throughput. Our proposed algorithms achieve about 83% of the optimal average node throughput. Moreover, the local tuning algorithm further improves average throughput.

Chapter 7. Conclusion

In this thesis, we focus on efficient resource allocation in the core network and radio access network of wireless cellular networks. Additionally, we address resource allocation in wireless mesh networks.

Efficient resource allocation results in reduced end-to-end delay and improved network throughput. We therefore measure the performance of the proposed resource allocation algorithms in terms of end-to-end delay and network throughput.

The evaluation results show that the efficient allocation of processing resources of the core network elements can reduce the end-to-end call delivery delay in wireless cellular networks. The evaluation results also show that the utilization of multiple frequencies and the efficient assignment of the frequencies so as to reduce the number of interfering links in relay networks can improve the throughput of multi-hop wireless cellular networks and wireless mesh networks. The frequency allocation combined with local tuning to reduce further interference in resulting relay networks can achieve the performance close to optimal.

7.1. Summary of Contributions

The major contributions of this thesis are outlined as follows:

- Resource allocation in the core network
    - We study three interworking approaches based on mobility management model to support seamless roaming across heterogeneous wireless networks; master-slave mobility management approach, federated mobility management approach, and unified mobility management approach. We compare the performance of these approaches conducting a comparative performance analysis. With network components modeled as an M/M/1 queuing system, we measure the average call delivery delay of the SIP terminated calls of each approach and compare them.
    - We propose an algorithm to achieve perfect load balancing across each network element and get better performance for the seamless roaming for IP-based services.

- Resource allocation in the radio access network
    - We define network models of multi-frequency, multi-hop wireless cellular networks and wireless mesh networks in which user devices dynamically form a relay network to establish a connection to the BS or GW mesh routers.
    - We propose a set of distributed algorithms for relay network formation. In case of multi-hop wireless cellular networks, these algorithms include GW discovery, path discovery to a GW, and frequency allocation. In case of multi-hop wireless mesh networks, these algorithms include path discovery to a relay mesh router, path selection, and frequency allocation. Since we assume wireless mesh networks that consist of nodes equipped with multiple radios, we also propose an

interface assignment algorithm.

- We propose a local tuning algorithm to reduce interference in resulting relay networks further.
- We evaluate the performance of these algorithms in terms of the throughput of the resulting relay network. We compare that with the performance which we can achieve using multiple frequencies with optimal assignment.
- For multi-hop wireless cellular networks, we also propose algorithms that provide the two-hop path diversity to reduce the bit error rate and increase the network goodput.

7.2. Future Research Directions

A great deal of work has recently focused on a method called *network coding* to achieve asymptotically optimal performance for wireless networks even with high packet loss rate. In this method coupled with multicasting, the data is coded across packets in order to withstand the loss of packets. Network Coding allows mixing of data at intermediate network nodes. A receiver collects these data packets and deduces from them the messages that were originally intended for the multicast members.

Network coding requires that routes should be diverse enough so that congestion in one location will not bring down a whole stream. Therefore, the performance of network coding is tightly coupled with the routing protocol in that high routing diversity results in high degree of data coding to achieve improved network throughput. Moreover, the network coding with an intelligent link channelization can result in much higher throughput gain.

Some multimedia applications may need to multicast data to mobile nodes in a cell of multi-hop wireless cellular networks or a set of mesh clients connected to a relay mesh router in wireless mesh networks. We plan to extend our research by designing a set of relay network formation algorithms that utilize network coding efficiently for down-stream data transmission.

References

[1] G. Aggelou and R. Tafazolli, "On the relaying capability of next generation GSM cellular networks," IEEE Wireless Communications, vol. 8, no.1, pp. 40–47, 2001.

[2] D. Aguayo, J. Bicket, S. Biswas, G. Judd, and R. Morris, "Link-level Measurements from an 802.11b Mesh Network," In Proc. of ACM SIGCOMM, 2004.

[3] I. Akyildiz, X. Wang, W. Wang, "Wireless Mesh Networks: A Survey," Elsevier Computer Networks Journal, 2005.

[4] M. Alicherry, R. Bhatia, L. Li, "Joint Channel Assignment and Routing for Throughput Optimization in Multi-radio Wireless Mesh Networks," In Proc. of ACM MOBICOM, 2005.

[5] F. Bai, N. Sadagopan, A. Helmy, "A Framework to Systematically Analyze the Impact of Mobility on Performance of Routing Protocols for Ad Hoc Networks," In Proc. of IEEE INFOCOM, 2003.

[6] P. Bender et al, "CDMA/HDR: a bandwidth-efficient high-peed wireless data service for nomadic users," IEEE Communications Magazine, vol. 38, pp.70-78, 2000.

[7] A. Bertossi, M. Bonuccelli, "Code assignment for hidden terminal interference avoidance in multihop packet radio network," IEEE/ACM Trans. on Networking, 1995.

[8] J. Bicket, D. Aguayo, S. Biswas, and R. Morris, "Architecture and Evaluation of an Unplanned 802.11b Mesh Network," In Proc. of ACM MOBICOM, 2005.

[9] D. Couto, et al., "A High-Throughput Path Metric for Multi-hop Wireless Routing," In Proc. of ACM MOBICOM, 2003.

[10] A. Das, H. Alazemi, R. Vijayakumar, S. Roy, "Optimization Models for Fixed Channel Assignment in Wireless Mesh Networks with Multiple Radios," In Proc of IEEE SECON, 2005.

[11] S. Das, C. E. Perkins, E. M. Royer, "Performance Comparison of Two On-demand Routing Protocols," In Proc. of IEEE INFOCOM, 2000.

[12] R. Draves, J. Padhye, B. Zill, "Routing in Multi-Radio, Multi-Hop Wireless Mesh Networks," In Proc. of ACM MOBICOM, 2004.

[13] Ephremides and Truong, "Scheduling broadcasts in multihop radio networks," IEEE Transaction on Communication, vol. COM-38, 1990.

[14] A. El Gamal and S. Zahedi, "Capacity of a Class of Relay Channels with Orthogonal Components," IEEE Transactions on Information Theory, vol. 51, no. 5, pp.1815-1817, 2005.

[15] M. Gerla, K. Tang and R. Bagrodia, "TCP performance in wireless multi-hop networks," In Proc. of IEEE WMCSA, 1999.

[16] J. Glossner, D. Iancu, E. Hokenek, and M. Moudgill, "A software-defined communications baseband design," IEEE Communications Magazine, vol. 41, no.1, pp.120–128, 2003.

[17] O. Haase, M. Xiong, and K. Murakami, "Multi-Protocol Profiles to Support User Mobility Across Network Technologies," In Proc. of IEEE MDM (Mobile Data Management), 2004.

[18] O. Haase, K. Murakami, and T.F. La Porta, "Unified Mobility Manager – Enabling Efficient SIP/UMTS Mobile Network Control," IEEE Wireless Communications Magazine, vol. 10, no. 4, pp. 66-75, Aug. 2003.

[19] A. Hac, "A Distributed Algorithm for Performance Improvement Through Replication and Migration," In Proc. of IEEE Computer Networking Symposium, 1986.

[20] A. Hac, and T. J. Johnson, "A Study of Dynamic Load Balancing in a Distributed System," In Proc. of ACM SIGCOMM Symposium on Communications, Architectures and Protocols, 1986.

[21] A. Hac,, and X. Jin, "Dynamic Load Balancing in a Distributed System Using a Sender-Initiated Algorithm," In Proc. of IEEE-CS and ACM SIGARCH Workshop on Instrumentation for Distributed Computing Systems, 1987.

[22] P. Herhold, E. Zimmermann, and G. Fettweis, "A Simple Cooperative Extension to Wireless Relaying," In Proc. of Int. Zurich Seminar on Communications, 2004.

[23] G. Holland and N. H. Vaidya, "Analysis of TCP Performance over Mobile Ad Hoc Networks," In Proc. of ACM MOBICOM, 1999.

[24] W. Hongyi, Q. Chunming, S. De, and O. Tonguz, "Integrated cellular and ad hoc relaying systems: ICAR," IEEE Journal on Selected Areas of Communications, vol. 19, no.10, pp. 2105–2115, 2001.

[25] A. Høst-Madsen and J. Zhang, "Capacity bounds and power allocation in wireless relay channel," IEEE Transactions on Information Theory, vol. 51, no. 6, pp. 2020–2040, 2005.

[26] H. Hsieh and R. Sivakumar, "Performance Comparison of Cellular and Multi-hop Wireless Networks: A Quantitative Study," In Proc. of ACM SIGMETRICS, 2001.

[27] R. Isukapalli, T. Alexiou, and K. Murakami, "Global Roaming and Personal Mobility with COPS Architecture in SuperDHLR," Bell Labs Technical Journal, vol.7, no.2, pp.3-18, 2002.

[28] A. Jalali et al., "Data throughput of CDMA-HDR a high efficiency-high data rate personal communication wireless system," In Proc. of IEEE VTC, 2000.

[29] D. Johnson and D. Maltz, "Dynamic Source Routing in Ad Hoc Wireless Networks," Mobile Computing, Chapter 5, pages 153-181, Kluwer Academic Publishers, 1996.

[30] L. Kleinrock, “Queuing Systems, vol. 2: Computer Applications,” Wiley Interscience, 1976.

[31] B. Ko, V. Misra, J. Padhye, D. Rubenstein, “Distributed Channel Assignment in Multi-radio 802.11 Mesh Networks”, Tech Report, Columbia Univ., 2006.

[32] M. Kodialam, T. Nandagopal, “Characterizing the Capacity Region in Multi-Radio Multi-Channel Wireless Mesh Networks,” In Proc. of ACM MOBICOM, 2005.

[33] G. Kramer, M. Gastpar, and P. Gupta, ”Cooperative strategies and capacity theorems for relay networks,” IEEE Transactions on Information Theory, vol. 51, no. 9, 2005.

[34] A. Kusuma, L. Andrew, “Minimum Power Routing for Multihop Cellular Networks,” In Proc. of IEEE GLOBECOM, 2002.

[35] P. Kyasanur, N. Vaidya, “Capacity of Multi-Channel Wireless Networks: Impact of Number of Channels and Interfaces,” In Proc. of ACM MOBICOM, 2005.

[36] J. Laneman and G. Wornell, “Exploiting Distributed Spatial Diversity in Wireless Networks,” In Proc. of Allerton Conf. Communications, Control, and Computing, 2000.

[37] J. Laneman, D. Tse, and G. Wornell, “Cooperative diversity in wireless networks: Efficient protocols and outage behavior,” IEEE Transactions on Information Theory, vol. 50, no. 12, pp. 3062–3080, 2004.

[38] K. Lee and A. Yener, “On Resource Allocation for the Multi-Band Relay Channel,” In Proc. of 39th Annual Conference on Information Sciences and Systems, 2005.

[39] J. Lennox, K. Murakami, M. Karaul, and T.F. La Porta, “Interworking Internet Telephony and Wireless Telecommunications Networks,” ACM Computer Communications Review, vol.31, no.5, pp.25-36, 2001.

[40] H. Li, D. Yu, H. Chen, “New approach to multihop - cellular based multihop network,” In Proc. of IEEE PIMRC, 2003.

[41] Y. Liang and V. Veeravalli, "Gaussian orthogonal relay channel: optimal resource allocation and capacity," IEEE Transactions on Information Theory, vol. 51, no. 9, pp.3284-3289, 2005.

[42] S. Liese, D. Wu, P. Mohapatra, "Experimental Characterization of an 802.11b Wireless Mesh Network," In Proc. of IEEE IWCMC, 2006.

[43] M. Livny, and M. Melman, "Load Balancing in Homogeneous Broadcast Distributed Systems," In Proc. of ACM Computer Network Performance Symposium, 1982.

[44] Lloyd and Ramanathan, "Efficient distributed algorithms for channel assignment in multihop radio networks," Journal of High-Speed Networks, vol. 2, no. 4, 1993.

[45] P. Lungaro, "Cost/Performance Trade-offs of Two Layers Ad Hoc Cellular Systems," In Proc. of 4th Swedish Workshop on Wireless Ad-hoc Networks, 2004.

[46] H. Luo, R. Ramjee, P. Sinha, L. Li, and S Lu, "UCAN: unified cellular and ad-hoc network architecture," In Proc. of ACM MOBICOM, 2003.

[47] Lyiu, "Design of a new operational structure for mobile radio networks," Ph.D Thesis, Clemson University, 1990.

[48] X. Ma and E. Lloyd, "Evaluation of a distributed broadcast scheduling protocol for multihop radio networks," In Proc. of IEEE MILCOM, 2001.

[49] M. McGlynn and S. Borbash, "Birthday protocols for low-energy deployment and flexible neighbor discovery in ad hoc wireless networks," In Proc. of ACM MobiHoc, 2001.

[50] R. Misra, C.R.Manda1, "Performance comparison of AODV/DSR on-demand routing protocols for ad hoc networks in constrained situation," In Proc. of IEEE ICPWC, 2005.

[51] K. Murakami, O. Haase, J. Shin, T. F. La Porta, "Mobility Management Alternatives for Migration to Mobile Internet Session-based Services," IEEE Journal on Selected Areas in Communications, vol. 22, no. 5, page 818 - 833, 2004.

[52] R. Nabar, H. Bölcskei, and F. Kneubühler, "Fading relay channels: Performance limits and space-time signal design," IEEE Journal on Selected Areas in Communications, vol. 22, no. 6, pp. 1099–1109, 2004.

[53] V. Namboodiri, M. Agarwal, and L. Gao, "A study on the feasibility of mobile gateways for vehicular ad-hoc networks," In Proc. of ACM VANET, 2004.

[54] J. Padhye, S. Agarwal, V. Padmanabhan, L. Qiu, A. Rao, B. Zill, "Estimation of Link Interference in Static Multi-hop Wireless Networks," Internet Measurement Conference, 2005.

[55] G. Patel, S. Dennett, "The 3GPP and 3GPP2 Movements Toward and All-IP Mobile Network," IEEE Personal Communications Magazine, vol. 7, no. 4, pp. 62-64, 2000.

[56] C. Perkins and E. Royer, "Ad-Hoc On Demand Distance Vector Routing," In Proc. of IEEE WMCSA, 1999.

[57] A. Pirzada, M. Portmann, and J. Indulska, "Evaluation of Multi-Radio Extensions to AODV for Wireless Mesh Networks," In Proc. of ACM MobiWac, 2006.

[58] D. Plasse, "Call Control Scenarios in the "All-IP" UMTS Core Network," In Proc. of IEEE PIMRC, 2000.

[59] K. Ramachandran, E. M. Belding, K. C. Almeroth, M. M. Buddhikot, "Interference-Aware Channel Assignment in Multi-Radio Wireless Mesh Networks," In Proc. of IEEE INFOCOM, 2006.

[60] S. Ramanathan, "A unified framework and algorithm for (T/F/C)DMA channel assignment in wireless networks," In Proc. of INFOCOM, 1997.

[61] A. Raniwala and T. Chiueh, "Architecture and Algorithms for an IEEE 802.11-Based Multi-Channel Wireless Mesh Network," In Proc. of IEEE INFOCOM, 2005.

[62] J. Reed, "Software Radio: A Modern approach to radio engineering," Prentice-Hall, 2002.

[63] T. Rouse, I. Band, S. McLaughlin, "Capacity and Power Investigation of Opportunity riven Multiple Access (ODMA) Networks in TDD-CDMA Based Systems," In Proc. Of IEEE ICC, 2002.

[64] A. Sendonaris, E. Erkip, and B. Aazhang, "User cooperation diversity – Part I: System Description," IEEE Transactions on Communications, vol. 51, no. 11, pp. 1927–1938, 2003.

[65] A. Sendonaris, E. Erkip, and B. Aazhang, "User cooperation diversity – Part II: Implementation Aspects and Performance Analysis," IEEE Transactions on Communications, vol. 51, no.11, pp. 1939–1948, 2003.

[66] A. Sheth, J. Larson, "Federated Database Systems for Managing Distributed, Heterogeneous, and Autonomous Databases," ACM Computing Surveys, vol. 22, no. 3, pp. 183-236, 1990.

[67] J. Shin, P. Kishen, T. F. La Porta. "Dynamic Multi-Frequency, Multi-Hop Wireless Cellular Network," In Proc. of IFIP MED-HOC-NET, 2006.

[68] J. Shin, K. Lee, A. Yener, T. F. La Porta, "On-demand Diversity Wireless Relay network," ACM/Kluwer Journal of Mobile Networks and Applications (MONET), vol. 11, no. 4, pp. 593–611, 2006.

[69] E. Silva, and M. Gerla, "Load Balancing in Distributed Systems with Multiple Classes and Site Constraints," In Proc. of ACM International Symposium on Computer Performance Modelling, Measurement and Evaluation**,** 1984.

[70] A. Subramanian, H. Gupta, S. R. Das, "Minimum-Interference Channel Assignment in Multi-Radio Wireless Mesh Networks," In Proc. of IEEE ICNP, 2005.

[71] A. Tantawi, and D. Towsley, "Optimal Static Load Balancing in Distributed Computer Systems," Journal of ACM, vol. 32, no. 2, pp. 445-465, 1985.

[72] B. Timus, "A Cost Model for Ad Hoc Extended Cellular Systems," In Proc. of 4th Swedish Workshop on Wireless Ad-hoc Networks, 2004.

[73] Y. Wang, and R. Morris, "Load Sharing in Distributed Systems," IEEE Transactions on Computers, vol. C-34, no.3, pp. 204-217, 1985.

[74] G. Willman, and Paul Kuhn, "Performance Modeling of Signaling System No. 7," IEEE Communications Magazine, vol. 28, no. 8, pp. 44-56, 1990.

[75] J. Zhou and R. Yang, "Parcels: Pervasive Ad-hoc Relaying for Cellular Systems," In Proc. of IFIP Med-Hoc-Net, 2002.

[76] W. Zhou, et al., "Comparative Study of Routing Metrics for Multi-Radio Multi-Channel Wireless Networks," In Proc. of IEEE WCNC, 2006.

[77] IEEE 802.11a Standard, standards.ieee.org/getieee802/download/802.11b-1999.pdf

[78] IEEE 802.11b Standard, standards.ieee.org/getieee802/download/802.11b-1999.pdf

[79] IEEE 802.15 Standard Group: <http://www.ieee802. org/15/>

[80] IEEE 802.16 Standard Group: <http://www.ieee802. org/16/>

[81] GSM/ANSI-136 Interoperability Team (GAIT), "Network Interworking between GSM MAP and ANSI-41 MAP," PN-4857, TR-46, 2000.

[82] ETSI ES 202 915, "Open Service Access (OSA); Application Programming Interface (API)," v1.2.1, 2003.

[83] 3rd Generation Partnership Project Technical Specification Group, "Services and System Aspects; IP Multimedia Subsystem (IMS); Stage 2 (Release 5)," ver. 5.5.0, 3GPP TS 23.228, 2002.

[84] TIA/EIA, "Cellular Radio Telecommunications Intersystem Operations," TIA/EIA ANSI-41-D, TIA, 1997.

[85] The Official Bluetooth Specification - Version 1.2 Core Document Number: MISC BLUETOOTH CORE Miscellaneous (Undefined and/or multiple publishers), 2003

[86] IEEE 802.15 Standard Group, Task Group 5 – WPAN Mesh Networking: <http://www.ieee802. org/15/pub/TG5.html>

[87] IEEE 802.16 WiMax Forum: <http://www.wimaxforum.org>

[88] Q. Dong et al., “Throughput Optimization and Fair Bandwidth Allocation in Multi-hop Wireless LANs,” In Proc. of IEEE Infocom, 2006.

[89] S. Lee et al., “The Case for a Multi-hop Wireless Local Area Network,” In Proc. of IEEE Infocom, 2004.

[90] A. Subramanian et al., “Interference Aware Routing in Multi-Radio Wireless Mesh Networks,” In Proc. of IEEE WiMesh, 2006.

[91] W Su et al., “Mobility prediction and routing in ad hoc wireless networks,” International Journal of Network Management, Vol. 11, 2001.

Vita

JaeSheung Shin received his B.S. and M.S. degree in Computer Science and Engineering from DongGuk University, Korea, in 1991 and 1993, respectively. He is currently working toward the Ph.D. degree in Computer Science and Engineering at the Pennsylvania State University, University Park. He is a member of the Networking and Security Research Center. Prior to joining Pennsylvania State University, he was with Electronics and Telecommunications Research Institute (ETRI) in Korea, since 1993. He has worked on the development of core network elements in 2G and 3G wireless cellular networks. His research interests include performance evaluation and design of seamless roaming schemes for 4G All-IP wireless cellular networks; dynamic multi-hop relay network formation including link channelization, diversity setup, and throughput optimization for multi-frequency, multi-hop wireless cellular networks and wireless mesh networks comprised of agile/cognitive radios; and routing and link channelization to support the network coding efficiently in multi-hop wireless networks.