A LOAD BALANCING AND RELAYING FRAMEWORK IN A-CELL NETWORKS

BY

YIK HUNG TAM

A thesis submitted to the School of Computing
In conformity with the requirement for
the degree of Master of Science

Queen's University
Kingston, Ontario, Canada
September 2004

Library and
Archives Canada

Published Heritage
Branch

395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque et
Archives Canada

Direction du
Patrimoine de l'édition

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence
ISBN: 0-612-99690-5
Our file Notre référence
ISBN: 0-612-99690-5

NOTICE:
The author has granted a non-exclusive license allowing Library and Archives Canada to reproduce, publish, archive, preserve, conserve, communicate to the public by telecommunication or on the Internet, loan, distribute and sell theses worldwide, for commercial or non-commercial purposes, in microform, paper, electronic and/or any other formats.

The author retains copyright ownership and moral rights in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

AVIS:
L'auteur a accordé une licence non exclusive permettant à la Bibliothèque et Archives Canada de reproduire, publier, archiver, sauvegarder, conserver, transmettre au public par télécommunication ou par l'Internet, prêter, distribuer et vendre des thèses partout dans le monde, à des fins commerciales ou autres, sur support microforme, papier, électronique et/ou autres formats.

L'auteur conserve la propriété du droit d'auteur et des droits moraux qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

In compliance with the Canadian Privacy Act some supporting forms may have been removed from this thesis.

While these forms may be included in the document page count, their removal does not represent any loss of content from the thesis.

Conformément à la loi canadienne sur la protection de la vie privée, quelques formulaires secondaires ont été enlevés de cette thèse.

Bien que ces formulaires aient inclus dans la pagination, il n'y aura aucun contenu manquant.

Canada

Abstract

Towards the end of the past decade, mobile communication has gained a significant commercial popularity. The number of mobile subscribers is expected to continue to increase. In addition to traditional voice communications services, mobile carriers will also be providing data communications services. Data rates in data communications services will also be increased to match the Quality of Service (QoS) requirements of different applications. In third Generation (3G) cellular systems, user data rates can be up to 2 Mbps. Although high data rates are achieved, fundamental capacity limitation of cellular networks still exists. Call requests are frequently blocked in hotspot areas, such as city centers and places of sport events. Load balancing in a cellular network helps to solve this problem by shifting the load from a hotspot cell to less loaded cells. Several load balancing schemes such as channel borrowing, coverage negotiation, and traffic relaying through limited mobility relaying stations, were proposed to solve this problem. However, these schemes are designed based on conventional mainstream cellular systems in which FDMA or TDMA technology is used. They are not flexible enough, or even practical, for 3G systems in which W-CDMA technology is used. While routing, and medium access schemes can greatly affect load balancing, these functions are isolated. Furthermore, most relaying schemes are based on contention-based medium access techniques in which radio resources are not fully utilized. Likewise, the existing routing protocols do not fully utilize the useful characteristics of CDMA cellular systems.

In this thesis, we propose a novel load balancing and relaying framework for 3G TDD W-CDMA cellular systems. This framework consists of a load balancing scheme called ALBA, a routing scheme called ACAR and a slot assignment scheme called E-DSSA. This framework is not only suitable for existing 3G CDMA cellular systems, but can also be applied to conventional 2G CDMA systems. Simulation results show that this framework reduces the call blocking ratios of hotspot areas, balances load among cells and increases system throughput.

Keywords: 3G Cellular Networks, TDD, W-CDMA, Load Balancing, Relaying, Routing, Slot Assignment, Mobile Ad hoc Networks, Multi-hop Cellular Networks.

Acknowledgements

First of all, I would like to thank the loving God for guiding me to this stage that I really wanted to be in.

I would like to express my deepest gratitude to my supervisors Dr. Ahmed Safwat and Dr. Hossam Hassanein. Without their expert knowledge, excellent supervision and guidance, encouragement and support, I could not have imagined how I can finish this challenging work in time.

Gratefully acknowledged is the financial support provided by Queen's University, the Natural Sciences and Engineering Research Council of Canada (NSERC), the Ontario Graduate Scholarship program, and the Communication Information Technology of Ontario (CITO).

I would like to thank my parents, my friends and relatives. Mom and Dad, thank you so much for your deepest and unconditional love you left for me to carry on my journey. Grandma, Grandpa, and big auntie, thank you for your deep and genuine love. Brian Siu, without your humongous unconditional help, caring and support, I would not be able to make this. Chun-ho and Angela, thank you for your love, caring, listening, and standing by me all the time. Pauline and your

families, thank you for making me feel home in Canada. Ben Kam, thank you for encouraging me to take this academic path that I would not think of. Wenying and Richard, thank you for your tremendous help and guidance in my undergraduate studies which gave me a strong foundation for my graduate studies. Benjamin, thank you for your expert technical support. Marianne, thank you for your help and great idea of coming to Canada. Uncles, aunties, cousins, Winnie, Ven, Edmond, Kevin, SKC friends, Shun, FAS friends, PIE friends, Ronnie, Ivy, Yee, Connie, Jenny, Benny, Agatha, Simon Yu, Chi-hung, UI colleagues, Adrianne, Mandy, Grace, Simon, IN25 buddies, Bonnie (Trent), John, and Logo Christian fellowship, thank you for your love, caring and support all the time – I love you all.

Finally, I would like to thank the staff, especially Debby, faculty members in the School of Computing, and my fellow researchers in the Telecommunications Research Laboratory (TRL), especially Kenan, Abdelhamid, and Mohammed, for their help, care, valuable suggestion, and the fun times we spent together.

Contents

List of Tables

List of Figures

List of Acronyms

1G	1st Generation Wireless Communication Systems
2G	2nd Generation Wireless Communication Systems
3G	3rd Generation Wireless Communication Systems
3GPP	3rd-Generation Partnership Project
4G	4th Generation Wireless Communication Systems
A-Cell	Ad hoc-Cellular network
ACK	ACKnowledgement
AODV	Ad-Hoc On-Demand Distance Vector
ALBAR	A-Cell Load BAlancing Relaying
ALBA	A-Cell Load BAlancing
ACAR	A-Cell Adaptive Routing
ARS	Ad hoc Relaying Station
BCR	Base-Centric Routing
BS	Base Station
CAC	Call Admission Control
CBR	Call Blocking Ratio
CDMA	Code Division Multiple Access
CN	Core Network

DSDV	Destination-Sequence Distance Vector
DSR	Dynamic Source Routing
DSSA	Delay-Sensitive Slot Assignment
E-DSSA	Extended – Delay Sensitive Slot Assignment
FDD	Frequency Division Duplex
FDMA	Frequency Division Multiple Access
FIFO	First In First Out
FM	Frequency Modulation
GPS	Global Positioning System
GSM	Global System for Mobile communications
HWN	Hybrid Wireless Network
iCAR	integrated Cellular and Ad hoc Relay
IS-136	Interim Standard 136
IS-95	Interim Standard 95
ISM	Industrial, Scientific, and Medical
LOS	Line of sight
MAC	Medium Access Control
MANET	Mobile Ad Hoc NETwork
MCN	Multi-hop Cellular Network
MT	Mobile Terminal
ODMA	Opportunity-Driven Multiple Access
PSTN	Public Switching Telephone Networks
PDC	Pacific Digital Cellular
QoS	Quality of Service
PARCelS	Pervasive Ad hoc Relaying for Cellular System

RNC	Radio Network Controller
RREP	Route REPly
RREQ	Route REQuest
SCN	Single-hop Cellular Network
TDD	Time Division Duplex
TDMA	Time Division Multiple Access
UMTS	Universal Mobile Telecommunication System
W-CDMA	Wideband Code Division Multiple Access

Chapter 1

Introduction

Mobile communications has become affordable and popular since the past decade. In 2001, there were approximately 630 million cellular and Personal Communication Service (PCS) subscribers. The number of users continues to increase and is expected to reach 3 billion in 2006 [31], which would be about 30% of the world population at that time.

Wireless communications has gone through three generations. In the 1st Generation (1G), analog frequency modulation techniques and Frequency Division Multiple Access (FDMA) air interface technology with Frequency Division Duplex (FDD) were used. In the 2nd Generation (2G), digital modulation and Time Division Multiple Access (TDMA) with FDD and narrowband Code Division Multiple Access (CDMA) with FDD techniques were used [31]. 2G systems also provided and still do provide, limited Internet browsing and Short Messaging Service (SMS). Examples of 2G standards includes Global System Mobile (GSM), IS-136, PDC, and IS-95. In 3G systems, Wideband CDMA (W-CDMA) technology is used. W-CDMA is a 3G wireless communication standard which is based on CDMA technology. This standard uses a wider frequency band compared to that of 2G CDMA cellular systems to achieve a higher data rate, which can be up to 2Mbps. This technology also allows mobile terminals to communicate with base station (BS) using the same frequency band simultaneously. Moreover, each terminal can have several connections at the same time. These connections are distinguished from each other by using different orthogonal spreading codes [21]. Examples of 3G systems are Universal Mobile Telecommunication Systems (UMTS) [14, 15] and cdma2000 [31].

Conventional or mainstream cellular systems are based on FDMA or TDMA which are bandwidthlimited [31] which means that the system capacity depends on the number of frequency bands available. 3G systems, on the contrary, use the interferencelimited CDMA technology [31] which means that the cell capacity depends on the degree of interference. In CDMA, one frequency can be reused for all cells in the system. In other words, all mobile terminals can use the same frequency to communicate with their corresponding base stations (BSs). Although the spectrum) is efficiently utilized, interference becomes a major issue that affects the quality of transmission and reception. When transmission power increases, interference increases and thus the transmission quality decreases. If the transmission quality, i.e. bit error ratio, of a call is below a predefined threshold, the call is dropped. This issue limits the number of users that can be served. It also raises the cell breathing effect and the near-far problem.

The cell breathing effect [31] describes the dependency between capacity and coverage. When the coverage of a cell increases, the capacity of the cell decreases, and vice versa. To increase the capacity of a cell without sacrificing the cell's coverage, an ad hoc multi-hop relaying protocol, called Opportunity-Driven Multiple Access (ODMA), was proposed in [9]. The idea of this protocol is to break a long path from the source mobile node to the BS into several small paths using intermediate mobile nodes such that the transmission distances are reduced [38]. Traffic is relayed from the source node to the BS by multi-hopping through the intermediate nodes. As the transmission distance is reduced, transmission power reduces. Consequently, interference reduces and cell capacity increases.

The near-far [31] problem occurs when two mobile terminals, one is far from the BS and the other nearby, simultaneously send signals to the BS using the same power levels. As signals attenuate when they travel through the air, signals from the closer terminal arrive at the BS stronger than those of the terminal far away. This results in the BS unaware of or unable to recognize the signal of the farther terminal. Power control schemes exist for handling this problem [14].

1.1 Motivation

Although many problems and issues related to 3G wireless systems have been addressed, capacity limitation, which is inherent in all cellular systems, still persists. In cellular networks, some cells may be heavily loaded (hot) and some cells are slightly loaded (cool). Call requests in a hot cell have a higher probability of being blocked. While calls are blocked in hot cells, radio resources in cool cells are idle. Load balancing and relaying can solve this problem.

> Load balancing in a network is the act of distributing the network load evenly across the network such that the performance of the system is enhanced.

Examples of load balancing schemes in cellular networks are channel borrowing [5], coverage negotiation [7], and traffic relaying using limited mobility ad hoc relay stations [6]. Channel borrowing is to borrow channels, i.e., frequency bands, from neighboring cool cells to serve a call request. This idea is more suitable for FDMA and TDMA systems rather than CDMA or W-CDMA systems. This is due to the fact that CDMA only uses a single frequency band, with no additional frequencies for borrowing. Coverage negotiation is also based on the bandwidth--limited assumption. The idea of this approach is to reduce cell sizes of hot cells and increase the cell size of neighboring cooler cells so that the area originally covered by the hot cell is covered by the cooler cells. However, this approach assumes a constant cell capacity, which is not the case in CDMA systems. Using ad hoc relay stations with limited mobility to relay traffic from a hot cell to a neighboring cooler cell works for any cellular system, but it is still costly and may not be flexible enough to deal with highly dynamic load situations in 3G systems. As many types of services, technically translated to a wide range of data rates, are provided in 3G systems, demand becomes quite unpredictable and hotspots may surface at any time and anywhere. In other words, the load status of a cell becomes dependent not only on the number of users within, but also on the types of services they request. Recently, another method to solve the capacity limitation problem has been proposed, namely the multi-hop cellular concept. The idea is to increase the separation among cells by shrinking cell size or removing adjacent BSs. This helps to avoid co-channel interference [31]. Thus, more frequencies can be reused. In other words, the capacity of a cell increases. The traffic from the mobile node in the non-covered region can be relayed to the BS through other mobile terminal. This concept can also be applied in 3G systems although the capacity increase is due to the cell breathing effects instead of avoiding co-channel interference. As a cell size decreases, more users or higher data rates are allowed. Although the multi-hop cellular concept helps to increases the capacity of a cell, the accessibility of mobile nodes to BSs is reduced and the instability of connections increases. In addition, issues such as under-utilization of cool cells and traffic congestion in hot cells have not been addressed.

From the above, it can be seen that existing load balancing and relaying schemes are neither practical nor flexible enough for 3G systems, and that there remains a need for a new load balancing relaying paradigm.

1.2 Thesis Objectives

The objective of this thesis is to design a load balancing relaying framework for 3G TDD Wideband CDMA cellular systems. Specifically, this research work aims to:

1. design a load balancing algorithm for balancing the load of multi-hop cellular systems.
2. design a routing scheme specific for CDMA multi-hop cellular systems.
3. improve existing slot assignment scheme for TDD W-CDMA multi-hop cellular systems.
4. release congestion of heavily loaded (hotspot) cells for reducing call blocking probability.
5. balance load among cells for better resource utilization.
6. increase throughput of the whole system.

As we mentioned earlier, most load balancing schemes do not fit well in 3G cellular systems. Multi-hop relaying or multi-hop cellular strategy seems to be a promising solution. However, the existing multi-hop cellular concept is based on short transmission ranges, which greatly reduces the reachability of mobile terminals. In addition, a contention-based medium access protocol is usually assumed for multi-hopping. This assumption further affects the utilization of resources and induces more delay, adversely affecting the performance of load balancing. To obtain the best load balancing relaying performance in 3G TDD W-CDMA multi-hop cellular systems, routing and medium access schemes have to be taken into consideration. We propose a load balancing relaying framework in which there are schemes for load balancing, routing and medium access. The framework integrates all three components to achieve an effective load balancing performance in 3G TDD W-CDMA cellular systems. The load balancing algorithm is suitable for 3G multi-hop cellular systems. The design of routing and modified slot assignment schemes utilizes the cell breathing characteristic of CDMA systems. The objectives of this framework are to release congestion, balance load among cells and increase system throughput. Releasing the congestion of cells will reduce the call blocking probability. Balancing load among cells helps to utilize the network resources so that fewer cells are under-utilized. In both venues, the system throughput increases.

In this thesis, we focus on studying the effect of our framework on the uplink [23] connection to the BS.

1.3 Thesis Organization

In the next chapter, we review some background information and related work on cellular networks, Mobile Ad hoc Networks (MANETs), and multi-hop cellular networks. Related works including medium access technologies, relaying schemes, routing protocols, and load balancing schemes of these networks are described. In Chapter 3, a novel load balancing relaying framework for 3G TDD W-CDMA cellular systems is introduced. This framework consists of a load balancing scheme, a routing scheme and a slot assignment scheme. In Chapter 4, the simulation model and performance metrics are explained along with the simulation parameters. and the performance results. In Chapter 5, conclusions are made and potential future work is discussed.

Chapter 2

Related work

Before explaining the design of our load balancing relaying framework, it is important to understand the characteristics of existing wireless systems, and the advantages and disadvantages of different schemes for medium access, relaying, routing, and load balancing. In this chapter, background information about wireless networks including cellular networks, Mobile Ad hoc Networks (MANETs) and multi-hop cellular networks is provided. Related work including multiple (medium) access techniques, relaying schemes, routing protocols and load balancing schemes of each of these networks are discussed.

2.1 Wireless Networks

Cellular networks, MANETs, and multi-hop cellular networks are all wireless networks. They have similarities and differences in theirs system architectures, medium access control, relaying, routing and load balancing. To facilitate the understanding of these areas, Table 1 summarizes the similarities and differences among them. It can be seen from the table that MANETs have no architecture (infrastructure) or relaying scheme. This is because MANETs do not rely on BSs or any other form of centralized access coordination to relay to. PARCelS and ODMA are two relaying schemes for cellular systems. Since relaying is actually a multi-hopping concept, they should also be categorized as a relaying scheme in multi-hop cellular networks. Detailed discussions on these areas are provided in the remaining part of the chapter.

	Cellular networks	**MANETs**	**Multi-hop cellular networks**
Systems	Centralized	Distributed	Centralized + distributed
Architecture	Conventional, 3G	-	MCN, HWN
Flexibility	Low	High	Medium
Limitation	BS capacity, coverage	MT resource	BS capacity + MT resource
Medium access	Contention-free e.g. FDMA, TDMA, CDMA	Contention e.g. IEEE 802.11MAC	Contention and/or Contention-free
Relaying	iCAR, PARCeLS, ODMA. Device: ARS, MT	-	A-Cell, PARCeLS, ODMA Device: MT
Routing	Hierarchical and flat routing	MANET routing	BCR + MANET routing
Load balancing	Aim: release BS congestion. Methods: Relaying, Channel borrowing, Flexible users, Coverage negotiation	Aim: release MT congestion Methods: Load-aware routing, multi-path routing, route de-coupling	Aim: release BS + MT congestion.

Table 1.1: Summary of similarities and differences of three types of wireless networks

Cellular networks are centralized networks which consist of base stations (BSs) and mobile terminals[1] (MTs). These networks are also called Single-Hop Cellular Networks (SCN). In conventional 1G, 2G, and 2.5G cellular networks, calls of MTs go through BSs, which are in turn connected to Mobile Switching Center (MSC). The purpose of MSC is to pass the calls to the Public Switching Telephone Networks (PSTN). Figure 2.1 shows the system architecture of a conventional cellular network. In 3G cellular systems, a Radio Network Controller (RNC) [14] is added for the purpose of radio resource management. It connects to a core network which is the gateway to the Internet and PSTN. Figure 2.2 illustrates the system architecture of 3G cellular systems. A cellular network has no energy consumption concern because it is wired. BS coordinates the MTs within its coverage for radio resource management. In addition, the presence

---

[1] The terms mobile stations, mobile terminals, mobile nodes, and nodes are used interchangeably throughout this document.

of RNC in 3G systems helps coordinate BSs. Topology information of the network and state information of the nodes can be obtained through the BSs. This information can be very useful for channel allocation, relaying, routing, and load balancing. Although cellular networks have these advantages, they are challenged with a high infrastructure cost, and coverage and capacity limitations.

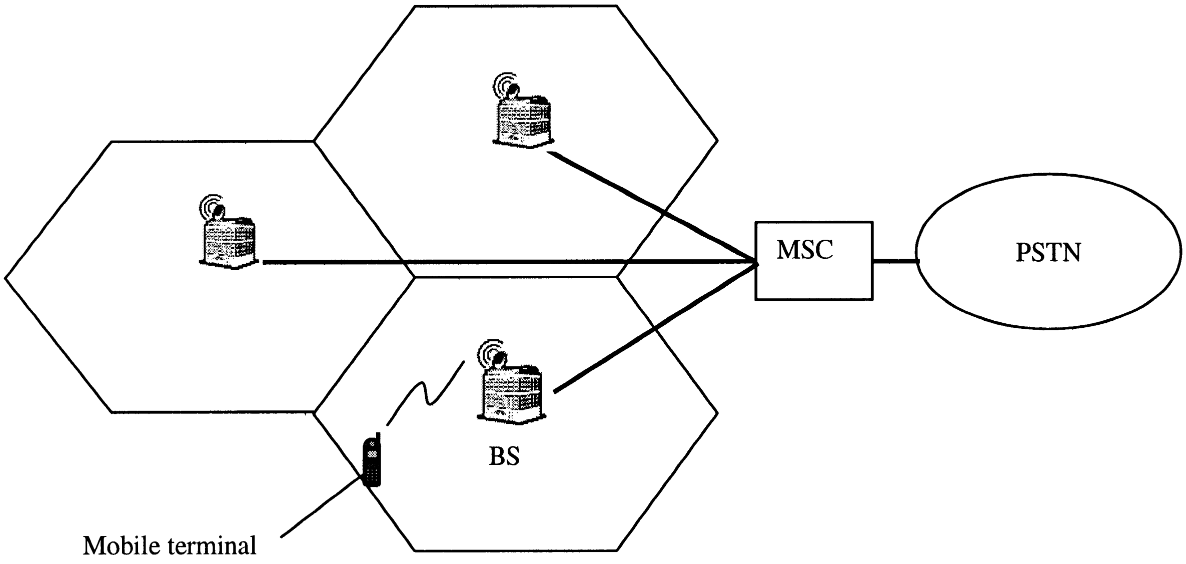


Figure 2.1: System architecture of conventional (1G, 2G, 2.5G) cellular systems

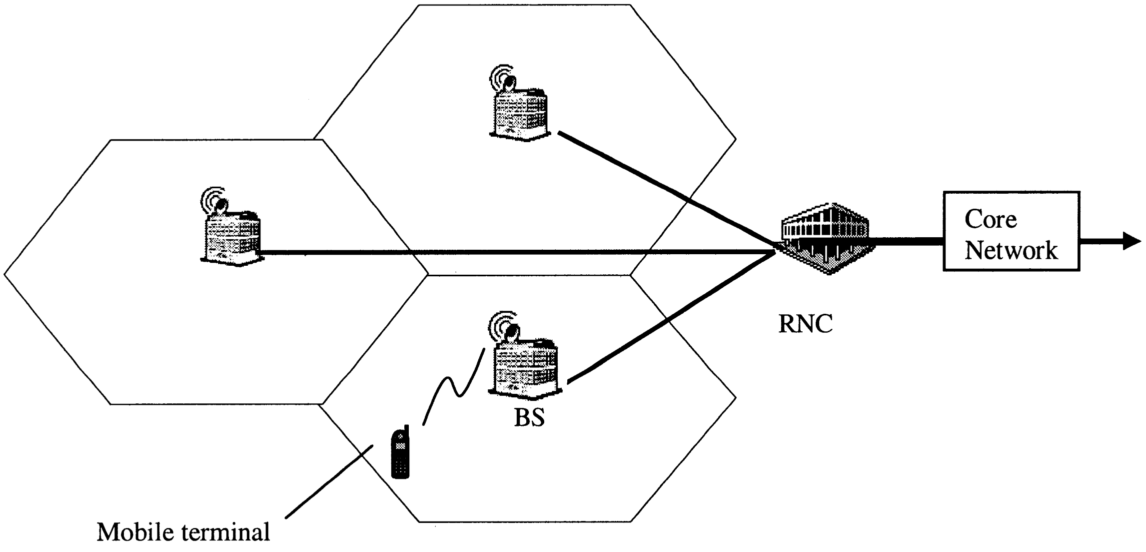


Figure 2.2: System architecture of 3G cellular systems

A Mobile Ad Hoc Network (MANET) [24], sometimes also called a wireless multi-hop ad hoc network, consists of a number of mobile nodes which communicate with each other over the wireless air interface in a peer-to-peer fashion. Communications between source and destination node is made through multi-hopping through other intermediate nodes. No existing network infrastructures or central administrations are available or required. Figure 2.3 illustrates an example of a MANET. Although node C is outside the transmission range of node A, node A can communicate with node C through the intermediate node B. These networks are very flexible and can be deployed anywhere and anytime. They also have no infrastructure cost. Military actions in a battlefield and emergency rescue operations are two typical examples of their usages. However, although MANETs have these benefits, the resources of these networks such as battery and bandwidth are limited. Frequent disconnections may occur due to mobility or battery depletion. If no route exists, a source node cannot communicate with a desired destination node. Multi-hopping also increases packet delay. Naturally, routing is a major issue in such networks. Routing protocols for MANETs are discussed in Section 2.2.2.

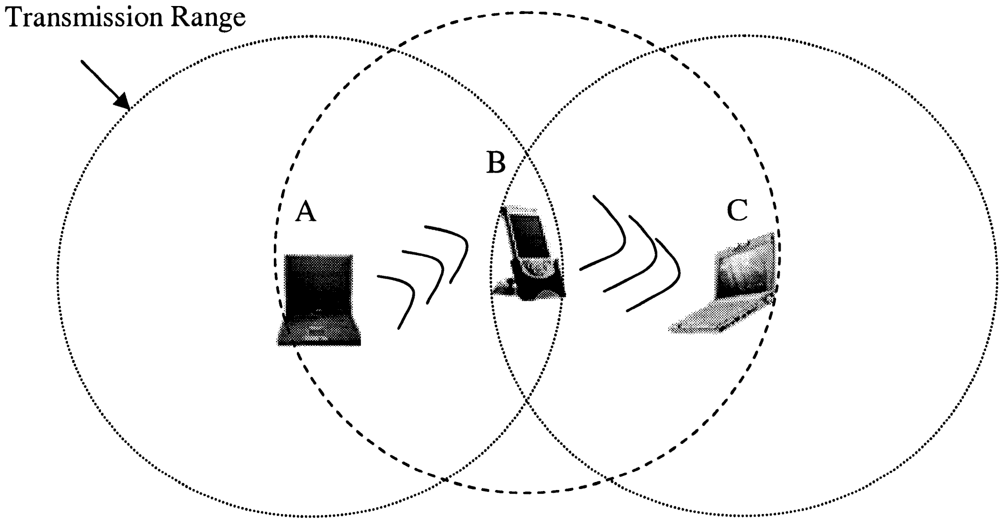


Figure 2.3: Mobile ad hoc networks

When the concept of MANET is incorporated with cellular systems, a multi-hop cellular network is formed. Such networks exploit a smaller number of BSs or lower transmission powers. Thus, infrastructure cost is reduced and more frequencies can be reused. The ad hoc relaying component of these networks allows peer-to-peer communication, which further reduces the load of cells. Overall system throughput is further increased. In addition, multi-hopping facilitates traffic relaying, which increases the coverage and capacity of cells. It would be even more attractive when this concept is applied in 3G cellular systems because the RNC in these systems can coordinate BSs, which helps for load balancing among BSs. Multi-hop Cellular Networks (MCN) [22] and Hybrid Wireless Networks (HWN) [4] are two types of multi-hop cellular networks. MCN is designed for densely connected networks. The idea of MCN is to reduce the coverage of BSs or the number of BSs. The former is called MCN-p and the latter is called MCN-b. Figure 2.4 and 2.5 respectively illustrate their topology. HWN is similar to MCN except that each mobile node has two modes: a cellular mode and an ad hoc mode, used for sparse and dense topologies, respectively. Although the multi-hop cellular concept has many advantages, the ad hoc relaying component inherits the limitations of MANETs such as limited battery, limited bandwidth, frequent disconnections, higher packet delay and routing overhead. MCN is also restricted to densely connected networks. In sparsely connected networks, more nodes will become unreachable.

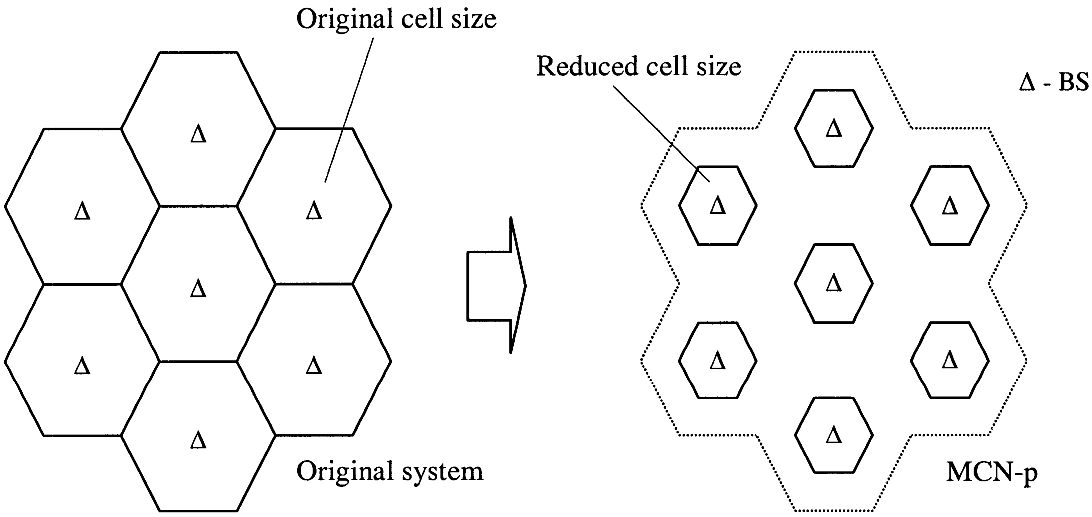
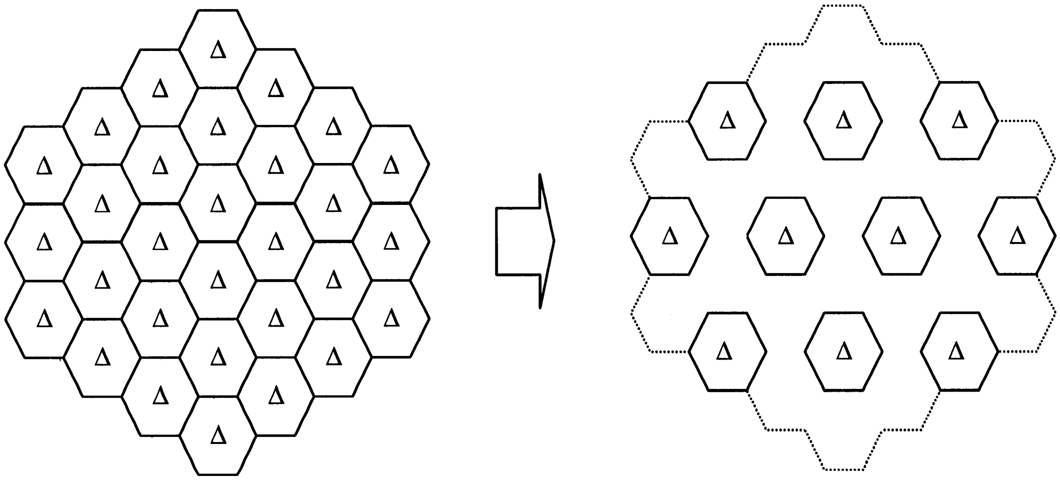


Figure 2.4: The MCN-p architecture



Original system has many base stations

MCN-b has fewer base stations

Figure 2.5: The MCN-b architecture

2.2 Medium Access

Medium access techniques in wireless networks can be grouped into two main categories: contention-free and contention-based.

2.2.1 Contention-Free Medium Access

Contention-free techniques provide a mobile terminal or a connection a dedicated channel in the form of frequency, time slot or code. Examples of contention-free medium access technique are Frequency Division Multiple Access (FDMA), Time Division Multiple Access (TDMA) and Code Division Multiple Access (CDMA) [31]. These techniques can be further categorized as bandwidth-limited and interference-limited technology. FDMA and TDMA are bandwidth-limited technologies. In FDMA, different channels are assigned different frequencies. The capacity of a cell depends on the number of channels that are available. TDMA is an overlaying technology on FDMA. Each frequency is divided into several time slots. Each user can be assigned one or more time slots. In this way, more users can be accommodated and higher data rates can be achieved. Figure 2.6b shows one frequency divided into 4 time slots. CDMA is an interference-limited technology that is based on code division. Each user or connection is assigned a different and fully orthogonal code (sequence) [21] to avoid interference. All users can send their signals simultaneously to the same destination on the same frequency if they are using different codes. The signal of each connection can be recovered because of the orthogonality of the codes, i.e., each signal can be completely distinguished. The load of a cell depends on the date rates, energy per bit per noise power density (E_b/N_0) [14], and traffic activity. When a cell allows more loading, a larger interference margin [14] is needed in the uplink. Thus, the coverage area reduces This is the phenomenon called cell breathing [14, 31]. Figure 2.6c illustrates the CDMA technique. When contention-free techniques are used, no channel conflict occurs, and radio resources are highly utilized. However, this requires a centralized administration. Therefore, it is best suited for cellular systems. In addition, FDMA or TDMA have no interference problems, only low frequency reuse. CDMA maximizes frequency reuse but bandwidth may not be fully utilized unless interference is fully handled. In 3G systems, CDMA technology with wide frequency band is used. Wide frequency band allows high data rate. W-CDMA is one of the 3G wireless standards.

In addition to these medium access techniques, there are two full duplex communication options: Frequency Division Duplex (FDD) and Time Division Duplex (TDD) [15, 10]. A full duplex communication allows simultaneous traffic in both directions. To achieve this, FDD requires two frequencies while TDD requires at least two time slots. Figure 2.7 shows the operation of these two duplex operations. In Figure 2.7a, frequencies f1 and f2 are respectively used for uplink and downlink communications of node A, in Figure 2.7b, time slots s1 and s3 are used for uplink and s2 and s4 are used for downlink direction. For CDMA multi-hop cellular systems, using TDD is more preferred to FDD as it is difficult with the latter to synchronize the two frequencies for each node. Further explanation can be found in [1].

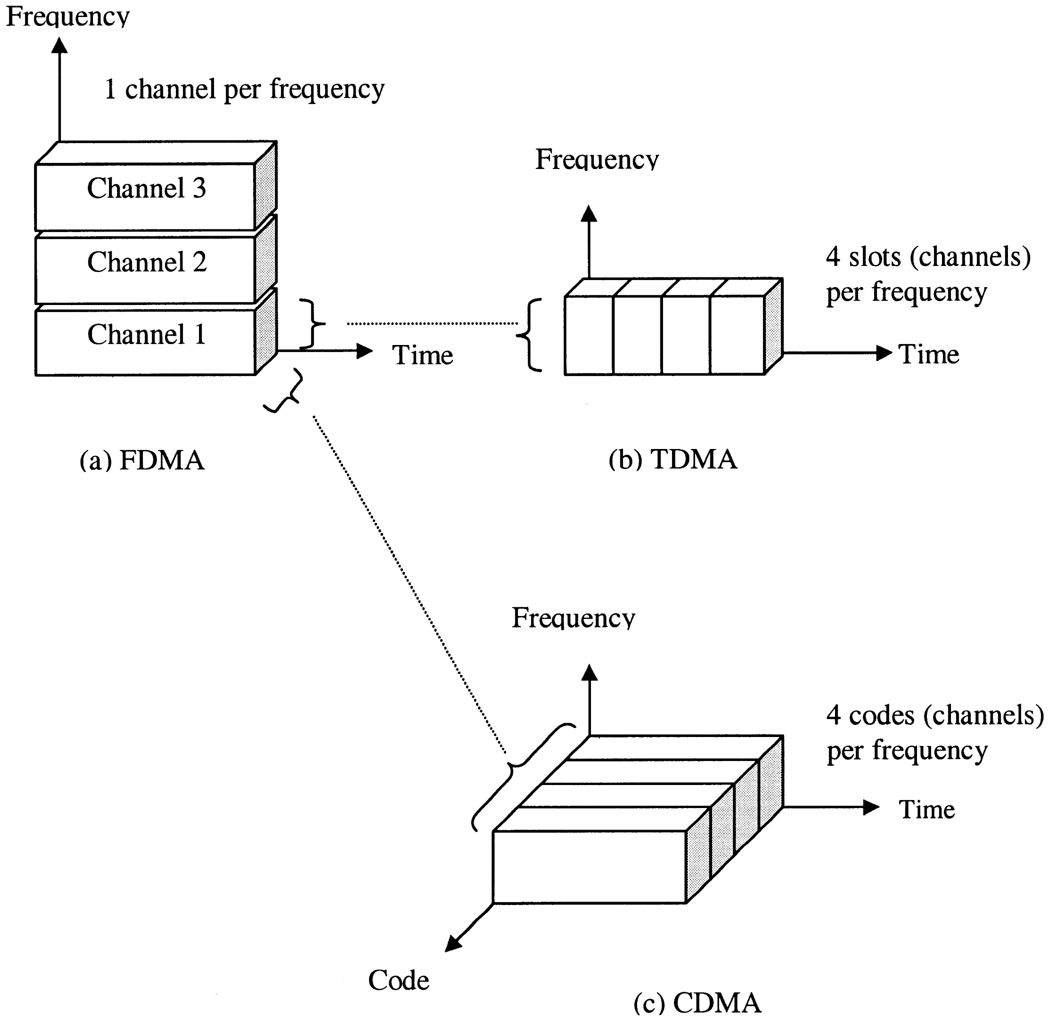


Figure 2.6: Three multiple access techniques: a) FDMA, b) TDMA, c) CDMA

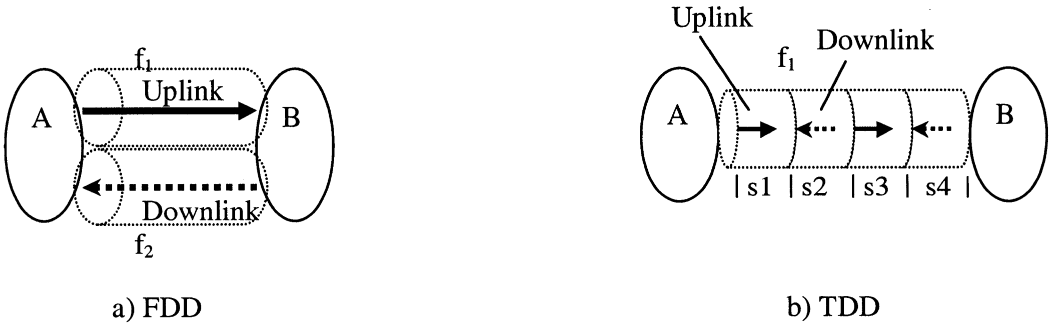


Figure 2.7: Operation of FDD and TDD communications

2.2.2 Contention-Based Medium Access

Contention-based techniques require mobile nodes to contend for the radio resource (medium). The IEEE 802.11 Medium Access Control (MAC) protocol [21, 37] is an example of contention-based medium access techniques. It has two modes of operation: Distributed Coordinate Function (DCF) and Point Coordinated function (PCF). DCF is based on Carrier Sense Multiple Access with Collision Avoidance (CSMA/CA). It uses physical channel sensing or virtual sensing. When physical channel sensing is used, a mobile node actually senses the medium. If the medium is idle, it sends packets; otherwise, the node defers transmission until the medium is idle. If a signal collision occurs, the node backs off for a period of time and tries again later. In virtual sensing, Request-To-Send (RTS), Clear-To-Send (CTS), and Acknowledge (ACK) signals, in addition to an ACK timer, are all used for coordinating the data transmissions. PCF is built on top of DCF with a coordinator is added. The coordinator polls the mobiles to see if they have data to send. PCF provides a contention-free period for the mobiles to assure their access. The advantage of contention-based techniques is that they require no central administration. They are suitable for distributed networks such as MANETs. However, this involves high overhead for coordination of mobile nodes and signal collision resolution.

2.2.3 Medium Access in Multi-hop Cellular Network

For multi-hop cellular networks, as they are a combination of cellular and MANET concepts, the medium access technique can be either contention-free, contention-based or a hybrid. Using contention-based medium access techniques for both cellular and ad hoc communications reduces the complexity of mobile devices and avoids the need for complex channel allocation schemes. MCN [21] is an example of such. It uses the IEEE 802.11 MAC for both cellular and ad hoc communications. Using a hybrid of medium access techniques is also a common strategy. Since both techniques are well-established and widely used, mobile devices can be easily modified to have both medium access abilities. The ad hoc interface also allows communications with other wireless systems such as WiFi and Bluetooth. Although there are several advantages of using a contention-based or a hybrid protocol, the presence of the former introduces contention, co-channel interference, and resource competition between cellular and non-cellular users. Sharing medium with other non-cellular users may intensify security issues. Furthermore, the infrastructure of the cellular network would not be fully utilized.

With a proper channel allocation scheme, using one contention-free medium access technique for both cellular and ad hoc communication is a better choice. This idea is realized by using the Delay Sensitive Slot Assignment (DSSA) [1] scheme in the A-Cell [35] relay architecture. The idea of the scheme is to allocate a time slot code pair to the source and relaying nodes for both cellular and ad hoc communications, such that the overall end to end delay of a packet is minimized. Directional antennas are used to increase spatial reuse. DSSA consists of two phases: an Elimination Phase and a Selection Phase. In the Elimination Phase, conflicting channels are eliminated to ensure that no neighbors of next hop nodes are transmitting on these channels and no neighbors of current node are receiving on these channels. In the Selection Phase, channels are selected such that the delay is reduced. Figure 2.8 illustrates DSSA with directional antennas. Node A can be assigned any channel except (slot 3, code 1) since node b and a are respectively transmitting and receiving on this channel. A channel with slot 3 and a code other than code 1 is preferred to minimize the slot waiting time. Thus, the packet delay is minimized. For example, (slot 3, code 2) is a good candidate channel. Simulation results show that DSSA increases throughput and decreases delay [1].

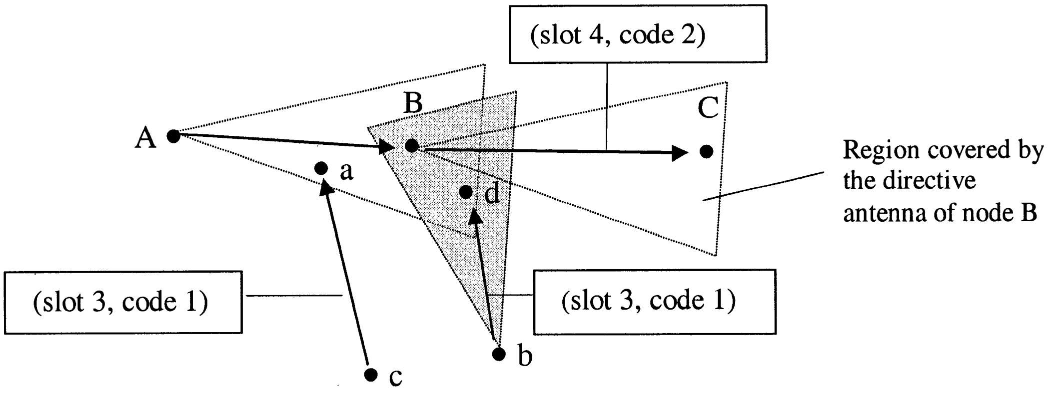


Figure 2.8: DSSA with directional antenna

Although the results show that DSSA has a good performance in terms of delay and throughput, a crucial condition has not been clearly specified. The condition is to restrict the current node from assigning the same channel as the channel that the successive node is transmitting on. Without this condition, sending and receiving signals collide at the current node. Figure 2.9 illustrate this consecutive channel conflict situation. Assume that A and B have already been assigned channels (slot 3, code 1) and (slot 4, code 2), respectively, for the route (A-B-C). A channel (slot 3, code 1) is proposed for node B for the route D-B-C. According to DSSA, if no neighbors of the next hop node (node C) are transmitting on this channel and no neighbors of the current node (node B) are receiving on this channel, (slot 3, code 1) can be assigned to node B. Since C and B have no neighbors, conditions are satisfied and (slot 3, code 1) will be assigned to node B. While node B continues to receive on (slot 3, code 1), node B starts transmitting on (slot 3, code 1). That is, node B itself has to receive and transmit on the same channel. If the hardware of the mobile is allowed to receive and transmit signal at the same time, signal collision occurs. If the mobile is not allowed to transmit and receive at the same time, then packets cannot be relayed.

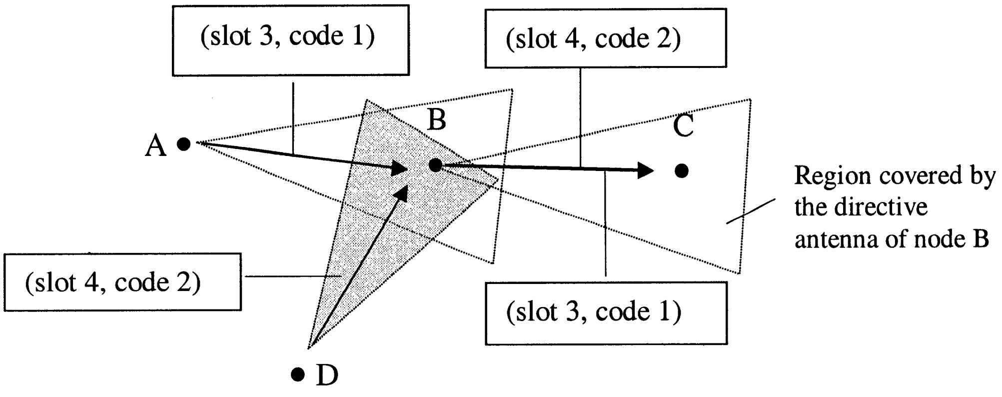


Figure 2.9: Situation of consecutive channel conflict

Contention-free medium access protocols have no signal collision, back-off requirement, co-channel interference, hidden terminal, or exposed terminal problems. Bandwidth is hence better utilized. However, these techniques require centralized control. Therefore, they are suitable for cellular systems, but not for MANETs which have no infrastructure or centralized control. MANETs usually are assumed to use IEEE 802.11 MAC protocol. In multi-hop cellular systems, since it has infrastructure, contention-free medium access protocol is preferred for better bandwidth utilization. This can also avoid co-channel interference, hidden terminal, and exposed terminal problems. Multi-hop cellular networks contain both centralized and distributed component. If the centralized component can help administrate the distributed component, contention-free protocols are preferred. In conclusion, contention-free medium access techniques with slot assignment [34, 36] scheme should be the direction for 3G or future generation multi-hop cellular networks.

2.3 Relaying

Relaying [11] is different from routing in wireless networks. Instead of finding paths, it is concerned with how to relay traffic from one cell to another cell or from a non-covered area to a cell. To facilitate relaying, any MANET routing protocol can be assumed. Table 1.1 in section 2.1

reveals some ideas about this. In the table, there is no relaying scheme for MANETs because these networks have no BSs. Some relaying schemes in cellular networks in which mobile terminals are used for relaying should also be considered as relaying schemes in multi-hop cellular networks because relaying is made through multi-hopping. ODMA [9,13] and PARCelS [43] show examples of how this can be achieved.

Since cellular networks are capacity limited, relaying in these networks is very useful for releasing traffic congestion in congested (hot) cells and balancing load among cells. This helps to reduce call blocking probability and increasing system throughput. Relaying can be made through ad hoc relaying stations or mobile terminals. Some existing relaying schemes or architectures are iCAR [6,42], PARCelS [43], ODMA [9], and A-Cell [35]. Brief descriptions of these follow.

Integrated Cellular and Ad Hoc Relay (iCAR) is an ad hoc relaying load balancing scheme using low cost limited mobility Ad hoc Relay Stations (ARS) to relay traffic and achieve load balancing. The idea is to place a number of low cost ARSs in hotspot areas to relay excessive traffic from congested (hot) cells to their neighboring medium hot or cool cells. Traffic is further relayed to outer cooler cells so that congestion of the hot cells is released, call blocking probability of these cells is reduced, and load is balanced among cells. Details of load balancing algorithms will be discussed in section 2.4. An ARS has two air interfaces: Cellular (C) interface and Relaying (R) interface. The C interface is used for ARS-to-mobile or ARS-to-BS communications. The frequency band which is the licensed band remains unchanged. The R interface is used for mobile-to-ARS and ARS-to-ARS communications. This interface uses Industrial, Scientific, and Medical (ISM) bands [31] which are free. Licensed band and ISM are different and would not cause interference. Routing is based on both hierarchical and flat routing [6]. Medium Access is based on IEEE 802.11 WLAN MAC protocols. Figure 2.10 illustrates the operation of iCAR. Traffic of mobile A is relayed to adjacent BS through ARSs.

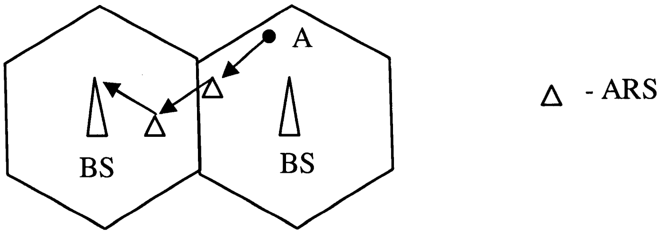


Figure 2.10: Relaying through Ad hoc Relay Stations

Pervasive Ad hoc Relaying for Cellular System (PARCelS) [43] is a network architecture that uses mobile nodes for relaying, without the need for ARSs. Thus, no additional cost of equipment, maintenance, or handling is required. When a BS is congested, mobile nodes search routes to other cells which have more free channels. Mobile nodes compute the best relay route. Desirable routes based on battery life, mobile speed and route length are sent to BS, which selects the best mobile nodes based on the location and the status of the destination BS for load balancing purpose. Figure 2.12 illustrates the traffic relaying through other MTs.

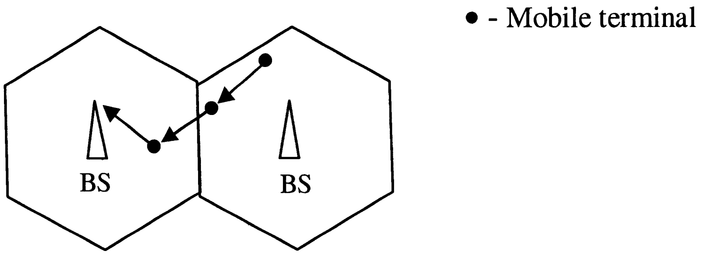


Figure 2.11: Relaying through other mobile terminals

Opportunity Driven Multiple Access (ODMA) [9] is a multi-hop relaying protocol for 3G cellular system. It is specifically designed for 3G cellular systems. The idea of ODMA is to break down a long single hop path into several small distance hops. Traffic is relayed through intermediate mobile terminals and the route is computed based on minimum total path loss. ODMA also reduces interference. Figure 2.12 show the operation of ODMA. Mobile nodes in low data rate region can transmit at high data rates using ODMA multi-hop relaying. Although ODMA is a possible solution for increasing the capacities of 3G systems, the routing issue has not been addressed in detail.

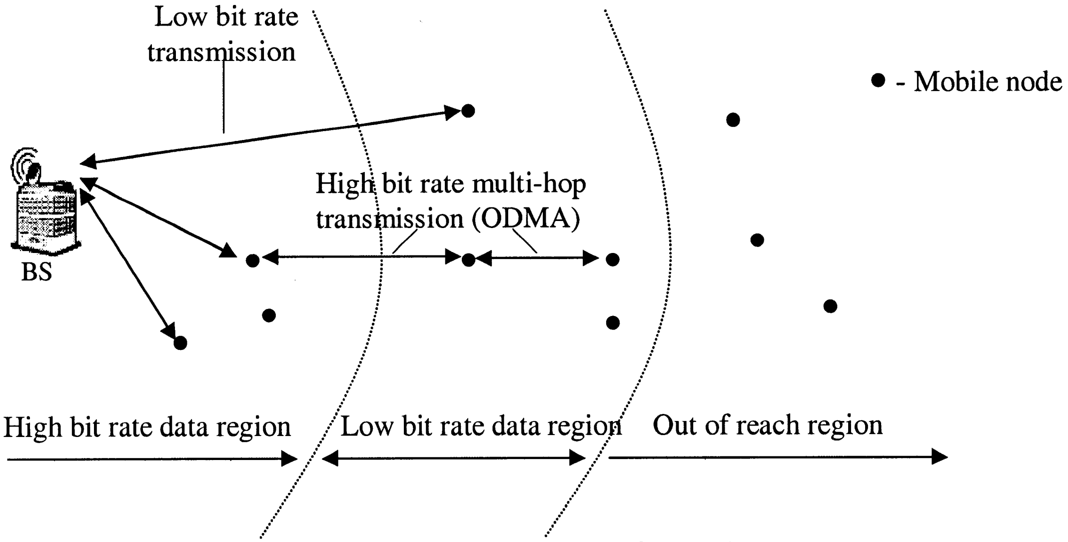


Figure 2.12: Operation of ODMA

Recently, a new multi-hop relay architecture has been proposed, namely the Ad hoc-Cellular (A-Cell) relay [35]. It is designed for TDD W-CDMA cellular systems. The idea of this architecture is similar to ODMA except that it uses directive antennas [25, 26] and the Global Positioning System (GPS). It also involves a contention-free medium access (channel allocation) model. On the contrary, ODMA does not provide channel allocation. Figure 2.13 shows the A-Cell relay architecture.

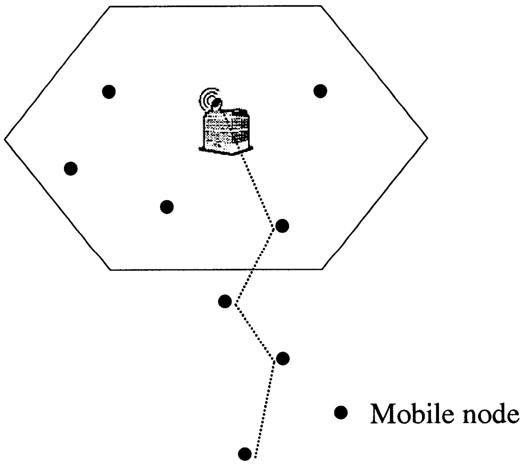


Figure 2.13: A-Cell relay networks architecture

The advantages of using ARSs are reliability and stability. Using ISM requires no extra licensing cost. It would not cause interference with the cellular band or consume cellular capacity. However, ARSs are still costly to implement. In addition, ARSs may not be effective in 3G systems in which traffic load is highly dynamic. Traffic of a cell is no longer strictly proportional to the number of users, but also to the types of services users request. Users may have several connections simultaneously for different purposes such as video conferencing and music streaming, consuming large amounts of cell capacity. The time and location of these users are difficult to predict. Congestion may occur at anytime in anywhere. However, utilizing the ISM band introduces co-channel interference. ARSs have to compete for the medium with other non-cellular users. The advantage of PARCelS is that it avoids the additional cost for relaying stations. It is also more flexible. However, it inherits all the disadvantage of MANETs. ODMA increases both cell coverage and capacity. However, it is complex and does not specify how routing is to be performed. A-Cell inherits the advantages of multi-hop cellular networks such as reduction in power consumption, enhancement of coverage and capacity. In addition, the usage of directive antennas and GPS reduces interference and increases spatial reuse. A-Cell also inherits the benefits of contention-free medium access protocols. Altogether, A-Cell stands as a promising solution for 3G and beyond wireless networks.

2.4 Routing

In the previous section, we described different relaying proposals. However, although relaying has many advantages, its efficiency is diminished without a good routing protocol.

In cellular networks, if relaying is made through limited mobility Ad hoc Relay Station (ARS), hierarchical routing and flat routing should be sufficient. However, if the relaying is made through mobile terminals, MANETs routing protocols may be required.

Routing is a major issue in MANETs because it directly affects the throughput and delay. There are two types of routing protocols: table-driven and demand-driven [40]. In table-driven routing, each mobile node maintains one or more routing tables to store routing information. The routing information is periodically updated throughout the network. Thus, each mobile node has consistent and up-to-date routing information of all nodes. However, periodical updates consume network resources. Examples of table-driven routing protocols are Destination Sequenced Distance Vector (DSDV) [29] routing, which is an extension of the distributed Bellman-Ford routing algorithm. In demand-driven routing, no route discovery is initiated unless there is a need. Periodic updates and event-triggered updates are partly or fully eliminated. Therefore, routing overhead is greatly reduced. Examples of demand driven, or on-demand, routing protocols are Dynamic Source Routing (DSR) [18] and Ad hoc On-Demand Distance Vector (AODV) [28]. On-demand routing protocol is preferred for MANETs because routing overhead is reduced. DSR has little overhead in route maintenance, but its latency of route discovery and overhead of source route are high. AODV has a better route maintenance strategy. The initial latency of route discovery is less, but it creates constant overhead even if there is no data transmission. In multi-hop cellular networks, a MANET routing protocol such as DSR and AODV or a specific routing protocol such as Base-Centric Routing (BCR) [16] can be used.

DSR uses source routing in which a complete route is in the header of a packet, i.e. routing tables are not required. Figure 2.14 shows the operation of DSR where source node A broadcasts a Route Request (RREQ) message that is flooded toward the destination D. Intermediate nodes add their node address to the message. When a RREQ reaches the destination, a source route (A-B-C-

D) is obtained. Destination node then sends Route Reply (RREP) message with the source route to the source. Source node can then send data packets with a complete route (path) in the header of the packet. Intermediate nodes route the packet based on the source route in the packet header. To improve the efficiency of route discovery, each node maintains a route cache to store the route information it learns from relayed packets. If a route is broken, a Route Error (RERR) message is sent to the source.

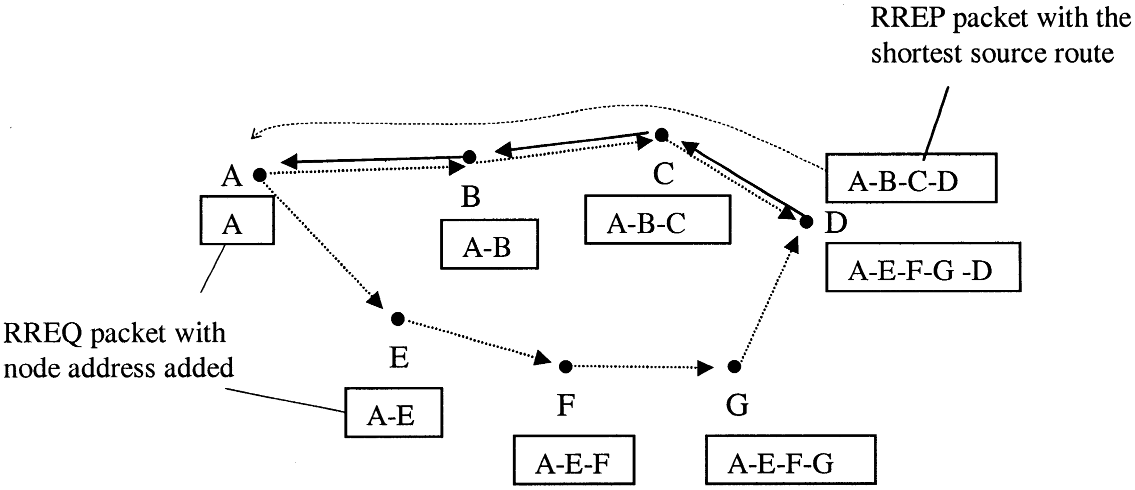
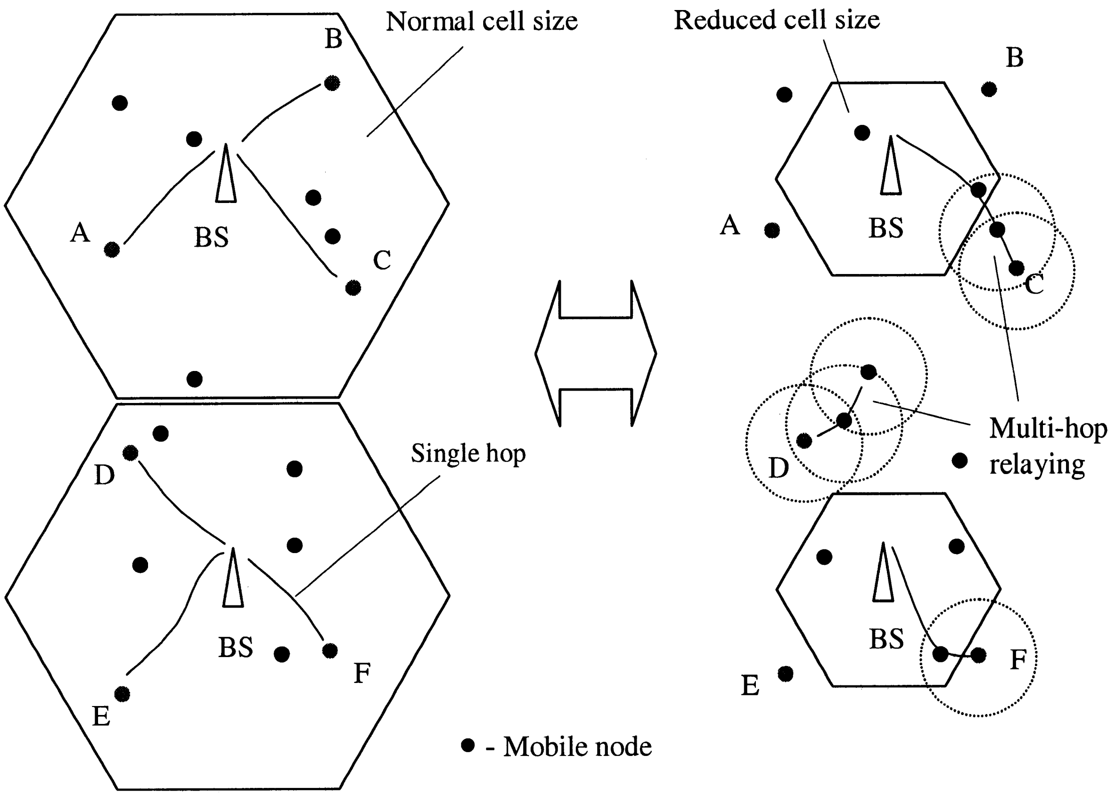


Figure 2.14: Operations of DSR

AODV is a combination of DSR and Distance Sequence Distance Vector routing (DSDV). It is similar to DSR except that it further employs hop-by-hop routing, sequence number (time stamp) and periodic hello messages (beacons) that are used in DSDV. AODV establishes route table entries at intermediate nodes dynamically. When a node initiates the route discovery process, it broadcasts a Route Request (RREQ) message with a sequence number (time stamp) to its neighbors. The message floods to the destination in a controlled manner. Each intermediate node creates a reverse path. In this manner, route table entries are established at each intermediate node. AODV relies on the routing table entries to propagate the Route Reply (RREP) back to the source. The source can then send packet to the destination. Node sends periodic Hello message to its neighbors for route maintenance.

Base-Centric Routing (BCR) [16] is a hybrid of table-driven and demand-driven protocols that is designed for Multi-hop Cellular Networks (MCN) [21]. In BCR, the BS keeps track of the network topology of its own cell by using table driven protocols. Each mobile terminal sends a table (list) of its neighboring nodes to the BS. Path or route is computed based on the topology information at the BS. If a mobile terminal needs a route, it sends route request to the BS. If a mobile terminal is out of the transmission range of a cell or cannot get a route from the BS, the mobile terminal discovers the route using AODV. Simulation results show that BCR has better performance than traditional MANET routing protocols in MCN. However, although BCR is more suitable for use in multi-hop cellular network, it can only be used in dense networks; otherwise, many nodes that can originally reach the BS become unreachable. The reason is due to the increased separation among BSs and the small transmission range. Figure 2.15 illustrates this situation. Nodes A, B, D, and E become unreachable after the cellular network is changed to a MCN. Although this problem cannot be solved in conventional multi-hop cellular network, it may be solved in 3G multi-hop cellular systems because the cell size and transmission range in these system are changeable. Another issue of BCR is that it chooses paths based on a small number of hop counts. This raises fairness and energy efficiency issues. Nodes on the route which has smaller hop count may always be chosen for relaying. They consume more energy and power outage becomes more likely. In addition, BCR is designed for conventional cellular systems; it does not acknowledge the characteristics of 3G cellular systems. In section 3, a new routing protocol is designed to address these issues.



a) Communication in cellular networks

b) Communication in MCN-p

Figure 2.16: Communications in a) cellular networks and b) MCN-p

2.5 Load balancing

Load balancing schemes in cellular networks, MANETs, and multi-hop cellular networks are quite different because of the unique characteristics of the networks involved. Load balancing in cellular networks basically attempts to share load among cells, i.e. BSs, to release congestion and reduce call blocking probability. Since there are no intermediate nodes, balancing the load of mobile terminals (MTs) is not necessary. The methods can be traffic relaying [6, 43], channel borrowing [5], capacity sharing [7] or flexible users directing [41]. Load balancing in MANETs is made through balancing the load among mobile terminals to release nodal congestion, increase packet delivery ratio, increase throughput, and reduce delay. Since there is no BS, balancing load among BSs is not necessary. The methods include load-aware routing [12], multi-path routing

[30], and route de-coupling [33]. Load balancing in multi-hop cellular network involves balancing the load among BSs or intermediate nodes or both. Balancing load among BS is more important because it reduces the call blocking probability. Balancing load among intermediate nodes helps avoiding congestion of the relaying route. In addition, load balancing in multi-hop cellular networks should involve selecting a node or path for relaying traffic among cells. Discussions of these load balancing schemes are made below.

2.5.1 Load Balancing in Cellular Networks

In cellular networks, load balancing by traffic relaying can be done through ARSs or MTs. Examples of relaying schemes are iCAR, PARCelS, and ODMA. Details of these relaying schemes have been described in previous section 2.2. Here, we will focus on the load balancing component of these schemes.

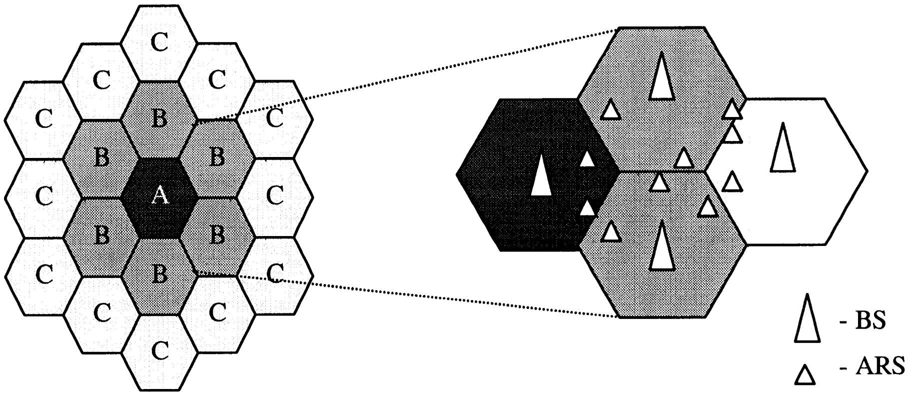


Figure 2.16: Three-tier cell system

The analytical model of load balancing of iCAR is based on a 3-tier cell system, as shown in Figure 2.16. Center cell A is a hot spot cell surrounded by medium hot cells (B cells). The outer C cells are cool cells. ARSs are placed in the cells. Call blocking probability of cells in each tier is computed using the traffic intensity, the number of channels, and the coverage of ARSs. A small portion of traffic of cell A is relayed to cell Bs if the traffic intensity of cell A is greater than or equal to the traffic intensity of cell B. A small portion of traffic of cell B is relayed to cell Cs if the traffic intensity of cell B is greater than or equal to the traffic intensity of cell C. The above relaying steps continue until any one of the following condition are fulfilled.

1. Call blocking probability of all 3-tier cell are equal.
2. The total relayed traffic from cell A to cell B and from cell B to cell C is greater than or equal to the total traffic intensity that can be relayed by ARSs
3. Call blocking probability of tier A cell is equal to call blocking probability of tier B cells and total relayed traffic from tier B cells to tier C cells is greater than the traffic intensity that can be relayed by the ARSs.

Several assumptions for this load balancing model were made. The traffic intensities of all tier B cells are equal. The traffic intensities of all tier C cells are equal. Traffic is relayed from cell A equally to neighboring B cells and from cell B equally to neighboring C cells. Traffic relaying among same tier cells does not occur. Performance evaluation is based on these assumptions. Obviously, these assumptions are too limited. Heterogeneous load environment has not been addressed. Neighboring cells of a hot cell can be medium hot or cool cells. The model is also too theoretical. Traffic intensity is based on call arrival rate and call holding time. Although this value can be calculated based on the average value of measurements, selection of source nodes for relaying cannot be easily made as the actual holding time of the nodes is not known. In 3G systems, the situation becomes even worse because the call or connection holding time is difficult to predict.

In PARCelS [43], when a BS is congested, mobile nodes search routes to other cells which have more free channels. Mobile nodes compute the best relay route. Desirable routes based on battery life, mobile speed and route length are sent to BS. BS selects the best mobile nodes based on the location and the status of the destination BS for load balancing purpose. A disadvantage in PARCelS is that the searching for routes involves high routing overhead. In fact, the cellular infrastructure is not fully utilized for route searching.

Channel borrowing [5] is another method to achieve load balancing. The idea is that hot cells borrow a number of channels from their neighboring cooler cells based on a structured borrowing mechanism. This mechanism requires that a hot cell can only borrow channels from its adjacent outer ring cells. A hot cell should also borrow enough channels to serve itself and its adjacent hot cells. Although this scheme works in conventional cellular networks, it does not apply to CDMA networks in which there is only one frequency for the whole networks. In other words, there are no additional channels (frequencies) that can be borrowed.

In [19], bandwidth migration and preemption was proposed to balance the load of multi-media cellular networks. The idea is based on on-line bandwidth migration and reservation. Network conditions are measured on-line (or in real-time) and preemption decisions of existing calls is made. Cells are classified as Peak status cell (P-cell), Potential Peak cell (PP-cell), and Safe cell (S-cell). A P-cell is a hot cell that needs to borrow bandwidth from other cells. A PP-cell has reserved bandwidth such that it will not lend or borrow bandwidth from other cells. An S-cell is a cool cell that can lend bandwidth to P-cells. Migration load is unified to reduce migration overhead. The idea of call preemption is that calls with low values are preempted for high value calls. The value of a call depends on the bandwidth requirements and priority of the call. Although results show that throughput increases and that call dropping and blocking probabilities decrease, this scheme is not suitable for CDMA systems because no additional bandwidth (frequency) can be migrated or borrowed in CDMA systems.

In [7], a cooperative (coverage) negotiation approach was presented. The idea is to change the cell size (coverage) based on the loading situation of cells using a certain antenna technology. The antenna pattern in a hot cell contracts to reduce the serving area while the antenna pattern in cool cells expands to serve a larger area. As the serving area reduces, the number of users needed to be served is also reduced. Thus, congestion is released and the call blocking probability is lowered. Expanded neighboring cool cells serves the users which are originally served by the hot cell. In other words, traffic of hot cell is shifted to these cooler cells. Load balancing among cells is achieved. The decision of cell size is based on the cooperative negotiation between congested cell and its neighboring cells. Figure 2.17 shows the operation of this approach. Hot cell A is shrunk while neighboring cooler cell expands to cover the area that is originally covered by cell A. Some users that are originally served by cell A are served by neighboring cool cells. In this approach, the complexity of antenna increases. The proposal is based on the assumption that the capacity of a cell remains unchanged when the size of the cell varies. However, this assumption is not applicable in CDMA systems in which capacity of a cell decreases as coverage increases. This approach is also not flexible enough to deal with the situation where adjacent cells of hot cells are also congested while the outer-most cells are non-congested. In this case, the decision of cell size and cell shape becomes very complex. In fact, this critical situation cannot be addressed by this approach.

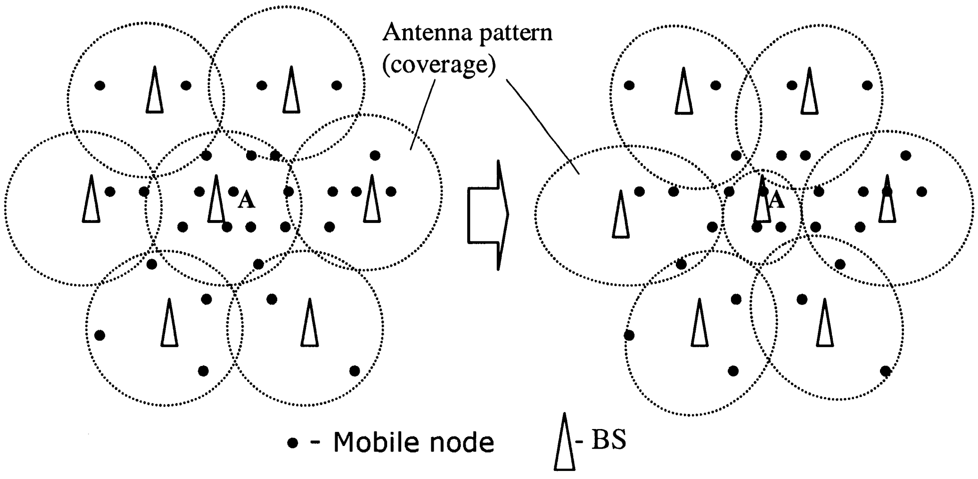


Figure 2.17: Operation of cooperative negotiation approach

In [41], directing flexible users in the overlapping regions of cells is proposed. These users have the flexibility to communicate with any of the cells which covers this region. When one cell is hot, new call arrivals in the overlapping region of the cell can be directed to one of the adjacent cells. In this manner, the load of the hot cell will not increase. Current calls in the overlapping region, which belongs to the hot cell, can also be shifted to connect to an adjacent cooler cell. Figure 2.18 illustrates this scheme. User A and B must communicate with BS 1 and BS 2 respectively. User C in the overlapping region can communicate with BS 1 or BS 2. The scheme is also too limited because it depends on the existence and size of the overlapping region. It also depends on the number of active users, i.e. source nodes, in the region. If the overlapping region does not exist or there is no user in the region, this scheme does not work.

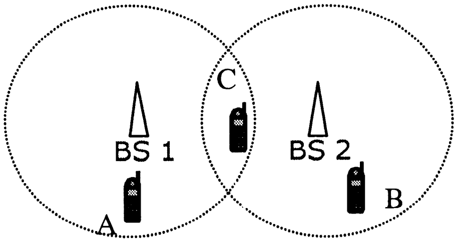


Figure 2.18: Flexible user in an overlapping region

2.5.2 Load Balancing in MANETs

Load balancing in MANETs is usually incorporated with the routing protocol. With a good load-aware routing protocol, traffic can be evenly distributed among mobile nodes such that the load can be balanced.

In [12], a load-aware routing protocol is proposed, namely Load-Balanced Ad hoc Routing (LBAR). The idea is to add load awareness to existing MANET routing protocols. The load of a MT is defined as a function of its activities and the activities of its neighboring nodes. The load information of each node along all possible paths from a source to its destination is sent to the destination during the route discovery process. The destination selects the best cost path based on the information received and sends this decision back to the source. The source then start sending signals (packets) through the selected path. Figure 2.19 illustrates this situation. Although A-B-C-D is shorter path for node A, the path A-B'-C'-D is chosen because the nodes on the path has less traffic or activities. Simulation results show that LBAR outperforms existing MANET routing protocols in packet delivery ratio and average end-to-end delay. Although this scheme is not applicable to cellular networks because cellular network has no intermediate nodes, it may help balance the load among relaying nodes in multi-hop cellular networks.

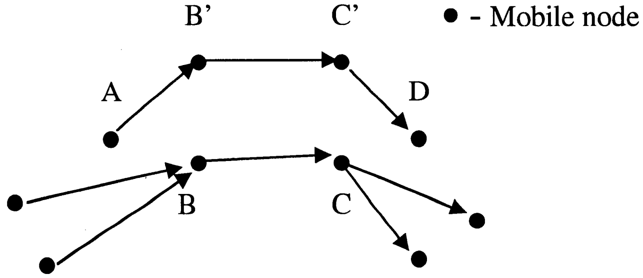
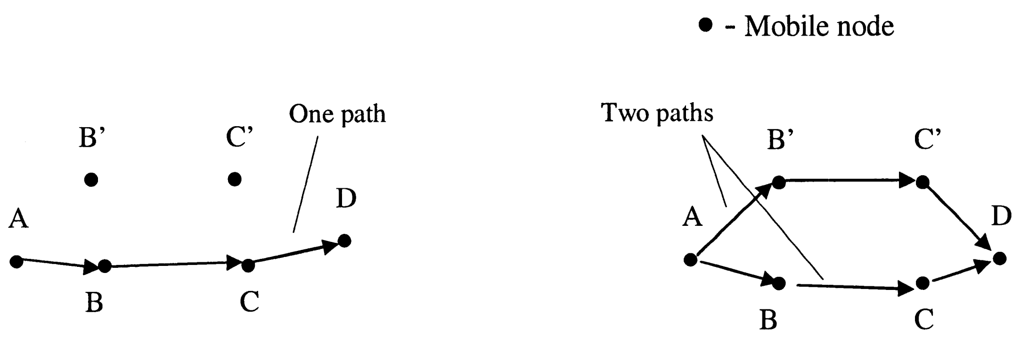


Figure 2.19: Operation of load-aware routing

In [30], a multi-path load balancing concept is proposed. Conventional MANET routing protocols, such as DSR or AODV, send packets through a single path. Using multi-path routing, packets can be sent from source to destination through several different paths such that traffic is distributed among nodes. This helps avoiding traffic concentration on a single route. Hence, load among nodes is balanced. Figure 2.20 illustrates single path routing and multi-path routing. In Figure 2.20a, packets of node A go though a single path (A-B-C-D). In Figure 2.20b, packets of node A can go through two paths: (A-B-C-D) and (A-B'-C'-D).



a) Single path routing

b) Multi-path routing

Figure 2.20: Illustration of a) single path routing and b) multi-path routing

Although multi-path routing protocols help reduce congestion, increase throughput and route resilience compared to single-path routing in MANET, interference due to the effect of medium contention increases. In [33], a network-aware MAC and routing protocol is proposed to address this issue. The idea is to select maximally zone disjoint routes to minimize route coupling. Route coupling occurs when two routes are close to each other such that they are contending for the same medium and, hence, network performance deteriorates. This effect is even worse when multi-path routing is used. In this protocol, each node keeps status information (active or inactive) of neighboring nodes. Then, each node has topology and activities information for computing the best next hop node. This scheme uses small, low cost and low power ESPAR directive antennas [25, 26] to further reduce the interference. Figure 2.21 shows how directive antennas are used to reduce route coupling.

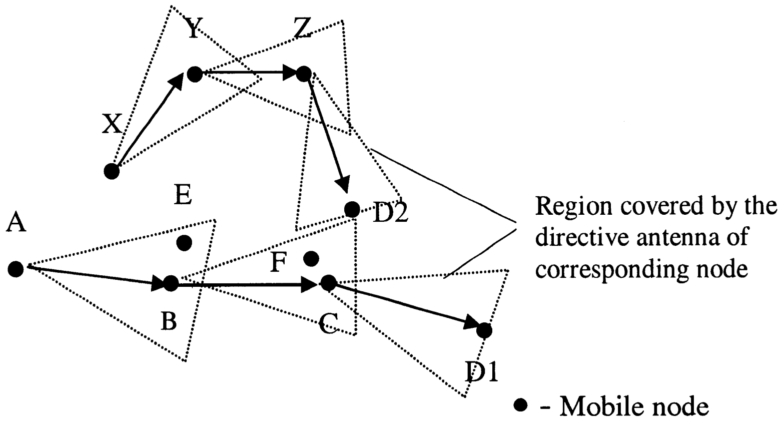


Figure 2.21: Route de-coupling by using directional antenna

The load balancing protocols above are designed for MANETs; they are not applicable in cellular networks. However, these protocols may be applicable in multi-hop cellular networks to balance the load of relaying MTs. Directive antennas help decouple multi-path routes and reduces interference. This idea is also applicable in CDMA multi-hop cellular systems: When interference is reduced, capacity increases. In addition, directional antennas increase spatial (channel) reuse, and reduce power consumption and interference.

2.5.3 Load Balancing in Multi-hop Cellular Networks

Very few load balancing schemes are specifically designed for multi-hop cellular networks. PARCelS can be considered as a load balancing scheme in these networks because load is relayed through MTs. Other load balancing schemes are designed for either single hop cellular networks or MANETs. The reason might be because the load balancing in multi-hop cellular networks involves two components: load balancing among cells (cellular component) and load balancing among MTs (MANET component). There are existing solutions for these two areas. It may be convenient to take the solutions from each area and combine them together. However, without coordination between these two components, the solution may not be applicable or effective. For example, when a cellular load balancing component is activated, which source MT should be chosen for re-routing? Applying the MANET component cannot solve this problem because it is only concerned with finding a route for a source MT. This should be added to the functionalities of the cellular component. Existing cellular load balancing schemes seldom address this issue. In 3G cellular systems, since the traffic generated by a MT can be very large, the issue of choosing MTs for relaying to achieve load balancing is non-trivial, and requires further involvement on the part of the cellular component.

2.6 Summary

Cellular networks and mobile ad hoc networks (MANETs) are two distinctive wireless systems. Cellular networks are centralized systems while MANETs are distributed and have no infrastructure. Their medium (multiple) access techniques, relaying schemes, routing protocols, and load balancing schemes are also different. In addition, within cellular systems, there are different medium access techniques: FDMA, TDMA, and CDMA. FDMA and TDMA are bandwidth-limited technologies and are mainstream techniques for conventional cellular systems. CDMA is interference-limited technology in which the cell breathing effect is present. In 3G wireless systems, Wideband CDMA (W-CDMA) technology is used which allows data rates up to 2 Mbps. The wide range of user data rates makes it difficult to predict the load of a cell.

The main issue in cellular networks is traffic congestion of a cell which causes call blocking. The main issue in MANETs is routing. A multi-hop cellular network is a combination of cellular networks and MANETs. It inherits the merits, complexity, and issues of both networks. Although the network is complex, it helps reduce the number of BSs, power consumption, interference, and increasing system throughput. It also facilitates relaying and load balancing for releasing congestion of a cell, and allows peer-to-peer communication within a cell for alleviating the load. Examples of multi-hop cellular networks are MCN, HWN, and A-Cell. In 3G systems, W-CDMA is used such that existing relaying schemes, routing schemes, load balancing schemes may not be effective or even applicable in this environment. Designing a load balancing relaying scheme for these systems is needed. However, it is a non-trivial task. We have to consider the medium access, relaying, routing, and load balancing issues in cellular networks and MANETs.

Contention-Free medium access techniques such as FDMA, TDMA, and CDMA, are better for bandwidth utilization because medium contention and back-off are not required. The techniques also avoid hidden and exposed terminal problems. However, these techniques require centralized control. Therefore, they are suitable for cellular networks but not for MANETs. In multi-hop cellular networks, MCN and HWN use the contention-based IEEE 802.11 MAC protocol for simplicity. A-Cell employs TDD W-CDMA technology which requires a slot assignment scheme such as DSSA. Simulation of DSSA show promising performance results in terms of delay and throughput. Although this scheme has not clearly specified the condition of restricting the current node from assigning the same channel as the channel that the successive node is transmitting on, contention-free medium access techniques, e.g. TDD W-CDMA, with slot assignment scheme should be the direction for the medium access technique in 3G or future generation multi-hop cellular networks.

Relaying is different from routing. Relaying usually refers to the way to relay traffic from one cell to another cell or from non covered nodes to a cell. Details of routing (or finding a route) are left to routing protocols. Relaying helps to release cell congestion and, thus, reduce call blocking probability. Relaying in cellular networks can be usually performed through ARSs or MTs. If ISM band and IEEE 802.11 MAC protocol are used, the effectiveness of relaying scheme may be affected because they introduce medium contention and co-channel interference. Security and billing issues also exist. If MANETs routing protocols are assumed, the vulnerability of routes is also introduced. In MANETs, relaying is not necessary because these networks have no BSs. Routing is in fact a major issue in these networks because it affects the throughput and delay of the systems. On-demand routing protocols are preferred because routing overhead is reduced. In multi-hop cellular networks, relaying is done through mobile terminals. PARCelS, ODMA, and A-Cell relay are examples. PARCels involve routing while ODMA and A-Cell relay does not. In PARCels, mobile terminals search for routes to the cells with more free channels. In ODMA, a single route is broken into several smaller hops to reduce transmission power and interference. In A-Cell relay architecture, directive antennas are used to further reduce interference, energy consumption and increase spatial reuse. Although the relaying concept has many advantages, without a good routing protocol, the efficiency of relaying is greatly affected.

There is no routing in cellular networks unless relaying schemes are used in these networks. In multi-hop cellular networks, MANET routing protocols are usually assumed. However, they are not efficient and do not fully utilize the infrastructure of the cellular systems. BCR is a routing protocol specifically designed for MCN and is suitable for densely connected networks only. If this protocol is applied in a sparsely connected network, more potential source nodes will not be able to reach BS because there are not enough multi-hop routes. This protocol also does not address fairness and energy issues. In addition, BCR is designed for conventional cellular system. It does not fully utilize the characteristic of 3G cellular systems. A routing protocol for 3G multi-hop cellular networks will be presented in section 3.

The roles of load balancing in cellular networks and MANETs are different, although their objectives are balancing load and releasing congestion. The objective of load balancing in cellular networks is to share load among cells (BSs) to release congestion and reduce call blocking probability in the networks. This can be made through traffic relaying, channel borrowing, coverage negotiation, or flexible users directing. Most schemes are based on mainstream cellular systems in which FDMA or TDMA are used. They may not apply or be flexible enough for 3G systems in which W-CDMA technology is used. For example, iCAR is costly and is not flexible enough to deal with the highly dynamic load environment of 3G systems. Channel (frequency or bandwidth) borrowing can not apply in 3G systems because there are no extra channels that can be borrowed. The coverage negotiation approach assumes constant cell capacity which is not applicable to CDMA systems. The idea of flexible users works but load migration is limited to the overlapping region of cells. Load balancing in MANETs is a totally different concept, the objective of which is to balance the load among MTs to release nodal congestion, increase throughput, reduce delay, and reduce energy consumption. Fairness and QoS provisioning should be considered. Call blocking probability is not a concern in these networks unless an admission control mechanism is added. Methods include load-aware routing, multi-path routing of packets, and route de-coupling routing.

In multi-hop cellular networks, few load balancing schemes have been developed. iCAR and PARCelS may be considered as load balancing schemes in these networks. However, these schemes are still in the context of conventional cellular systems. It may not be suitable for 3G environments. A new load balancing paradigm is needed for 3G systems.

In the next chapter, we propose a load balancing and relaying framework specifically for 3G wireless environments, where a load balancing scheme is designed for balancing load among BSs. Load balancing among relaying nodes is handled by a load-aware routing scheme. A slot assignment scheme is also included. This framework is aimed at future generation wireless networks [2, 8].

Chapter 3

A-Cell Load Balancing Relaying (ALBAR) Framework

As we stated in the summary of Chapter 2, a new load balancing paradigm is required for 3G wireless systems. We herein propose a new load balancing and relaying framework for 3G TDD W-CDMA cellular systems. This framework consists of three components: a load balancing scheme, a load-aware routing scheme, and a slot assignment scheme. This framework is designed not only for balancing among base stations, but also for balancing the load among relaying nodes. Since the main objective of load balancing in cellular networks is to balance the load among cells to reduce the call blocking probability, we focus on investigating the effect of our framework on load balancing among cells (or BSs). Investigating the effect of load balancing on relaying nodes will be part of our future work.

3.1 Framework Overview

The A-Cell Load Balancing Relaying Framework (ALBAR) is a centralized dynamic load balancing relaying framework for 3G TDD Wideband CDMA (W-CDMA) cellular systems in which a 3G base station and the Radio Network Controller (RNC) are present. Location-aware relaying is based on the Ad hoc Cellular (A-Cell) relay architecture [35]. Routing is based on the characteristic of dynamic cell size (cell breathing effect [31]) in 3G cellular systems. The W-

CDMA medium access technique is used. The slot assignment scheme is based on the idea of Delay Sensitive Slot Assignment (DSSA), with a minor modification. GPS is used to provide the location information of each node to facilitate routing. This reduces the routing overhead. Directional antennas are used to reduce power consumption, interference, and more importantly to increase spatial channel reuse.

The three main components of the framework are:

1. A-Cell Load Balancing (ALBA) Scheme.
2. A-Cell Adaptive Routing (ACAR) Scheme.
3. Extend-Delay Sensitive Slot Assignment (E-DSSA) Scheme.

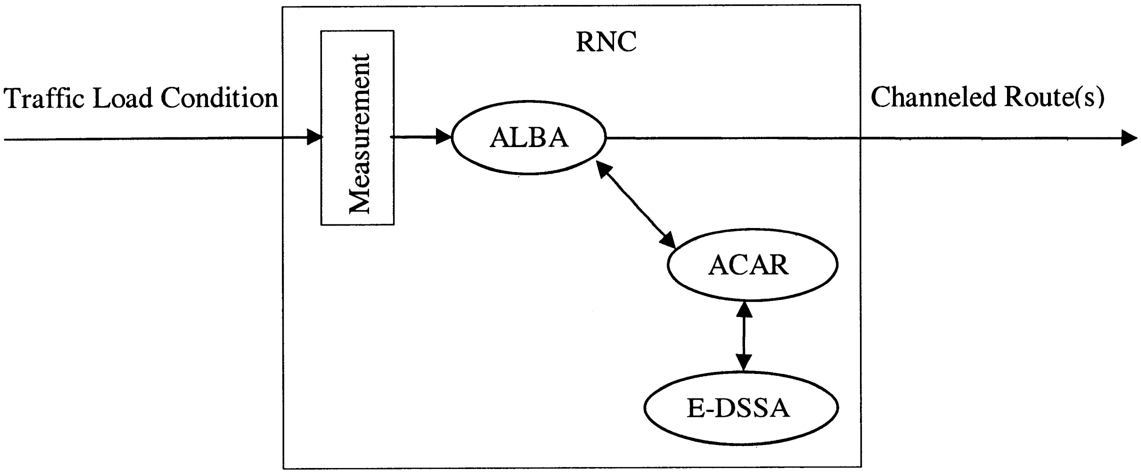


Figure 3.1: ALBAR Framework Architecture

All three components are located at the RNC as it has the topology and channel information of all BSs and mobile nodes which are within the coverage of the BSs. Figure 3.1 shows the architecture of the framework. ALBA monitors the load situation of the networks by measuring the traffic load condition periodically. Whenever load deviation among cells in the networks is greater than a predefined threshold, ALBA starts performing load balancing. ALBA selects

Source Cells, Source Nodes and Target Cells for traffic load migration. Selection of the cells is based on their load situations. Selection of the Source Nodes is based on their Migration Priority (MP). Source Nodes with higher MP are selected to be migrated. Since load migration is not possible unless valid routes are present, ALBA has to interact with ACAR to obtain valid routes. In fact, ACAR first finds a route and then interacts with E-DSSA for finding channels for the route. ALBA continues to select Source Nodes and cells for load migration until load deviation among cells is less than the threshold or no further load migration can be done, i.e. all cells are tried. RNC then sends signals to the corresponding BSs to update their channel information. The BSs send signals to the corresponding Source Nodes and Relaying Nodes to update their channel tables and routing tables. Detailed descriptions of each component with algorithms and examples are given in the next section.

3.2 Description of the Framework

The three components of this framework are described below.

3.2.1 A-Cell Load Balancing (ALBA) Scheme

ALBA is a dynamic load balancing scheme for multi-hop ad hoc cellular system. It is located at the RNC. The scheme operates in one of two states: The Load Monitoring State and the Load Balancing, or migration, State. Figure 3.2 illustrates the state diagram in ALBA. ALBA Scheme is illustrated in Figure 3.4. Before we start to describe this scheme, we define some terminologies:

1. **Load** (*L*) is the traffic of a base station (BS) or a node. The traffic can be received traffic or transmitted traffic or both. The traffic is in term of packets per second or Load Factor [14]. Load Factor is the fraction of the capacity of a cell.
2. **Source Cell** (C_{Src}) is a cell that is selected to shift its load to the other cell.
3. **Target Cell** (C_{Trg}) is a cell that is selected to receive load from a Source Cell.
4. **Source Node** (*Src*) is a mobile node that generates traffic.

5. **Relaying Node** (**Nd_{relay}**) is a mobile node that relays traffic.
6. **Time Slot** is a time interval (slot) of a frame for signal transmission.
7. **Code** is a spreading code [39] that is used for spreading in CDMA systems.
8. **Channel** is a Time Slot Code pair (*TS, C*). Example: (4, 2) represents Time Slot 4 and Code 2.
9. **Channeled Route** is a route (or path) on which each node can be assigned a Channel.
10. **Channel Pool** is a pool of available channels in the BS.
11. **Channel Table** is a table for storing channel information for the connections of Source Nodes. This table is located at the Source Nodes and the Relaying Nodes.
12. **Channel Allocation Table** is a table storing channel information for the connections of Source Nodes. This table is located at the BS.
13. **Routing Table** is a table storing routing information.
14. **Connection** is a communication link between a Source Node and a BS. A Source Node can have several connections with one BS at any given time.
15. **Neighboring Load Deviation** (**d_n**) is the load difference between two neighboring cells.
16. **Neighboring Load Deviation Threshold** (**d_{nThres}**) is a load threshold value above which load migration planning between Source Cell and Target Cell can start.
17. **Network Load Deviation** (**D_N**) is the load difference between the highest load cell and the least load cell in the network.
18. **Network Load Deviation Threshold** (**D_{NThres}**) is the load threshold value for triggering load migration.
19. **Load Sampling Period** (**T_s**) is the time interval for checking load situation of the network.
20. **Migration Priority** (***MP***) is the priority level of a Connection of a Source Node for load migration. Higher priority of a connection has a higher chance to be selected for load migration.
21. **QoS Class** is the Quality of Service (QoS) class of a connection. For examples, Conversational, Streaming, Interactive, and Background are four common QoS Classes [14].

22. **Call Blocking Ratio** (***CBR***) is the ratio of the number of calls blocked to the number of calls requested.

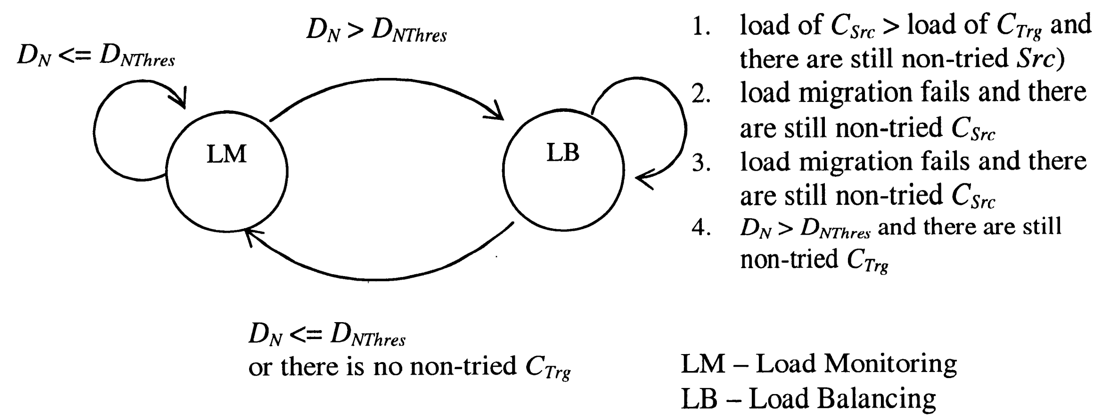


Figure 3.2: State diagram for ALBA

In the Load Monitoring State, ALBA periodically checks the load status of the networks through RNC every Load Sampling Period (T_s). If Network Load Deviation (D_N) is greater than Network Load Deviation Threshold (D_{NThres}), this state transfers to the Load Balancing State for load migration.

In the Load Balancing State, ALBA performs load migration. ALBA selects a Target Cell, a Source Cell, and a Source Node of the Source Cell. The Target Cell is the least load cell in the network. The Source Cell is a cell neighboring the Target Cell with the highest load above the Target Cell's load. For example, in Figure 3.3, cell C is chosen as the Target Cell and cell B is chosen as the Source Cell. To proceed further, the load difference between the Source Cell and

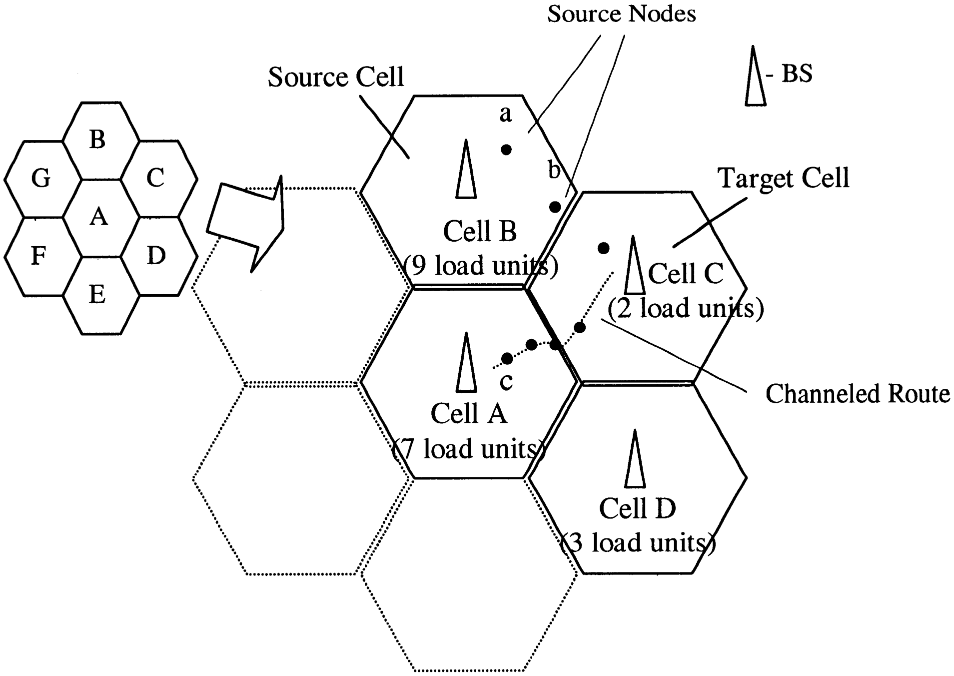


Figure 3.3: Load migration in a seven-cell scenario

the Target Cell must be greater than the Neighboring Load Deviation Threshold (d_{nThres}); otherwise, no load can be migrated. The Call Blocking Ratio (*CBR*) of the Source Cell also needs to be greater than or equal to the CBR of the Target Cell to avoid unfairness. This is because the load of the Target Cell may be lower than the load of the Source Cell even though the Target Cell may have a higher *CBR*. Allowing load migration in this situation will cause more traffic congestion at the Target Cell at a later time.

Next, we can select a Source Node which belongs to the Source Cell. The Source Node is not necessarily located in the Source Cell because the Source Node may have been relayed from another neighboring cell or even from the Target Cell at an earlier time. Selection of a Source Node is based on the Migration Priority (*MP*) level of the Source Node. The level can be calculated based on a cost function (equation 1).

$$\text{Migration Priority } P_i = k_1 * C_i + k_2 * f(d_i) \qquad (1)$$

$$\text{Where} \quad \begin{array}{rl} f(d_i) = 4 & | \; d_i =< R \\ = 3 & | \; R < d_i =< 1.25\,R \\ = 2 & | \; 1.25\,R < d_i =< 1.5\,R \\ = 1 & | \; 1.5\,R < d_i \end{array}$$

Where P_i is the Migration Priority level of a Connection of Source Node i; C_i is the QoS class of the connection; d_i is the distance between the Source Node and the BS of Target Cell; $f(d_i)$ is a mapping function to map the distance d_i to a distance factor which is an integer from 1 to 4; and R is the cell radius. Since there are four QoS classes, we map the distance d_i into four values so that, at a certain location of the Source Node, the value contributed from the QoS class of the connection breaks even with the distance factor contributed from the distance between the Source Node and the BS. When d_i is greater than 1.5R, which is mid-way between the cell border and the BS of the Source Cell, the distance factor is 1. Constant k_1 and k_2 are weighing factors for C_i and $f(d_i)$, respectively. There are four QoS classes: Conversational (1), Streaming (2), Interactive (3), and Background (4). Equation (1) implies that if two Source Nodes have the same distance and k_1 and k_2 are equal, the Source Node with a Connection of Conversational class will have the least *MP*. When considering the *MP* for a QoS class C_i, since relaying for load migration introduces a longer delay and instability in a route, it is better to avoid choosing the connection of Conversational class for load migration. In other words, the connection of Background traffic class has a higher priority to be migrated. When considering the *MP* for a distance d_i, the shorter the distance to the Target Cell BS, the higher the *MP* is. The reason behind thus is that shorter distances may have fewer hops and shorter transmission ranges. Thus, packet delay, route power consumption, and cell interference are all decreased.

ALBA Scheme

A) Load Monitoring State

```
1. Checks the load status of the network through RNC in every Load Sampling Period ($T_s$)
2. If Network Load Deviation ($D_N$) > $D_{NThres}$, then
3.     Transit to Load Balancing State
```

B) Load Balancing State

```
1.  Do
2.      Do select * (next) least load cell as Target Cell ($C_{Trg}$)
3.          Do select † (next) highest load (higher than $C_{Trg}$) non-tried neighboring cell of
            the $C_{Trg}$ as $C_{Src}$

4.              If Neighboring Load Deviation > $d_{nThres}$ and $CBR$ of $C_{Src}$ >= $CBR$ of $C_{Trg}$,
5.                  Do select non-tried $Src$ with highest MP for neighboring load migration.
6.                      Perform ACAR to find a Channeled Route from the $Src$ to the $BS$ of
                        $C_{Trg}$.

7.                  While (load of $C_{Src}$ > load of $C_{Trg}$ and there are still non-tried $Src$)
8.          While (neighboring load migration fails and there are still non-tried $C_{Src}$)

9.      While (neighboring load migration fails and there are still non-tried $C_{Trg}$)

10. While ($D_N$ > $D_{NThres}$ and there are still non-tried $C_{Trg}$)

11. ALBA sends signals to corresponding BS(s) through RNC to update their Channel
    Allocation Tables and Channels Pools.
12. BS(s) sends signals to the nodes on both old and new route(s) to update their Channel
    Table(s) and Routing Table(s). (Order of signal sending can be simultaneously or least
    loaded cell first.)

12. Transit to Load Monitoring State.
```

* - If several least load cells are tied, randomly select one which is not an adjacent cell of previous C_{Trg} if possible.

† - If several highest load cells are tied, randomly select one of these cells which is not a previous C_{Trg} if possible.

Figure 3.4: ALBA Scheme

Once a Source Node is Selected, ALBA calls ACAR to find a Channeled Route to relay the Connection of the Source Node to the Target Cell. ACAR itself performs path (route) finding. It first gets the topology and state information of all the relaying nodes in the corresponding cells.

It computes a route based on this information. If a possible route is found, ACAR calls E-DSSA to find a channel for each node on the route. If each node on the route can be assigned a channel, the route is a Channeled Route and load migration for the Source Node is a success. ACAR replies to ALBA with a successful signal and the information of the Channeled Route. If any node on the route cannot be assigned a channel, the route is considered invalid. Then, ACAR continues to compute a valid route for this Source Node until a Channeled Route is found or all Relaying Nodes are tried. Please recall that this computation takes place at the RNC. No signal is being sent to BS or mobile nodes at this stage. Whether the load migration for this Source Node is a success or not, ALBA continues to select another Source Node for load migration until the load of the Source Cell is not greater than the load of the Target Cell or all Source Nodes in the Source Cell are tried. If no Channeled Routes are found for any of the Source Nodes, the load migration between the Source Cell and the Target Cell fails. For example, in Figure 3.3, Source Node a in Source Cell B cannot find a route to Target Cell C. If node b also cannot find a Channeled Route to cell C, the load migration between these two cells fails. In this case, ALBA selects the next highest load neighboring cell, higher than the load of cell C, of cell C as a new Source Cell, i.e. cell A. Previous procedures of finding a Source Node and a Channeled Route are repeated. If load migration between cell A and cell C still fails, ALBA continues to try the next highest load neighboring cell, higher than the load of the Target Cell, as the new Source Cell until all neighboring cells are tried or load migration is a success. If load migration still fails, ALBA selects the next least load cell as the new Target Cell, cell D. Previous procedures of trying Source Cells and Source Nodes are repeated. If load migration is still a failure, ALBA continues to find a new Target Cell until all cells are tried or load migration is success. Figure 3.5 shows an extreme scenario where no route exists between any pair of the cells. Thus, no load can be migrated. ALBA first tries cell C as the Target Cell and Cell A, B, and D as the Source Cell one by one. Since no load can be migrated, ALBA tries cell B as a new Target Cell and cell G, and A as the Source Cell one by one. Since no load can be migrated, ALBA tries cell D as a new Target Cell and so no. ALBA continues to try until cell F is tried.

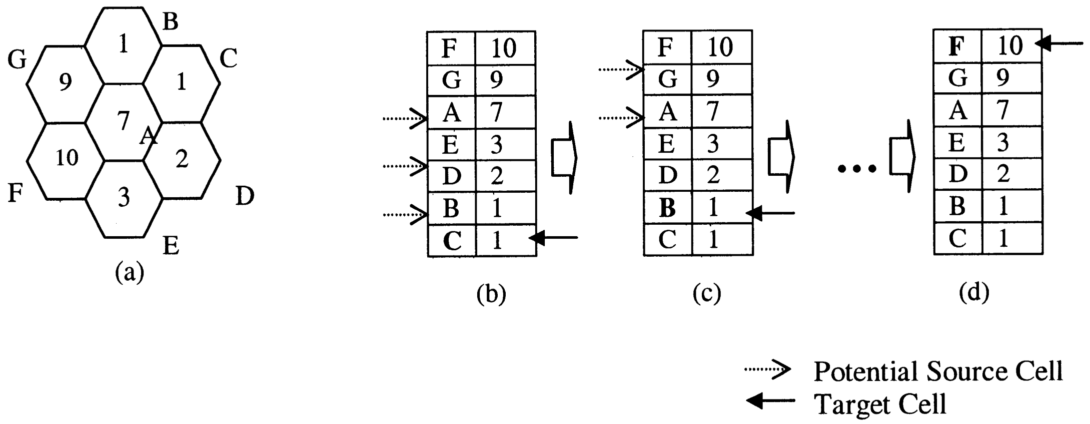


Figure 3.5: Load migration failure scenario. a) Seven cells with impossible load migration. b) 1st iteration. c) 2nd Iteration. d) Last iteration.

If load migration between the Source Cell and the Target Cell is a success, ALBA reviews the new load distribution of the network. If the Network Load Deviation (D_N) is still greater than D_{NThres}, ALBA selects another set of Target Cell, Source Cell and Source Node for load migration. These procedures continue until D_N is less than D_{NThres} or no further load migrations can be done, i.e. all cells are tried.

After the load migration planning is finished, ALBA sends signal through the RNC to the corresponding BSs to update their Channel Allocation Tables and Channels Pools. BSs send signals to the nodes on both old and new routes to update their Channel Tables and Routing Tables.

The Load Sampling Period (T_s) and Network Load Deviation Threshold (D_{NThres}) are two important parameters which affect the sensitivity of ALBA. If there are hotspots in the network, a small T_s and a small D_{NThres} should be used. Small T_s increases the frequency of load checking so that ALBA can react to unbalanced load situation promptly. Small D_{NThres} helps to obtain a higher degree of load balance in the network. This way, radio resources are better utilized and unnecessary call blockings can be avoided. If there are no hotspots in the networks, relaxing T_s and D_{NThres} reduces the frequency of load balancing operations.

The value of Neighboring Load Deviation (d_{nThres}) should be chosen such that the flipping of load migration will not occur. The minimum value of d_{nThres} should be greater than the load of one connection (load unit) and smaller than the load of two connections. For example, in Figure 3.6, if a Source Cell and a Target Cell have 52 and 51 load units, respectively, d_{nThres} should not be 1; otherwise, after load migration, the Target Cell has one load unit more than the number of load units of Source Cell. Then the Source Cell may be chosen as a Target Cell in the next iteration. Consequently, the load migration process between these two cells will continue forever.

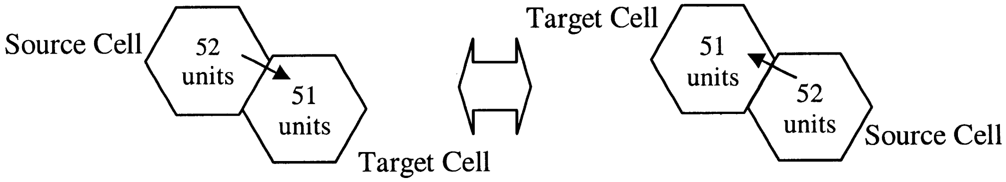


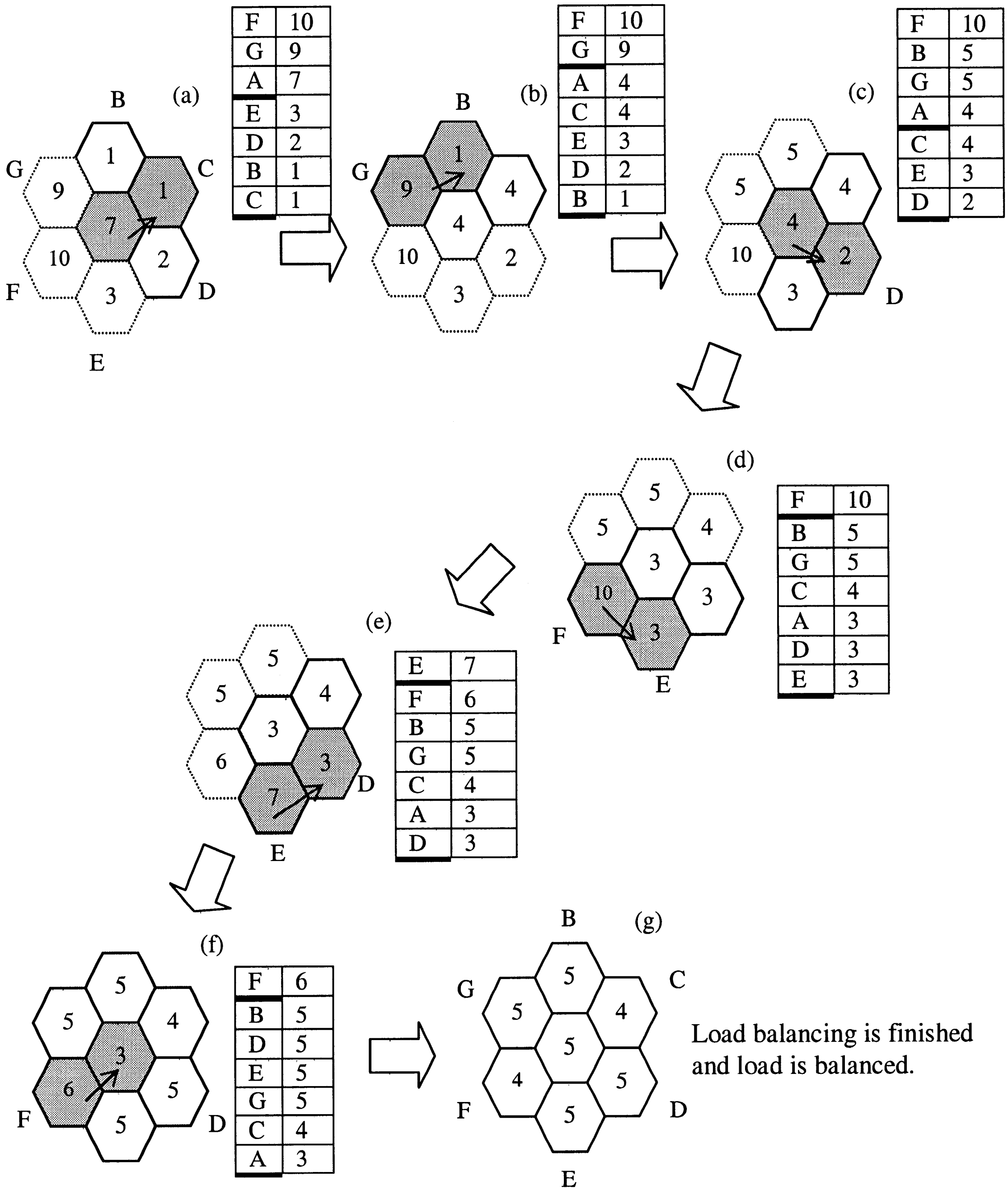
Figure 3.6: Load migration flipping situation

Illustration of ALBA scheme

We will describe two examples to illustrate the operations of ALBA for different network load situations. Example 1 is a heterogeneous traffic load environment based on seven cells (Cell A, B, C, D, E, F, and G) (see Figure 3.7). Example 2 is a hotspot environment based on a 3-tier cell system (see Figure 3.8). In these two examples, all cells are assumed to have enough source nodes and to be free to shift their load to and from their neighboring cells. Call Blocking Ratio (*CBR*) is assumed to be proportional to the load of a cell. That is, the higher load cell has a higher *CBR*. Network Load Deviation Threshold D_{NThres} and Neighboring Load Deviation Threshold d_{nThres} are set at 1.1 load unit. These two examples show that load can be balanced among all cells in a system.

Example 1

This example demonstrates how ALBA balances loads in a heterogeneous load environment. Some descriptions are followed with the line number in the ALBA scheme for easy reference. Figure 3.4 shows ALBA Scheme. In Figure 3.7a, since Network Load Deviation is greater than D_{NThres} which is 1 load unit, Load Balancing State is activated. Cell B and C are two least load (1 load unit) cells. Cell C is randomly chosen as a Target Cell (Line 2). Cell A, B, and D are neighboring cells of C. Cell A has the highest load (7 load units) (higher than the load of cell C) among these three cells and is chosen as the Source Cell (Line 3). Since Neighboring Load Deviation between cell A and cell C is greater than d_{nThres} and the CBR of cell A is greater than the CBR of cell B, load migration starts for these two cells. It is assumed that Source Nodes are available and Channeled Routes exists for all Source Nodes in cell A. ALBA selects the Source Node with the highest Migration Priority for load migration. Whether this migration is a success or not, it continues to select another Source Node for load migration until all Source Nodes are tried (Line 7). If cell A migrates 3 load units to cell C, both cells will have 4 load units. The new network load situation will now be reviewed (Line 10) (also see Figure 3.7b). Since the Network Load Deviation is still greater than the threshold D_{NThres} and potential Target Cells exist, load migration continues. Cell B has the least load and is chosen as a Target Cell. Cell G has the highest load among Cell B's neighbors and is chosen as a Source Cell. Previous iterations are



(a)

F	10
G	9
A	7
E	3
D	2
B	1
C	1

(b)

F	10
G	9
A	4
C	4
E	3
D	2
B	1

(c)

F	10
B	5
G	5
A	4
C	4
E	3
D	2

(d)

F	10
B	5
G	5
C	4
A	3
D	3
E	3

(e)

E	7
F	6
B	5
G	5
C	4
A	3
D	3

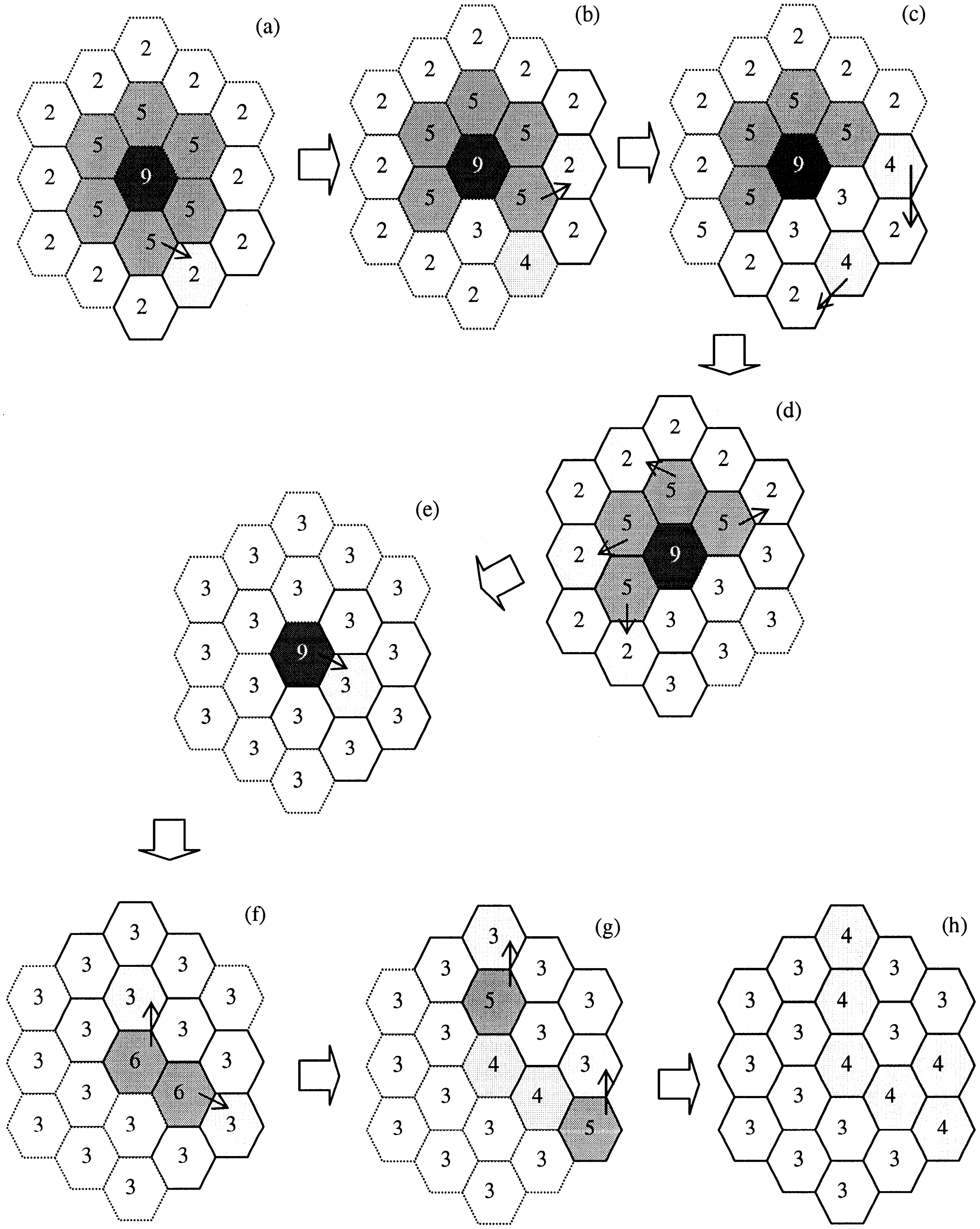
(f)

F	6
B	5
D	5
E	5
G	5
C	4
A	3

Load balancing is finished and load is balanced.

Figure 3.7: Load balancing of ALBA in heterogeneous load environment

repeated. Cell G migrates 4 load units to cell B. Iterations continue until the Network Load Deviation is not greater the threshold D_{NThres} or no load can be shifted (Line 10). Figure 3.7g represents the balanced load situation.



Load balancing is finished and load is balanced.

Figure 3.8: Load balancing of ALBA in hotspot environment

Example 2

This example demonstrates how ALBA balances loads in a hotspot environment. In Figure 3.8a, a 3-tier cell environment is shown. The center cell is a hot cell with 9 load units. Each second tier (medium hot) cell has 5 load units. Each third tier (cool) cell has 2 load units. Traffic is free to be relayed among neighboring cells. The same iterative procedures as in the previous example are used. First, a cool cell (boundary cell with 2 load units) is chosen as a Target Cell. The neighbor with the highest load (5 load units) is chosen as a Source Cell. Load is migrated from the Source Cell to the Target Cell. The same iterations are performed for the other pairs of cells. After several iterations, the load of medium hot cells and cool cells is balanced (see Figure 3.8e). The load of the hot cell is then migrated to its neighboring cells. After several iterations, the load of this system is balanced (see Figure 3.8h).

3.2.2 A-Cell Adaptive Routing (ACAR) Scheme

ACAR is a centralized on-demand load-aware routing scheme specifically designed for 3G cellular systems. ACAR is located at the RNC. Similar to existing MANET routing protocols [17, 18, 20, 24], it has two mechanisms: Routing Discovery and Route Maintenance. Figure 3.10 shows the ACAR Scheme. Figure 3.11 shows the table and message Format, conditions, and assumptions for the scheme. Before we describe the scheme, in addition to the terminologies, **Load(L)**, **Source Cell (C_{Src})**, **Target Cell (C_{Trg})**, **Source Node (Src)**, **Relaying Node (Nd_{relay})**, **Time Slot**, **Code**, **Channel,** and **Channeled Route,** defined in the previous section 3.2.1, we define some other terminologies below.

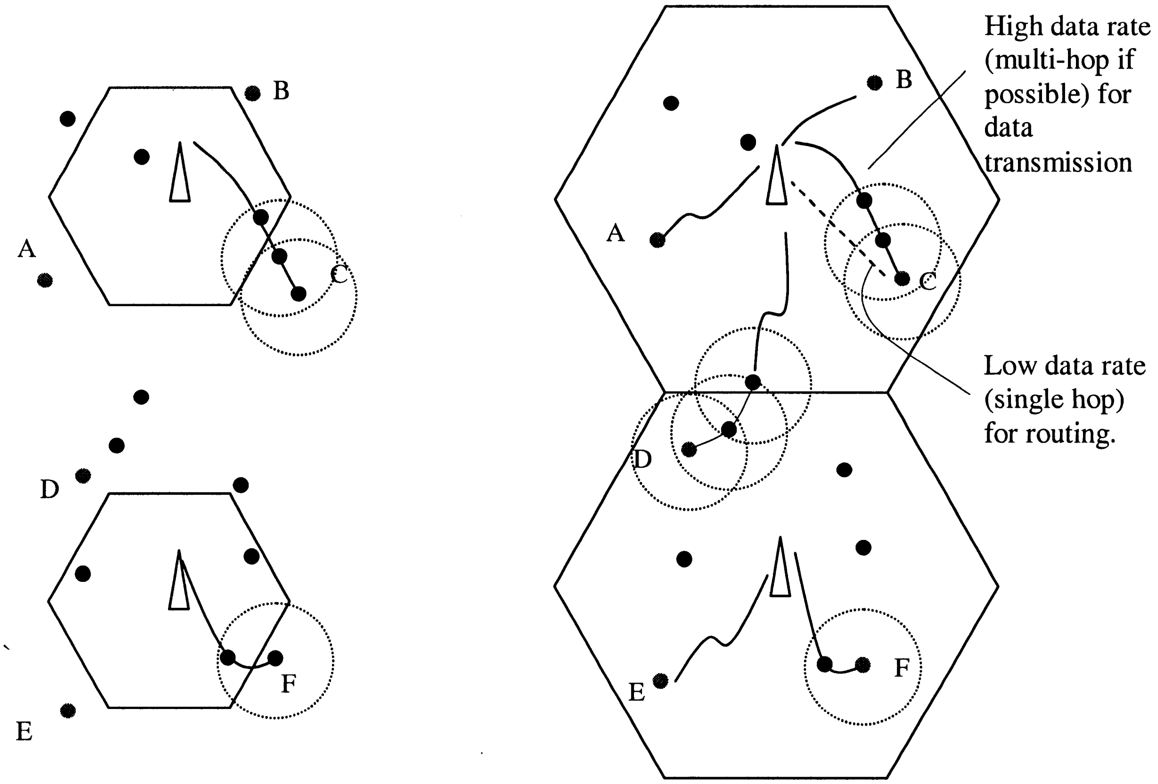
Definitions:

1. **Route** is a path from a Source Node to a destination. Destination can be a mobile node or BS.
2. **Last Hop Node** (Nd_{last}) is the nearest (last hop) node to the BS on a route. It can be a Source Node or Relaying Node.
3. **Second Last Hop Node** ($Nd_{2ndlast}$) is the node which is second nearest to the BS on a route.

4. **Reliable Relaying Node** ($Nd_{reliable}$) is a Relaying Node which is considered reliable for relaying purpose.
5. **Normal Transmission Range** (R_{normal}) is the transmission range for normal operation of a cell.
6. **Relaying Transmission Range** (R_{relay}) is the transmission range for relaying purpose. Last Hop Node is not restricted to this range.
7. **State Information** (*SI*) is the status information of a mobile node including location, speed, remaining battery, and load.
8. **State Information Request** (***SREQ***) is a message for requesting State Information.
9. **State Information Reply** (***SREP***) is a reply message containing State Information.
10. **Maximum Hop Count** (Hop_{max}) is the maximum number of hops that is allowed for a route.
11. **Intra-cell Routing Zone** is a region containing all possible routes from a Source Node to the BS in same cell.
12. **Inter-cell Routing Zone** is a region containing all possible routes from a Source Node to the BS of Target Cell.
13. **Separation Threshold** (Sep_{Thres}) is the distance between $Nd_{2ndlast}$ and Nd_{last} below which the $Nd_{2ndlast}$ replaces the Nd_{last}.
14. **Inter-cell Separation Threshold** (**InterSep**$_{Thres}$) is the distance between Nd_{last} and Target Cell BS above which the route should not be chosen for relaying.
15. **Energy Consumption** (E_0) is the current energy consumption of a connection based on path loss.
16. **Current Load** (L_0) is the current load of a node.
17. **Battery** (B_0) is the current remaining battery of a node.
18. **Speed** (S_0) is the current speed of a node.
19. **Energy Consumption Threshold** (E_{Thres}) is the energy consumption above which a node is considered as unreliable for relaying.

20. **Load Threshold** (***L_{Thres}***) is the load above which node is considered as unreliable for relaying.
21. **Remaining Battery Threshold** (***B_{Thres}***) is the remaining battery below which a node is considered as unreliable for relaying.
22. **Speed Threshold** (***S_{Thres}***) is the speed above which a node is considered as unreliable for relaying.
23. **Route Broken** (***RBROKE***) is a message for showing route breakage.
24. **Node Unstable** (***UNSTABLE***) is a message for showing that a node is unstable. Reasons of instability can be that the remaining battery level is below B_{Thres}, the speed exceeding S_{Thres}, or the load above L_{Thres}.
25. **Route Request** (***RREQ***) is a route request from a party, e.g. ALBA and Call Admission Control (CAC).
26. ***Route Reply (*RREP**) is a route reply from ACAR.
27. **Neighboring List** is a list of a mobile node storing the addresses of neighboring nodes.
28. **Connection id** is the identification number of a connection.

As we mentioned in section 2.2, the method of gaining cell capacity in a conventional multi-hop cellular network is to increase the separation among cells so that more channels (frequencies) can be reused. Methods to increase the separation include shrinkage of a cell or removing adjacent BSs. In a 3G system, although capacity gain is also achieved by shrinkage of a cell, the benefit of this shrinkage is to reduce transmission power and, thus, reduce interference. As interference reduces, the capacity increases. Therefore, cell size in 3G systems does not necessarily need to be small to increase frequency reuse. Making use of this cell size flexibility, i.e. the cell breathing [31] effect, route discovery and route maintenance within the maximum coverage limit of a cell can be done in a single hop. Figure 3.9 shows the comparison of possible connections in MCN-p and ACAR. Low data rates with large coverage can be used to communicate topology and control information directly between a mobile node and a BS. This greatly reduces route discovery latency, resource consumption, and routing overhead. High data rates with short transmission range and multi-hopping are used for data communications. If a multi-hop route is unavailable, the Source Node communicates directly with the BS in a single hop. Therefore, no potential call is denied. The flexibility and the infrastructure of 3G W-CDMA systems can be fully utilized. This idea is especially useful in highly loaded cells with sparse node densities.



High data rate (multi-hop if possible) for data transmission

Low data rate (single hop) for routing.

a) MCN-p

b) ACAR

Figure 3.9: Comparison of the routing behavior in MCN-p and ACAR

In ACAR, if a mobile node is within the normal coverage of a cell (BS), control messages or topology information can be communicated directly between the node and the BS. Multi-hopping for route discovery or route maintenance is not required in this region. Since ACAR uses an on-demand strategy, periodic hello messages or updates are not required. If MTs are outside the coverage of a certain cell, a MANET routing protocol such as AODV or DSR is assumed. Neighboring List is an optional feature to ensure the connectivity among MTs in a region in which communication barriers among mobile nodes exist. A node that can communicate with the BS may not be able to communicate with its neighbors because of these barriers. This neighboring list is also obtained on-demand. Global Positioning System (GPS) is employed to facilitate route discovery and route maintenance.

ACAR facilitates both intra-cell and inter-cell routing. Intra-cell routing is triggered when ACAR receives a Route Request (*RREQ*) from a Call Admission Control (CAC) [14] mechanism for a new call arrival. CAC is located at the RNC to control the admission of a call based on the capacity of a cell. Intra-cell routing is used for setting up a route from a Source Node to the BS of the cell of the Source Node. Inter-cell routing is triggered when ACAR receives a *RREQ* from ALBA for the purpose of load balancing. Inter-cell routing is used for setting up a route from a Source Node of a Source Cell to the BS of the Target Cell. Information of Source Cell, Target Cell, and Source Node are provided by ALBA.

Figure 3.10 shows the Route Discovery steps of ACAR which consists of six phases: Information Collection, Selection, Route Computation, Path Refinement, Channel Allocation, and Route Reply.

1) Information Collection Phase (Line 2 to Line 5)

The objective of this phase is to collect topology information and State Information (SI) of MTs in a region of interest. Sending a State Information Request (*SREQ*) to the neighboring cells is optional because there may be a trade-off between the higher chance for finding a Channeled Route for the Source Node and the higher complexity in managing Relaying Nodes in different cells. This trade-off will be part of our future work. When ACAR receives a *RREQ* with the information of the Source Node and the BS of the Source Node (Line1) from CAC, it sends a *SREQ* to the BS and (optional) the neighboring cells of the BS. When ACAR receives *RREQ* from ALBA, it sends a *SREQ* to all BSs in the networks. When ACAR receives Route Broken (*RBROKEN)* or Node Unstable (*UNSTABLE*) message from the RNC for Route Maintenance, it sends *SREQ* to the corresponding BSs related with the route. The BSs send the *SREQ* to all relaying nodes within their coverage. The Relaying Nodes send their State Information Reply (*SREP*) to their BSs (Line 4). The BSs pass the information to ACAR for processing.

2) Selection Phase (Line 6 to Line 7)

The objective of this phase is to select Reliable Relaying Nodes in the region of interest (Line 6). Relaying Nodes are chosen based on their State Information and predefined thresholds. The thresholds are Energy Consumption Threshold (E_{Thres}), Remaining Battery Threshold (B_{Thres}), Speed Threshold (S_{Thres}), and Load Threshold (L_{Thres}). $f(E_O, B_O, S_O, L_O)$ is a function to decide whether a Relaying Node is reliable or not. A node is chosen as a reliable node if its energy consumption, speed, and load does not exceed E_{Thres}, S_{Thres}, and L_{Thres}, respectively, and its remaining battery is above B_{Thres}.

3) Route Computation Phase (Line 8 to Line 15)

The objective of this phase is to compute the route for relaying. Reliable Relaying Nodes in the routing zone of interest are selected for computation. For call arrival, the Intra-cell Routing Zone is used. For load balancing, the Inter-cell Routing Zone is used.

If no Reliable Relaying Nodes exist, the single hop route is the default route for new call arrival. If Reliable Relaying Nodes exist, the RNC computes the distance of every node, including the Source Node and Relaying Nodes, to the BS of the Source Node. Source Node and Relaying Nodes are assumed to use Relaying Transmission Range (R_{relay}) while Last Hop Node (Nd_{last}) uses Normal transmission Range (R_{normal}). Only the nearest node to the BS communicates directly with the BS. Based on the distance, Relaying Transmission Range (R_{relay}) and Neighboring Lists (optional), the connectivity among nodes can be obtained (Line 9). The connectivity among nodes is used to find a shortest path [32] from the Source Node to the BS of the Target Cell.

For load balancing, there is an additional condition. If the distance between Nd_{last} of a path and the BS of the Target Cell is greater than the Inter-cell Separation Threshold ($InterSep_{Thres}$), the route is rejected (Line 12). This condition is used to avoid using a route in which Nd_{last} is too far away from the Target Cell BS. This requires higher transmission powers, which results in higher interference and decreases cell capacity (data rates).

If there are barriers that affect the connectivity among nodes in the routing zone, the Neighboring List of each node should be obtained to ensure the connectivity among neighboring nodes. To obtain a neighboring list, each node broadcasts their node address to their neighbors using R_{relay}. Neighboring List is included in the SREP.

If a shortest path is found, ACAR proceeds to the Path Refinement Sub-phase. If the path is not found, ACAR takes out the Nd_{last} and computes the shortest path again until a path is found or all Reliable Relaying Nodes are taken out.

4) Path Refinement (Line 17 to Line 20)

The objective of this phase is to refine the computed path . If both the Nd_{last} and the Second Last Hop node ($Nd_{2ndlast}$) can reach the BS using R_{relay}, the $Nd_{last}$is replaced by the $Nd_{2ndlast}$ to avoid unnecessary delay. This is based on the assumption that R_{relay} is a near-optimal distance such that further reduction of the distance will have little improvement in capacity (data rate). If the number of hops of the path is greater than the Maximum Hop Count, excessive nodes are taken away (Line 18). If the separation between the Nd_{last} and the $Nd_{2ndlast}$ is closer than Separation Threshold (Sep_{Thres}), the Nd_{last} is replaced by the $Nd_{2ndlast}$ (Line 20). After the Path refinement phase is finished, the path is considered as a potential route.

5) Channel Allocation (Line 21 to Line 24)

The objective of this phase is to assign Channel (Time Slot Code pair) to each node on the route. ACAR calls E-DSSA to do this task (Line 21). If each node on the path can be assigned a Channel, Channel allocation is a success and this path is called a Channeled Route. If any node on the route cannot be assigned a channel, this route is invalid and Channel allocation fails. Accordingly, nodes which are not able to assign a channel to their successors are taken out and the Route Computation Phase is repeated.

6) Route Reply Phase (Line 25 to Line 27)

ACAR sends *RREP* to the corresponding Parties, either CAC or ALBA. For load balancing, ACAR replies ALBA with a channel allocation success or failure signal. In both cases, ACAR updates the Load L_0 and energy consumption (E_0) of the corresponding Relaying Nodes, preparing for the next iteration for the *RREQ* from ALBA. ALBA may send ACAR another Source Node or another set of Source Node, Target Cell and Source Cell for route finding.

ACAR Scheme

A) Route Discovery

```
ACAR received RREQ (with Src and BS) from a party (CAC, ALBA or RNC).

/* Information Collection Phase */
ACAR sends SREQ to the BS and related BS(s) through RNC.
The BS(s) broadcasts the SREQ to all Relay Node ($Nd_{relay}$) within their coverage.
Corresponding $Nd_{relay}$ replies to the BS(s) with SREP.
BS(s) pass the SREPs to ACAR through RNC for processing.

/* Selection Phase */
Select $Nd_{reliable}$ based on $E_{Thres}$, $B_{Thres}$, $S_{Thres}$, $L_{Thres}$ in Routing Zone.
    $f(E_O, B_O, S_O, L_O)$ = reliable   | if $E_O < E_{Thres}$ and $B_O < B_{Thres}$ and $S_O < S_{Thres}$ and $L_O < L_{Thres}$
                          = unreliable | otherwise.

/* Route Computation Phase */
Do
    Do compute connectivity among the Src and $Nd_{reliable}$ based on $R_{relay}$
    and Neighboring List (optional).
        Choose the closest $Nd_{reliable}$ to the Target Cell BS as direct link to the BS.
        Find the shortest Path from the Src to the BS.
        If load balancing and distance between $Nd_{last}$ and the BS > $InterSep_{Thres}$,
            Path is NOT found.
            Take out the $Nd_{last}$ (excluding Src).
    While (Path is NOT found and there are still $Nd_{reliable}$ (s))
    If Path is found

        /* Path Refinement Sub-phase */
        If both $Nd_{last}$ and $Nd_{2ndlast}$ can reach BS through $R_{relay}$, $Nd_{2ndlast}$ replaces $Nd_{last}$.
        If number of hops > $Hop_{max}$, take out excess nodes starting from $Nd_{last}$.
        If separation between $Nd_{last}$ and $Nd_{2ndlast}$ < $Sep_{Thres}$ and both nodes are in the BS
            $Nd_{2ndlast}$ replaces $Nd_{last}$.

        /* Channel Allocation Sub-phase */
        Perform E-DSSA to find a Channel for each node on the Path.
        If Channel allocation is NOT success for any of the nodes,
            Take out $Nd_{reliable}$ (s) which cannot assign channels to its successive node.
While (Channeled allocation fails and there are still $Nd_{reliable}$ (s))

/* Route Reply Phase */
ACAR sends RREP (Success with Channeled Route information or Failure) to the party.
If the party is ALBA,
    update $L_O$ and $E_O$ of corresponding $Nd_{reliable}$ for next ALBA's RREQ.
```

Remarks:

1. State Information of non-covered Nd_{relay} is sent through the covered Nd_{relay}.
2. The party is responsible for sending signals to BS(s) to update their Channel Assignment Table(s) and Channel Pool(s) through RNC. BS(s) send signals to update the Channel Table(s) and Routing Table(s) of corresponding nodes.

(Continue in next page)

Figure 3.10: ACAR Scheme

(Continue from previous page)

B) Route Maintenance

1. If *Route Broken (*RBROKEN*) or † Node Unstable (*UNSTABLE*),
2. Corresponding relaying nodes sends *RBROKE* or *UNSTABLE* message to its BS.
3. BS sends signal to ACAR through RNC to reinitiate Route Discovery process for the route.

* Route Broken occurs when a Nd_{relay} does not receive packets from its previous node within a period of time (T_{out}). Possible reasons are out of range, communication barriers, or power outage.

† Node Unstable occurs when a Nd_{relay}'s B_0, S_0, E_0, or L_0 reaches the corresponding thresholds (B_{Thres}, S_{Thres}, E_{Thres}, L_{Thres}).

Figure 3.10: ACAR Scheme (Continued)

Table and Message Format:

1. Routing Table - connection id, next hop address
2. State Information (*SI*) - node address, location, speed, remaining battery, load, energy consumption, and neighboring list (optional).
3. *RREQ* Information (I_{RREQ}) – (*Src id*, *BS id*).
4. *RREP* Information (I_{RREP}) – (*Src id*, *Channel*, Nd_{relay} *1 id*, *Channel*, Nd_{relay} *2 id*, *Channel*,…)
5. Message – node address, event type (*RREQ*, *RREP*, *SREQ*, *SREP*, *RBROKE*, or *UNSTABLE*), content (*SI*, I_{RREQ}, or I_{RREP}).

Conditions:

1. Only Last Hop Node communicates with the BS directly.
2. All nodes, except Last Hop Node, use Relaying Transmission Range for communication.
3. A Source Node can also be Last Hop Node if there is no relaying nodes are available (Single hop situation).

Assumption:

Relaying Transmission Range is a near-optimal range such that the increase in a node or cell capacity does not increase significantly even a node is closer to its receiving node.

Figure 3.11: Table and Message Format, conditions, and assumptions for ACAR scheme

Illustration of ACAR scheme

Intra-cell routing and Out-of-Coverage routing are performed upon a new Call Arrival. There are three cases: Single hop, Multi-hop, and Out-of-coverage cases. Figure 3.12 illustrates these cases. Detailed explanations are as below.

Single hop (direct communication case) –

When node A places a call request, the CAC in RNC sends a *RREQ* to ACAR. ACAR sends a *SREQ* to the corresponding BS. The BS sends *SREQ*s to all relaying nodes (B, C, D, E, F, G, I, and J) in the cell. All Relaying Nodes send State Information Reply (*SREP*) through the BS to ACAR. ACAR selects Reliable Relaying Nodes. Assume that node C is the only unreliable node because of its low battery capacity or high speed. The Reliable Nodes are B, D, E, F, G, I, and J. Node D is selected as Last Hop Node. After first iteration, no route is found because A cannot reach node D using Relaying Transmission Range (R_{relay}). Node D is taken out and route computation is repeated. Since no reliable nodes can be reached by A using R_{relay}, A communicates directly to the BS.

Multi-hop case -

For Source Node H, the shortest path is H-F-E-D-BS. Assume both D and E can reach BS using $\mathrm{R_{relay}}$, node D is eliminated during the Route Refinement Phase. The final path is H-F-E-BS. ACAR calls E-DSSA to assign Channel to each node on the path.

Out-of-Reach case -

Assume node M places a call request. Node M has found a path to reach node J using MANET routing protocol. Since the nodes (K, L and M) are Out-of-Coverage, their Channel Tables must be sent to the BS through multi-hopping to J. Node J relays their information to the BS. The BS passes the information to ACAR for processing.

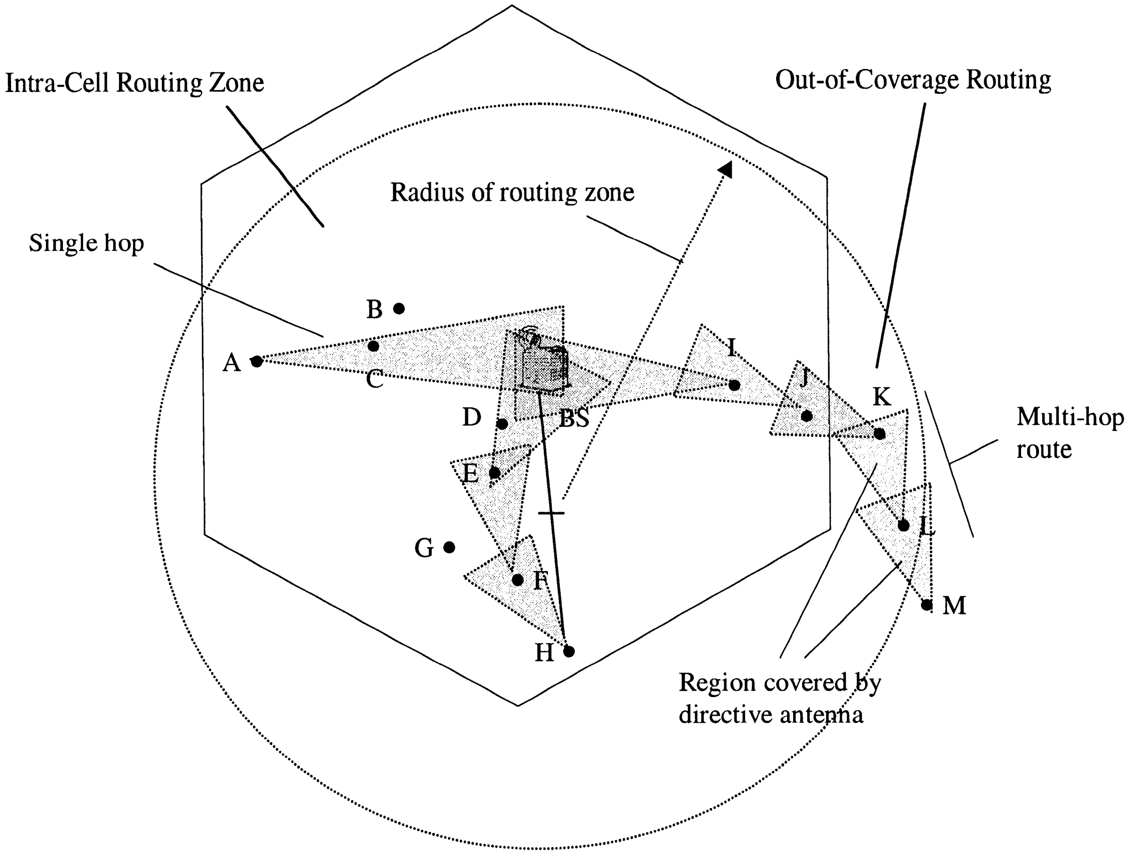


Figure 3.12: Intra-cell routing and Out-of-Coverage routing of ACAR

For load balancing, inter-cell routing is used. Figure 3.13 illustrates this situation. Assume node H is selected as a Source Node for load migration. BS_{Trg} is the BS of the Target Cell. ACAR gets the State Information of all Relaying Nodes in the Source Cell, the Target Cell and their neighboring cells (optional). ACAR computes the shortest path from H to BS_{Trg}. Assume that a route is found, ACAR then calls E-DSSA to allocate a Channel to each node on the route. Assume channel assignment is a success, then the route (H-N-O-P-Q-R-BS_{Trg}) is a Channeled Route. ACAR replies to ALBA with the channeled Route. ALBA sends signals to the BS_{Src} and the BS_{Trg}, and their neighboring BSs (optional) through RNC to update their Channel Assignment Table and Channel pools. The BSs send signals to the Source nodes and the Relaying Nodes on old route (H-F-E- BS_{Src}) and new route (N-M-O-P-Q- BS_{Trg}) to update their Channel Tables and Routing Tables.

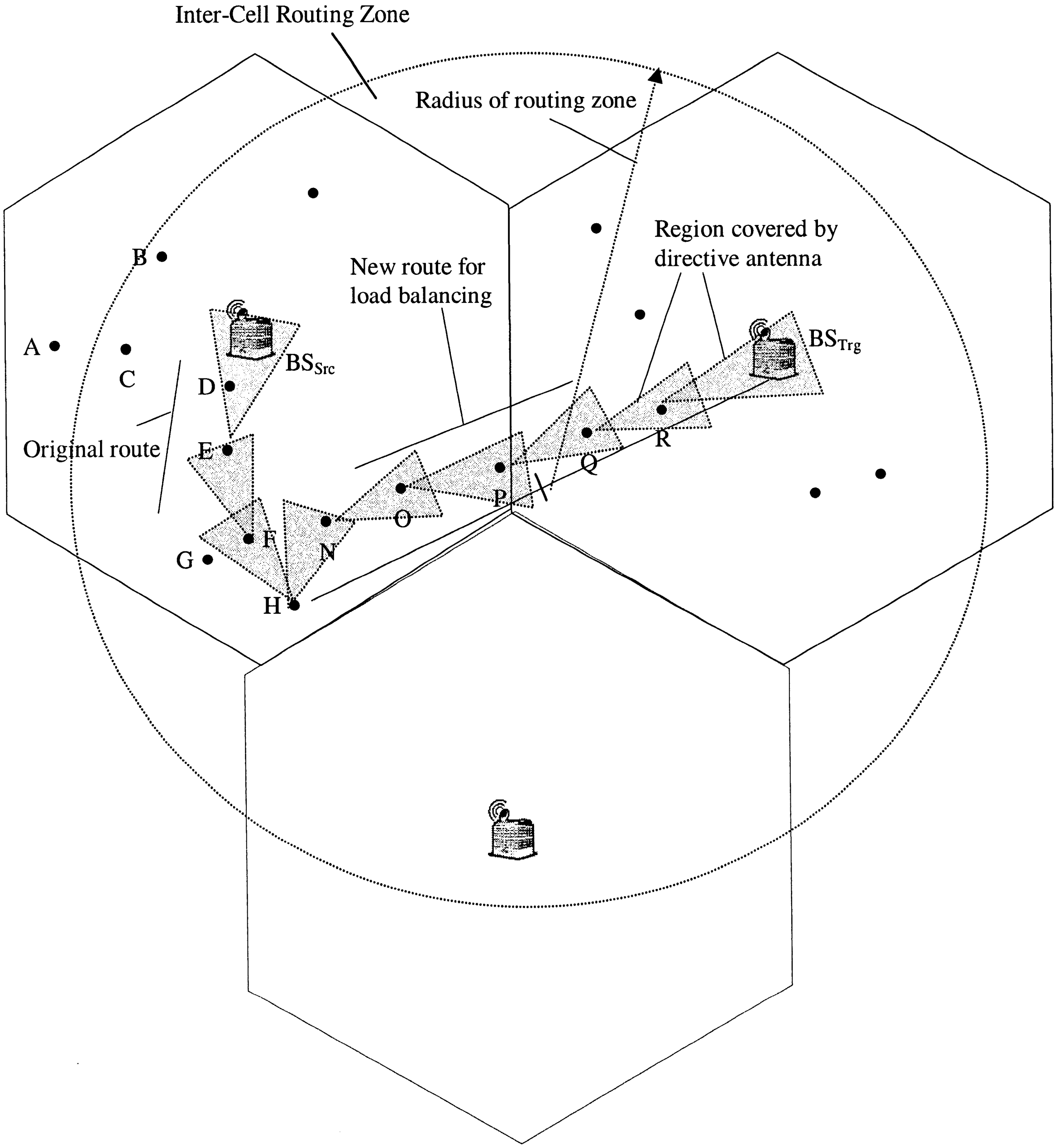


Figure 3.13 Inter-cell routing of ACAR

3.2.3 Extended-Delay Sensitive Slot Assignment (E-DSSA) Scheme

E-DSSA is a slot assignment scheme which is inherited from Delay Sensitive Slot Assignment (DSSA) [1]. E-DSSA has all features of DSSA such as employing directive antennas and Global Positioning System (GPS), except that constraints are added for consecutive channels and Last Hop Node interference. The consecutive channels constraint is used to avoid signal interference as mentioned earlier in section 2.2.3. The idea is that a mobile should not send and receive data on the same channel at the same time. The Last Hop Node Interference constraint is used to avoid assigning a Channel to a node (Current Node) such that the Last Hop Nodes on other routes is transmitting on that Channel towards the next hop node of the Current Node. This may happen because the Last Hop Node may transmit by the Normal Transmission Range (R_{normal}) instead of Relaying Transmission Range (R_{relay}). Figure 3.14 illustrates the E-DSSA scheme. Before we describe the scheme, we define some terminologies as follows. The terms **Time Slot, Code, Channel, Channel Pool, Channel Table,** and **Channel Allocation Table** are as defined in section 3.2.1. The terms **Last Hop Node** (Nd_{last}), **Normal Transmission Range** (R_{normal}), and **Relaying Transmission Range** (R_{relay}) are as defined in section 3.2.

Definition:

1. **Current Node** (Nd_{curr}) is the mobile node to which a channel is going to be assigned.
2. **Previous Node** (Nd_{prev}) is the mobile node that is the next hop of the current node.

In Figure 3.14, when E-DSSA receives a channel allocation request with a Path (Route) from ACAR, it starts assigning a Channel to the node (Current Node) starting from the Last Hop Node to the BS (Line 2). E-DSSA selects an available Channel (TS_m, C_n) from the Channel Pool of the BS. If no node in the line of sight from the Last Hop Node (Current Node) to the BS is receiving on this Channel, Channel assignment for the Current Node is a success; otherwise, E-DSSA continues the search for another available Channel for the Current Node until a channel is found or all the available channels are tried (Line 7).

If the Last Hop Node is assigned a channel, E-DSSA starts to find a Channel for the Second Last Hop Node (Line 9). Please note that the Last Hop Node becomes the Previous Node of the Second Last Hop Node. The Second Last Hop Node becomes the Current Node. A channel is selected with a time slot which is successive and closest to the time slot of the channel previously assigned to the Last Hop Node (Previous Node). In this case, a Channel with TS_{m-1} is selected. The objective of this strategy is to reduce the Time Slot waiting time, i.e. the delay, of a packet. The Channel has to satisfy the following conditions before it can be accepted.

1. The Current Node itself is not receiving on this Channel. This condition applies when a mobile terminal has the ability to send and receive signals at the same time.
2. The Current Node itself is not receiving on the same Time Slot as the Time Slot of this Channel. This condition applies when a mobile terminal cannot send and receive signal at the same time.
3. The neighbors of the Previous Node are not transmitting on this Channel.
4. The Last Hop Nodes on the other routes are not transmitting on this Channel towards the Previous Node.
5. The neighbors of the Current Node are not receiving on this Channel.

If all the conditions above are satisfied, the Channel is assigned to the Current Node. If the Channel assignment is not a success because condition (1) is not satisfied, E-DSSA finds another Channel with the same Time Slot of the previous Channel, but in different Code. This process continues until a channel is found or all the available Codes are tried (Line 16).

E-DSSA Scheme

A) Channel Allocation

```
Input a Path (Route) from ACAR
Assign Channel starting from the Last Hop Node (Current Node ($Nd_{curr}$)) to the BS
Do
    Do select a non-tried available Channel ($TS_m$ $C_n$) from Channel Pool of BS
    and the capacity of this slot has not been used up.
    (Capacity can be used up before the available Codes are exhausted)
        If $Nd_{curr}$ is not receiving on this * Channel (or **Time Slot) for other route
        † and No node on the LOS of this $Nd_{curr}$ to the BS receives on this Channel,
            Channel assignment of $Nd_{curr}$ is success.
    While (Channel assignment fails and there are still non-tried available Channels).

    If Channel assignment of $Nd_{curr}$ is success,
        Do select successive node (new $Nd_{curr}$) (previous $Nd_{curr}$ becomes $Nd_{prev}$)
            Do select an available Channel with the NEXT highest Time Slot
            and the capacity of this time slot has not been used up.
                Do
                  If $Nd_{curr}$ is not receiving on this *Channel or **Time Slot for other
                  route,
                    and neighbors of the $Nd_{prev}$ are not transmitting on this Channel,
                  † and $Nd_{last}$ are not transmitting on this Channel towards the $Nd_{prev}$,
                    and neighbors of the $Nd_{curr}$ are not receiving on this Channel,
                        Channel assignment of $Nd_{curr}$ is success.
                    Else
                        Select an available Code that has not been chosen.
                While (Channel assignment fails and there are still non-tried Codes)
            While (Channel assignment fails and there are still non-tried Channel)

        While (Channel assignment is a success and there are still successive nodes)

While (Channel assignment of $Nd_{curr}$ fails and there are still non-tried Channels in
        Channel Pool of BS)
Reply ACAR with Channel assignment result (success or failure).
```

* - Consecutive channel constraint (if MT can send and receive signal simultaneously)
** - Time Slot constraint (if MT cannot send and receive signal simultaneously)
† - Last Hop Node Interference constraint

B) Channel De-allocation

1. Input a Route
2. De-allocate the Channels of each node on the Route.

Table Format:

1. Channel Allocation Table - (Relaying Node address, Channel, Source Node address).
2. Channel Table – (Channel, Source Node address).

Figure 3.14: E-DSSA Scheme

If Channel assignment still fails, E-DSSA tries another available Channel with the next highest Time Slot until a channel is found or all the available Channels are tried (Line 17).

If Channel assignment is a success for the Current Node, E-DSSA starts to find a Channel for the next successive node on the route (Line 18), and so on.

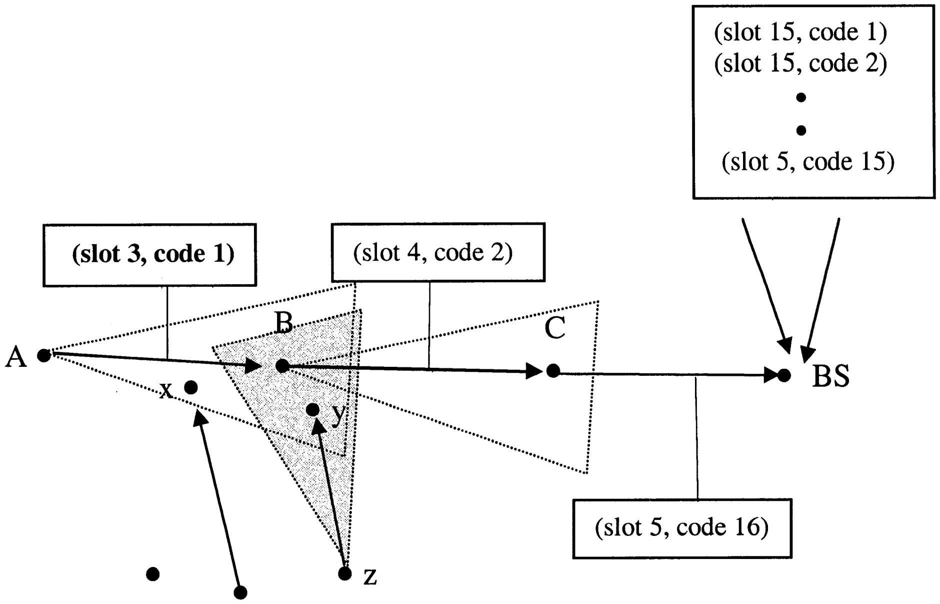
If any node on the route cannot be assigned a Channel, the Channel allocation steps starting from the Last Hop Node is repeated until channel assignment for the whole route is a success or all the available Channels of Channel Pool of BS are tried.

Illustration of E-DSSA scheme

The following is an example to illustrate the operation of E-DSSA. Figures 3.15 shows how E-DSSA handles Channel conflict, consecutive channel conflict, and Last Hop Node interference situations respectively.

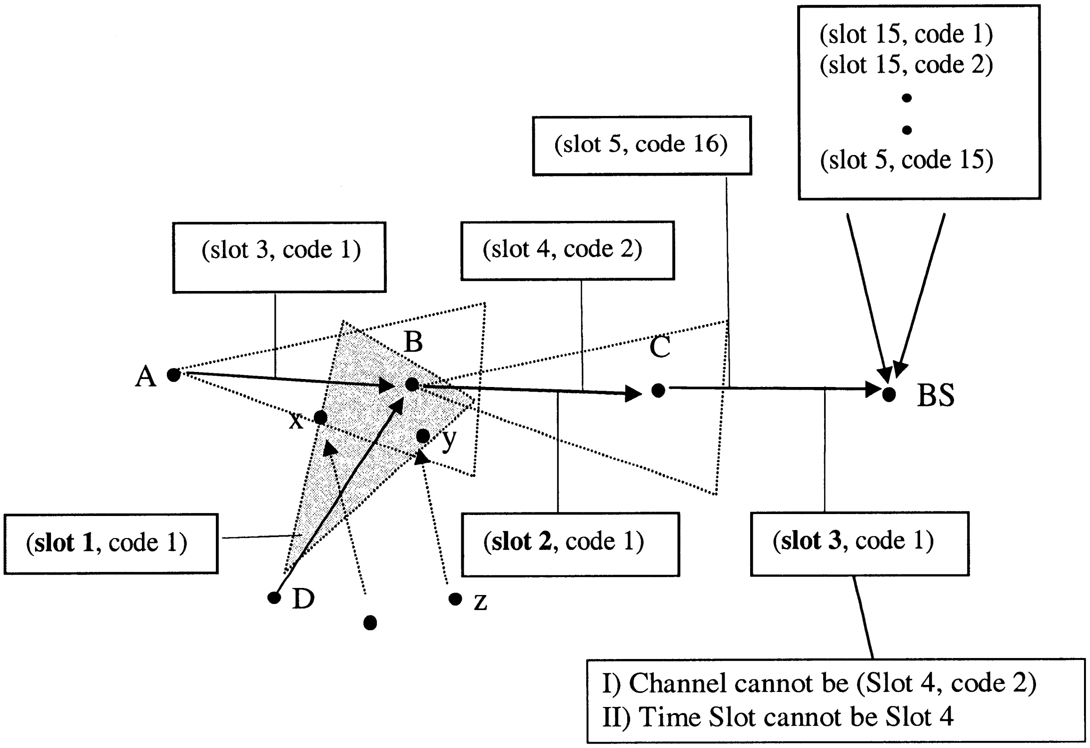
Assume that the capacity of a cell is used up when codes are used up. In Figure 3.15a, assume that all channels of the slots (from slot 15 to slot 5) are used up except code 16 of slot 5. Channel (slot 5, code 16) and Channel (slot 4, code 1) are successfully assigned to node B and C respectively. Node A is proposed to be assigned a Channel (slot 3, code 1) for minimizing the delay of packet. Node x is neighbor of A. Node z is a neighbor of B. Assume that node x is not receiving on this Channel and node z is not transmitting on this Channel. No Channel conflict (interference) occurs. The Channel (slot 3, code 1) can be assigned to node A.

In Figure 3.15b, when a new call request from node D arrives, path D-B-C-BS is the shortest path. In the Channel Pool of BS, Channel (slot 4, code 2) is available to be assigned to node C to serve the call request. There is no Channel conflict because no neighbors of BS are transmitting on this Channel and no neighbors of C are receiving on this channel. However, since C is receiving on this Channel, consecutive channel conflict might occur. This channel can not be assigned to C. Then, Channel (slot 3, code 1) of next highest Time Slot is chosen. No Channel conflict or consecutive channel conflict occurs. This Channel is assigned to node C. Channel (slot 2, code 1) causes no Channel conflict or consecutive channel conflict. This Channel is assigned to node B to minimize delay. Channel (slot 1, code1) is assigned to node D.

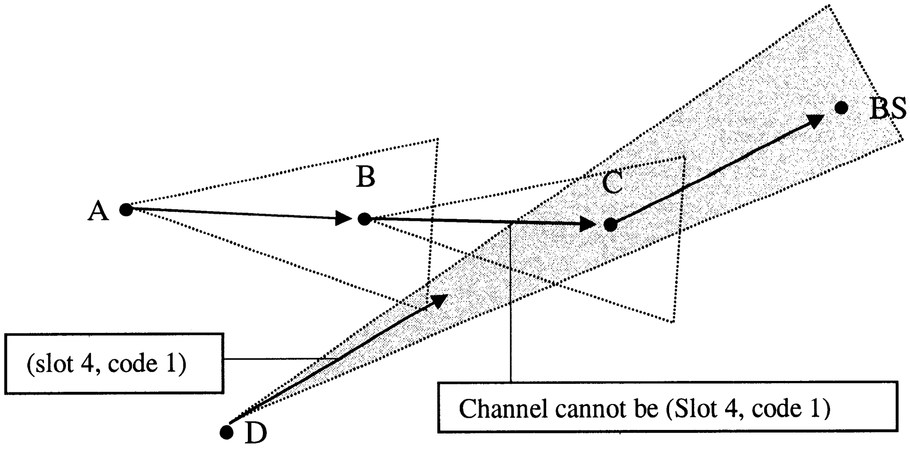


a) Avoiding Channel conflict

In Figure 3.15c, when a new call request from node E arrives, node E itself is the Last Hop Node that can transmit its signal within R_{normal}. If a channel (slot 4, code 1) is proposed to assign to node D, it has to ensure that node C is not using this channel for reception; otherwise, a channel conflict at node C occurs. This situation also applies when node B is proposed to assign a channel and node E has already been assigned a channel. The candidate channel for node B should not be the same as the Channel that are used by node D.



b) Avoiding (I) consecutive channel conflict if a MT can send and receive signal simultaneously or avoiding (II) Time Slot conflict if a MT cannot send and receive signal simultaneously.



c) Avoiding Last Hop Node Interference.

Figure 3.15: Three scenarios of E-DSSA with directional antenna. a) Avoiding Channel conflict, b) Avoiding consecutive channel or Time Slot conflict, c) Avoiding Last Hop Node Interference.

3.3 Limitations

This framework is centralized and is executed by the RNC. However, the major limitation of this framework is that if there is no relaying node, the effect of load balancing will be greatly diminished. In this case, only the flexible users at the overlapping region among cells can shift their load for load balancing. Another limitation is that our framework is designed for uplink communication only. Implementation for downlink communication requires further investigation. In addition, since this framework employs directive antennas and uses shorter transmission distances, the interference and power control issues also require further study.

3.4 Summary

In this chapter, we designed a novel load balancing and relaying framework, namely A-cell Load Balancing Relaying Framework (ALBAR) for 3G TDD W-CDMA multi-hop cellular systems. Within this framework, we designed a load balancing scheme (ALBA), a routing scheme (ACAR), and a slot assignment scheme (E-DSSA). All three components are executed by the RNC. ALBA is a centralized, dynamic load balancing scheme. It can handle hotspots as well as heterogeneous load environments. It calls ACAR to find Channeled Routes for facilitating load migration. First, ACAR computes the shortest path and calls E-DSSA to find channels for nodes on the shortest path that it found. The iterations of load migration continue until the Network Load Deviation (D_N) is no greater than the Network Load Deviation Threshold (D_{NThres}) or all Target Cells are tried.

ACAR is a centralized on-demand load-ware routing scheme. Load-awareness distributes traffic evenly among relaying nodes to avoid congestion. Using ACAR, routing can be performed efficiently. ACAR also fully utilizes cell breathing - such that the cell size is not necessarily small. This allows location and state information of the mobile nodes within the maximum possible coverage area to be sent to the BS in a single hop. Routing overhead is greatly reduced. No potential call requests from the mobile terminal. In addition to facilitating load balancing, ACAR is also responsible for Intra-cell multi-hop routing for new outgoing calls. E-DSSA is an extension of DSSA. In addition to the merits of channel conflict resolution and the minimization of packet delay attained by DSSA, E-DSSA avoids consecutive channel conflict and Last Hop Node Interference. In the next Chapter, we will present a simulation model to evaluate the performance of this framework.

Chapter 4

Performance Evaluation

In this chapter, we will describe the simulation model and performance metrics used for the performance evaluation of our framework. A commercial professional modeling and simulation tool for networking called OPNET [27] is used. Simulation results show that call blocking ratio of the heavily loaded cell is reduced, load deviation among cells is decreased, system throughput is increased, and end-to-end delay is slightly increased.

4.1 Simulation Model and Performance Metrics

4.1.1 Simulation Model

Our simulation model is a three-cell one. Figure 4.1 shows the layout of the cells. This model is used for studying the effectiveness of our load balancing relaying routing framework (ALBAR) which includes ALBA, ACAR, and E-DSSA. Call blocking ratio of the three cells, load deviation among cells, maximum load deviation of the system, system throughput, and end-to-end delay of packets are investigated.

This model is tougher than a conventional 3-tier cell model because there is only one neighboring cell that the load of the hot cell can be migrated to. A conventional 3-tier cell system is shown in Figure 3.8. In the conventional model, there are six neighboring cells around the hot cell such that relaying nodes at each interface of the hot cell can be fully utilized, and load can be shifted in all directions. While the channels for the intra-cell routing are pointing towards the centre of the hot cell, most channels for the inter-cell relaying are pointing away from the centre of the hot cell. Thus, less channel conflicts occur and more channels can be re-used. This helps getting better performance results. However, the conventional model cannot be applied to some areas where only one adjacent cell can be placed. Moreover, simulation based on the conventional 3-tier cell model also requires simulating more than 3000 mobile nodes which is beyond the capability of ordinary equipment running OPNET. Therefore, we choose a three-cell model to evaluate our framework. Similar results are expected if the 3-tier cell model has been used.

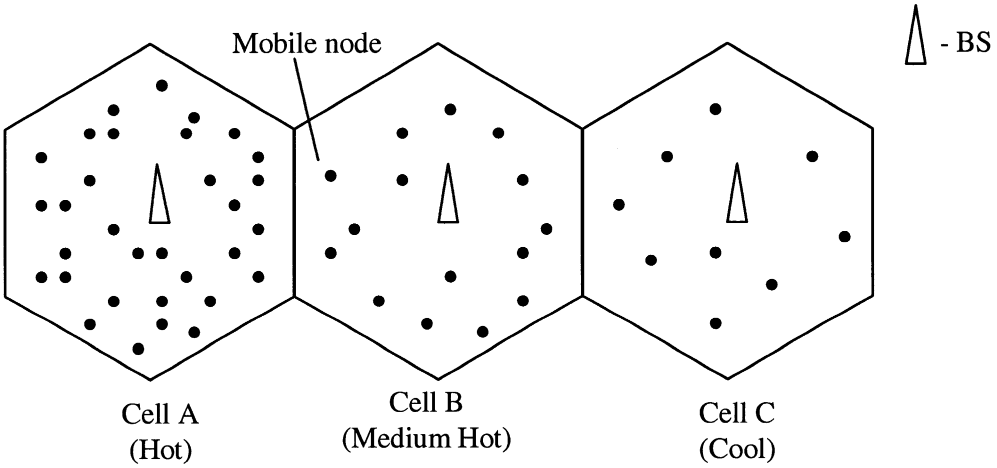


Figure 4.1: Three-cell model

4.1.2. Parameters

The followings are the major parameters and conditions of the model. The load condition of cell A, B, and C are hot, medium, and cool, respectively.

	Cell A (hot)	Cell B (medium)	Cell C (cool)
No. of Source nodes	100	70	60
No. of relaying nodes (**Scenario 1**)	**0/ 40/ 80/ 120/ 160**	160	120
No. of relaying nodes (**Scenario 2**)	160	**0/ 40/ 80/ 120/ 160**	120
Load balancing (**Scenario 1 and 2**)	**with / without**		
Distribution of nodes	Uniform		
Call arrival rate (Poisson distribution)	0.1 calls/sec.		
Call duration (exponential distribution)	5 sec.		
Simulation duration	30 sec.		

Other Parameters

Parameter	**Value**
Cell Capacity	1 Mbps
Relaying Node Capacity	1 Mbps
Normal Transmission Range	1000 m
Relay Transmission Range	250 m
Duplex method	TDD
Medium (multiple) Access	W-CDMA
Frame size	10 ms
Number of Time Slots per Frame	15
Number of available Codes per Time Slot	16
Data rate per code	13.8 Kbps
Number of Uplink Slots	13
Number of Downlink Slots	2
Packet arrival rate per connection	21 packets/sec. (constant rate)
Packet size	662 bits
QoS Class (C_i)	1 (Conversational)
QoS Class Weight Factor (k_c)	0.5
Distance Weight Factor (k_d)	0.5
Queue size (buffer)	750 packets
Neighboring Load Deviation Threshold (d_{nThres})	22 packets/sec.
Network Load Deviation Threshold (D_{NThres})	22 packets/sec.
Load Sampling Period	1 sec
Separation Threshold (Sep_{Thres})	30 m
Inter-cell Separation Threshold ($InterSep_{Thres}$)	900 m
Mobility	No

Intra-cell Route Zone	2500 m in diameter
Inter-cell Route Zone	3000 m in diameter
Antennas Beam Angle	45°
Load Threshold (L_{Thres})	1000 packets/sec.
Speed Threshold (S_{Thres})	3 km/hr
Maximum Hop Counts (Hop_{max})	7

Assumptions:

1. Perfect power control.
2. Perfect physical medium. (Fading effect is negligible.)
3. Each Source Node has one connection.
4. Energy consumption of every node is below Energy Consumption Threshold (E_{Thres}).
5. Remaining battery of every node is above Remaining Battery Threshold (B_{Thres}).

Simulator: OPNET [27]

A 90% confidence level with confidence intervals $[\overline{X}-10\%\overline{X}, \overline{X}+10\%\overline{X}]$ was used in the simulation, see appendix A.

4.1.3. Performance Metrics

1. **Call Blocking Ratio** - Ratio of the number of calls blocked to the number of call requested.
2. **Load Deviation** – Load difference between any two cells or Neighboring Load Deviation in the network (packets per second).
3. **Maximum Load Deviation** – Load difference between highest load cell and least load cell or Network Load Deviation in the network (packets per second).
4. **Throughput** – The rate of packet received of all BSs in the network (packets per second).
5. **End-To-End Delay** - Delay of a packet from a source node to its destination (BS).

The following are two scenarios for examining the effect of the number of relaying nodes in a hot cell or a medium hot cell on Call Blocking Ratio of different cells, Load Deviation and Maximum

Load Deviation, Throughput, End-to-End Delay of the system. Since Network Load Deviation can be much higher than Neighboring Load Deviation, if load balancing is triggered based only on Neighboring Load Deviation, the network load may not be balanced.

4.1.4 Scenarios

This load balancing and relaying framework is based on the A-Cell relay architecture in which channel assignment is involved. To my best knowledge, no existing load balancing scheme is based on this relay architecture. In other words, no existing load balancing scheme can be used for comparison. In our framework, load balancing is carried out through relaying. Without relaying, only the traffic of the source nodes in the overlapping regions of cells can be manipulated, and the load balancing effect is greatly diminished. Obviously, the number of relaying nodes and the locations of the relaying nodes will affect the effectiveness of load balancing. To investigate the effect caused by relaying nodes, we design the following two scenarios.

Scenario 1

- Varying the number of relaying nodes in cell A (0, 40, 80, 120, 160)
- Number of relaying nodes in cell B and cell C are 160 and 120 respectively
- Conditions: with and without load balancing

Scenario 2

- Varying the number of relaying nodes in cell B (0, 40, 80, 120, 160)
- Number of relaying nodes in cell A and cell C are 160 and 120 respectively
- Conditions: with and without load balancing

4.2 Simulation Results

Simulation results including Call blocking Ratio, Average Load Deviation, Average Maximum Load Deviation, Average Throughout, and Average End-to-End Delay are shown in Figure 4.2 to Figure 4.13.

4.2.1 Call Blocking Ratio

Figure 4.2 shows that without load balancing Call Blocking Ratio (CBR) of cell A is approximately 19 to 23%. When the load balancing scheme (ALBA) is enabled, CBR of cell A decreases to 16%. Also, the CBRs of cell B and cell C slightly increase. The sudden drop in CBR, even though there is no Relaying Nodes in cell A, is because the load of the Source Nodes in cell A near to the border of cell B can still be relayed through the Relaying Nodes in cell B. This effect does not apply in scenario 2. If no Relaying Nodes are in cell B, only little load can be migrated though the Source Nodes that are at the overlapping regions of both cell (see Figure 4.3). When a portion of the traffic load of cell A is relayed to cell B, more capacity is available in cell A to accept more calls. Hence, the CBR of cell A decreases. As cell B receives load from cell A, a portion of the capacity of cell B is consumed by the migrated load. Capacity of cell B hence decreases and the calls in cell B have a higher probability to be blocked. Thus, the CBR of cell B increases. This effect also happens in cell C because a portion of the load of cell B is relayed to cell C. This demonstrates that when load balancing scheme is enabled, load is migrated not only from cell A to cell B, but also from cell B to cell C. In other words, the load balancing effect penetrates all the cells. When the number of relaying nodes of cell A increases, the CBR of cell A further decreases. This is because the presence of more Relaying Nodes in cell A increases the chance for the Source Nodes in cell A in getting routes to the Target Cell B. Thus, more loads can be relayed to cell B so that the CBR of cell A is further decreased.

Figure 4.3 shows that scenario 2 (varying the number of Relaying Nodes in cell B) has similar effect as scenario 1 (varying the number of Relaying Nodes in cell A) except that the rate of reduction in CBR of cell A is more regular. Unlike in the previous scenario, there is no sudden drop of CBR. This is because at the beginning of the simulation, the number of Relaying Nodes within cell B is not large enough such that only a small amount of traffic can be relayed to cell B. Thus, the reduction in CBR in cell A is relatively smaller. In scenario 1, since there are already 160 Relaying Nodes in cell B at the beginning, once load balancing is enabled, the load of the

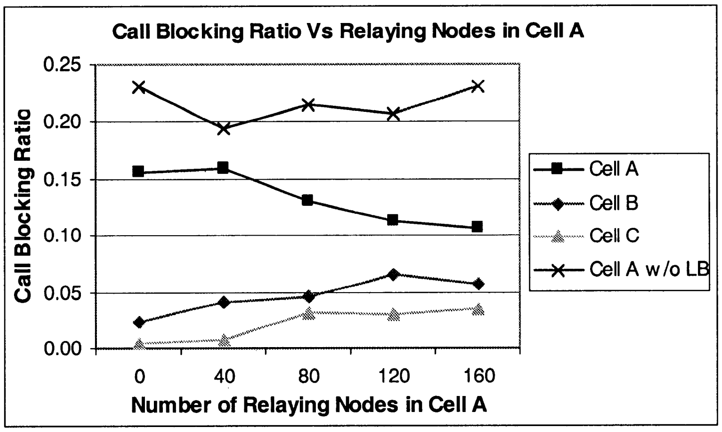


Figure 4.2: Comparison of Call Blocking Ratio among different cells with and without load balancing under various numbers of relaying nodes in cell A

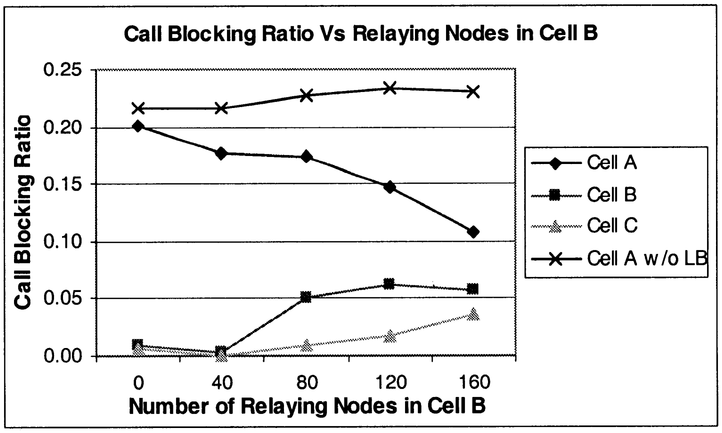


Figure 4.3: Comparison of Call blocking Ratio among different cells with and without load balancing under various numbers of relaying nodes in cell B

Source Nodes in cell A which are near to the border of cell B can be relayed immediately to cell B using the abundant Relaying Nodes in cell B. In other words, the presence of Relaying Nodes in cell B provides a faster reaction in load migration.

4.2.2 Load Deviation

Figure 4.4 and Figure 4.5 respectively show the load deviation among cells in scenario 1 and scenario 2 are reduced when load balancing is enabled. When the number of Relaying Nodes in cell A or cell B increases, the load deviation is further reduced. This is because more routes can be found for load migration between cell A and cell B. Figure 4.4 shows a graduate drop of load deviation. It also shows that the load deviation drops relatively more quickly when the number of relaying is beyond 40. It is because when the number of relaying nodes is below 40, few relaying node fall into the region near the border of cell B. When the number of relaying nodes is sufficiently large, i.e. beyond 40, there is a higher chance that the relaying node falls near to the border of cell B. These relaying nodes can be used to relay traffic for several connections. Thus, the load deviation starts to drop relatively more quickly. In Figure 4.5, load deviation does not drop until the number of Relaying Nodes in cell B is greater than 40. This demonstrates that the Relaying Nodes in cell B are more effective in load migration than the Relaying Nodes in cell A. When the number of Relay Nodes in cell B is beyond 120, the rate of reduction slows down because most Source Nodes has already shifted their load to their neighboring cell. These reasons also apply to the results of Maximum Load Deviation in Figure 4.6 and 4.7.

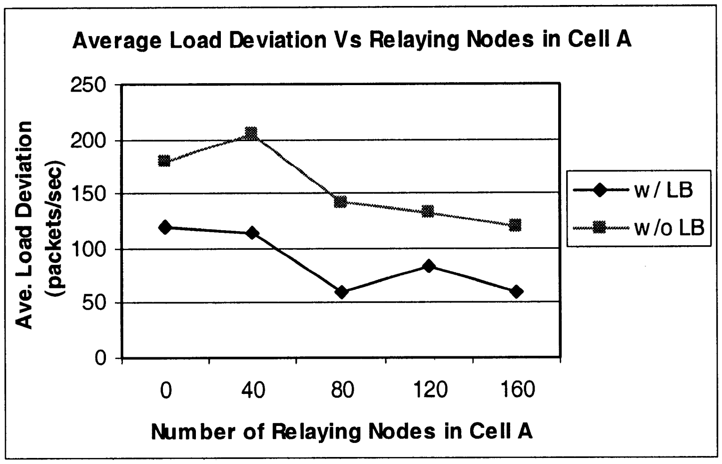


Figure 4.4: Comparison of average load deviation among cells with and without load balancing under various numbers of relaying nodes in cell A

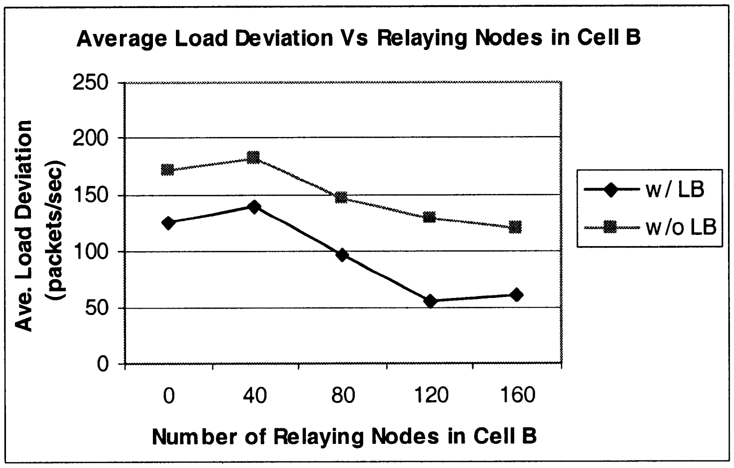


Figure 4.5: Comparison of average load deviation among cells with and without load balancing under various numbers of relaying nodes in cell B

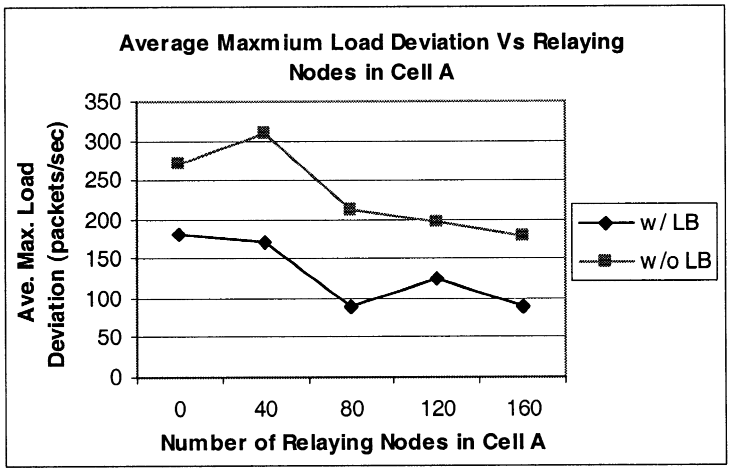


Figure 4.6: Comparison of average maximum load deviation of the system with and without load balancing under various numbers of relaying nodes in cell A

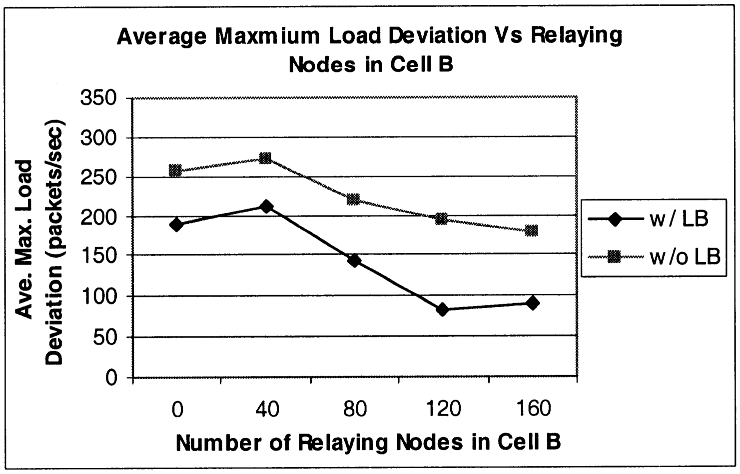


Figure 4.7: Comparison of average Throughput of the system with and without load balancing under various numbers of relaying nodes in cell B

4.2.3 Throughput

Both Figure 4.8 and Figure 4.9 show that the average BS (system) Throughput is increased when load balancing is enabled. Since the load balancing effect reduces call blocking ratio, more calls can be served and the system throughput is increased.

Figure 4.8 shows that the throughput decreases gently when the number of Relaying Nodes increases. Although more Relaying Nodes increases the chance of load migration, the number of hops per route also increases. The increase of hop count introduces more delay for a packet. In other words, fewer packets arrive at their BSs in same period of time. Thus, throughput is relatively lower.

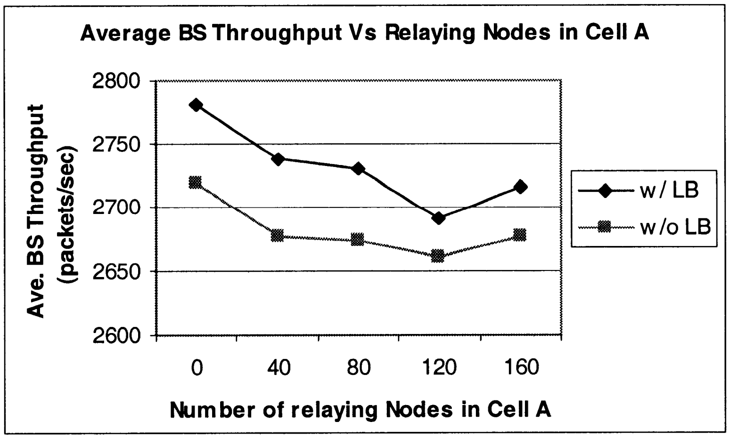


Figure 4.8: Comparison of average Throughput of the system with and without load balancing under various numbers of relaying nodes in cell A

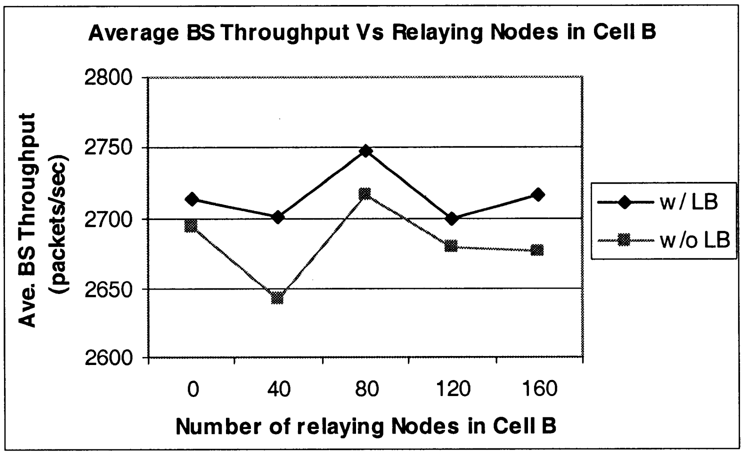
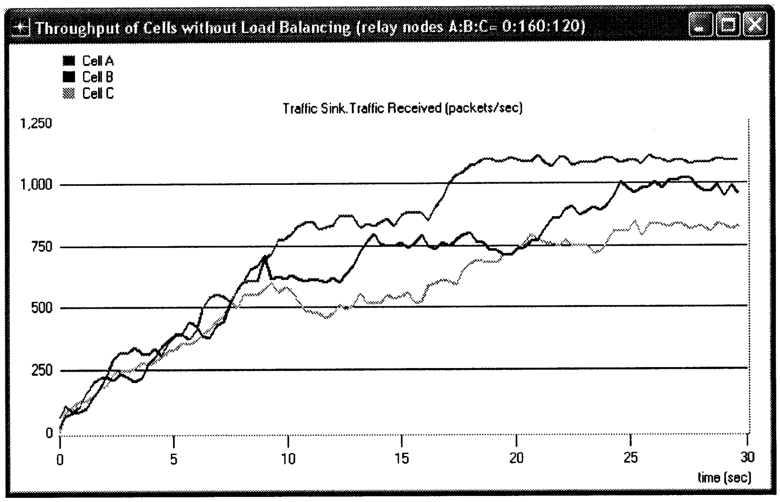


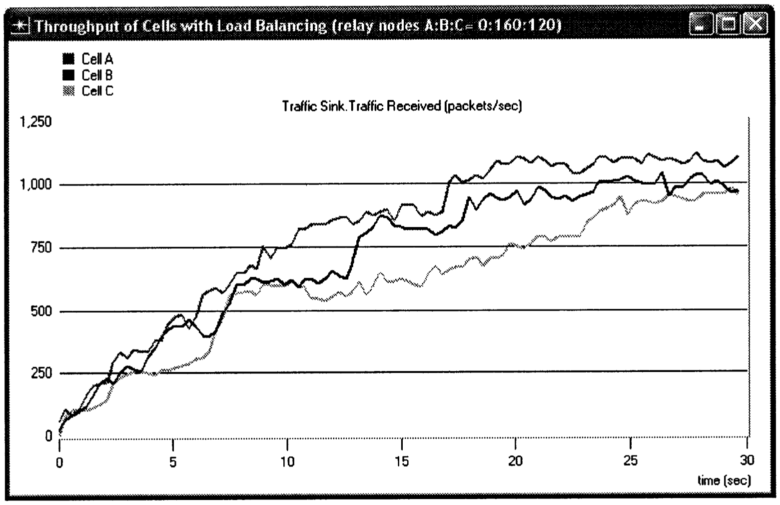
Figure 4.9: Comparison of average Throughput of the system with and without load balancing under various numbers of relaying nodes in cell B

In Figure 4.9, the throughput does not decrease with the increase in the number of Relaying Nodes in cell B. This is because the Relaying Nodes that are already in cell A dominate the packets' delay. Source Nodes in cell A start communicating with the BS using multi-hopping from the beginning, increasing the delay and, hence, the initial throughput in Figure 4.9 is lower than the initial throughout in Figure 4.8.

Figures 4.10a and 4.10b show the variation of throughput of cell A, B, and C with time when load balancing is enabled and disabled, respectively. The number of Relay Nodes in cell A, B, and C are 0, 160, and 120 respectively. When comparing these two figures, it can be seen that the load difference among the three cells is smaller when load balancing is enabled. The effect is more appealing when the state of the simulation becomes steady at simulation time 17 seconds and beyond. Figures 4.11a and 4.11b shows that the effect of load balancing on load deviation reduction is even better. This is because more Relaying Nodes are present in cell A. The number of Relay Nodes for Figures 4.11a and 4.11b in cell A, B, and C are 160, 160, and 120, respectively.



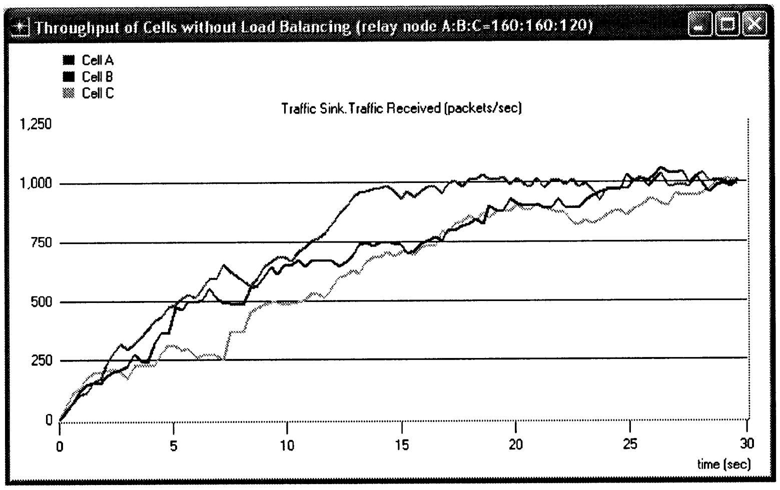
a) Without load balancing



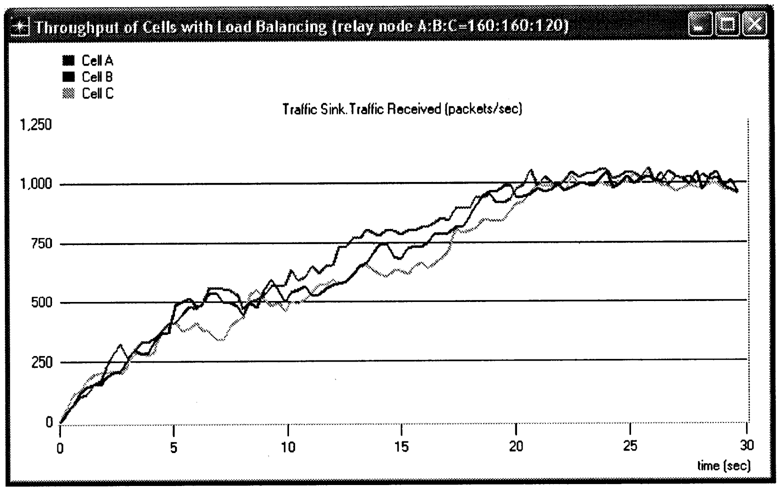
b) With load balancing

Figure 4.10: Throughput of three cells with Relay Nodes in Cell A:B:C = 0:160:120.

a) Without load balancing, b) With load balancing



a) Without load balancing



b) With load balancing

Figure 4.11: Throughput of different cells with Relay Nodes in Cell A:B:C = 160:160:120. a) Without load balancing, b) With load balancing

4.2.4 End-to-End Delay

Figure 4.12 shows that the average End-to-End Delay of packets is higher when load balancing is enabled. This is because the average number of hops in inter-cell routing, i.e. load balancing is enabled, is higher than the average number of hops in intra-cell routing, i.e. load balancing is not enabled.

When the number of Relaying Nodes in cell A increases, the delay in both load balancing and without load balancing situations increases. This is because the increase in the number of Relaying Nodes in cell A increases the number of hops in both intra-cell and inter-cell routing. When the number of Relaying Nodes is more than 120, the increase of delay slows down because most routes reach the Maximum Hop Count.

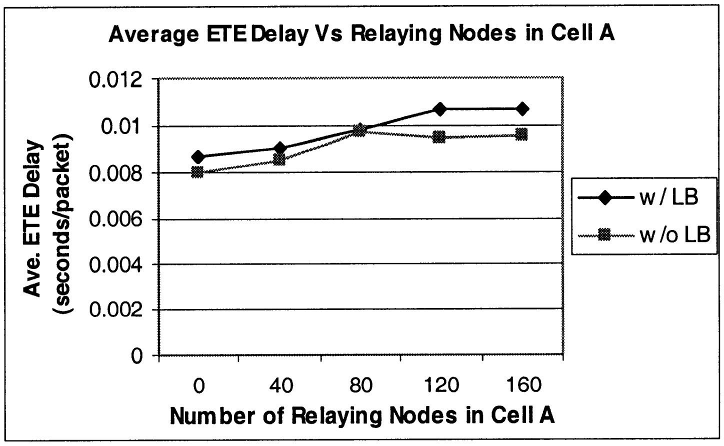


Figure 4.12: Comparison of average End to End Delay of packets with and without load balancing under various numbers of relaying nodes in cell A

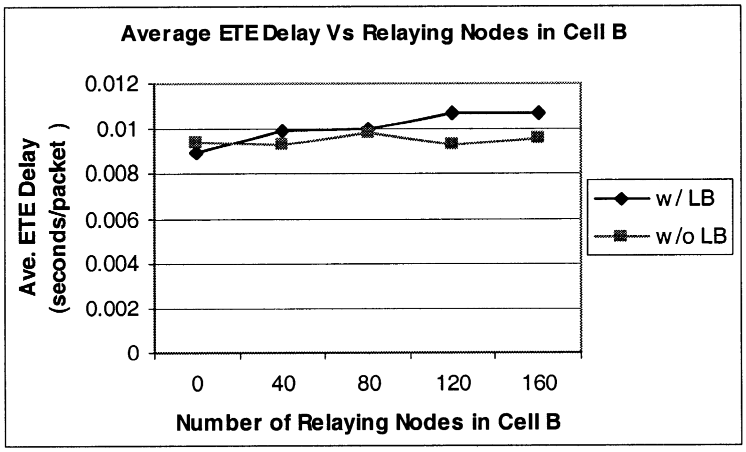


Figure 4.13: Comparison of average End to End Delay of packets with and without load balancing under various numbers of relaying nodes in cell B

Figure 4.13 shows that the average delay is higher when load balancing algorithm is enabled except for the time when the number of Relaying Nodes in cell B is about zero. At that time, a large number of Relaying Nodes are already present in cell A and too few Relaying Nodes are in cell B. The Source Nodes in cell A start communicating with the BS using multi-hopping which causes high packet delay. Since cell A has the largest number of Source Nodes in the system, the average delay in cell A dominates the average delay of the system. Thus, a higher system delay is obtained. When load balancing is enabled, Source Nodes near to the boundary of cell A are relayed to cell B. Since there are fewer Relaying Nodes in cell B at that time, the number of hops of the intra-cell relaying routes is smaller. In other words, the Source Nodes, which are originally using more hops to communicate with their BS, uses fewer hops to communicate with their new BS. Thus, the delay is smaller than the delay without load balancing at that time. The slope of the curve in Figure 4.13 is gentler than the slope of the curve in Figure 4.12. This also reflects that the delay in cell A is dominated the delay in the system.

When the number of Relaying Nodes in cell B is beyond 20, the delay with load balancing enabled is higher than the delay without load balancing. This result is consistent with the result in previous Figure 4.12.

Although delay seems longer when load balancing is enabled, if the capacity (data rate) gain due to the reduction in transmission distance is taken into account, the actual delay should be shorter and system throughput would be higher.

4.3 Summary

Simulation results show that ALBAR framework reduces the Call Blocking Ratio of hot cells, balances load among cells, and increases system Throughput. End-to-end delay only slightly increases. The Call Blocking Ratio in a hot cell is reduced because the load in the hot cell is relayed to its neighboring medium hot cell through the Relaying Nodes. The load of the medium hot cell is also relayed to its neighboring cool cell such that the medium hot cell has more available capacity to accept the load from the hot cell. When the number of relaying nodes in the hot cell or the medium hot cell increases, the Call Blocking ratio of the hot cell is further reduced. Results show that the Relaying Nodes inside the neighboring cells of a hot cell are more effective for load balancing than the Relaying Nodes inside the hot cell. The load deviation among cells with load balancing is lower than that of without load balancing. The increase of system (BSs) Throughput is because of the decrease of the Call Blocking Ratio of the hot cell. In other words, more call requests are accepted. Thus, the system throughput increases. The slight increase of the end-to-end delay is because relaying requires more hops. f the capacity gained due to the reduction in transmission distance is taken into account, the actual delay is expected to be smaller and the system throughput would be even higher.

Chapter 5

Conclusions and Future Work

In this thesis, we developed a load balancing and relaying framework, called A-cell Load Balancing Relaying Framework (ALBAR) for 3G TDD W-CDMA multi-hop cellular systems. This framework includes a load balancing scheme called A-Cell Load Balancing (ALBA), a routing scheme called A-Cell Adaptive Routing (ACAR), and a slot assignment scheme called Extended-Delay Sensitive Slot Assignment (E-DSSA). This framework integrates load balancing, routing and channel allocation functionalities to provide a more realistic solution for the 3G multi-hop cellular environment. This framework is adaptive and can be applied to different environments in terms of network topology, load distributions, and relaying route availability. It is based on the contention-free medium access technique W-CDMA so that radio resources are better utilized. The usage of licensed band instead of ISM band avoids co-channel interference and channel competition with other non-cellular users.

ALBA reacts to load unbalance situations promptly, i.e. in an on-demand fashion – hence reducing the load balancing overhead. It is designed to suit the dynamic load situation of 3G cellular systems, yet can also be applied to conventional cellular systems. The sensitivity of the load balancing effect can be adjusted by setting the values of Network Load Deviation Threshold and Load Sampling Period.

ACAR fully utilizes the dynamic cell size characteristic (cell breathing effect) of CDMA cellular systems. It is suitable for both densely and sparsely connected multi-hop cellular networks. Route discovery can be performed in a single hop for all nodes which are within the coverage of a cell simultaneously. Routing overhead is greatly reduced and call requests within the maximum coverage of a cell will not be denied. Similar to ALBA, ACAR employs an on-demand strategy. Therefore, the routing overhead is further reduced.

E-DSSA is an extension of the DSSA scheme. It inherits the merits of DSSA that it resolves channel conflicts. Directive antennas are used to reduce interference and increase spatial reuse. In addition, E-DSSA avoids the consecutive channel conflicts that occur when a mobile terminal is designed to be able to send and receive signals simultaneously. It also avoids the time slot conflict that happens when a mobile terminal is designed not to send and receive signals simultaneously. E-DSSA addresses the dynamic cell size characteristic of CDMA cellular system by using the Last Hop Node Interference Constraint.

This framework is especially useful for 3G cellular systems in which load situations are highly dynamic. This framework can be backward compatible with 2G CDMA systems on the condition that BSs are coordinated. Simulation results show that the framework reduces the call blocking ratio of a hotspot cell significantly. Load deviation among cells and maximum load deviation are also reduced. System throughput increases. End-to-end delay of packet increases only slightly. If the capacity gain due to the reduction in transmission distance is taken into account, a higher system throughput and a lower end-to-end delay of packet will possibly be obtained.

In the future, a load-aware Call Admission Control (CAC) scheme for multi-hop cellular systems can be developed and incorporated into our framework. The frequency of the execution of the load balancing routine can be reduced. Thus, the overhead of sending and receiving information and processing can be reduced. We will also add a Quality of Service (QoS) provisioning component to this framework. A detailed comparison of ACAR with other multi-hop cellular routing protocols can be performed, and the applicability of this framework to other networks such as mobile ad hoc networks and sensor networks investigated.

Bibliography

[1] M. Al-Riyami, A. M. Safwat and H. S. Hassanein, "A Slot Assignment Scheme for TDD W-CDMA Multi-hop Cellular Networks," *Proceedings of the 22nd Biennial Symposium on Communications,* Queen's University, Canada, May 2004.

[2] A. Bria et al, "4th-Generation Wireless Infrastructures: Scenarios and Research Challenges," *IEEE Personal Communications,* pp. 25-31, Dec. 2001.

[3] J. Broch, D. A. Maltz, D.B. Johnson, Y.C. Hu and J. Jetcheva, "A Performance Comparison of Multi-Hop Wireless Ad Hoc Network Routing Protocols," *Proceedings of the MOBICOM '98,* pp. 85-97, 1998.

[4] R. Chang, W. Chen and Y. Wen, "Hybrid Wireless Network Protocols," *IEEE Trans. Vehicular Technology*, vol. 52, no. 4, pp. 1099-1109, July 2003.

[5] S. K. Das, A.K. Sen and R. Jayaram, "A Structured Channel Borrowing Scheme for Dynamic Load Balancing in Cellular Networks," *IEEE ICDCS*, 1997.

[6] S. De, O. Tonguz, H. Wu and C. Qiao, "Integrated Cellular and Ad hoc Relay (iCAR) Systems: Pushing the Performance Limits of Conventional Wireless Networks," *35th Annual Hawaii International Conference on System Sciences (HICSS),* 2002.

[7] L. Din, J. Bigham, L. Cuthbert, P. Nahi and C. Parini, "Intelligent Cellular Network Load Balancing Using Cooperative Negotiation Approach," *IEEE WCNC,* 2003.

[8] R. Esmailzadeh, M. Nalagawa and A. Jones, "TDD-CDMA for the 4th Generation of Wireless Communications," *IEEE Wireless Communications*, Aug. 2003.

[9] 3rd Generation Partnership Project (3GPP); Technical Specification Group Radio Access Network; "Opportunity Driven Multiple Access," Sophia Antipolis, Valbbonne, 1999. (3G TR 25.925)

[10] 3rd Generation Partnership Project (3GPP); Technical Specification Group Ratio Access Network; "Physical channels and mapping of transport channels onto physical channel (TDD)," 2003. (3G TS 25.221 V3.6.0)

[11] T. Harrold and A. Nix, "Intelligent Relaying for Future Personal Communication Systems," *IEEE Colloquium on Capacity and Range Enhancement Techniques for Third Generation Mobile Communications and Beyond,* Digest No. 00/003, Feb. 2000.

[12] H. Hassanein and A. Zhou, "Routing with load balancing in wireless Ad hoc networks," *Proceedings of the 4th ACM international workshop on Modeling, analysis and simulation of wireless and mobile systems (MSWIM)*, July 2001.

[13] O. B. Holm, T. Friisø and T. Haslestad, "Improving UTRA Capacity with ODMA," *IST Mobile Communications Summit*, 2002.

[14] H. Holma and A. Toskala, "*WCDMA for UMTS, Radio Access for Third Generation Mobile Communications*," revised edition, John Wiley & Sons, 2001.

[15] H. Holma and A. Toskala, "*WCDMA for UMTS Radio Access for Third Generation Mobile Communications*," John Wiley & Sons, 2001.

[16] Y.C. Hsu and Y.D. Lin, "Base-Centric Routing Protocol for Multi-hop Cellular Networks," *IEEE GLOBECOM*, 2002.

[17] A. lwate, C. C. Chiang, G. Pei, M.Gerla and T. W. Chen, "Scalable Routing Strategies for Ad-hoc Wireless Networks," *IEEE Journal on Selected Areas of Communications,* vol. 17, no. 8, pp. 1369-1379, Aug. 1999.

[18] D. B. Johnson and D. A. Maltz, "Dynamic Source Routing in Ad Hoc Wireless Networks," *Mobile Computing*, Kluwer Academic Publishers, 1996.

[19] S. Kim and P.K. Varshney, "Adaptive Load Balancing with Preemption for Multimedia Cellular Networks," *IEEE WCNC*, pp. 1680-1684, 2003.

[20] Y.-B. Ko and N. H. Vaidya, "Location-Aided Routing (LAR) in mobile ad hoc networks," *ACM/IEEE International Conference on Mobile Computing and Networking (MobiCom)*, Oct. 1998.

[21] A. Leon-Garcia and I. Widjaja, "*Communication Networks*," McGrawHill, 2000.

[22] Y. D. Lin and Y.C. Hsu, "Multihop Cellular: A New Architecture for Wireless Communications," *IEEE INFOCOM'2000*, Tel Aviv, Israel, Mar. 2000.

[23] E. G. Lindin, F. Gunnarsson and F. Gustafsson, "Uplink Load Estimation on WCDMA," *IEEE Wireless Communications and Networking Conference*, New Orleans, USA, March 2003.

[24] S. Murthy and J. J. Garcia-Luna-Aceves, "An Efficient Routing Protocol for Wireless Networks," *ACM Mobile Networks and Applications Journal, Special issue on Routing in Mobile Communication Networks,* 1(2):183 197, 1996.

[25] T. Ohira and K. Gyoda, "Electronically steerable passive array radiator antennas for low-cost analog adaptive beamforming," *Phased Array Systems and Technology*, 2000.

[26] T. Ohira and K. Gyoda, "Design of electronically steerable passive array radiator (ESPAR) antennas," *Antennas and Propagation Society International Symposium*, 2000.

[27] OPNET Technologies Inc., www.opnet.com , Sept. 2004.

[28] C. Perkins and E. Royer, "Ad-hoc On-demand Distance Vector Routing," *Proceedings of the* 2^{nd} *IEEE Workshop on Mobile Computing Systems and Applications*, pp. 90-100, Feb. 1999.

[29] C. E. Perkins and P. Bhagwat, "Highly Dynamic Destination-Sequenced Distance-Vector routing (DSDV) for Mobile Computers," *ACM Proc. SIGCOMM '94,* pp. 234-244, Oct. 1994.

[30] P. Pham and S. Perreau, "Multi-path routing protocol with load balancing policy in mobile ad hoc network," *Mobile and Wireless Communications Network*, pp. 48-52, 2002.

[31] T. S. Rappaport, "*Wireless Communications, Principles and Practice*," 2nd edition, Prentice Hall, 2002.

[32] K. H. Rosen, "*Discrete Mathematics and Its Applications*," 4th edition. McGraw-Hill, 1999.

[33] S. Roy, D. Saha, S. Bandyopadhyay, T. Ueda and S. Tanaka, "Directional antenna: A network-aware MAC and routing protocol for effective load balancing in ad hoc wireless networks with directional antenna," *Proceedings of the fourth ACM international symposium on Mobile ad hoc networking & computing(MobiHoc)*, June 2003

[34] A. Safwat, "ACA: Channel Assignment in Ad hoc, 4G and Beyond Wireless Networks with Directional Antennas," *Proceedings of IEEE International Conference on Communications (ICC)* 2004, to appear.

[35] A. Safwat, "A-Cell: A Novel Multi-hop Architecture for 4G and 4G+ Wireless Networks," *Proceedings of IEEE Vehicular Technology Conference (VTC),* 2003.

[36] A. Safwat, "On CAC, DCA, and Scheduling in TDD Multi-hop 4G Wireless Networks," *Proceedings of IEEE International Performance, Computing, and Communications Conference (IPCCC),* 2004, to appear.

[37] W. Stallings, "*Data and Computer Communications*," 7th edition, Prentice Hall, 2000.

[38] D. Tacconi, C. U. Saraydar and S. Tekinay, "Short paper session: Improving UMTS throughput by utilizing ad hoc routing," *Proceedings of the 6th ACM international workshop on Modeling analysis and simulation of wireless and mobile systems*, Sept. 2003.

[39] A. S. Tanenbaum, "*Computer Networks*," 4th edition, Prentice Hall, 2002.

[40] C. K. Toh, "*Ad hoc Mobile Wireless Network, Protocols and Systems*," 1st edition, Prentice Hall, 2001.

[41] J.S.H. van Leeuwaarden, S. Aalto and J. Virtamo, "Load balancing in cellular networks using first policy iteration," *5th Finnish teletraffic seminar*, Finland, Feb. 2002.

[42] H. Wu and C. Qiao, "Modeling iCAR via multi-dimensional Markov Chains," *Mobile Networks and Applications,* vol. 8, June 2003

[43] J. Zhou and Y. Yang, "PARCelS: Pervasive Ad-hoc Relaying for Cellular Systems," *Proceedings of Med-Hoc-Net,* Sardegna, Italy, Sep. 2002.

Appendix A

Confidence Intervals

Since simulation results are estimates, the confidence level of these estimates has to be specified. With a certain confidence level, the confidence intervals can be calculated. Confidence intervals of the mean value of simulation results can be used to estimate the accuracy of the results. Considering the results of n independent simulation runs: $X_1, X_2, \ldots, X_n$, for the same experiment. The sample means, $\overline{X}$, of the results is given by

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

The variance of the distribution of the sample values, S_x^2 is given by

$$S_x^2 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n-1}$$

The standard derivation of the sample mean is given by

$$\frac{S_x}{n}$$

Under the assumption of independence and normality, the sample mean is distributed in accordance to the Student's t distribution. Then the confidence interval of the 100(1-α)% confidence level for the sample mean of the simulation runs is given by

$$\overline{X} \pm \varepsilon,$$

where

$$\varepsilon = \frac{S_x t_{\alpha/2,n-1}}{\sqrt{n}}$$

and $t_{\alpha/2,N-1}$ is the value of the *t*-distribution with n-1 degrees of freedom with probability $\alpha/2$.