iCAR : an Integrated Cellular and Ad hoc Relaying System

by

Hongyi Wu

(May 10, 2002)

A dissertation submitted to the
Faculty of the Graduate School of State
University of New York at Buffalo
in partial fulfillment of the requirements for the
degree of

Doctor of Philosophy

Department of Computer Science and Engineering

iCAR : an Integrated Cellular and Ad hoc Relaying System

by

Hongyi Wu

(May 10, 2002)

Major Professor: Chunming Qiao, Ph.D.

A dissertation submitted to the
Faculty of the Graduate School of State
University of New York at Buffalo
in partial fulfillment of the requirements for the
degree of

Doctor of Philosophy

Department of Computer Science and Engineering

To my wife, parents and sister

ACKNOWLEDGEMENTS

I would like to express my gratitude to my advisor. Dr. Chunming Qiao, for his outstanding guidance and continuous support without which this work could not have been possible. It was always a pleasure to work with him. His deep and broad knowledge in Computer Science, his never-ending enthusiasm for new and interesting ideas, and his complete dedication to his research work have been, and will always be, an example and inspiration for me.

I would like to thank my thesis committee members, Dr. Ozan Tonguz and Dr. Shambhu Upadhyaya, for their encouragement and advices. I would like to acknowledge Dr. Sudhir Dixit in Nokia Research Center for his support to my research work. My thanks also go to my thesis outside reader, Dr. Sung Ju Lee, for his cogent comments.

In addition, I would like to thank Swades De and Evsen Yanmaz for the thoughtful discussions we had in the last a couple of years.

Finally, I would like to thank my wife, my parents and my sister for their constant encouragement and support.

Contents

List of Figures

List of Tables

ABSTRACT

The cellular concept was introduced for wireless communication to address the problem of having scarce frequency resource. It is based on the sub-division of geographical area to be covered by the network into a number of smaller areas called cells. Frequency reuse in the cells far away from each other increases system's capacity. But at the same time, the cell boundaries prevent the channel resource of a system to be fully available for users. No access to *Data Channels* (or DCHs) in other cell by the *mobile host* (or MH) limits the channel efficiency and consequently the system capacity.

In this dissertation, we propose a new wireless system architecture based on the integration of cellular and modern ad hoc relaying technologies, called iCAR. It can efficiently balance traffic loads and share channel resource between cells by using *Ad hoc relaying stations* (ARSs) to relay traffic from one cell to another dynamically. This not only increases the system's capacity cost-effectively, but also reduces transmission power for mobile hosts and extends system coverage. We analyze the system performance in terms of the call blocking probability and queuing delay using multi-dimensional Markov chains for the new call requests and the call dropping probability for handoff requests, and verify the analytical results via simulations. Our results show that with a limited number of ARSs and some increase in the signaling overhead (as well as hardware complexity), the call blocking/dropping probability in a congested cell as well as the overall system can be reduced. We also propose a seed-growing approach for ARS placement, and discuss the upper bound on the number of seed ARSs needed in the system. In order to quantitatively evaluate ARS placement strategies, we introduce the concept of a new performance metric called quality of (ARS) coverage (QoC) for the comparison of various ARS placement strategies, and propose three rules of thumb as guidelines for cost-effective ARS placement in iCAR. Furthermore, we propose the signaling and routing protocols for establishing QoS guaranteed connections for IP traffic in iCAR. In particular, we discuss how a relaying route between a MH and a BTS in a nearby cell can be established via ARSs, and evaluate the performance of the protocols in terms of request rejection rate and signaling overhead through simulations. In addition, we propose a novel concept called "managed mobility" and address the ARS mobility management in iCAR.

Chapter 1

Introduction

In the last decade, with the unprecedented increase in demand for personal mobility and dependence on personal communications, both the number of subscribers and the amount of wireless traffic have surged at an exploding speed. With the advent of Internet, especially the wireless access to the Internet, wireless data traffic is expected to exacerbate the demand for bandwidth. The carriers and infrastructure providers now face a major challenge in meeting the increased bandwidth demand of mobile Internet users.

The cellular concept [1, 2, 3] was introduced for wireless communication to address the problem of having scarce frequency resource. It is based on the sub-division of geographical area to be covered by the network into a number of smaller areas called cells. Frequency reuse in the cells far away from each other increases system's capacity. But at the same time, the cell boundaries prevent the channel resource of a system to be fully available for users. This is because in order to avoid potential channel interference resulted from frequency reuse, a *mobile host* (or MH) in a cellular system can use only the *Data Channels* (or DCHs) of the current serving base transceiver station (or BTS), which is a subset of the data channels available in the system. No access to DCHs in other cell by the MH limits the channel efficiency and consequently the system capacity. More specifically, when a call request arrives in a cell which has no free DCHs, this call will be blocked or dropped although there are free DCHs in other cells in the system. Moreover, the presence of *unbalanced and bursty traffic* (e.g. wireless data traffic) will exacerbate the problem of having limited capacity and no access to channels in other cells in existing cellular systems. As a significant number of calls may be blocked and dropped due to localized congestion even though the traffic load does not reach the maximum capacity of the entire system, and the locations of congested cells (called *hot spots*) vary from time to time (e.g., downtown areas on Monday morning, or amusement parks in Sunday afternoon), it is difficult, if not impossible, to provide the guarantee of sufficient resource in each cell in a cost-effective way. In fact, increasing bandwidth of a cellular system (e.g., the number of DCHs in each cell) can increase the system capacity but not the efficiency to deal with the time-varying unbalanced traffic.

At the same time, various efforts in providing various access services such as wireless LANs [4, 5], ad hoc networks, Bluetooth [6, 7] and home RF [8, 9] networks, are further stimulating the growth of wireless traffic and the requirement for an ubiquitous wireless infrastructure. More specifically, continued proliferation of these services will call for interoperability between heterogeneous networks such as *ad hoc* and cellular systems. In addition, such an interoperability will create even heavier traffic in cellular systems as more and more traffic from wireless LAN's, ad hoc networks and Bluetooth devices, will be carried by the cellular infrastructure.

For the reasons cited above and the fact that the traffic in future cellular systems will be more bursty and unevenly distributed than conventional voice traffic, it is anticipated that *congestion* will occur in peak usage hours even in the next generation (e.g., third generation or 3G) systems, despite of its increased capacity. By congestion, we mean that in some cells, data channels (DCHs) are less frequently available than the minimum acceptable level and as a result, the grade of service (GoS) in those cells has deteriorated below a prescribed threshold level (e.g., the call blocking probability in those cells becomes above 2%). Note that, however, control channels (CCHs) for signaling (or paging) *may* still be accessible by all mobile hosts (MHs) in a congested cell.

In this work, we address the important problem of how to evolve from the existing, heavily-invested cellular infrastructure to next generation wireless systems that scale well with the number of mobile hosts. We propose to integrate the cellular infrastructure with modern Ad hoc relaying technologies to achieve dynamic load balancing among different cells in a cost-effective way. The basic idea of the proposed iCAR (integrated Cellular and Ad hoc Relay) system is to place a number of *Ad hoc Relay Stations* (or ARS's) at strategic locations, which can be used to relay signals between MHs and base stations [10, 11, 12]. By using ARSs, it's possible to divert traffic in one (possibly congested) cell to another (non-congested) cell. This helps circumvent congestion, and make it possible to *maintain* (or hand-off) calls involving MHs (especially a high-priority call) that are moving into a congested cell, or to accept new call requests involving MHs that are in a congested cell. There are many other benefits of the proposed iCAR system. For example, the ARSs can, in a flexible manner, extend cellular system's coverage (similar to the wireless routers used in the Rooftop system[13]), and provide interoperability between heterogeneous systems (by connecting ad hoc networks and wireless LANs to Internet for example). Additional benefits include enhanced reliability (or fault-tolerance) of the system, and potential improvement in MHs' battery life and transmission rate. We evaluate the performance of the proposed iCAR system in terms of call blocking probability via both analysis and simulations [14, 15, 16]. Our results indicate that an iCAR system with a limited number of ARSs is able to efficiently balance the traffic load among cells, and moreover, overcome the barriers imposed by the cell boundaries, which in turn, leads to significantly lower call blocking probability than a corresponding cellular system.

In order for the proposed iCAR system to provide the benefits described earlier, one need not only to make many modifications to existing cellular and Ad hoc networking techniques, but also invent novel approaches for dealing with the unique problems and characteristics of the integrated system. In this work, we propose these approaches/modifications, and evaluate their feasibilities and trade-offs (between cost and effectiveness).

The rest of this dissertation is organized as follows. Chapter 2 introduces the background of iCAR including the traditional cellular systems and modern ad hoc networks. In Chapter 3, we present the system architecture of the proposed iCAR system. Chapter 4 discusses the analytical and simulation model as well as the results for iCAR performance. Chapter 5 addresses the ARS placement issues. Chapter 6 presents the signaling and routing protocols to establish a QoS guaranteed connections. Chapter 7 introduces the strategies for ARS mobility management. Finally, Chapter 8 concludes this dissertation.

Chapter 2

Background and Related Works

This chapter introduces the background and motivations of our research, and the works related to the iCAR system.

2.1 Background

We first present an overview of the traditional wireless systems and the modern ad hoc networks, as well as their latest development trend and design challenge.

2.1.1 An Overview of Wireless Mobile Systems

The first mobile telephone service was introduced in 1946 by At&T. On June 17, 1946, a driver in St. Louis, Mo., pulled out a handset from under his car's dashboard, placed a phone call and made history. It was the first mobile telephone call [17]. Within a year, mobile telephone service was offered in more than 25 American cities. These mobile telephone systems were based on Frequency Modulation (FM) transmission. Most of these systems used a single powerful transmitter to provide coverage of up to 50 miles or more from the base [18]. The mobile telephone transmitter needed to have a line of sight to the BTS. It also implied that the few available radio channels were locked up over a large area by a small number of users.

Demand for mobile telephone service grew quickly and stayed ahead of the available capacity in many of the large urban cities. For example, 2,000 subscribers in New York City shared just 12 channels in 1965, and typically waited 30 minutes to place a call. It was wireless, but with ”strings” attached [17]. In the late 1960s and the early 1970s Bell Labs introduced the first cellular telephone system to alleviate the problem of spectrum congestion. The term ”cellular” refers to dividing the service area into many small regions (cells) each served by a low-power transmitter, with automatic call handoff from one cell to another and reuse of frequencies within a city. It is enormously expensive to build a system with thousands of cells right from the beginning. However, large-radius cell can evolve gracefully into small-radius cells over a period of time using cell-splitting. When the traffic reaches a point that the existing cell can no longer support the users with required grade of service, the cell is subdivided into small cells, and thus provides virtually unlimited capacity.

In the late 1970s, the first generation cellular system was standardized in the United States, called Advanced Mobile Phone System (AMPS) [19]. It used the frequency band around 850MHz and had 666 or 832 channels in a cell. The data channels were analog and used FM, while the control channels were digital and used Frequency Shift Keying (FSK) [20] modulation. 30 KHz were allocated to each data channel, which resulted in 10 kbps data transmission rate. Other standards were developed by other countries later, such as Total Access Communications System (TACS) in United Kingdom, Nordic Mobile Telephone (NMT) in Scandinavia, Nippon Telephone & Telegraph (NTT) in Japan, C450 in West Germany, etc.

The first generation cellular system were designed for business customers. With the increased demand of cellular telephones and the problem of incompatible standards in different countries in Europe, the Conference of European Posts and Telegraphs (CEPT) formed a study group to develop a new digital cellular standard, later called Global System for Mobile communication [21, 22] (GSM), that would cope with the ever burgeoning demands on European mobile networks in 1982. In 1989, the responsibility was transferred to the European Telecommunication Standards Institute (ETSI) [23], and phase I of the GSM specifications which is based on Time Division Multiple Access (TDMA) [24, 25, 26] were published in 1990. GSM systems use the frequency band around 900MHz. One RF channel occupies 200KHz band and has 8 speech channels with 270 kbs bit rate. Later on, American and Asian countries developed their own second generation system. For example, IS-54(based on TDMA) and IS-95(based on CDMA [27, 28])were developed in the United States, while Japanese Digital Cellular (JDC) standards were developed in Japan.

The explosive growth of Internet, and in particular, the introduction of IP version 6 [29, 30] (resulting a huge address space and a phenomenal increase in the number of mobile users and wireless nodes that all have their own globally unique IP addresses) has stimulated the interest in the development of packet switching data services in existing and future cellular systems. The 2.5G system, GPRS [31, 32] (General Packet Radio Service), is based on the GSM system and provide low-rate packet-switched data service of up to 64 Kbps through serving GPRS support node (SGSN) and gateway GPRS support node (GGSN).

However, voice and low-rate data services are insufficient in a world where high-speed Internet access is taken for granted. The third generation (3G) [33, 34] systems are designed to offer flexible multimedia services to users on-demand anywhere, and at any time. The main characteristics of 3G systems, known collectively as IMT2000 [35, 36], are a single family of compatible standards that have the following characteristics:

- Used worldwide
- Used for all mobile applications
- Support both packet-switched (PS) and circuit-switched (CS) data transmission
- Offer high data rates up to 2 Mbps (depending on mobility/velocity)
- Offer high spectrum efficiency

IMT2000 is a set of requirements defined by the International Telecommunications Union (ITU). As previously mentioned, IMT stands for International Mobile Telecommunications, and 2000 represents both the scheduled year for initial trial systems and the frequency range of 2000 MHz (WARC92: 18852025 MHz and 21102200 MHz). All 3G standards have been developed by regional standards developing organizations (SDO's). In total, proposals for 17 different IMT2000 standards were submitted to ITU in 1998. Evaluation of the proposals was completed at the end of 1998, and negotiations to build a consensus among differing views were completed in mid 1999. All 17 proposals have been accepted by ITU as IMT2000 standards. The specification for the Radio Transmission Technology (RTT) was released at the end of 1999. The most important

IMT2000 proposals are the Universal Mobile Telecommunications System (UMTS) [37, 38, 39] with Wideband-CDMA [40, 41] as the successor to GSM, CDMA2000 as the interim standard 95 (IS95) successor, and time divisionsynchronous CDMA (TDSCDMA) (universal wireless communication136 [UWC136]/EDGE) as TDMAbased enhancements to DAMPS/GSMall of which are leading previous standards toward the ultimate goal of IMT2000 [42].

Looking back into the history of wireless cellular systems, we see the main driver of the development is the increasing capacity demand. New technologies are invented and new frequency bands are allocated to support more and more mobile users and wireless traffics. Meanwhile, the increased system capacity and transmission data rate will stimulate new applications, which consequently result in additional bandwidth requirement.

2.1.2 Modern Ad hoc Networks (MANET)

The cellular system is an extension of wired networks, as only one hop (from a MH to a BTS) of a connection is wireless. It is strongly supported by the very matured techniques and has served customers for many years. But in the next generation of wireless communication systems, there will be a need for the rapid deployment of independent mobile users e.g. establishing survivable, efficient, dynamic communication for emergency/rescue operations, disaster relief efforts, and military networks. Such network scenarios cannot rely on centralized and organized connectivity, and can be conceived as applications of Mobile Ad hoc Networks (MANET) [43]-[54] which is a dynamic multi-hop wireless network established by a group of mobile hosts on a shared wireless channel. Hosts that are in close proximity can hear each other and are said to be neighbors. Since the nodes are mobile, the network topology may change rapidly and unpredictably over time. The network is decentralized, where all network activity including discovering the topology and delivering messages must be executed by the nodes themselves. Each host is potentially a router and it is possible to dynamically establish routes by connecting a sequence of neighboring hosts from a source to a destination in the ad hoc network.

The main challenges in the design and operation of Ad hoc networks stem from the possibility of rapid movements of the mobile hosts as well as the lack of a centralized control and management entity as in cellular system. Nodes can be connected dynamically in an arbitrary manner. Links of the network vary timely and are based on the proximity of one node to another node. They are also subject to frequent disconnection during node's mobility. In addition, wireless links have significantly lower capacity than the wired links. They are affected by several error sources that result in degradation of the received signal and high bit error rate.

All of the special features or the requirements of ad hoc networks mentioned above make designing of the routing protocol a challenging task. First and foremost, an effective ad hoc routing protocol should perform acceptably in a dynamic, low bandwidth environment . There are essentially two different strategies for Ad hoc routing. Since bandwidth and power consumption are scarce, precomputation of all routes may not be feasible. Thus it may be useful to use an on-demand approach to routing (e.g. [47, 49, 48, 55]), where routes are not computed until there is data which needs to be sent. This has the advantage of using resources more efficiently but the disadvantage of adding a route construction delay to data transmission as well as the overhead for route searching. When the network is large and traffic intensity is high, the connections may not be set up due to the long latency and congested channel (by the route discovery traffic). Proactive protocols (e.g. [46, 50, 56] which precompute routes avoid the extra latency but suffer from limited scalability because of the need to maintain routing entries for all other nodes. In traditional wired and wireless networks, the solution is usually to introduce hierarchy to the routing protocol. But maintaining hierarchical structure in a dynamic ad hoc network is much more difficult.

With some limitations, e.g. low mobility of MHs, the Ad hoc network will be easier to implement and achieve higher performance. One successful example is Bluetooth [7]. Originally developed by Ericsson (and now supported by other vendors as well), Bluetooth is a specification for low-cost, low-power, short-range radio links between cell phones, pagers, laptops, PDAs and other portable devices. Bluetooth was designed to replace the cables required to connect these kinds of devices and can also connect virtually any peripheral device including printers, desktop PCs, fax machines and keyboards. The first generation of Bluetooth permits exchange of data up to a rate of 1 Mbps per second, even in areas with high interference. It transmits and receives via a short-range radio link using a globally available frequency band (2.4 GHz ISM band).

2.2 Related Works

In order to meet the demand of increasing number of subscribers, the system needs to be redesigned. One obvious solution would be to allocate more frequency for the cellular system. While this is being done, it is important to realize that there is only a limited amount of frequency bandwidth that can be used. As we push frequency transmission above the giga-hertz range, device cost begins to increase rapidly. The bottom line is that frequency bandwidth is a very limited and scarce resource, and some alternative approaches of increasing the system capacity should be sought. This section presents the basic concepts of these approaches as well as their advantages and disadvantages.

2.2.1 Cell Splitting

The cellular concept takes the advantage of the fact that a communication channel, or a band of frequencies, can be used simultaneously by many callers if these callers are spaced physically far enough apart that their calls do not interfere with one another. In a cellular system, the co-channel interference is a function of the frequency reuse ratio $q = D/R = \sqrt{3N}$ where D is distance between two BTSs using the same frequency, R is the center-to-vertex distance, and N is the number of cells in a frequency reuse cluster. The larger the N, the higher the system capacity is.

Cell splitting is actually a built-in feature in the cellular systems. As the economic considerations made the concept of creating full systems with many small areas impractical, system operators developed the idea of cell splitting. It is done by reducing R. More specifically, when a service area becomes full of users, it further divides a cell to yield more (smaller) cells to deal with an increased number of uses in a given coverage area. The splitting of cell areas by adding new BTSs provides for an increasing amount of channel reuse and, hence, increasing subscriber serving capacity [57, 58, 59]. In this way, urban centers can be split into as many areas as necessary to provide acceptable service levels in heavy-traffic regions, while larger, less expensive cells can be used to cover remote rural regions (see Figure 2.1).

However, decreasing the cell radii imply that cell boundaries will be crossed more often. This will result in more handoffs per call and higher processing load per subscriber. In addition, it can be very costly to install new BTSs in each of these smaller cells, especially in very crowded downtown

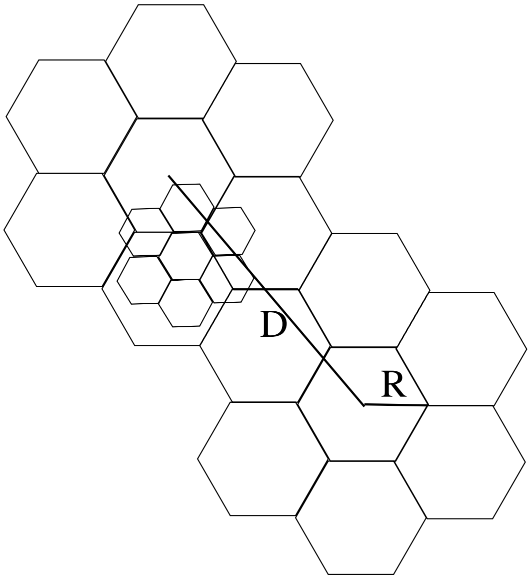


Figure 2.1: Cell Splitting

areas of big cities like NYC or LA (in such geographical areas, the cost of the so called “right-of-way” might be more expensive than the hardware cost of the BTSs).

2.2.2 Cell Sectorization

Sectorization is another technique to increase the system capacity. It uses a directional antenna to reduce the cochannel interference. In this scheme, each cell is divided into three or six sectors and uses three or six directional antennas at the BTS. Each sector is assigned a set of channels (frequencies). Because of the use of directional antennas, the number of interfering cells is reduced from six to two. This can consequently increase the channel reuse rate and the system capacity [60, 61]

There are two important factors that influence the system performance when using sectorization technique. The first is the number of sectors per cell. Intuitively,the more sectors in a cell, the less interference in the system. However, too many sectors at a cell will increase the interference to other cells and require excessive handoffs. Furthermore, the more sectors in a cell increase the base station cost. Therefore, most base stations in current cellular systems have three to six sectors at a cell [62, 63]. The second important factor for sectorization is the beamwidth of directional antenna

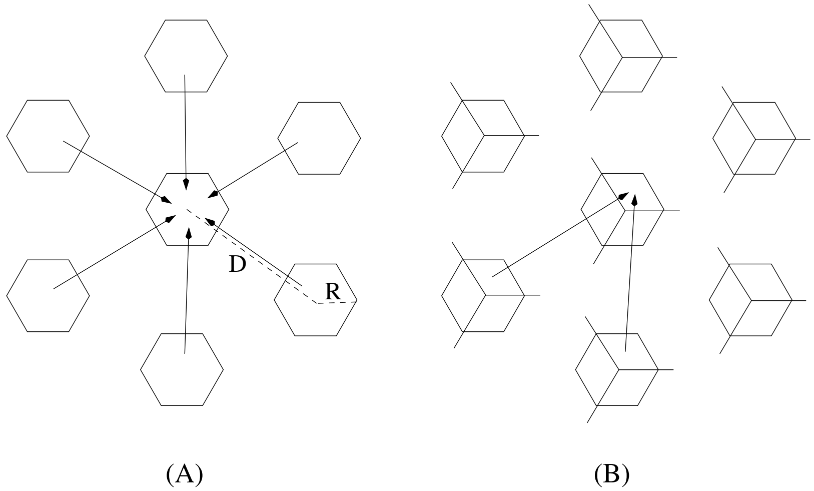


Figure 2.2: The co-channel interference is reduced by sectorization. (A) Without Sectorization; (B) With Sectorization.

used in each sector. Commonly used directional antennas in sectorized cellular systems have a 3 dB beamwidth of 60^o to 120^o . Two common 3-sector techniques used in current mobile systems studied in [63] are the wide-beam trisector cell (WBTC) and the narrow-beam trisector cell (NBTC). A WBTC is defined as a cell composing of three $100^o - 120^o$ directional antennas. An NBTC, on the other hand, is a cell with 3 sectors and each of which is served by a $60^o - 70^o$ directional antenna. It has been shown that the NBTC system performs better than the WBTC system. On the other hand, a six sector cellular system using six 60^o antennas at a cell is proposed to improve the capacity of the Global System for Mobile communications (GSM).

2.2.3 Dynamic Channel Borrowing

Since frequency channels are a scarce resource in a cellular mobile system, many schemes have been proposed to assign frequencies to the cells such that the available spectrum is efficiently used and thus the frequency reuse is maximized. These schemes can be broadly classified as fixed [64, 65, 66], dynamic [67, 66] and flexible [68, 66]. In a fixed assignment (FA) scheme, a set of channels is permanently allocated to each cell, which can be reused in another cell, sufficiently distant, so that the co-channel interference is tolerable. Such a pair of cells is called co-channel cells. In one type of FA scheme, clusters of cells, called compact pattern, are formed by finding the shortest distance between two co-channel cells, and each cell within a compact pattern is assigned a different set of frequencies. The advantage of an FA scheme is its simplicity which is the primary reason why it is adopted in most of the existing systems. But the disadvantage is that if the number of calls exceeds the number of channels assigned to a cell, the excess calls are blocked. This problem can be partially alleviated by channel borrowing methods, in which a channel is borrowed from one of the neighboring cells in case of blocked calls provided that it does not interfere with the existing calls. The disadvantage of channel borrowing is that the borrowed channel has to be locked in those co-channel cells of the lender which are non-co-channel cells of the borrower in order to avoid interference.

The motivation behind all basic channel assignment strategies is the better utilization of the available frequency spectrum with the consequent reduction of the call blocking probability in each cell, very few of them deal with the problem of non-uniformity traffic demand in different cells which may lead to a gross imbalance in the system performance. In the directed retry with load sharing scheme [69], it is assumed that the neighboring cells overlap and the users in the overlapping region are able to hear transmitters from the neighboring cells almost as well as in their own cell. Whenever the cell starts getting overloaded, some of those users handoff to the neighboring cells. The main drawbacks of this scheme include increased number of handoffs and co-channel interference, and also the load sharing is dependent on the number of users in the overlap region. In the channel borrowing without locking (CBWL) scheme, Jiang and Rappaport [70] proposed channel borrowing when the set of channels in a cell gets exhausted, but to use the borrowed channels under reduced transmission power to avoid co-channel interference. A serious drawback of the strategy is that not all users are always in the right zone to use the borrowed channels. Additionally, the fact that only a fraction of the channels in all the neighboring cells are available for borrowing, can severely affect the system performance in a highly overloaded system. The load balancing with selective borrowing (LBSB) scheme [71], attempts to alleviate the preceding problems by selectively borrowing channels before the available channel set in a cell is exhausted. A cell is classified as ‘hot’, if its degree of coldness (defined as the ratio of the number of available channels to the total number of channels allocated to that cell) is less than or equal to some threshold value. Otherwise the cell is ‘cold’. The LBSB scheme proposes to migrate a fixed number of channels from cold cells to a hot one through a centralized channel borrowing algorithm run periodically by an MSC server in charge of a group of cells. Aided by a channel allocation strategy within each cell, it has been shown in that the centralized LBSB achieves almost perfect load balancing and leads to a significant improvement over fixed assignment, simple borrowing, directed retry and CBWL schemes in case of an overloaded cellular system. However, the disadvantage of the LBSB strategy is that it is a centralized scheme and hence, too much depends on the central server in the MSC. The further work of LBSB as well as analysis and simulation results was propose in [72].

2.2.4 Cell Breathing and Sector Synthesis

In wireless communication, the cell boundaries overlap into each other’s cellular regions. The overlapping is important for mobiles near the cell boundary to perform soft handoff and to counteract fluctuations of receiving signal power. One way to achieve load sharing is for the heavily loaded cells to handoff some of its users in the overlapping region to less heavily loaded neighbors. So the size of a cell shrinks as the load increases, and expands as the load decreases. The phenomenon is also referred to as cell breathing [73]. In a CDMA system, as more handsets enter a cell, ambient RF noise and link loss increase, which causes the signal to degrade, and the cell size has to be reduced to decrease the active users and interference. Although this allow the system to dynamically adjust the cell coverage, it cannot increase the system capacity. Moreover, an unhappy result is that a user near the cell fringe might not be able to make or maintain a call [74].

Similar to cell breathing, Sector Synthesis lets the system control cell site sectorization for increased CDMA capacity and improved network performance. It provides CDMA service providers with flexible tuning options for controlling interference, creating dominant servers, managing handoff activity, and dealing effectively with nonuniform and time-varying traffic distributions. Working within a three-sector configuration, operators can adjust sector azimuth (orientation) pointing angles in 30-degree increments; select from sector beamwidths of 60, 90, 120, 180 and 240 degrees; and change gain settings to expand or contract coverage in highly localized areas. With Sector Synthesis capabilities, operators can create antenna patterns specifically designed for local traffic patterns and terrain without repeatedly climbing towers to mount custom antennas [75].

2.2.5 ODMA: Opportunity Driven Multiple Access

ODMA is an intelligent relaying protocol that sits upon the WCDMA radio sub-system. The protocol breaks difficult radio paths into a sequence of shorter hops which enables lower transmit powers or higher data rates to be used. It is the goal of the protocol to chose the least cost route through the relaying system when the relays are moving and the radio paths are dynamically changing.

Relaying is a widely used technique for radio packet data transmission both in commercial and military, systems but it has so far not been widely used in Cellular systems. In ODMA, any mobile hosts may serve as Relay Nodes and relay information between two other Nodes. Relays can not only add flexibility to a communication system, but also overcome the shadowing problem and potentially extend the range of high data rate coverage, which in turn results in high capacity.

A feasibility study conducted by the Alpha and Epsilon concept groups concluded that WCDMA can support relaying and the ODMA protocol with negligible increase in mobile complexity or cost. Simulations have shown that relaying has the potential to improve coverage and flexibility and may also increase capacity by lowering transmission powers and associated inter-cell interference. To relay information requires the use of radio resources such as codes. In a conventional structure resources are used once per cell, however in ODMA, the resources can be used many times within the basic cell area. This is because transmissions are lower power and so interference has only localized effect [76].

However,note that, ODMA relays on the MHs to be the intermediate nodes and set up the relaying routes, and thus may have disadvantages such as security (authentication, privacy), billing, routing, reliability, mobility management (of the MHs), and so on.

2.2.6 MACA: An Efficient Channel Allocation Scheme in Cellular Networks

In [77], the authors proposed a new channel allocation scheme, i.e. mobile-assisted connection-admission (MACA) algorithm, to achieve load balancing in a cellular network. In MACA, an ad hoc overlay network is added on the fixed-infrastructure cellular network. Channels assigned to this ad hoc network can be used to help the fixed-infrastructure to achieve load balancing. Specifically, a user in a congested cell may set up a multi-hop relaying route through other users, using the ad hoc channels, to nearby non-congested cell. Thus, the call blocking/dropping probability may be reduced.

In a wireless network with MACA, a mobile unit which would like to be an agent will broadcast a ”free” signal through the signaling channel, in which the information such as signal power, agent’s ID and traffic in the cell of this agent is included. This ”free” signal can only be received by the mobile units within the coverage of the ad hoc channel. Normally, only the mobile units in cold cells (cells which have free RF channels) can act as agents, but in special cases, such as the case where crossing-cell MACA is needed, the mobile units in the hot cells can also act as agents.

A user (an idle user or an active user) will collect these ”free” signals to build an agent table. The agents are picked by the connection stability with this mobile user, which is similar to the associativity-based routing (ABR) [55] scheme in an ad hoc network. When this user needs to use MACA, it will pick up one agent from the agent table, find a proper ad hoc channel (the ad hoc channel without co-channel interference), and build the connection with this agent. Once the agent gets the MACA request, it will send back an ACK to the user. Simultaneously, the agent will send a request to its base station. Once an RF channel is assigned to the agent, it will inform the user to start using MACA. The user can finally communicate with the fixed infrastructure.

A user keeps an agent table in which it caches all the reachable agents’ information. When the agent used in MACA is not suitable any more, normally because the quality of service (QoS) measured by the received bit-error rate (BER) in the ad?oc channel degrades to a certain value caused by the motion of both the agent and the user, the mobile user will send MACA request to another agent picked from its agent table. The new MACA link is built and the old one is released.

The base station has to keep the information about all its users and agents. When a user who is using MACA moves out of all the agents’ coverage or a newcomer to the hot cell cannot reach any agent in any cold cell, the base station will use the ”channel switching” scheme, i.e., it will ask one of its active users at the edge of the cell to use MACA and re? lease the local RF channel. This user is picked by its handoff approach; thus, it can use the channel in the foreign cell directly when it

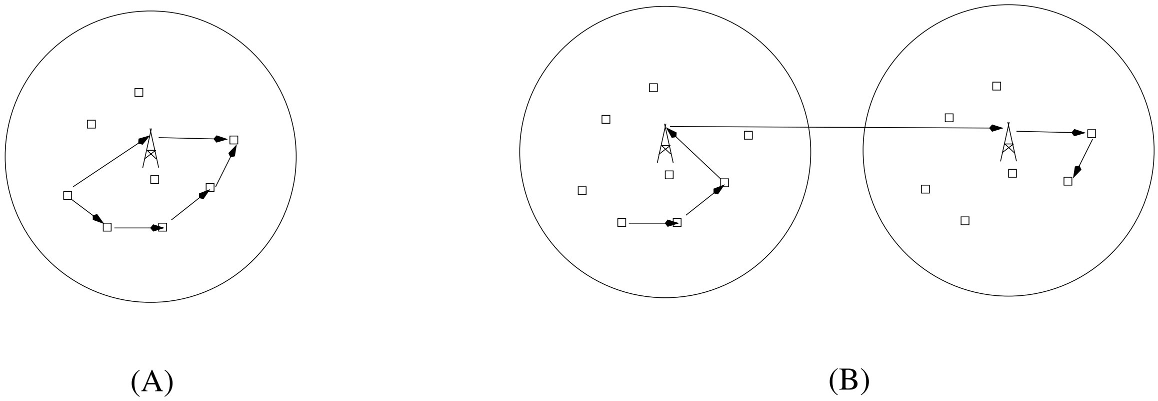


Figure 2.3: The multihop cellular network. (A) Intra-cell communication; (B) Inter-cell communication.

crosses the boundary. When the call is finished, or there is a local RF channel released and ACR is used, the MACA user will release both the ad hoc channel and the RF channel in the agent's cell.

The MACA concept is quite similar to our proposed iCAR system (to be discussed later in Chapter 3). However, it relays on the MHs to be the intermediate nodes and set up the relaying routes, and thus share many disadvantages in terms of security (authentication, privacy), billing, routing, reliability, mobility management (of the MHs), and so on with mobile Ad hoc networks.

2.2.7 Multihop Cellular

In [78], the authors proposed a new architecture, called **Multihop Cellular Network (MCN)**, as a viable alternate to the conventional single hop cellular networks. More specifically, a MH in MCN can reach the BTS in the *same cell* via a multihop route involving other MHs. Thus, MCN has several merits: (1) the number of bases or the transmission ranges of both mobile stations and base can be reduced, (2) connections are still allowed without base stations, (3) multiple packets can be simultaneously transmitted within a cell of the corresponding Singlehop Cellular Network (SCN), and (4) paths are less vulnerable than the ones in ad hoc networks because the bases can help reduce the wireless hop count.

In SCN, base and mobile stations in the same cell are always mutually reachable in a single hop. When having packets to send, mobile stations always send them to the BTS within the same cell. If the destination and the source are in the same cell, the BTS directly forwards packets to the destination. If the destination is in a different cell, the BTS forwards them to the base of the cell where the destination resides. The BTS of the latter cell then forwards packets to the destination in a single hop.

The architecture of MCN (See Figure 2.3) resembles that of SCN except that BTS and mobile stations are not always mutually reachable in a single hop. Similar to Ad hoc networks, a key feature of MCN is that mobile stations can directly communicate with each other if they are mutually reachable, which is not allowed in the conventional cellular system. If the source and the destination are in the same cell, other mobile stations can be used to relay packets to the destination, which achieves multihop routing within a cell. If not in the same cell, packets are sent to the BTS first, probably in multiple hops, and then be forwarded to the BTS of the cell where the destination resides. Packets can then be forwarded to the destination, probably in multiple hops again.

In the multihop cellular systems approach, relaying is performed by MHs. Thus, similar to ODMA, this approach will face problems such as security (authentication, privacy), billing, and mobility management (of the MHs) with mobile Ad hoc networks. In addition, the main goal of the multihop cellular systems is to reduce the number of BTSs or the transmission power of each BTS, but it can no longer guarantee a full coverage of the area. In fact, even in the ideal case where every MH in an area uncovered by any BTS can find a relaying route (through other MHs), the multihop approach will neither increase the system capacity nor decrease the call blocking/dropping probability, unless a large percentage of the calls are intra-cell calls (i.e., calls whose source and destination are in the same cell), which usually is not the case in practice.

2.2.8 A Hierarchical Routing Protocol

In [79], the authors presented a hierarchical structure for wireless mobile systems with a fixed backbone. In order to access the backbone, all MHs *have to* go through a Mobile Base Station (which can be thought of as a cluster head). The major contribution of this work is the routing algorithm which balances the cost of location-update and path-finding operations by partitioning the terminals and mobile base stations to produce a virtual topology. Based on the virtual topology, each network entity stores a fraction of the network topology information and maintains the routing efficiency.

2.2.9 Cellular IP

Hosts connecting to the Internet via wireless interface are likely to change their point of access frequently. A mechanism is required that ensures that packets addressed to moving hosts are successfully delivered with high probability. A change of access point during active data transmission or reception is called a handoff. During or immediately after a handoff, packet losses may occur due to delayed propagation of new location information. These losses should be minimized in order to avoid a degradation of service quality as handoffs become more frequent.

Cellular IP [80, 81, 82] is a protocol that provides mobility and handoff support for frequently moving hosts. It is intended to be used on a local level, for instance in a campus or metropolitan area network. Cellular IP can interwork with Mobile IP [83, 84, 85] to support wide area mobility, that is, mobility between Cellular IP Networks.

The following is an overview of the operation of Cellular IP. BTSs periodically emit beacon signals. MHs use these beacon signals to locate the nearest Base Station. A Mobile Host can transmit a packet by relaying it to the nearest Base Station. All IP packets transmitted by a Mobile Host are routed from the BTS to the Gateway by hop-by-hop shortest path routing, regardless of the destination address. Cellular IP Nodes maintain Routing Cache. Packets transmitted by the Mobile Host create and update entries in each Node's Cache. An entry maps the Mobile Host's IP address to the interface through which the packet entered the Node. The chain of cached mappings referring to a single Mobile Host constitutes a reverse path for downlink packets addressed to the same Mobile Host. As the Mobile Host migrates, the chain always points to its current location because its uplink packets create new mappings and old mappings are automatically cleared after a soft state timeout. After a migration, before the old mappings are cleared, a Node can temporarily have mappings for the same Mobile Host to multiple interfaces. (This causes the chain to temporarily have a fork.) IP packets addressed to a Mobile Host are routed by the chain of cached mappings referring to the said Mobile Host. To prevent its mappings from timing out, a Mobile Host can periodically transmit control packets. Control packets are regular IP packets with empty payloads. Mobile Hosts that are not actively transmitting or receiving data but want to be reachable for incoming packets, let their Routing Cache mappings time out but maintain Paging Cache mappings. IP packets addressed to these Mobile Hosts will be routed by Paging Caches. Paging Caches have a longer timeout value than Routing Caches and are not necessarily maintained in every Node.

Chapter 3

An Overview of The iCAR System

In this chapter, we describe the basic operations and principle benefits of the new architecture. To simplify the following presentation, we will focus on cellular systems where each BTS is controlled by an Mobile Switching Center (or **MSC**) (which is sometimes referred to as Mobile Telephone Switching Office as well) [18, 86]. The major differences between BTSs and the proposed ARSs are as follows. Once a BTS is installed, its location is fixed since it often has a wired interface to an MSC (and a backbone network). On the other hand, an ARS is a *wireless* communication device deployed by a network operator. It has its own controller with a much lower complexity and fewer functionality than that needed for a BTS. In addition, it may have a limited mobility (which, unlike that of an MH, is under the control of an MSC), and can communicate *directly* with an MH, a BTS or another ARS through air interfaces.

An example of relaying is illustrated in Figure 3.1 where MH X in cell B (congested) communicates with the BTS in cell A (or BTS A, which is non- congested) through two ARSs (there will be at least one ARS along what we call a *relaying route*). Note that each ARS has two air interfaces, the C (for cellular) interface for communications with a BTS and the R (for relaying) interface for communicating with an MH or another ARS. In the following discussion, we will assume that the C interface operates at around 1900 MHz (PCS), and the R interface uses an unlicensed band at 2.4 GHz though our concept also applies when different bands are used (for example, 850 MHz for the C interface as in 2G systems or 2 GHz for 3G systems). The R interface (as well as the

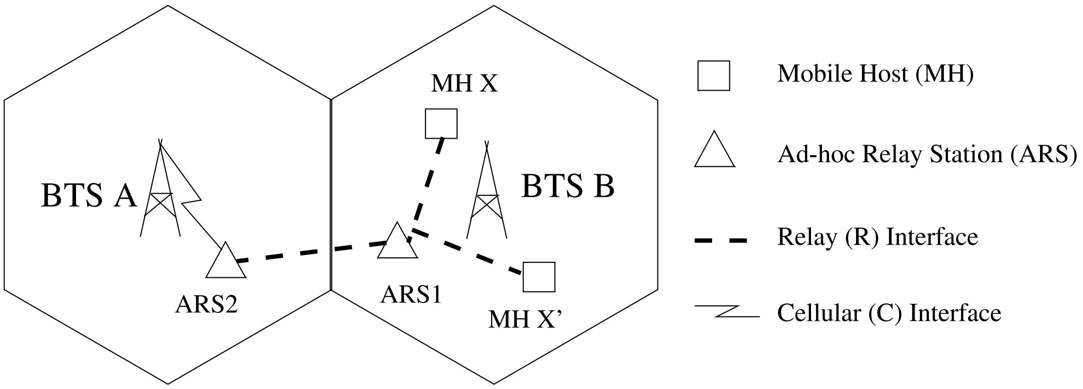


Figure 3.1: A relaying example where MH X communicates with BTS A through two Ad- hoc Relaying Stations (ARSs) (it may also communicate with MH X' through ARS 1)

medium access control (MAC) protocol used) is similar to that used in wireless LANs or Ad hoc networks (see for example [47, 49, 45]). Note that because multiple ARSs can be used for relaying, the transmission range of each ARS using its R interface can be much shorter than that of a BTS, which means that an ARS can be much more smaller and less costly than a BTS. At the same time, it is possible for ARSs to communicate with each other and with BTSs at a higher data rate than MHs can due to limited mobility of ARSs and specialized hardware (and power source).

Among the ARSs involved in relaying, we may call an ARS which directly communicates with an MH (e.g. ARS 1 in Figure 3.1) a *proxy*, and an ARS which directly communicates with a BTS (e.g. ARS 2 in Figure 3.1) a *gateway* (an ARS can serve as both a proxy and a gateway at the same time as illustrated in Figure 3 (a)). When and *only when* an ARS serves as a gateway, it uses the C and R interfaces concurrently. Other ARSs along a relaying route use the R interface only. This means that an ARS does not use any DCH unless it is serving as a gateway between an MH and a BTS, in which case, a DCH will be allocated to the ARS dynamically by a MSC).

Note that, to enable relaying, a MH also needs to have the R interface to communicate with an ARS, in addition to having the C interface used to communicate directly with a BTS under the normal situation (i.e. without relaying). Although it is possible to treat an MH just as an ARS, that is, to use the MH (and its R interface) to relay signals between another MH and a BTS as in Ad hoc networks or the so-called *Opportunity Driven Multiple Access* (ODMA) proposal (see http://www.etsi.org), issues such as security (authentication, privacy), billing, and unpredictable movement of the MHs make such an approach difficult to implement. In fact, the main challenges in the design and operation of Ad hoc networks stem from the possibility of rapid movements of the MHs as well as *the lack of a centralized control and management entity* [87]. In the proposed system, ARSs (approximate) locations and their (potential) movement are *under the control of MSC's*; and as a result, a relaying route with satisfactory QoS parameters (if it exists) can be established more quickly, and once established, maintained with a higher degree of stability, making our approach more suitable for real time applications.

3.1 Congestion-Induced (CI) Relaying

A principle benefit of the proposed integration of the cellular and Ad hoc relaying technologies is that both the blocking probability of new calls to/from a congested cells, and the call dropping probability during hand-offs to a congested cell, can be drastically reduced via what we call congestion-induced (or CI) relaying. This is illustrated below.

3.1.1 New Call

In an existing cellular system, if MH X is involved in a new call (as a caller or callee) but it is in a congested cell B, the new call will be blocked. In the proposed next generation wireless system with integrated cellular and relaying technologies, the call does not have to be blocked. More specifically, MH X which is in the congested cell B, can *switch over* to the R interface to communicate with an ARS in cell A, possibly through other ARSs in cell B (see Figure 3.1 for an example). We call this strategy that establishes a relaying route between MH X (in a congested cell) and a BTS in a nearby non- congested cell *primary relaying*.

With primary relaying, MH X can communicate with BTS A, albeit indirectly (i.e. through relaying). Hereafter, we will refer to the process of changing from the C interface to the R interface (or vice versa) as switching-over, which is similar to (but different from) frequency-hopping [86]. Of course, MH X may also be relayed to another nearby non-congested cell other than cell A. Finally, a relaying route between MH X and its corresponding (i.e., caller or callee) MH X' may also be established, (in which case, both MHs need to switch over from their C interfaces to their R interfaces), even though the probability that this happens could be very low.

If primary relaying is not possible because, for example in Figure 3.1, ARS 1 is not close enough to MH X to be a proxy (and there are no other nearby ARSs), one may resort to *secondary relaying* so as to *free up* a DCH from BTS B for use by MH X. Two basic cases are illustrated in Figure 3.2 (a) and (b), respectively, where MH Y denotes any MH in cell B which is currently involved in a call. More specifically, as shown in Figure 3.2 (a), one may establish a relaying route between MH Y and BTS A (or any other cell). In this way, after MH Y switches over, the DCH used by MH Y can now be used by MH X. Similarly, as shown in Figure 3.2 (b), one may establish a relaying route between MH Y and its corresponding MH Y' in cell B or in cell C, depending on whether MH Y is involved in an intra-cell call or an inter-cell call. Note that, given that cell B is congested which means that there are a lot of on-going calls (or candidates for MH Y), the chance that case (b) in Figure 2 could occur should be better than that a relaying route between MH X and MH X' can be established using primary relaying (as in Figure 1). In addition, although the concept of having such an MH-to-MH call via ARSs only (i.e. no BTSs are involved) is similar to that in Ad hoc networking, a distinct feature (and advantage) of the proposed integrated system is that an MSC can perform (or at least assist in performing) critical call management functions such as authentication, billing, locating the two MHs and finding and/or establishing a relaying route between them, as mentioned earlier. Such a feature is also important to ensure that switching-over of the two MHs (this concept is not applicable to Ad hoc networks) is completed fast enough so as not to disconnect the on-going call involving the two MHs or not to cause severe QoS degradation (though the two MHs may experience a "glitch" or jitter).

If neither primary relaying, nor basic secondary relaying as shown in Figure 3.2 (a) and (b) works, the new call may still be supported. More specifically, assume that there is a relaying route, which can be either primary or secondary relayed, between MH X and ARS, say G (for gateway),in a nearby cell C which unfortunately is *congested*. As shown in Figure 3.2(c), one may apply any of the two basic secondary relaying strategies described above in the congested cell C (i.e. in a *cascaded* fashion) *as if* ARS G is being "handed-over" (see discussion below). Hence, if a relaying route between an MH (say MH Z) in cell C and either another BTS in a non-congested cell or MH Z' can be established, ARS G can be allocated the DCH previously used by MH Z in cell C, and in

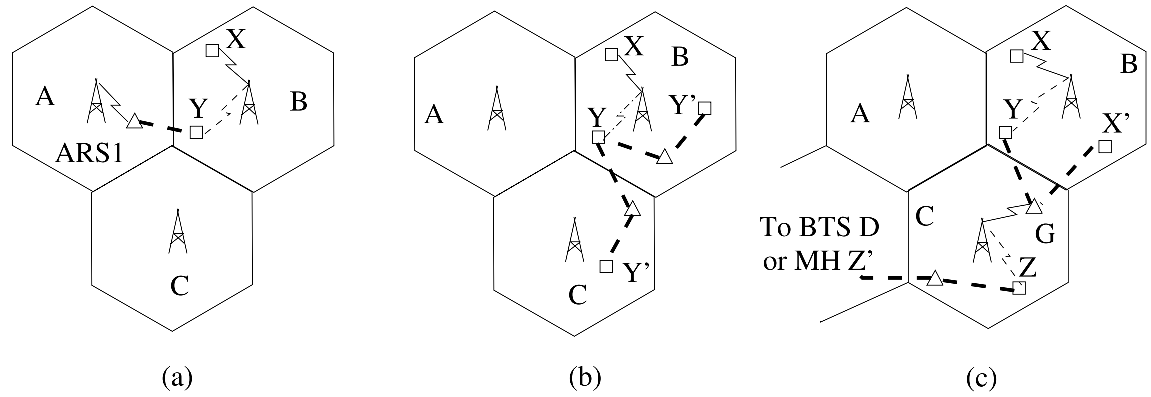


Figure 3.2: Secondary relaying to free up a channel for MH X. (a) MH Y to BTS A, (b) MH Y to MH Y', or (c) cascaded secondary relaying (i.e. MH Y to BTS C and MH Z to either MH Z' or BTS D).

turn MH X can be allocated the DCH previously used by MH Y in cell B if the route between MH X and ARS G is set up by secondary relay.

3.1.2 Handoff Call

In an existing cellular system, if an MH X involved in a call moves from cell A to cell B, a request for hand-off will be sent as soon as the power level from BTS A received by MH X goes below a certain threshold (and that from BTS B is becoming higher). A successful hand-off will take place, usually within a few hundred milliseconds (depending on the moving speed of the MH) before the received power from BTS A reaches an unacceptable level [18, 86].

If cell B is congested, the hand-off request may be queued (that is, the call may be blocked) for a short period of time, e.g., up to a few tens of milliseconds as long as the received power is still above the unacceptable level. If the congestion in cell B persists, that is, there are still no DCHs available in cell B after this short period of blocking time, the call will be dropped.

In the proposed integrated system, MSC may apply the primary relaying strategy to establish a relaying route between MH X to a BTS in a nearby non-congested cell (similar to Figure 3.1) or the secondary relaying strategies and cascaded relay to free up DCHs in cell B for use by MH X (similar to In this subsection, we will analyze the performance of primary relaying based on the multi-dimensional Markov chain model.Figure 3.2). In this way, the handoff call can take place successfully.

Note that by applying the relaying strategies (primary and secondary) to establish a relaying route between an MH in a congested cell B and a BTS in another cell (not necessarily an immediate neighbor), new calls involving MHs in cell B and hand-off calls involving MHs moving into cell B can now be supported, it is as if cell B has "borrowed" some DCHs from other cells. In other words, the capacity of cell B has been effectively increased, thus eliminating (or at least alleviating) congestion.

3.2 Noncongestion-Induced (NCI) Relaying

Clearly, relaying can also be used to *pro-actively balance load* among different cells by transferring calls from a heavily-loaded cell to other lightly-loaded, and possibly remote cells (for example, two cells such as B and D in Figure 3.2 (c), between which there are no relaying routes available). This is one of the main advantages of the proposed approach over channel borrowing via cell sectorization whereby a cell can only borrow a pre- determined set of channels (into one of its sectors) from its immediate neighbors [18, 88, 89].

Note that when no cells are currently congested, and relaying is used to, for example, balance load as described above, we might call this type of relaying *noncongestion-induced* (or **NCI**). NCI-relaying is also useful to overcome so-called shadows where no coverage by a BTS is available because either there are buildings surrounding an MH, which completely block signals from a nearby BTS to the MH, or no BTSs are close enough to the MH. This is illustrated in Figure 3.3 (a) and (b) respectively. In Figure 3.3 (a), an MH behind a building (or buildings) may still receive the signal from a BTS due to multi-path propagation of radio signals, though the signal could be very weak [86]. In such a case, NCI- relaying may improve the signal strength and other QoS performances. As an added benefit of relaying, either NCI or CI, one may reduce the power consumption of an MH since the distance between the MH and the proxy ARS can be much shorter than that between the MH and the BTS. More specifically, given that the typical transmission ranges of a cell and a ARS are 2km (at 1900 MHz) and 500m (at 2.4 GHz) respectively, the maximum path losses are 104.04 dB and 94.02 dB, respectively (according to the free space propagation model where the

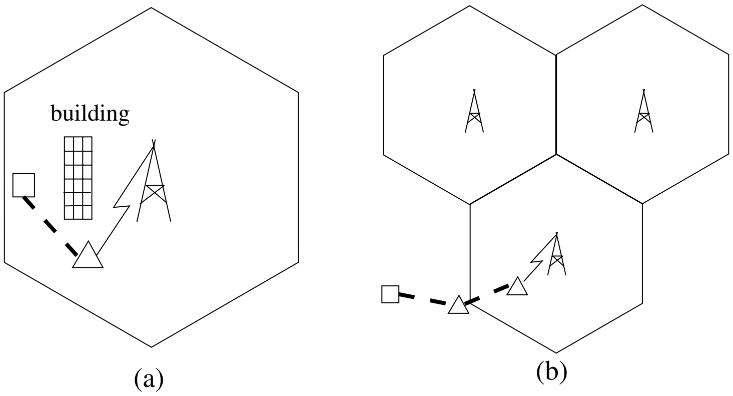


Figure 3.3: Non-congestion-induced (NCI) relaying to overcome shadows

path loss is equal to $32.44 + 20log(frequency) + 20log(distance))$. This means that MHs using relay consume almost 10 times less power in Watts.

Chapter 4

Performance Evaluation of iCAR

In this chapter, we evaluate the performance of the iCAR system. We establish two theorems serving as the principles, and show the performance improvement of iCAR in terms of the call blocking/dropping probabilities and queuing delay via analysis and simulations.

4.1 Principles

We first discuss the principles for the performance improvement of the iCAR system over a conventional cellular system assuming that the entire system can be covered by ARSs so that an MH in a cell can reach the BTS in any cell in the system via relaying. In this section, we consider a *loss system*, in which each cell is under the following two assumptions.

Assumption 1 : *All free channels are fully available to calls from the sources (i.e., MHs).*

Assumption 2 : *A call which meets congestion is discarded immediately.*

We present the following two theorems to show that iCAR will outperform the conventional cellular system. The first theorem states the best performance that a conventional cellular system can achieve.

Theorem 1 *Assume that the total traffic in an n-cell system is T Erlangs, then the (system wide) call blocking probability is minimized when the traffic in each cell is T/n Erlangs.*

Proof : Let the number of DCHs in each cell be M and assume that the traffic intensity is T_i in each cell i where $T = \sum_{i=1}^{n} T_i$. The probability of all the channels in cell i being busy is given by the following Erlang B formula [90].

$$B_i(M;T_i) = \frac{T_i^M/M!}{\sum_{i=0}^{M} T_i^i/i!} \tag{4.1}$$

For the n-cell system, the average blocking probability for the entire system is

$$B = \frac{\sum_{i=1}^{n} B_i \times T_i}{T} \tag{4.2}$$

Since $T = \sum_{i=1}^{n} T_i$, we may write $T_n = T - \sum_{i=1}^{n-1} T_i$. In other words, there are only $n-1$ independent T_i's. In order to compute the minimum value of B, we compute all the partial derivatives of B and set them to be 0, that is,

$$\frac{\partial B}{\partial T_i} = 0 \qquad (1 \le i \le n-1) \tag{4.3}$$

We omit the details but we can obtain the critical points (or the solutions to the above equations) as

$$T_1 = T_2 = ... = T_n = \frac{T}{n} = \bar{T} \tag{4.4}$$

By computing the second order partial derivatives of B which forms a matrix, and verifying that its determinant is larger than 0 at the above critical points, we have shown that the blocking probability reaches its minimum value when the traffic is evenly distributed.

∎

This theorem shows that as a result of being able to distribute traffic evenly in the system, the call blocking probability of a conventional cellular system will be minimized.

Note that, unlike a conventional system, where channel borrowing is limited by co-channel interference, an ideal n-cell iCAR system where an MH can be relayed to any BTS can be treated as a single *super* cell system with n times of DCHs. Given the same total traffic T Erlangs, the call blocking probability in the super cell is lower than that of a conventional cellular system even when the traffic is evenly distributed among the n cells. More formally, we have the following theorem.

Theorem 2 *For a given total traffic in a system, and a fixed number of DCHs in each cell, an ideal iCAR has a lower blocking probability than any conventional cellular systems (including a perfectly load-balanced one).*

Proof : Given that the iCAR system may be treated as a super cell with $T = n\overline{T}$ Erlangs and nM channels (where n is the number of cells in the system, $\overline{T}$ and M are the average traffic intensity and DCHs in each cell respectively), the blocking probability is

$$B(nM; n\overline{T}) = \frac{(n\overline{T})^{nM}/(nM)!}{\sum_{i=0}^{nM}(n\overline{T})^i/i!} \tag{4.5}$$

According to Theorem 1, the minimum blocking probability of any conventional cellular system with $\overline{T}$ Erlangs and M DCHs in each cell is

$$B(M; \overline{T}) = \frac{(\overline{T})^{M}/(M)!}{\sum_{i=0}(\overline{T})^i/i!} \tag{4.6}$$

We prove that $B(nM; nX) < B(M; X)$ for any $n, X > 1$ by showing that $\frac{1}{B(nM;nX)} > \frac{1}{B(M;X)}$ as follows.

$$\begin{aligned} \frac{1}{B(M;X)} &= \frac{M!\sum_{i=0}^{M} X^i/i!}{X^M} \\ &= \sum_{i=0}^{M} \frac{M!}{X^{M-i}i!} \\ &\overset{Let j=M-i}{=} \sum_{j=0}^{M} \frac{M(M-1)...(M-j+1)}{X^j} \end{aligned} \tag{4.7}$$

Similarly,

$$\begin{aligned} \frac{1}{B(nM;nX)} &= \sum_{j=0}^{nM} \frac{nM(nM-1)...(nM-j+1)}{(nX)^j} \\ &= \sum_{j=0}^{nM} \frac{M(M-\frac{1}{n})...(M-\frac{j+1}{n})}{X^j} \end{aligned} \tag{4.8}$$

Since every term in the above equation is positive,

$$\begin{aligned} \frac{1}{B(nM;nX)} &> \sum_{j=0}^{M} \frac{M(M-\frac{1}{n})...(M-\frac{j+1}{n})}{X^j} \\ &> \sum_{j=0}^{M} \frac{M(M-1)...(M-j+1)}{X^j} \\ &= \frac{1}{B(M;X)} \end{aligned} \tag{4.9}$$

∎

Note that the above two theorems serve as a proof of principle that iCAR can perform better than any conventional cellular systems. However, what has been implicitly assumed is that, in the ideal iCAR system, not only there are a sufficient number of ARSs, but also there is no bandwidth shortage along any relaying route such that any number of calls can be relayed through an ARS.

4.2 New Call Blocking Probability Analysis via Multi-dimensional Markov Chains

In this section, we evaluate the performance of iCAR with limited number of ARSs via a multi-dimensional Markov chain model, which is different and yields more accurate results than the analytical model presented in [15] to achieve load-balance. Readers are also referred to [91, 92] and other literatures for additional work on the performance analysis of wireless systems. Again, our discussion is under the Assumption 1 & 2.

The reduction in call blocking probability in iCAR stems largely from its ability to allow the MHs to access the channels that are not in the current cell and balance loads among cells via relaying. For simplicity, we assume that there is no bandwidth shortage along any relaying routes, and there is an ARS placed at each shared border of two cells, and its coverage is limited so that there is no overlap between any two ARSs. In other words, a MH will not be covered by more than one ARS. For the cells which have multiple ARSs, we assume all of them have the same coverage (see Figure 4.1 (a)). The ARS coverage in terms of the percentage of a cell covered by ARSs is denoted by $0 < p \leq 1$. We model the iCAR system using multi-dimensional Markov chains. For both primary and secondary relaying, we will first derive an approximation for a multi-cell system with low computing complexity, and then illustrate the general accurate solutions via a two-cell system (see Figure 4.1 (b)).

4.2.1 Primary Relaying

In this subsection, we will analyze the performance of primary relaying based on the multi-dimensional Markov chain model.

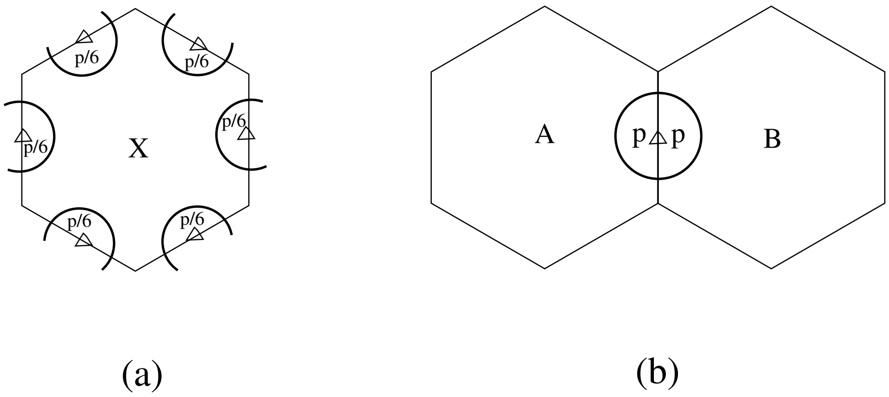


Figure 4.1: (a) The ARSs placement in cell X which has six neighbors. Each ARS covers $p/6$ percent of the total area. (b) A two cell system used for analysis. An ARS is placed at the border of two cells and covers p percent of the total area for each cell.

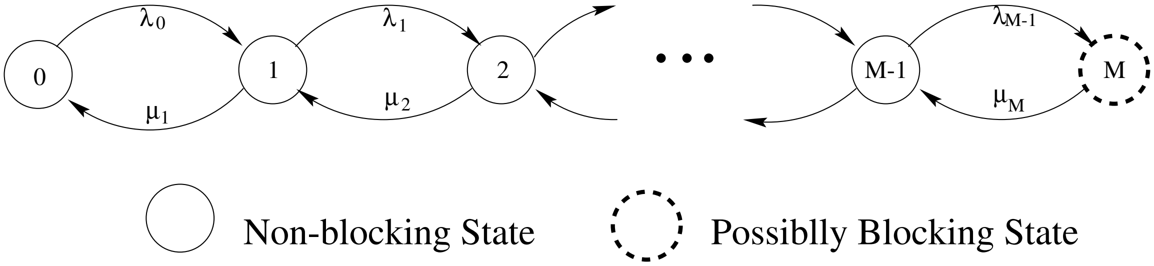


Figure 4.2: State diagram to obtain approximate modeling of primary relaying.

An approximate modeling

To obtain the approximate performance of primary relaying, we assume that when considering a cell (such as X in Figure 4.1 (a)), the traffic intensity and blocking probability of the six neighboring cells don't change as a result of relaying (this assumption will be nullified in the accurate modeling in Section 4.2.1). Let M be the number of data channels in a cell, the state diagram is shown in Figure 4.2, where state j means that there are j busy channels in the cell, λ_j and μ_j are the birth rate and death rate at state j, respectively. When $0 \leq j < M$, a state j will change to $j+1$ if a call arrives in cell X. Similarly, when a call finishes in cell X ($j > 0$), the state j will change to $j-1$.

Denote by $Q(j)$ the steady state probability that the system is at state j. According to the state diagram, we can write the following state equations.

For non-boundary states: $0 < j < M$

$$(\mu_j + \lambda_j) \cdot Q(j) - \lambda_{j-1} \cdot Q(j-1) - \mu_{j+1} \cdot Q(j+1) = 0 \tag{4.10}$$

For boundary states:

$j = 0$:

$$\lambda_0 \cdot Q(0) - \mu_1 \cdot Q(1) = 0 \tag{4.11}$$

$j = M$:

$$\mu_M \cdot Q(M) - \lambda_{M-1} \cdot Q(M-1) = 0 \tag{4.12}$$

In addition,

$$\sum_{j=0}^{M} Q(j) = 1 \tag{4.13}$$

In order to simplify the problem further, we use a few classic assumptions, which are also used to derive the Erlang B formula. More specifically, we assume the probability of a new call coming is independent of the number of busy sources, i.e., $\lambda_j = \lambda$ for some λ; and also, the death rate is proportional to the number of busy sources, i.e., $\mu_j = j\mu$ for some μ. Note that the above state diagram and equations are indeed the same as those for a conventional cellular system. By plugging the assumptions into Equations 4.10 through 4.13, we can obtain $M + 1$ equations. Solving them, we get $Q(j) = \frac{T^j/j!}{\sum_{i=0}^{j} T^i/i!}$ in which $T = \frac{\lambda}{\mu}$. .

Recall that, by using primary relaying, a call will be blocked if and only if it arrives when the cell X is at state M and the corresponding MH is not covered by ARS, or even if it is covered by ARS, the reachable neighboring BTS is also congested. Assuming that the average blocking probability of all six neighboring cells is b, the blocking probability in cell X with primary relaying is (approximately)

$$b_{X\_Primary} = Q(M) \times [(1-p) + p \times b] = \frac{T^M/M!}{\sum_{i=0}^{M} T^i/i!} \times [(1-p) + p \times b] \tag{4.14}$$

Note that, the equation can be adapted to a cell when there are unevenly distributed ARS coverage along the borders (e.g. p_k instead of $p/6$ where $\sum p_k = p$) and there are $l \leq 6$ neighboring cells, which may have different traffic intensity and thus blocking probability b_k, by replacing the $p \times b$ with $\sum_{k=0}^{l} p_k \cdot b_k$.

An accurate modeling

In the above approximate model, we ignored the effect of relaying on neighboring cells. In order to obtain accurate results, we need to keep track the number of active channels in not only the cell to be considered but also all of its reachable neighboring cells. Here, we discuss the general solution via a two-cell system (see Figure 4.1 (b)). The state diagram is shown in Figure 4.3 where a state (i, j) represents the state that there are i busy channels in cell A and j busy channels in cell B. $\lambda_{A(i,j)}$ and $\lambda_{B(i,j)}$ are the birth rate of new calls in cell A and cell B at state (i, j), respectively, while $\mu_{A(i,j)}$ and $\mu_{B(i,j)}$ are the death rate in cell A and B.

When $0 \leq i, j < M$, the states and their transitions are the same as those for a conventional cellular system. More specifically, a state (i, j) will change to $(i, j + 1)$ if a call arrives in cell B, or to $(i + 1, j)$ if a call arrives in cell A. Similarly, when a call finishes in cell A or B, the state (i, j) will change to $(i - 1, j)$ or $(i, j - 1)$. When $i = M$ and a new call arrives in cell A, it will be blocked in the conventional cellular system. But by using primary relaying, the call may be relayed to cell B if the MH is covered by ARS and cell B is not congested ($j < M$). In other words, when a new call is generated in cell A at state (M, j) where $j < M$, the state may change to $(M, j + 1)$ with a probability p (see the arrows at the top of Figure 4.3). Similarly, as a result of relaying a call from cell B to cell A, state (i, M) where $i < M$ may change to $(i + 1, M)$ (see the arrows at the right side of Figure 4.3).

Denote by $Q(i, j)$ the steady state probability that the system is at state (i, j). For given M, p, $\lambda_{A(i,j)}$, $\lambda_{B(i,j)}$, $\mu_{A(i,j)}$ and $\mu_{B(i,j)}$[1], one can obtain $Q(i, j)$ by solving a set of $(M + 1) \times (M + 1)$ equations, one for each state (similar to those in Sec 4.2.1). In the two-cell system with primary relaying, a call will be blocked if (1) the current state is (M, M), or (2) the current state is (M, j) or (i, M) and the corresponding MH is not covered by ARS. More specifically, the blocking probabilities of cell A and B with primary relaying are

$$b_{A\_Primary} = Q(M, M) + \sum_{j=0}^{M-1} Q(M, j) \cdot (1 - p) \tag{4.15}$$

[1] It's often reasonable as well to assume $\lambda_{A(i,j)} = \lambda_A$, $\lambda_{B(i,j)} = \lambda_B$, $\mu_{A(i,j)} = i\mu$ and $\mu_{B(i,j)} = j\mu$ for some μ.

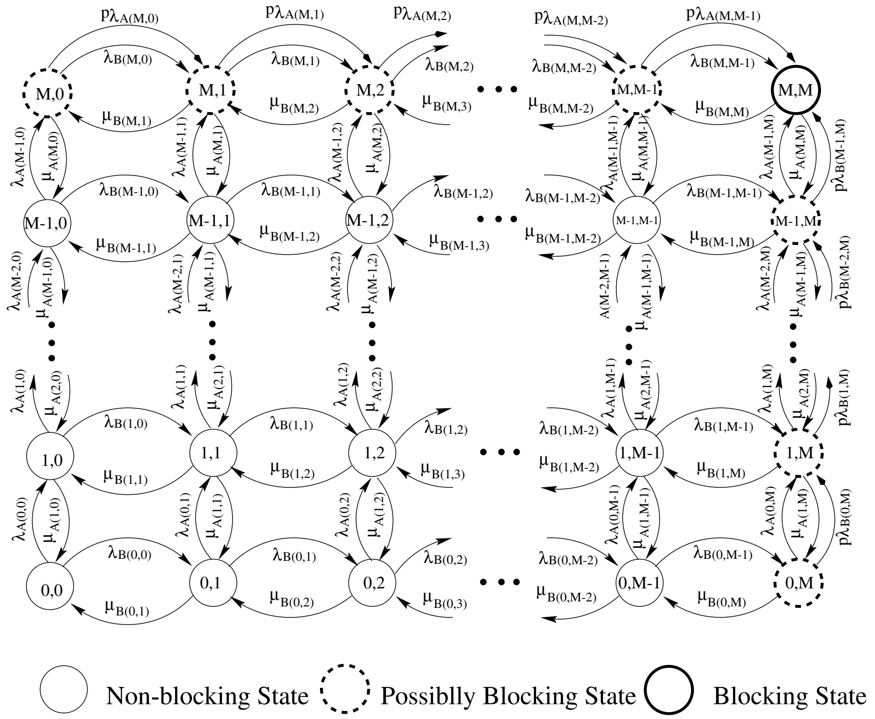


Figure 4.3: State diagram to obtain an accurate modeling of primary relaying in a two-cell system.

$$b_{B\_Primary} = Q(M, M) + \sum_{i=0}^{M-1} Q(i, M) \cdot (1 - p) \tag{4.16}$$

Note that, the equation can also be adapted in case the ARS coverage in cell A and B are p_A and p_B, respectively (with corresponding changes in the transitions in Figure 4.3).

For a $k-$cell system, the general solution needs a $k-$dimensional state diagram. When k is large, it becomes quite complicated and time consuming to construct the state diagram and solve the corresponding equations. But, if the traffic load in the system is only reasonable high (but not too high), the arrival rate of relayed calls in a cell is much lower than the arrival rate of the *native* calls which is generated by the MHs within the cell. Then, we can analyze a cell separately from other cells in the system, and thereby simplifying a $k-$dimensional chain to a one-dimensional chain as we discussed in Sec 4.2.1 (See also Sec 4.2.3 for more discussion).

4.2.2 Secondary Relaying

To analyze the performance of secondary relaying, we need to keep track not only the number of active channels in each cell as we did in Sec 4.2.1, but also the number of active MHs which are covered by ARS and directly using a cellular channel. This can be accomplished by using a two dimensional state diagram to model *each cell*. One for active MHs covered by ARSs and without relaying, and the other for all active MHs. Again, we first show the approximate approach under simplified assumptions.

An approximate modeling

Based on the similar assumptions we have discussed in Sec 4.2.1 for primary relaying, we draw the simplified state diagram as shown in Figure 4.4. A state (i, j) $(i \leq j)$ in Figure 4.4 means that there are j busy channels and i of them can be released via relaying (i.e. the corresponding MHs are covered by ARS). Similar to Figure 4.2, let $\lambda_{i,j}$ be the birth rate at state (i, j). Then, $p\lambda_{i,j}$ is the arrival rate of calls covered by ARSs, while $(1 - p)\lambda_{i,j}$ is the arrival rate of calls not covered by ARSs if MHs are evenly distributed in each cell. $\mu_{i,j}$ is the death rate of active MHs covered by ARS at state (i, j), and $\bar{\mu}_{i,j}$ is the death rate of active MH not covered by ARS at state (i, j). When $j < M$ and a new call comes in cell X at state (i, j), it will change to $(i + 1, j + 1)$ if the corresponding MH is covered by ARS, or change to $(i, j + 1)$ if it is not covered by ARS. When $j > 0$ and a call finishes in cell X at state (i, j), it will change to $(i - 1, j - 1)$ if the corresponding MH is covered by ARS and was directly using a DCHs to access the system (which implies $i > 0$), or change to $(i, j - 1)$ otherwise. When $j = M, i > 0$ and a new call comes in cell X at state (i, j), it may change to $(i - 1, M)$ if either primary relaying (with a probability of $(1 - p) + pb$) or secondary relaying (with a probability of $1 - b^j$) successes. Let $Q(i, j)$ be the probability that the system is at state (i, j), and b is the average blocking probability of neighboring cells. According to the state diagram, we can write the following state equations.

For non-boundary states: $0 < i \leq j < M$ (refer to state $(1, M - 1)$ in Figure 4.4.)

$$(\mu_{i,j} + \bar{\mu}_{i,j} + \lambda_{i,j}) \cdot Q(i, j) - p \cdot \lambda_{i-1,j-1} \cdot Q(i - 1, j - 1) -$$
$$(1 - p) \cdot \lambda_{i,j-1} \cdot Q(i, j - 1) - \bar{\mu}_{i,j+1} \cdot Q(i, j + 1) - \mu_{i+1,j+1} \cdot Q(i + 1, j + 1) = 0 \quad (4.17)$$

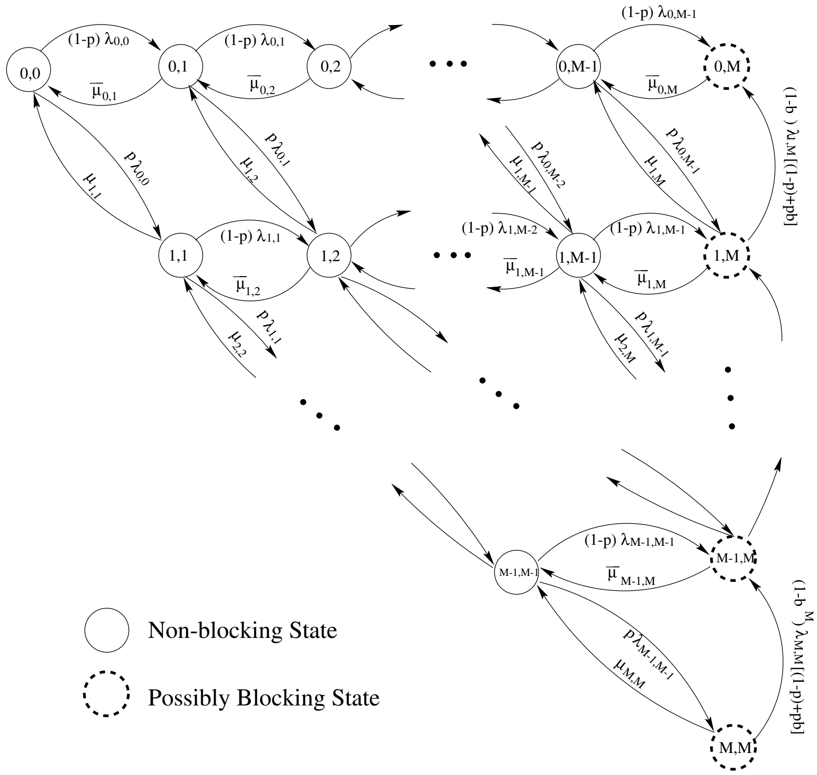


Figure 4.4: State diagram to obtain approximate modeling of secondary relaying.

For boundary states: $i = j = 0:$

$$\lambda_{0,0} \cdot Q(0,0) - \bar{\mu}_{0,1} \cdot Q(0,1) - \mu_{1,1} \cdot Q(1,1) = 0 \tag{4.18}$$

$i = 0,\ j = M$:

$$\begin{aligned} &\bar{\mu}_{0,M} \cdot Q(0,M) - (1-p) \cdot \lambda_{0,M-1} \cdot Q(0,M-1) - \\ &\quad (1-b) \cdot \lambda_{1,M} \cdot [(1-p) + p \times b] \cdot Q(1,M) = 0 \end{aligned} \tag{4.19}$$

$i = j = M:$

$$\begin{aligned} &(\mu_{M,M} + (1-b^M) \cdot \lambda_{M,M}) \cdot [(1-p) + p \times b] \cdot Q(M,M) - \\ &\quad p \cdot \lambda_{M-1,M-1} \cdot Q(M-1,M-1) = 0 \end{aligned} \tag{4.20}$$

$i = 0,\ 0 < j < M:$

$$\begin{aligned} &(\bar{\mu}_{0,j} + \lambda_{0,j}) \cdot Q(0,j) - (1-p) \cdot \lambda_{0,j-1} \cdot Q(0,j-1) - \\ &\quad \mu_{1,j+1} \cdot Q(1,j+1) - \bar{\mu}_{0,j+1} \cdot Q(0,j+1) = 0 \end{aligned} \tag{4.21}$$

$0 < i < M,\ j = M:$

$$\begin{aligned}(\mu_{i,M} + \bar{\mu}_{i,M} + (1-b^i)\cdot\lambda_{i,M}[(1-p)+p\times b])\cdot Q(i,M)-\\(1-p)\cdot\lambda_{i,M-1}\cdot Q(i,M-1) - p\cdot\lambda_{i-1,M-1}\cdot Q(i-1,M-1)-\\(1-b^{i+1})\cdot\lambda_{i+1,M}\cdot[(1-p)+p\times b]\cdot Q(i+1,M) = 0\end{aligned} \tag{4.22}$$

$0 < i = j < M:$

$$\begin{aligned}(\mu_{j,j} + \lambda_{j,j})\cdot Q(j,j) - p\cdot\lambda_{j-1,j-1}\cdot Q(j-1,j-1)-\\\bar{\mu}_{j,j+1}\cdot Q(j,j+1) - \mu_{j+1,j+1}\cdot Q(j+1,j+1) = 0\end{aligned} \tag{4.23}$$

In addition,

$$\sum_{i=0}^{M}\sum_{j=0}^{M} Q(i,j) = 1 \tag{4.24}$$

Similar to the case for primary relaying, we assume (1) the probability of a new call coming is independent of the number of busy source, i.e. $\lambda_{i,j} = \lambda$; (2) the death rate is proportional to the number of busy sources, i.e. $\mu_{i,j} = i\mu$, and $\bar{\mu}_{i,j} = (j-i)\mu$. By Plugging these value into Equations 4.17 through 4.24, we can get $(M+1)(M+2)/2$ equations. Solving them, we get $Q(i,j)$ for $0 \leq i \leq j \leq M$.

Since a new call will be blocked if and only if (1) the current state is (i, M), and (2) primary relaying is failed, and (3) secondary relaying is failed either (none of the i MHs which are covered by ARS, can find a non-congested reachable cell), the approximation of blocking probability in cell X after secondary relaying is

$$b_{X\_Secondary} = \sum_{i=0}^{M} Q(i,M)\times b^i \times [(1-p)+p\times b] \tag{4.25}$$

As in the case for primary relay, it is also possible to extend Equation 4.25 when cell X is surrounded by less than 6 neighbors, which have different traffic intensity and blocking probabilities.

An accurate model

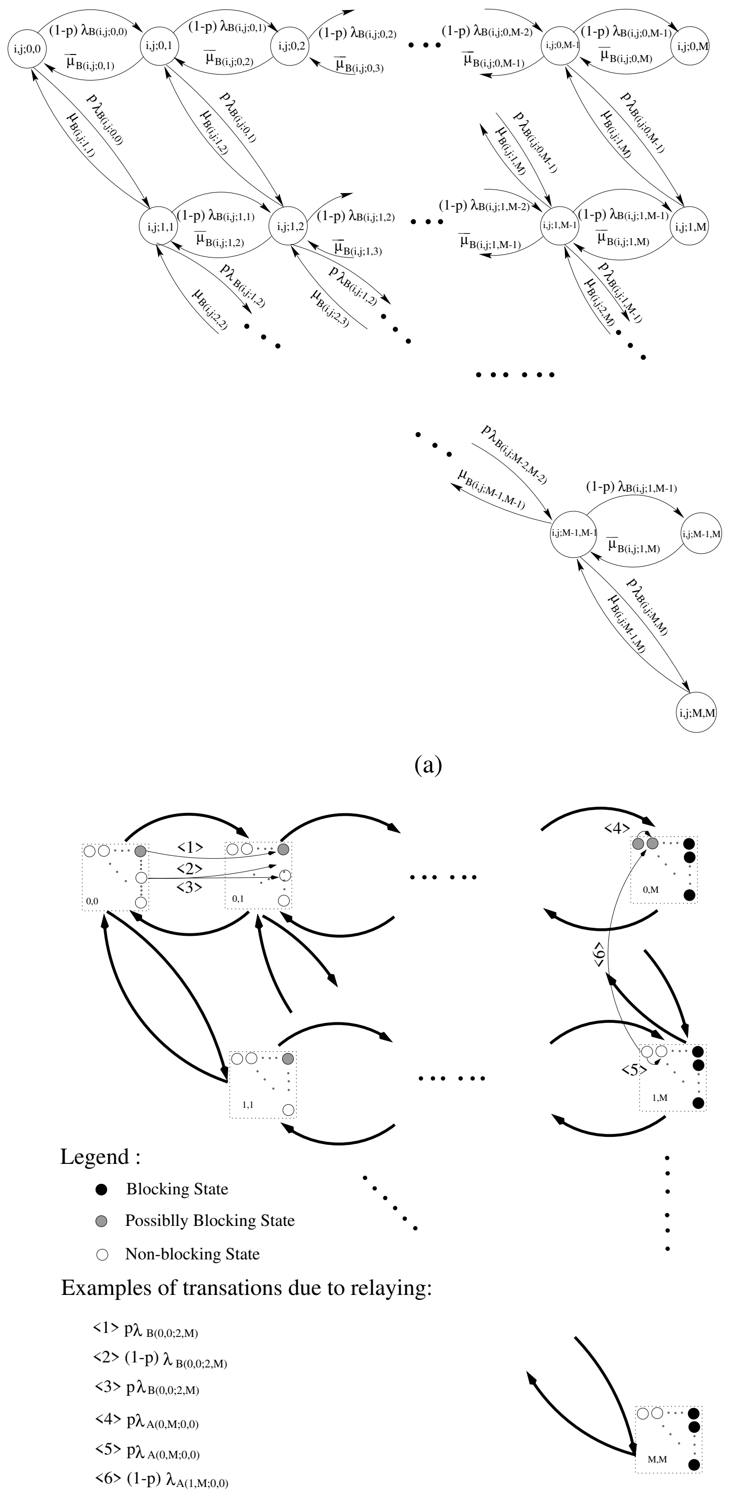
The state diagram which has $4-$dimensions to take the effect of relaying on the neighboring cell (in a two cell system) into consideration is sketched in Figure 4.5, where a state $(i, j; s, t)$ means that there are j and t active MHs (each using a DCH) in cell A and B respectively, of which $i \leq j$ and $s \leq t$ are covered by ARS respectively. $\lambda_{A(i,j;s,t)}$ and $\lambda_{B(i,j;s,t)}$ are the birth rate of new calls at state $(i, j; s, t)$ in cell A and cell B at state $(i, j; s, t)$, respectively. Similar to the approximate approach, $p\lambda_{A(i,j;s,t)}$ and $p\lambda_{B(i,j;s,t)}$ are the arrival rates of calls covered by ARSs, while $(1-p)\lambda_{A(i,j;s,t)}$ and $(1-p)\lambda_{B(i,j;s,t)}$ are the arrival rates of calls not covered by ARSs. $\mu_{A(i,j;s,t)}$ and $\mu_{B(i,j;s,t)}$ are the death rate of active MHs which are covered by ARSs in cell A and B. $\bar{\mu}_{A(i,j;s,t)}$ and $\bar{\mu}_{B(i,j;s,t)}$ are the death rate of active MHs which are not covered by ARSs.

Figure 4.5 (a) shows a subset of $\frac{M(M+1)}{2}$ states, and the transitions among them due to call arrival/departure in cell B when cell A has j active channels and i of them can be released via relaying. For instance, when $t < M$ and a new call comes in cell B at state $(i, j; s, t)$, it will change to $(i, j; s+1, t+1)$ if the corresponding MH is covered by ARS, or change to $(i, j; s, t+1)$ if it is not covered by ARS. When $t > 0$ and a call finishes in cell B at state $(i, j; s, t)$, it will change to $(i, j; s-1, t-1)$ if the corresponding MH is covered by ARS and $s > 0$, or change to $(i, j; s, t-1)$ otherwise.

If we treat the two-dimensional diagram in Figure 4.5 (a) as a cluster (i, j), we can construct the state diagram for the entire 2-cell system as shown in Figure 4.5 (b) where different clusters represent different i and j combinations. The two thick arrows between a pair of clusters represent two groups of transitions between all the corresponding states in the two clusters. For example, the thick arrow from cluster $(0, 0)$ to cluster $(0, 1)$ includes the $\frac{(M+1)(M+2)}{2}$ transitions from $(0, 0; 0, 0)$ to $(0, 1; 0, 0)$, from $(0, 0; 0, 1)$ to $(0, 1; 0, 1)$, ..., from $(0, 0; s, t)$ to $(0, 1; s, t)$, ..., and from $(0, 0; M, M)$ to $(0, 1; M, M)$. Since s and t are fixed, and only i and j can vary, the group transitions in each thick arrow are actually very similar to those intra-cluster transitions shown in Figure 4.5 (a) where i and j are fixed and only s and t can vary.

In addition to the transitions depicted by the thick arrows, there are other transitions between the two states due to relaying as follows.

- when $t = M$, $s = 0$, $j < M$ and a new call comes to cell B, the state may change from $(i, j; 0, M)$ to $(i, j+1; 0, M)$ with a probability of p via primary relaying. (see transition 1 in Figure 4.5 (b) for example).



(a)

(b)

Figure 4.5: State diagram to obtain an accurate modeling of secondary relaying in a two-cell system.

- when $t = M$, $s > 0$, $j < M$ and a new call comes to cell B, the state may change from $(i, j; s, M)$ to $(i, j + 1; s, M)$ with a probability of p via primary relaying. (see transition 3 in Figure 4.5 (b) for example). If primary relaying fails, the state will change to $(i, j + 1; s - 1, M)$ (see transition 2 in Figure 4.5 (b) for example).
- when $j = M$, $i = 0$, $t < M$ and a new call comes to cell A, the state may change from $(0, M; s, t)$ to $(0, M; s, t + 1)$ with a probability of p via primary relaying. (see transition 4 in Figure 4.5 (b) for example).
- when $j = M$, $i > 0$, $t < M$ and a new call comes to cell A, the state may change from $(i, M; s, t)$ to $(i, M; s, t + 1)$ with a probability of p via primary relaying. (see transition 5 in Figure 4.5 (b) for example). If primary relaying fails, the state will change to $(i - 1, M; s, t + 1)$ (see transition 6 in Figure 4.5 (b) for example).

Let $Q(i, j; s, t)$ be the probability that the system is at state $(i, j; s, t)$, for given M, p, $\lambda_{A(i,j;s,t)}$, $\lambda_{B(i,j;s,t)}$, $\mu_{A(i,j;s,t)}$ and $\mu_{B(i,j;s,t)}$, we can obtain $Q(i, j; s, t)$ by solving a set of equations, one for each state (although this might be time-consuming which is why we may use the approximate model described earlier in Sec 4.2.2). In a system applying secondary relaying[2], a call will be blocked if (1) $j = t = M$, or (2) a new call comes to cell B at state $(i, j; 0, M)$ with $j < M$ and the corresponding MH is not covered by ARS, or (3) a new call comes to cell A at state $(0, M; s, t)$ and the corresponding MH is not covered by ARS. More specifically, the blocking probabilities of cell A and B with secondary relaying are

$$b_{A\_Secondary} = \sum_{i=0}^{M} \sum_{s=0}^{M} Q(i, M; s, M) + \sum_{t=0}^{M-1} \sum_{s=0}^{t} Q(0, M; s, t) \cdot (1 - p) \tag{4.26}$$

$$b_{B\_Secondary} = \sum_{i=0}^{M} \sum_{s=0}^{M} Q(i, M; s, M) + \sum_{j=0}^{M-1} \sum_{i=0}^{t} Q(i, j; 0, M) \cdot (1 - p) \tag{4.27}$$

4.2.3 Numeric Results

By plugging in reasonable values of parameters in Equations 4.14, and 4.17 through 4.25, we obtain numeric results to show the performance improvement in terms of new call blocking probability by

[2] When using secondary relaying, it implies that primary relaying is also used.

using the iCAR system. More specifically, we consider a 19-cell system shown in Figure 4.19 and assume that there are $M = 50$ DCHs in each cell. The traffic intensities in cells A, tier B and tier C cells are T_A, T_B and T_C Erlangs respectively with an average holding time 120 $seconds$. The blocking probabilities of the three tier cells without relaying are denoted as B_A, B_B and B_C. When we consider cell A, the average blocking probability of neighboring cells is B_B. When we consider tier B cells, the average blocking probability of neighboring cells is$\frac{1}{6}B_A + \frac{2}{6}B_B + \frac{3}{6}B_C$. We will study three scenarios as follows.

Scenario 1: vary the traffic intensity of the entire system In this scenario, we assume the traffic intensity to be location-dependent. More specifically, it decreases at a rate of 0.8 from one tier of cells to another, which means that $T_B = 0.8T_A$ and $T_C = 0.8T_B$. Assuming that T_A increase from about 41 $Erlangs$ to about 53 $Erlangs$, T_B and T_C also increase accordingly. The results for cell A and tier B cells are shown in Figure 4.6 (a) and (b), respectively. As we can see, with any increase of traffic intensity, the blocking probability in cell A will exceed the acceptable level(usually 2%), and can be as high as about 15% when $T_A = 53$ $Erlangs$. With relaying, especially secondary relaying, we can significantly reduce the new call blocking probability in both cell A and tier B cells, and therefore increase the system capacity.

We also plot the simulation results (to be discussed later in Section 4.5 in the Figure 4.6 as a comparison. They were obtained from a system similar to the one used here, and with the same value of M and p (see Section 4.5 for more details of simulation). When traffic intensity is not very high ($T_A < 50$ $Erlangs$), the analysis results match with simulation results very well for both primary and secondary relaying, in both cell A and tier B cells. When $T_A > 50$ $Erlangs$, the difference between analysis results and simulation results on the blocking probability in cell A with secondary relaying increases. Such a difference is due to the fact that we have assumed that neighboring tier B cells are not affected by the relayed traffic in the simplified analytical model. Since when $T_A > 50$ $Erlangs$, cell A is heavily congested with a blocking probability higher than 10% without relaying and even with secondary relaying the blocking probability is above 2%, it is likely that a wireless system won't operate under such a heavy traffic load. Therefore, the simplified analysis model is good enough within a reasonable operating range.

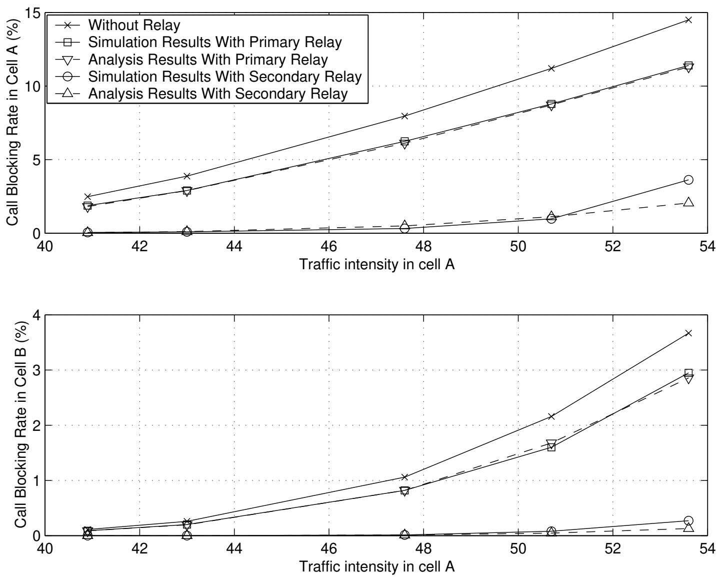


Figure 4.6: Scenario 1: blocking probability in cell A and cell B

Scenario 2: vary the traffic intensity in cell A and tier B cells In this scenario, we study the performance of iCAR with different traffic intensity in cell A and tier B cells. We first fix the traffic intensity in tier B and tier C cells and increase T_A. The blocking probability of cell B's and C's without relaying are assumed to be 2% and 1%, which correspond to $T_B = 40.25$ $Erlangs$ and $T_C = 37.90$ $Erlangs$ respectively. The traffic intensity in cell A (T_A) increases from 40.25 $Erlangs$ (which corresponding to about 2% blocking probability in cell A without relaying) to as high as 49.25 $Erlangs$. The blocking probability of cell A and cell B due to relaying are shown in Figure 4.7 (a) and (b). Similar to Figure 4.6, with the increase in traffic intensity, the blocking probability in cell A due to secondary relaying is much lower than that without relaying.

In Figures 4.7 (c) and (d) show the results when we fix the traffic intensity of cell A and tier C cells, and increase T_B. As we can see, the blocking probability of cell A is not affected by the increasing traffic intensity in tier B cells, although B_B increases with T_B.

Scenario 3: vary the ARS coverage p In this scenario, we fix T_A, T_B and T_C. The blocking probability of cell A, B's and C's without relaying are assumed to be 5%, 2% and 1%, which correspond to $T_a = 44.5$ $Erlangs$, $T_B = 40.25$ $Erlangs$ and $T_C = 37.90$ $Erlangs$, respectively. The ARS coverage p increases from 0.1 to 0.68 which is the maximum ARS coverage so that the ARSs

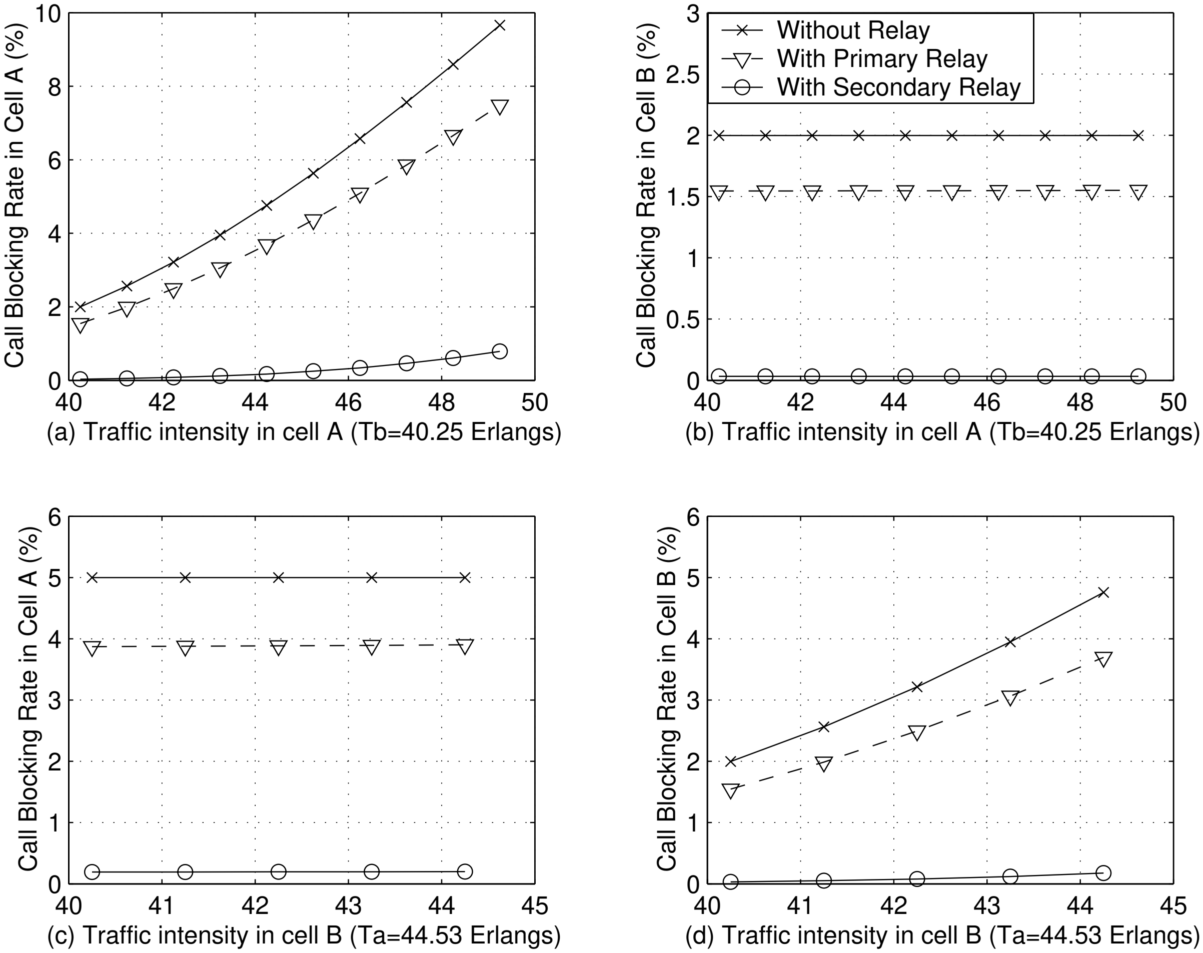


Figure 4.7: Scenario 2: blocking probability in cell A and B

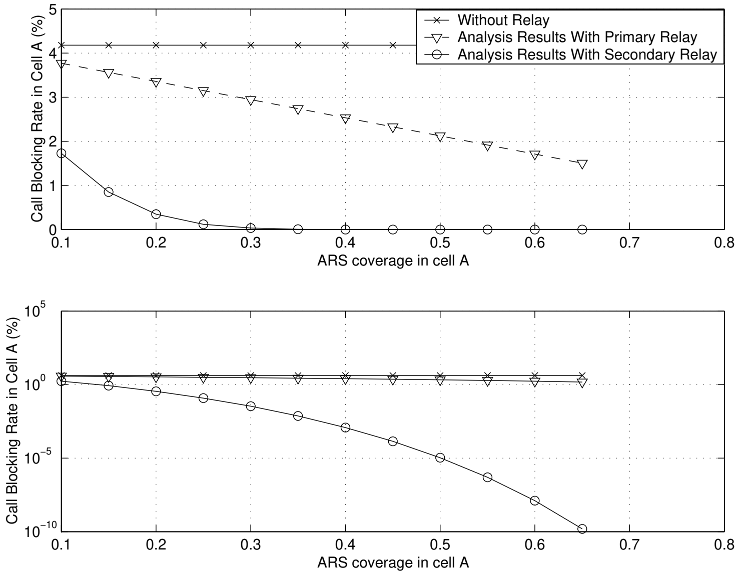


Figure 4.8: Scenario 3: blocking probability in cell A

don't overlap. The results are shown in Figure 4.8. We plot the results using both the normal scale (upper) and log-scale (lower) for clarity. With the increase of ARS coverage, the blocking probability of primary relaying decreases linearly, while the blocking probability of secondary relaying decreases exponentially. As can be seen, by using secondary relaying and with a large enough ARS coverage, the hot-spot in iCAR can be effectively eliminated.

4.3 New Call Queuing Delay Analysis via Multi-dimensional Markov Chains

In this section, we consider an iCAR system with queuing capability, and analyze the waiting time and the waiting probability of a new call request. Our discussion will be based on the Assumption 1 discussed in Section 4.1 and another assumption as follows.

Assumption 3 : *The number of calls in progress simultaneously is at most M. Calls arriving when all M channels are occupied form a queue (with infinite buffer size) and wait in the order of their arrival for free channels (i.e., a Fist In Fist Out (FIFO) queue).*

In an iCAR system, a new call request will wait if there is not DCH available at the BTS and

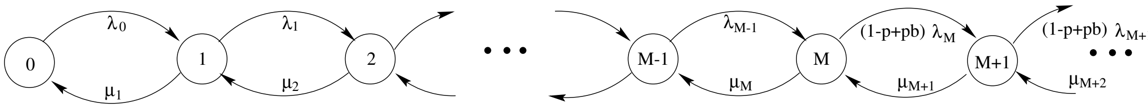


Figure 4.9: State diagram to obtain approximate modeling of primary relaying (delay analysis).

both primary and secondary relaying are failed. However, a request via relaying will not be queued, i.e., it will be rejected immediately if there is no free DCH.

Similar to the analytical model discussed in Section 4.2, we assume that there is no bandwidth shortage along any relaying routes, and one seed ARSs is placed at each shared border of two cells. The ARS coverage in terms of the percentage of a cell covered by ARSs is denoted by $0 < p \leq 1$. M is the number of data channels in a cell. For simplicity, we only consider the approximate model. More specifically, we assume that when considering a cell (such as X in Figure 4.1 (a)), the traffic intensities of the six neighboring cells are equal and don't change as a result of relaying. According to Erlang C formula [90], the probability that all channels are busy in a neighboring cell of cell X (e.g., cell Y) at an arbitrary instant is,

$$b = \frac{\frac{T_Y^M}{M!} \cdot \frac{M}{M-T_Y}}{1 + T_Y + \frac{T_Y^2}{2!} + ... + \frac{T_Y^{M-1}}{(M-1)!} + \frac{T_Y^M}{M!} \cdot \frac{M}{M-T_Y}} \tag{4.28}$$

where T_Y is the traffic intensity of the cell Y.

4.3.1 Primary Relaying

The state diagram for primary relaying is shown in Figure 4.9, where state j means that there are j calls being served or waiting in the queue, λ_j and μ_j are the birth rate and death rate at state j, respectively. When $0 \leq j < M$, a state j will change to $j + 1$ if a call arrives in cell X. Similarly, when a call finishes in cell X ($j > 0$), the state j will change to $j - 1$. When the current state is $j \geq M$, a new call request will be relayed to the neighboring cell if the corresponding MH is covered by ARSs and the neighboring cell has free DCHs (with a probability of $p(1 - b)$). Otherwise, the request will be put into the queue, i.e., state j will change to state $j + 1$ (with a probability of $(1 - p + pb)$).

Denote by $Q(j)$ the steady state probability that the system is at state j. According to the state diagram, we can write the following state equations.

$j = 0$ **:**

$$\lambda_0 \cdot Q(0) - \mu_1 \cdot Q(1) = 0 \tag{4.29}$$

$0 < j < M$**:**

$$(\mu_j + \lambda_j) \cdot Q(j) - \lambda_{j-1} \cdot Q(j-1) - \mu_{j+1} \cdot Q(j+1) = 0 \tag{4.30}$$

$j = M$ **:**

$$[\mu_M + (1 - p + pb)] \cdot Q(M) - \lambda_{M-1} \cdot Q(M-1) - \mu_{M+1} Q(M+1) = 0 \tag{4.31}$$

$j > M$ **:**

$$[\mu_j + (1 - p + pb)] \cdot Q(j) - (1 - p + pb)\lambda_{j-1} \cdot Q(j-1) - \mu_{j+1} Q(j+1) = 0 \tag{4.32}$$

In addition,

$$\sum_{j=0}^{\infty} Q(j) = 1 \tag{4.33}$$

Similar to the analytical model discussed in Section 4.2, we assume the probability of a new call coming is independent of the number of busy sources, i.e. $\lambda_j = \lambda$ for some λ; and also, the death rate is proportional to the number of busy sources, i.e. $\mu_j = j\mu$ if $j < M$, and $\mu_j = M\mu$ if $j \geq M$, for some μ. Solving the above state equations, we can obtain the probability of each state ($Q(j)$, $j \geq 0$), and accordingly compute the call waiting time and waiting probability.

The probability that exactly k calls end during the time t is given by the Poisson distribution with the parameter μM [90]. Thus, given the current state to be $j \geq M$, the probability that the waiting time of a new call is longer than time t, or in other words, the probability of $j - M$ or less calls terminating during the time t, is

$$P_j(t) = \sum_{k=0}^{j-M} \frac{(\mu M t)^k}{k!} e^{-\mu M t},\ j \geq M \tag{4.34}$$

The summation through all j yields the probability of delay exceeding t for an incoming call in the iCAR system with primary relaying:

$$W_p(t) = \sum_{j=M}^{\infty} Q(j) P_j(t) \tag{4.35}$$

Accordingly, the average waiting time of an incoming call is given by

$$\overline{W}_p = \int_0^\infty tW_p(t)dt \tag{4.36}$$

4.3.2 Secondary Relaying

Figure 4.10 shows the state diagram for the secondary relaying. A state (i, j) $(i \leq j)$ in Figure 4.4 means that there are j calls being served or waiting in the queue and i of the calls being served can be released via relaying (i.e. the corresponding MHs are covered by ARS). Similar to that discussed in Section 4.2, let $\lambda_{i,j}$ be the birth rate at state (i, j). Then, $p\lambda_{i,j}$ is the arrival rate of calls covered by ARSs, while $(1 - p)\lambda_{i,j}$ is the arrival rate of calls not covered by ARSs, if MHs are evenly distributed in each cell. $\mu_{i,j}$ is the death rate of active MHs covered by ARS at state (i, j), and $\bar{\mu}_{i,j}$ is the death rate of active MH not covered by ARS at state (i, j).

When $j < M$ and a new call comes in cell X at state (i, j), it will change to $(i + 1, j + 1)$ if the corresponding MH is covered by ARS, or change to $(i, j + 1)$ if it is not covered by ARS. When $M \geq j > 0$ and a call finishes in cell X at state (i, j), it will change to $(i - 1, j - 1)$ if the corresponding MH is covered by ARS and was directly using a DCHs to access the system, or change to $(i, j - 1)$ otherwise.

When $j \geq M$ and a new call comes in cell X at state (i, j), it may change to $(i - 1, j)$ if primary relaying fails but secondary relaying successes (with a probability of $(1 - b^i)(1 - p + pb)$). Otherwise, if both primary and secondary relaying fail, the state (i, j) will change to state $(i, j + 1)$ with a probability of $b^i(1 - p + pb)$. When a call ends, the state (i, j) may change to three possible states: (1) if the MH corresponding to the call ended (denoted by MH_d) is covered by ARSs and the MH corresponding to the first call request in the queue (denoted by MH_f) is not covered by ARSs, the state (i, j) will change to state $(i - 1, j - 1)$; (2) if both MH_d and MH_f are not covered by ARSs, or both MH_d and MH_f are covered by ARSs, the state (i, j) will change to state $(i, j - 1)$; (3) if MH_d is not covered by ARSs but MH_f is covered by ARSs, the state (i, j) will change to state $(i + 1, j - 1)$.

Let $Q(i, j)$ be the probability that the system is at state (i, j), we can write the following state equations according to the state diagram.

$i = j = 0$ **:**

$$\lambda_{0,0} \cdot Q(0,0) - \bar{\mu}_{0,1} \cdot Q(0,1) - \mu_{1,1} \cdot Q(1,1) = 0 \tag{4.37}$$

$i = 0,\ 0 < j < M$**:**

$$\begin{gathered}(\bar{\mu}_{0,j} + \lambda_{0,j}) \cdot Q(0,j) - (1-p) \cdot \lambda_{0,j-1} \cdot Q(0,j-1) - \\ \mu_{1,j+1} \cdot Q(1,j+1) - \bar{\mu}_{0,j+1} \cdot Q(0,j+1) = 0\end{gathered} \tag{4.38}$$

$i = 0,\ j = M$**:**

$$\begin{gathered}[\bar{\mu}_{0,M} + (1-p+pb)] \cdot Q(0,M) - (1-p) \cdot \lambda_{0,M-1} \cdot Q(0,M-1) - \\ (1-b) \cdot \lambda_{1,M} \cdot [(1-p) + p \times b] \cdot Q(1,M) - \\ (1-p)\mu_{1,M+1}Q(1,M+1) - (1-p)\bar{\mu}_{0,M+1}Q(0,M+1) = 0\end{gathered} \tag{4.39}$$

$i = 0,\ j > M$**:**

$$\begin{gathered}[\bar{\mu}_{0,j} + (1-p+pb)] \cdot Q(0,j) - (1-p+pb) \cdot \lambda_{0,j-1} \cdot Q(0,j-1) - \\ (1-b) \cdot \lambda_{1,j} \cdot [(1-p) + p \times b] \cdot Q(1,j) - \\ (1-p)\mu_{1,j+1}Q(1,j+1) - (1-p)\bar{\mu}_{0,j+1}Q(0,j+1) = 0\end{gathered} \tag{4.40}$$

$i = j = M$ **:**

$$\begin{gathered}(\mu_{M,M} + \lambda_{M,M} \cdot [(1-p) + p \times b]) \cdot Q(M,M) - p \cdot \lambda_{M-1,M-1} \cdot Q(M-1,M-1) - \\ p\mu_{M,M+1}Q(M,M+1) - p\bar{\mu}_{M-1,M+1}Q(M-1,M+1) = 0\end{gathered} \tag{4.41}$$

$i = M,\ j > M$**:**

$$\begin{gathered}(\mu_{M,j} + \lambda_{M,j} \cdot [(1-p) + p \times b]) \cdot Q(M,j) - p\mu_{M,j+1}Q(M,j+1) - \\ p\bar{\mu}_{M-1,j+1}Q(M-1,j+1) - b^{M}(1-p+pb) \cdot \lambda_{M,j-1} \cdot Q(M,j-1) = 0\end{gathered} \tag{4.42}$$

$0 < i < M,\ j = M$**:**

$$\begin{gathered}(\mu_{i,M} + \bar{\mu}_{i,M} + \lambda_{i,M}[(1-p) + p \times b]) \cdot Q(i,M) - (1-p) \cdot \lambda_{i,M-1} \cdot Q(i,M-1) - \\ (1-b^{i+1}) \cdot \lambda_{i+1,M} \cdot [(1-p) + p \times b] \cdot Q(i+1,M) - \\ p\bar{\mu}_{i-1,M+1}Q(i-1,M+1) - [p\mu_{i,M+1} + (1-p)\bar{\mu}_{i,M+1}]Q(i,M+1) - \\ p \cdot \lambda_{i-1,M-1} \cdot Q(i-1,M-1) - (1-p)\mu_{i,M+1}Q(i+1,M+1) = 0\end{gathered} \tag{4.43}$$

$0 < i < M,\ j > M$:

$$
\begin{aligned}
(\mu_{i,j} + \bar{\mu}_{i,j} + \lambda_{i,j}[(1-p) + p \times b]) \cdot Q(i, M) - \\
(1 - b^{i+1}) \cdot \lambda_{i+1,j} \cdot [(1-p) + p \times b] \cdot Q(i+1, j) - \\
p\bar{\mu}_{i-1,j+1} Q(i-1, j+1) - (1-p)\mu_{i+1,j+1} Q(i+1, j+1) - \\
b^i(1 - p + pb)\lambda_{i,j-1} Q(i, j-1) - [p\mu_{i,j+1} + (1-p)\bar{\mu}_{i,j+1}] Q(i, j+1) = 0
\end{aligned}
\tag{4.44}
$$

$0 < i = j < M$:

$$
\begin{aligned}
(\mu_{j,j} + \lambda_{j,j}) \cdot Q(j, j) - p \cdot \lambda_{j-1,j-1} \cdot Q(j-1, j-1) - \\
\bar{\mu}_{j,j+1} \cdot Q(j, j+1) - \mu_{j+1,j+1} \cdot Q(j+1, j+1) = 0
\end{aligned}
\tag{4.45}
$$

$0 < i \leq j < M$:

$$
\begin{aligned}
(\mu_{i,j} + \bar{\mu}_{i,j} + \lambda_{i,j}) \cdot Q(i, j) - p \cdot \lambda_{i-1,j-1} \cdot Q(i-1, j-1) - \\
(1-p) \cdot \lambda_{i,j-1} \cdot Q(i, j-1) - \bar{\mu}_{i,j+1} \cdot Q(i, j+1) - \mu_{i+1,j+1} \cdot Q(i+1, j+1) = 0
\end{aligned}
\tag{4.46}
$$

In addition,

$$\sum_{i=0}^{M} \sum_{j=0}^{\infty} Q(i, j) = 1 \tag{4.47}$$

Similar to the case for primary relaying, we assume (1) the probability of a new call coming is independent of the number of busy source, i.e. $\lambda_{i,j} = \lambda$; (2) the death rate is proportional to the number of busy sources, i.e., $\mu_{i,j} = i\mu$, and $\bar{\mu}_{i,j} = (j - i)\mu$ if $j < M$, or $\bar{\mu}_{i,j} = (M - i)\mu$ if $j \geq M$. By Plugging these value into Equations 4.37 through 4.47 and solving them, we get $Q(i, j)$ for $0 \leq i \leq j$.

Similar to the analysis discussed in Section 4.3.1 for primary relaying, we can compute the probability of $j - M$ or less calls terminating during the time t ($P_j(t)$) by using Equation 4.33. Thus, the probability of delay exceeding t for an incoming call is given by

$$W_s(t) = \sum_{i=0}^{M} \sum_{j=M}^{\infty} Q(i, j) P_j(t) \tag{4.48}$$

Accordingly, the average waiting time of an incoming call can be computed by

$$\overline{W}_s = \int_0^{\infty} t W_s(t) dt \tag{4.49}$$

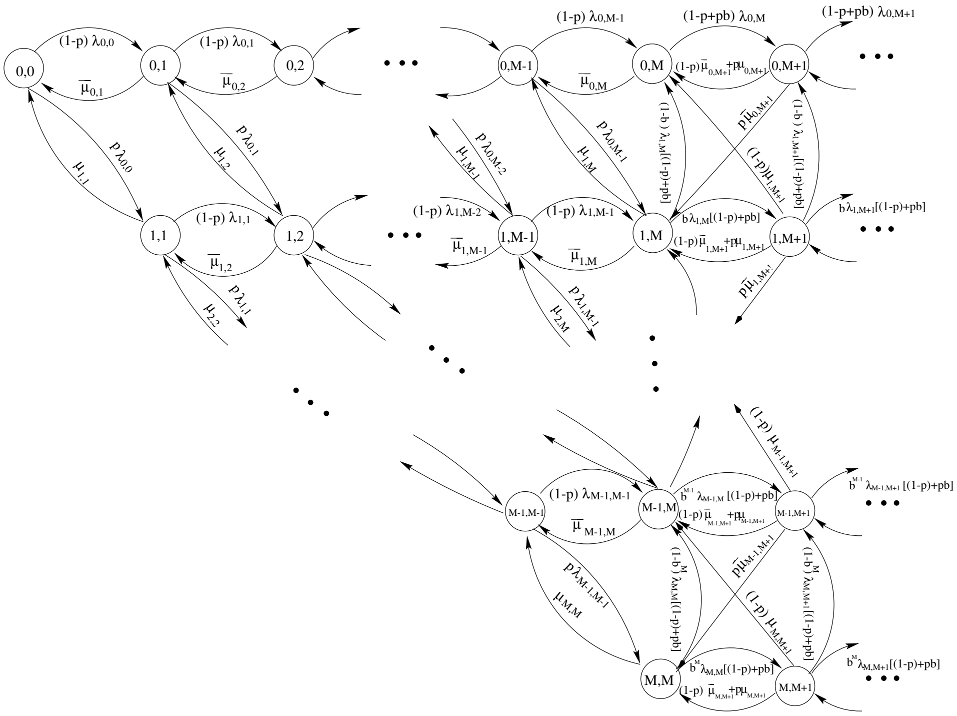


Figure 4.10: State diagram to obtain approximate modeling of secondary relaying (delay analysis).

4.3.3 Numeric Results

In this subsection, we present the numeric results of new call delay in iCAR. We consider a system similar to that discussed in Section 4.2.3. Figure 4.11 shows the average delay of a call in the iCAR system. As one expects, the queuing delay increases with the traffic intensity. The primary relaying can reduce the average delay significantly. By using secondary relaying, the average new call delay becomes no longer than 0.1 $seconds$. Similarly, as shown in Figure 4.12 and 4.13, the primary and secondary relaying significantly reduce the probability that a new call experiences a delay exceeding a given time t, under various traffic intensities and t values.

4.4 Handoff Performance Analysis

In this section, we will introduce an analytical model based on random variable and probability theories for the handoff calls in iCAR. We will first analyze the handoff call dropping probability (in Section 4.4.1 to 4.4.3), considering the system with Assumptions 1 & 2 (i.e., where a handoff

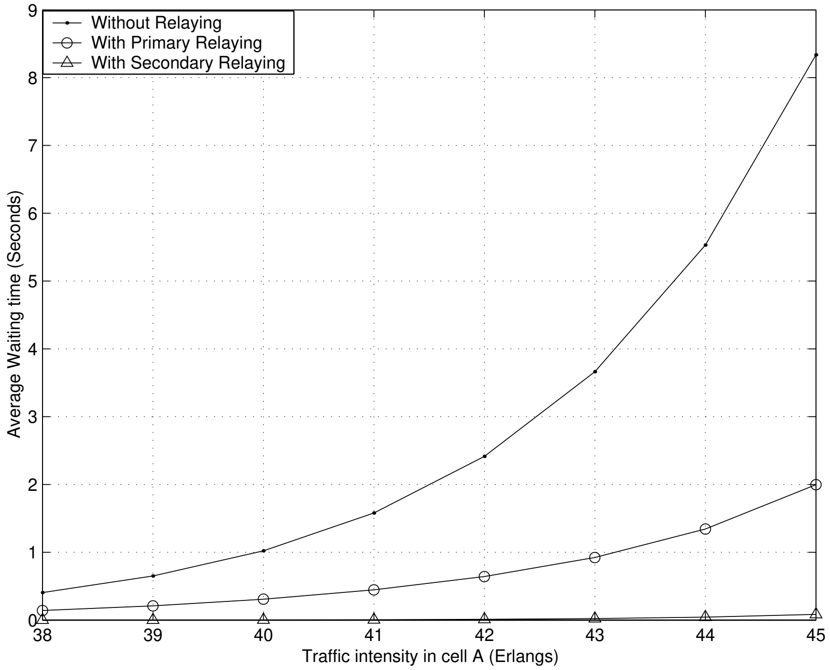


Figure 4.11: Average new call delay.

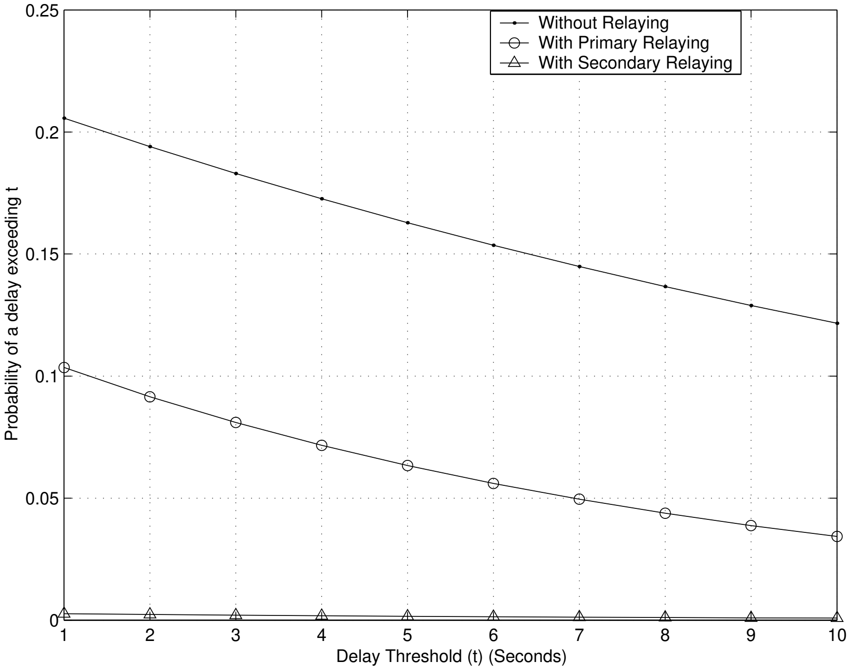


Figure 4.12: Delay probability v.s. delay time t.

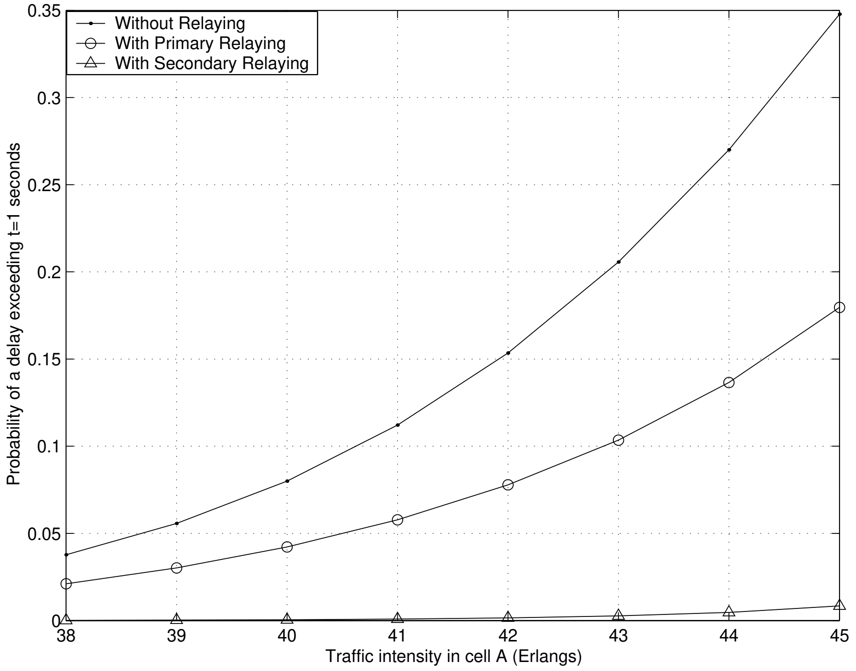


Figure 4.13: Delay probability v.s. traffic intensity.

request will be blocked immediately without queuing when there is no channel available), and then discuss the handoff delay (in Section 4.4.4) in the iCAR system with Assumptions 1 & 3. The readers are also referred to [93] for the handoff analysis in a conventional cellular system.

For simplicity, we assume there is no priority given to the handoff requests. In other words, there are no channels reserved for the handoff calls. We consider a system where one ARS is placed at each shared border of two cells, and assume p_A to be the *area* coverage of the ARSs which is a fraction of the cell area covered by the ARSs, and p_L to be the *line* coverage of the ARSs which is a fraction of the cell border covered by the ARSs.

We define a random variable T_M with an exponential distribution to denote the unencumbered duration of a new call or a handoff call. The density function of T_M is

$$f_{\mathbf{T_M}}(t) = \begin{cases} \mu e^{-\mu t}, & t > 0 \\ 0, & otherwise \end{cases} \tag{4.50}$$

where $\frac{1}{\mu}$ is the mean value of T_M. In addition, we assume that the velocity and the moving direction of an MH are uniformly distributed random variables but remains constant in a cell. More specifically, their density functions are

$$f_{\mathbf{V}}(v) = \begin{cases} \frac{1}{V_{max}}, & 0 \leq v \leq V_{max} \\ 0, & otherwise \end{cases} \tag{4.51}$$

where V_{max} is the maximum velocity of an MH, and

$$f_{\mathbf{\Theta}}(\theta) = \begin{cases} \frac{1}{\pi}, & 0 \leq \theta \leq \pi \\ 0, & otherwise \end{cases} \tag{4.52}$$

Note that, although the moving direction of the MH corresponding to a new call may be from 0 to 2π, we can consider the interval of $[0, \pi]$ only, because of the symmetry. Moreover, the moving direction for a handoff call is assumed to be from 0 to π (i.e., we assume that the active MH will not move back to the cell where it was located).

4.4.1 Probability That A Given Handoff Attempt Fails

We first discuss the probability that a given handoff attempt fails, given the assumptions 1 & 2. There are two types of handoff in iCAR, i.e., BTS-to-BTS handoff and ARS-to-BTS handoff. In the former, a connection *without relaying* is handed over from one BTS to another, while in the latter, a connection *via relaying* is handed over from an ARS to a BTS. We denote P_{Bj}^{o} and P_{Bj}^{r} to be the blocking probability of a cell j in a conventional cellular system (i.e., without relaying) and an iCAR system (i.e., with relaying), respectively, and M to be the number of DCHs in each cell.

ARS-to-BTS handoff

Given the assumption of no priority for the handoff attempts, the probability that an handoff attempt from an ARS to a BTS j will be rejected is

$$P_{A-B} = P_{Bj}^{r} \tag{4.53}$$

BTS-to-BTS handoff

For a handoff attempt from BTS_i to BTS_j, the probability that it fails in a conventional cellular system is equal to the blocking probability of cell j (without relaying), i.e., P_{Bj}^{o}.

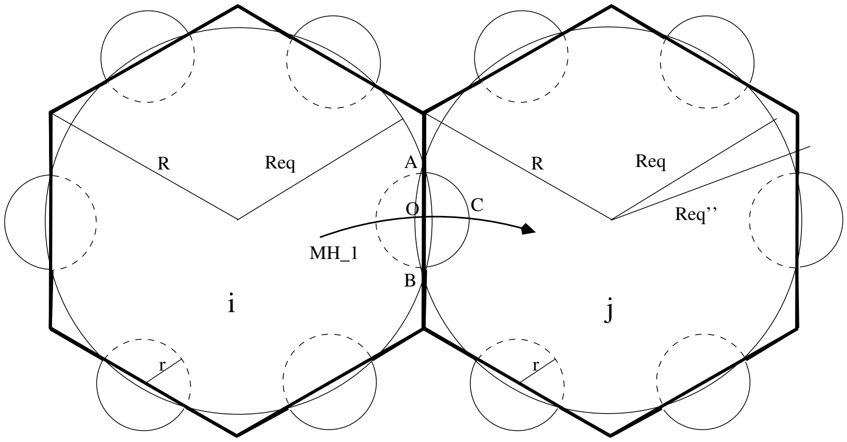


Figure 4.14: An example of call handoff.

In the iCAR system, when an MH crosses the shared border of two cells, it may be covered by ARSs with a probability p_L (see Figure 4.14). If the MH associated with the handoff attempt is not covered by an ARS, the probability of this attempt being rejected is equal to the blocking probability of cell j (with relaying), i.e., P^r_{Bj}. On the other hand, if the MH involved in the handoff is covered by an ARS (i.e., crossing line AOB), it will try a normal BTS-to-BTS handoff and succeed if there are free DCHs available in cell j. Otherwise, it will still use the DCH of BTS_i via relaying through the ARS until one DCH of BTS_j is released so that the MH may use the released DCH, or the call is finished, or the MH moves out of the coverage of the ARS. We define a random variable T_R to be the time duration of an MH travelling within the coverage of the ARS after crossing the cell border (i.e., the time of the MH travelling from a point on the line AOB to a point on the curve ACB). In other words, the MH has the additional time of up to T_R to complete the handoff process, and we call this period the *handoff buffer time* in iCAR. Because of the handoff buffer time, the handoff attempt will be rejected only when

1. all DCHs in cell j are busy (even with relaying) at the moment when the MH crosses the shared border of the two cells, with a probability of P^r_{Bj};
2. and the remaining call holding time is longer than the handoff buffer time, with a probability of $P_r\{T_M > T_R\}$;
3. and there is no DCH in cell j to be released, i.e., none of the on-going calls in cell j is finished and no active MH moves out from cell j within the handoff buffer time, with a

probability of $[P_r\{T_M > T_R\} \cdot (1 - P_N)]^M$, where P_N is the probability that a non-blocked new call requires at least one handoff before completion, which will be discussed later in Section 4.4.2.

Based on the above discussion, the probability that a BTS-to-BTS handoff attempt will be rejected in iCAR is

$$P_{B-B} = (1 - p_L) \cdot P^r_{Bj} + p_L \cdot P^r_{Bj} \cdot P_r\{T_M > T_R\} \cdot [P_r\{T_M > T_R\} \cdot (1 - P_N)]^M \qquad (4.54)$$

where the probabilities of $T_M > T_R$ is given by

$$P_r\{T_M > T_R\} = \int_0^\infty [1 - F_{T_M}(t)] f_{T_R}(t) dt \qquad (4.55)$$

In Equation 4.54, M and p_L are the known iCAR system design parameters. The call holding time T_M, being exponentially distributed with a mean value of $\frac{1}{\mu}$, is also known. P^r_{Bj} is obtained from the analytical model developed by Wu et. al. [15, 16]. But the distribution of T_R has to be determined.

In order to obtain the density function of T_R, we consider the ARS at the shared border of cell i and cell j, and first derive the density function of the random variable d, which is the distance that an MH travels before it moves out of the coverage of an ARS (i.e., the distance from a point on line AOB to a point on the curve ACB as shown in Figure 4.15).

Let us denote r to be the transmission range of an ARS and x to the random variable representing the distance from an MH on the line AOB to the origin O. Assuming that an MH has equal probability to appear at any position on line AOB,

$$f_{\mathbf{x}}(x) = \begin{cases} \frac{1}{r}, & 0 \le x \le r \\ 0, & otherwise. \end{cases} \qquad (4.56)$$

From Figure 4.15, we have

$$r^2 = d^2 + x^2 - 2dxcos\theta \qquad (4.57)$$

Since d is a function of two random variables x and θ, we can derive the density function of d (i.e., $f_{\mathbf{D}}(d)$) by defining an auxiliary variable $w = x$, so that $\theta = arccos(\frac{d^2+w^2-r^2}{2dw})$ and $x = w$.

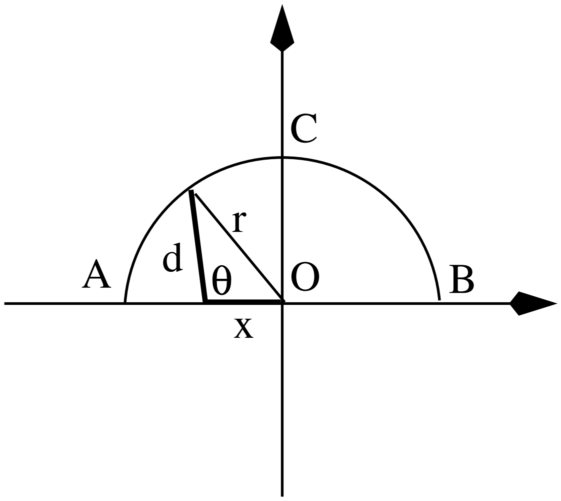


Figure 4.15: T_R analysis

Accordingly, the Jacobian transformation is

$$\begin{aligned} J^{-1} &= \begin{vmatrix} \frac{\partial\theta}{\partial d} & \frac{\partial\theta}{\partial w} \\ \\ \frac{\partial x}{\partial d} & \frac{\partial x}{\partial w} \end{vmatrix} = \left|\frac{\partial\theta}{\partial d}\right| \\ &= \left|\frac{w^2-d^2-r^2}{d\sqrt{4d^2w^2-(d^2+w^2-r^2)^2}}\right| \\ &= \frac{d^2+r^2-w^2}{d\sqrt{4d^2w^2-(d^2+w^2-r^2)^2}} \end{aligned} \tag{4.58}$$

and yields the joint density function of d and w

$$\begin{aligned} f_{\mathbf{DW}}(d,w) &= J^{-1} f_{\mathbf{x\Theta}}(w, arccos(\frac{d^2+w^2-r^2}{2dw})) \\ &= J^{-1}\cdot\frac{1}{r}\cdot\frac{1}{\pi} \quad (|d-r| \le w \le r) \end{aligned} \tag{4.59}$$

Hence, the density function of d is given by

$$f_{\mathbf{D}}(d) = \int_{|d-r|}^{r} f_{dw}(d,w)dw \tag{4.60}$$

The handoff buffer time is given by

$$T_R = \frac{D}{V} \tag{4.61}$$

with the corresponding density function

$$\begin{aligned} f_{\mathbf{T_R}}(t) &= \int_0^{Vmax} |v| \, f_{DV}(tv, v) dv \\ &= \int_0^{Vmax} v f_D(tv) f_V(v) dv \end{aligned} \qquad (4.62)$$

where $f_D(\cdot)$ and $f_V(\cdot)$ are obtained from Equations 4.61 and 4.62, respectively.

Finally, we may substitute the expression for f_{T_R} into Equation 4.55 to obtain $P_r\{T_M > T_R\}$, and compute P_{B-B} using Equation 4.54.

4.4.2 Probability That A Handoff Attempt Occurs

In this subsection, we will discuss the probability that a handoff attempt occurs.

BTS-to-BTS handoff

The probability that a BTS-to-BTS handoff attempt occurs may be obtained in a similar way as that introduced in [93] for a conventional cellular system. More specifically, denoting T_n to be the random variable representing the time for which an MH resides in the cell where the call is originated, and T_h to be the random variable representing the time for which an MH resides in the cell where the call is handed off, we may obtain the probability that a non-blocked new call requires at least one handoff before completion (P_N), and the probability that a handoff call requires another handoff before completion (P_H) as follows.

$$P_N = P_r\{T_M > T_n\} = \int_0^{\infty} [1 - F_{T_M}(t)] f_{T_n}(t) dt \qquad (4.63)$$

$$P_H = P_r\{T_M > T_h\} = \int_0^{\infty} [1 - F_{T_M}(t)] f_{T_h}(t) dt \qquad (4.64)$$

Approximating the cell (which is modelled as a hexagon) to be a circle with the same coverage (see the circle with radius R_{eq} in Figure 4.14), we may obtain the estimation of the distribution function of T_n and T_h in a similar way to that we used to obtain T_R. The only difference is that, in the case to obtain T_R, the MH can only appear *on* a *line* (i.e., line AOB in Figure 4.14), however for the case to obtain T_n and T_h, the MH may appear at any position *within* or *on* the *circle* with radius R_{eq}, respectively. Thus, the details are omitted.

ARS-to-BTS handoff

The probability that an ARS-to-BTS handoff happens in a cell (P_A) is

$$P_A = P_{AH} \times P_R \tag{4.65}$$

where P_{AH} is the probability that an ARS-to-BTS handoff attempt may happen given the call is supported via relaying, and P_R is the probability that a call is supported via relaying. Similar to P_N,

$$P_{AH} = P_r\{T_M > T_{AH}\} = \int_0^\infty [1 - F_{T_M}(t)] f_{T_{AH}}(t) dt \tag{4.66}$$

in which T_{AH} is a random variable of the time duration of an MH travelling within the coverage of the ARS, assuming it starts a call via relaying at any position within the ARS coverage. Its distribution function may be obtained in a similar way to that of T_n.

In each cell, there is a one-to-one mapping between the calls supported via relaying and the calls that would be blocked without relaying but be accepted because of using primary or secondary relaying. Thus, we may estimate P_R as $B_B^o - B_B^r$, where B_B^o and B_B^r are the average blocking probabilities without and with relaying in a cell, and then, compute the probability that an ARS-to-BTS handoff happens (P_A).

4.4.3 Call Dropping Probability In iCAR

In this subsection, we will derive the call dropping probability of an iCAR system based on the above discussions. We assume that all cells in an iCAR system have the same average traffic intensity and the same average call blocking probability (P_B^r). Thus, the probability that a non-blocked new call is dropped in the L-th cell, i.e. it

1. succeeds in the first $L-1$ BTS-to-BTS handoff attempts (with a probability of $P_N P_H^{L-2}(1-P_{B-B})^{L-1}$),

2. and succeeds in the ARS-to-BTS handoffs in the first $L-1$ cells (with a probability of $P_A^{L-1}(1 - P_{A-B})^{L-1}$),

3. but fails on the L-th BTS-to-BTS handoff attempt (with a probability of P_{B-B}),

4. or even though it succeeds on the L-th BTS-to-BTS handoff attempt, it fails on the ARS-to-BTS handoffs in the L-th cell (with a probability of $(1 - P_{B-B})P_A P_{A-B}$),

is

$$P_{FH}^L = [P_{B-B} + (1 - P_{B-B})P_A P_{A-B}] \cdot [P_N P_H^{L-1}(1 - P_{B-B})^{L-1} P_A^{L-1}(1 - P_{A-B})^{L-1}] \quad (4.67)$$

Accordingly, the probability that a non-blocked new call will be dropped is,

$$P_{FH} = \sum_{L=1}^{\infty} P_{FH}^L \quad (4.68)$$

and the dropping probability of an iCAR system P_d^r is

$$P_d^r = (1 - P_B^r) \cdot P_{FH} \quad (4.69)$$

4.4.4 Handoff Delay Analysis

In this subsection, we briefly introduce the analysis of the handoff delay in the iCAR system, given the Assumptions 1 & 3. Assuming no priority for the handoff attempts, the probability that the delay of an ARS to BTS handoff request exceeds a time t is the same as the probability that the delay of an incoming new call exceeds t, i.e., $W^h(t)_{A-B} = W_s(t)$ (see the details of the analysis of $W_s(t)$ in Section 4.3).

Similar to Equation 4.54, we can compute the probability that the delay of a BTS to BTS handoff request exceeds a time t.

$$W^h(t)_{B-B} = (1 - p_L) \cdot W_s(t) + p_L \cdot W_s(t) \cdot P_r\{T_M > T_R\} \cdot [P_r\{T_M > T_R\} \cdot (1 - P_N)]^M \quad (4.70)$$

Accordingly, we can compute the average ARS to BTS handoff delay and the average BTS to BTS handoff delay.

$$\overline{W}_{A-B}^h = \int_0^{\infty} tW^h(t)_{A-B} dt \quad (4.71)$$

$$\overline{W}_{B-B}^h = \int_0^{\infty} tW^h(t)_{B-B} dt \quad (4.72)$$

4.4.5 Numeric Results

In this subsection, we present the numeric results of the handoff performance in an iCAR system, where each cell has the same average traffic intensity varying from 40 to 50 $Erlangs$, and the same number of data channels ($M = 50$). We assume the center-to-vertex distance of a cell is $R = 2000$ meters, and thus $R_{eq} = \sqrt{\frac{6\sqrt{3}}{4\pi}} \times R = 1820$ meters. The ARS transmission range r is assumed to be 500 meters, which results in $p_L = 0.5$ and $p_A = 0.23$.

We first compute the call dropping probability of the systems with and without relaying under different MH mobilities, where the average call holding time is assumed to be fixed at 120 sec. As shown in Figure 4.16, the iCAR system has a much lower call dropping probability than that of a conventional cellular system, because the iCAR system has not only a lower call blocking probability but also the handoff buffer time for the handoff attempts. As we expected, the call dropping probability increases with the MH moving speed, because the higher MH mobility results in higher probability that an active MH may move out of the coverage of a BTS or an ARS, and consequently the higher probability that a handoff attempt occurs (i.e., P_H, P_N, and P_{AH}). More specifically, when the MH moving speed increases from $1.5m/s$ to $15m/s$, the call dropping probability increases by about 10 times in both the conventional cellular system and iCAR. In addition, the call dropping probability also increases with the traffic intensity, because higher traffic intensity results in higher blocking probability, which affects both the new calls and the handoff calls.

Unlike the call blocking probability, the call dropping probability may vary a lot under different average call holding time values (i.e., $1/\mu$), even though the traffic intensity is the same. As shown in Figure 4.17, a higher call holding time results in a higher call dropping probability, because increasing the call holding time will increase the probability that a handoff attempt occurs, and ultimately the call dropping rate.

Figure 4.18 shows the improvement of the iCAR system over a conventional cellular system (without relaying) in terms of the handoff delay. As we can see, the iCAR system can significantly reduce the probability of a handoff request delay exceeding a time t (e.g., $t = 1$ sec).

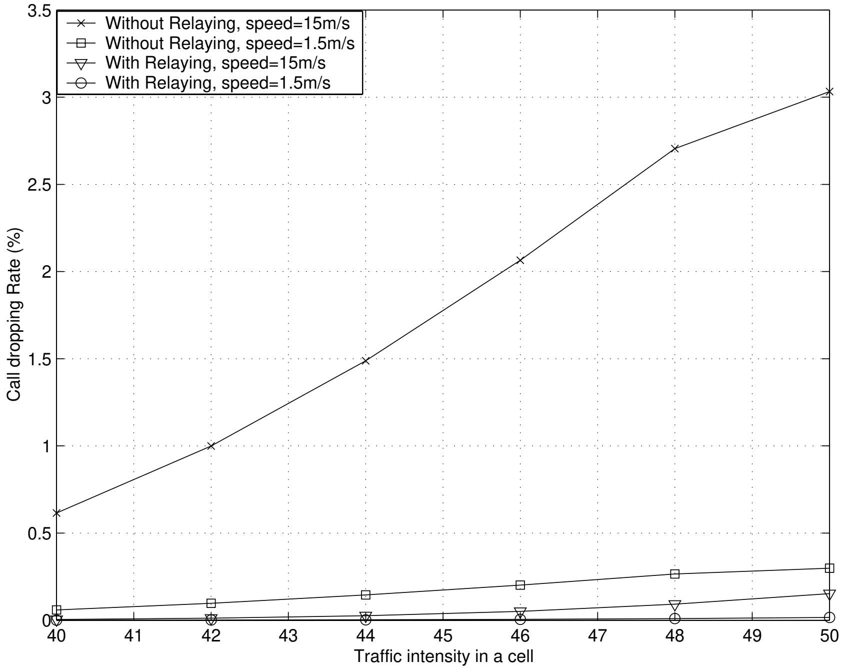


Figure 4.16: Call Dropping rates for different MH moving speeds. $\mu = 1/120$.

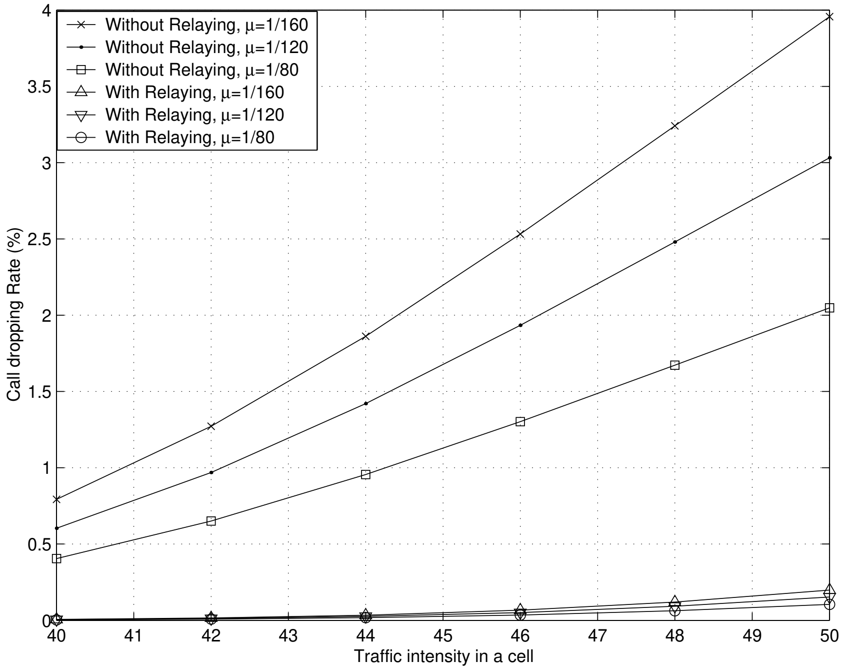


Figure 4.17: Call Dropping rates for different μ values. MH moving speed is $15m/s$.

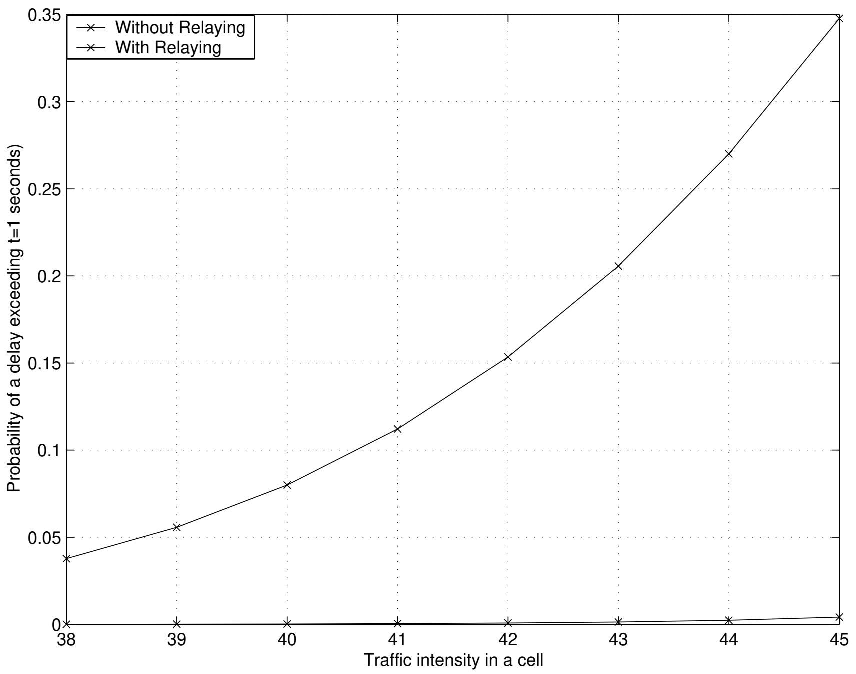


Figure 4.18: Handoff Delay.

4.5 Simulation

To obtain performance results under more realistic assumptions, we have also developed a simulation model. We partition the system with unbalanced traffic and scattered hot spots into sub-systems. In this simulation, we study only one sub-system (see the area inside the dashed rectangle in Figure 4.19). In addition, the results obtained from the simulation are under the assumption of no queuing.

4.5.1 Simulation Model

The average call arrival rate and holding time are two factors determining the traffic load (measured in *Erlangs*) in a cell. To facilitate our simulation of different traffic intensities, we keep the average call generation rate fixed, and vary the average call holding time (note that we could have varied the call generation rate instead). The holding time is a random variable with cut negative exponential distribution. Table 4.1 (b) gives an example of mapping from average holding time to traffic intensities we get from the simulation.

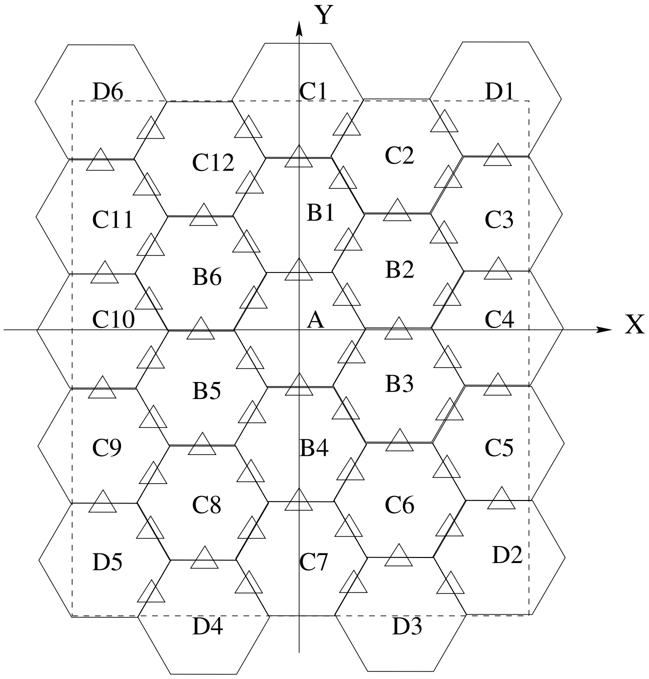


Figure 4.19: Simulation environment

There are $5 \times 5 = 25$ BTSs and 56 ARSs placed at each shared border of two cells in the simulation model. We assume that the longest transmission range of a BTS is 2 *Km* and an ARS (which is placed at each shared border of two adjacent cells) covers an area whose radius is $500\ m$. This results in the ARS coverage of $p = 0.23$. Each BTS has 50 cellular band channels (i.e.,. $M = 50$), and by default, each ARS can handle up to 3 cellular band channels using a proper multiplexing technique. In order to obtain good statistical results, over $25,000$ MHs are simulated which are initially placed in the system with uniform distribution. Table 4.1 (a) lists the parameters used in the simulation.

The simulations were performed using GloMoSim [94]. In addition to the operations in a conventional cellular system (including handoffs from one BTS to another BTS), we implement primary, secondary, cascaded relaying and various other handoffs (e.g., from a BTS to an ARS and from an ARS to a BTS). As mentioned in footnote 3, when we talk about the performance of secondary relaying, it implies that both primary and secondary relaying are implemented. Similarly, cascaded relaying actually include primary and secondary relaying. The call dropping/blocking probability, throughput, and additional signaling overhead introduced by relaying are the main met-

Cell Radius (R)	2 Km
Cell Number	25
ARS Radius (r)	500 m
ARS Number	56
MH Number	25600
Simulation Area	12Km x 15.6Km
DCH at each BTS	50
DCH at each ARS	3
Average MH Call Generation Rate	1 per hour

(a)

Average Holding Time (s)	Traffic Intensity (Erlangs)
110	40.9
120	43.0
130	47.6
140	50.7
150	53.6
160	59.4
170	62.4
180	66.3

(b)

Table 4.1: (a) Default Simulation Parameters. (b) Mapping From Average Holding Time To Traffic Intensity In Cell A With No Mobility And Evenly Distributed MHs.

rics used to evaluate the performance of both cell A and the entire sub-system[3]. The *random waypoint model* wherein an MH selects a random speed, moves for 8 seconds, stays there for 2 seconds and then starts to move again, is used to simulate different mobilities to study their effects on handoffs [95] and call dropping probabilities. The movement of MHs is limited within the dashed square area (which only has a few additional cell D's to simplify the simulation model). The moving direction is random from 0^o to 360^o. The absolute speed value is a random number within a range between 0 meter per second (m/s) and a specified maximum speed. In order to obtain converged results, we run the simulation for 10 hours for each traffic intensity and MH mobility combination before collecting the results. The MHs in the system generate over $250,000$ calls during this period.

4.5.2 Call Blocking Probability

A new call is blocked if there is no free DCH available when it is generated. Figure 4.20 shows the results for call blocking probability in cell A with stationary MHs. Without any relaying, as expected, the call blocking probability which increases with traffic intensity, is very close to that shown in the Erlang B table (which verifies that the simulation model is reasonable).

We observe from Figure 4.20 that there is a good match between the analysis results (which were

[3] Call blocking/dropping probability and throughput are obtained assuming abundant control bandwidth, i.e., a sufficient number of signaling channels.

obtained and presented in [15]) and the simulation results with primary, and secondary relaying. Minor differences may be attributed to the fact that in analysis we try to balance load by relaying traffic even if there is no instantaneous blocking in that cell, whereas in simulation relaying is attempted on a call-by-call basis whenever there is blocking.

With primary relaying, the call blocking probability can be reduced but not by much. When traffic load is not very high (average holding time is less than $110 seconds$), primary relaying can reduce the blocking probability to an acceptable level (e.g. less than 2%).

Secondary relaying reduces the call blocking probability much further. More specifically, the acceptable maximum blocking probability is normally 2%. By applying relaying, the capacity of cell A can increase from 40.255 *Erlang* (with holding time of 110s) to 51.816 *Erlang* (with holding time of more than 140s), which implies that the cell can take several hundred additional calls per hour and still keep the blocking probability below 2%.

Our simulation also reveals that among over $13,000$ calls generated in cell A, no more than ten of them can successfully establish cascaded relaying route. This is because after primary and secondary relaying, most of the ARSs in cell A and tier B cells have already been used to relay calls from cell A to B_i and from B_i to C_i respectively, and the active MHs using a DCH in cell A and B_i are most likely not covered by an ARS, and hence either one cannot find an active MH in cell A for a secondary relaying from A to B (as the first step in cascaded relaying), or even if such an MH is found in cell A, one cannot find an active MH in cell B to complete the cascaded relaying. This is why the curves for cascaded relaying in Figure 4.20 (and all following figures) almost overlaps with that for secondary relaying, implying that the cascaded relaying is not very helpful.

Figure 4.21 shows the impact of the relaying bandwidth (i.e. the number of cellular band channels each ARS can handle) on the performance. Although a higher traffic intensity may require more relaying bandwidth in order to achieve the lowest possible blocking probability in cell A, at most 3 cellular band channels need to be handled by each ARS for relaying purposes. Since cell A is the most congested cell (which needs to relay the largest amount of traffic), this number of channels is also enough for ARSs in cell B's and C's. This helps explain why the analytical results (which are based on the assumption that an ARS can handle as many cellular band channel as necessary)

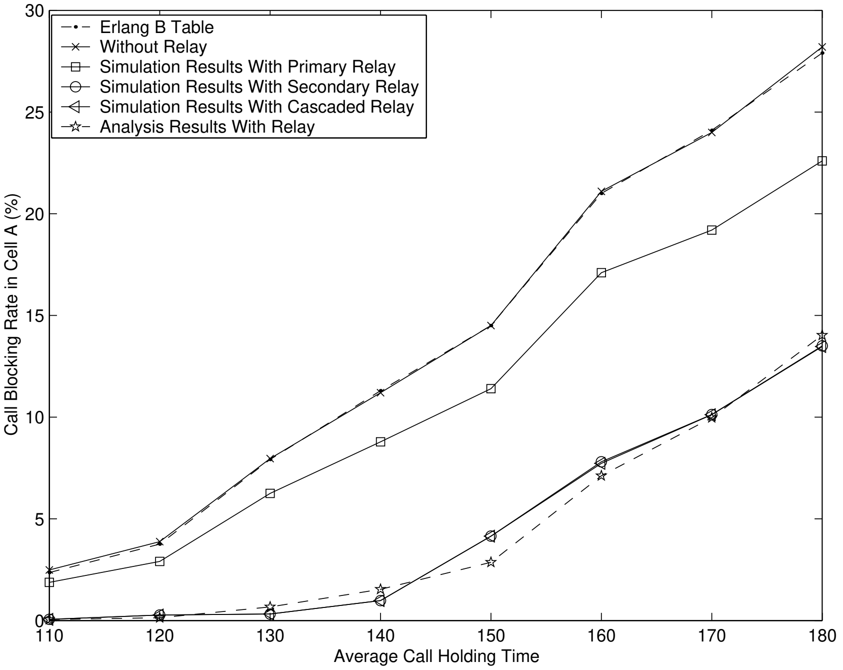


Figure 4.20: Blocking probability in cell A

agree so well with the simulation results.

Figure 4.22 shows the blocking probability of the entire sub-system. It is much lower than the results in cell A because all other cells have lower load than A. As one can see from the figure, the results due to relaying are fairly good. In particular, the system-wide blocking probability decrease although the blocking probability in other low-load cells may increase slightly because of the extra traffic relayed from the hot spot cell A. This agrees with Theorem 1 and Theorem 2 presented in Section 3, which prove that the iCAR system has the lowest blocking probability. Similar to the results in cell A, secondary relaying significantly reduces the call blocking probability, but cascaded relaying is only marginally useful. Though the results are not shown, mobility does not have any significant effect on the blocking probability in cell A or in the sub-system.

4.5.3 Call Dropping Probability

A call is dropped when the active MH moves into a congested cell. In this simulation, we assume that there are no DCHs reserved for handoff calls, and the handoff calls have no special priority [96]. Figure 4.23 shows the dropping probability vs. the maximum MH moving speed. With a higher MH

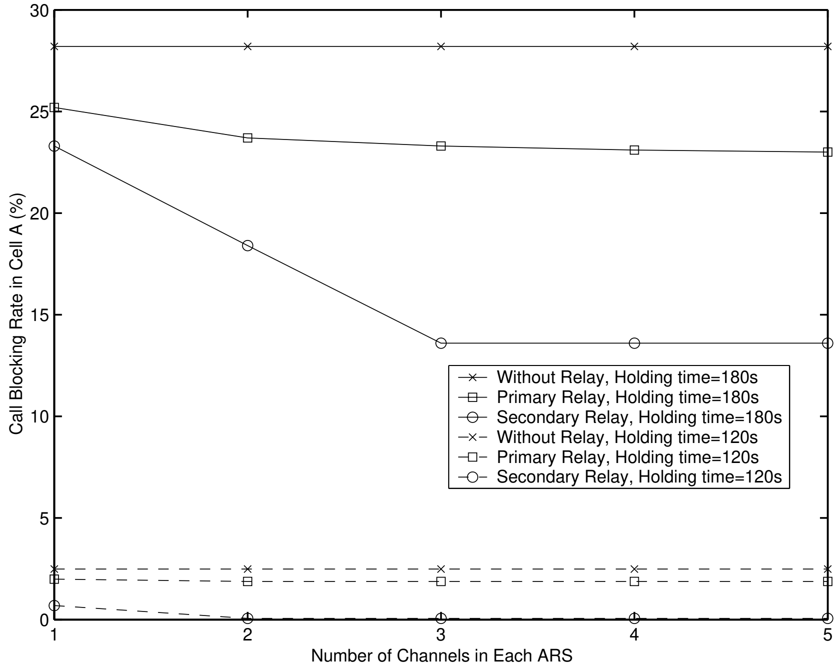


Figure 4.21: Blocking probability versus number of relaying channels in cell A

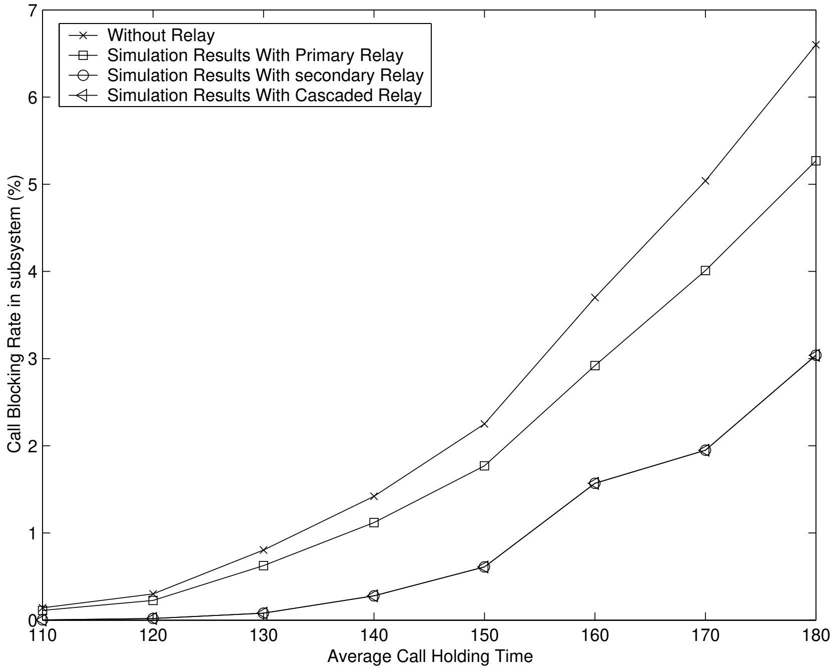


Figure 4.22: Blocking probability in the entire sub-system

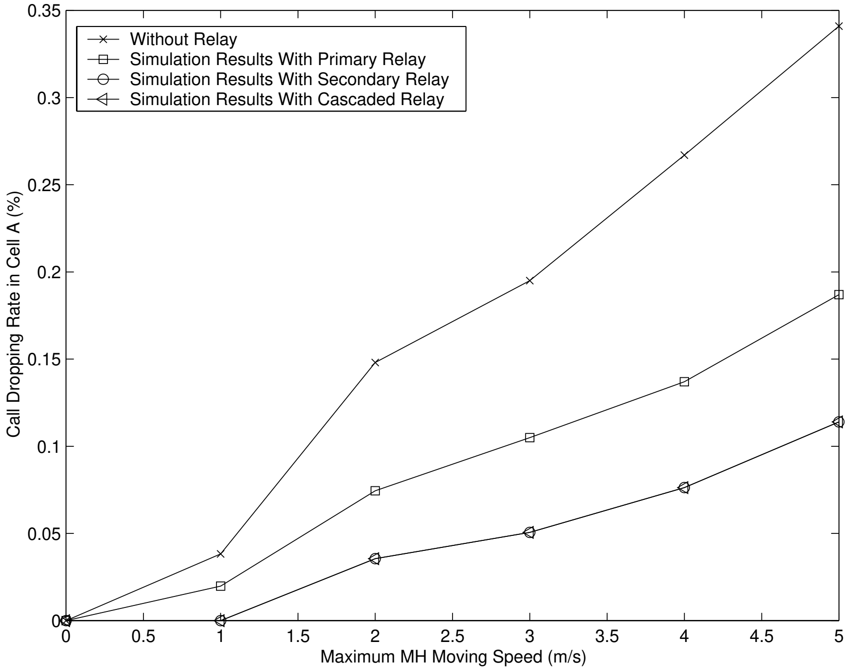


Figure 4.23: Dropping probability in cell A with average holding time=120 Sec.

mobility, the dropping probability increases sharply (recall that this is not the case for the blocking probability). In addition, when comparing with Figure 4.20, we see that primary relaying performs very well for handoff calls. For example, only about 20% blocked calls are saved by primary relaying. But for handoff calls, the primary relaying can reduce the dropping probability as much as 50%. There are two reasons for the good performance of primary relaying in handoffs. First, when a call is handed off from cell X to cell Y (which is congested), it is almost guaranteed that cell X has at least one free DCH (which is released by this MH). Second, handoffs always happen at boundaries of cells, where we put the ARSs. Since a cell is modelled as a hexagon, from Table 4.1 (a), we can see that a large portion of the boundaries of a cell is covered by the ARSs. In addition, secondary relaying reduces the dropping probability further to a certain level. But due to similar reasons to those mentioned in the previous subsection, cascaded relaying is not more helpful than secondary relaying.

4.5.4 Throughput

In our simulation, we assume that the transmission and reception buffer size to be zero. In other words, if a call is blocked or dropped, all the packets to be transmitted will be discarded immediately. We compare the throughput of the iCAR system with that of a cellular system (without relaying) by computing the throughput ratio, which is defined to be the ratio of received data over the data to be transmitted. This ratio is inversely proportional to the blocking/dropping probability. Figure 4.24 shows the results in cell A. In general, a higher traffic load results in a lower throughput ratio because of the limited capacity. When the traffic load in cell A is low enough (with an average holding time of less than 140 seconds), we can obtain above 99% throughput ratio by applying relaying. Under a higher traffic load, one can still improve the throughput by as much as 15%. For reasons similar to those discussed in subsection 4.2, cascaded relaying results in minor performance improvement. Though the results are not shown, we note that, for the overall subsystem, one can keep the throughput ratio as high as about 97%. Furthermore, with a higher MH moving speed, throughput decreases but not as dramatically as the increase in the dropping probability with the MH moving speed. This is because most of the packets are discarded during call blocking or in other words, the blocking probability dominates the throughput performance.

4.5.5 Signaling Overhead

An undesired side effect due to relaying is the signaling extra overhead. In addition to ARSs, three system components, MSC, BTS, and MH, have to send and receive more signaling packets than the case without relaying. In simulation, we study the ratio of additional amount of signaling traffic due to primary, secondary, and cascaded relaying over the basic amount of signaling traffic without relaying.

A simple signaling protocol (similar to that described in 6.3.1) is implemented in the simulator. Our results (though not shown) indicate that the relaying doesn't add much burden to MSC. More specifically, primary relaying results in only one percent more overhead. Even in the case when one applies all three kinds of relaying, the additional overhead is at most 20 percent. This is reasonable because MSC doesn't get involved much in relaying operations.

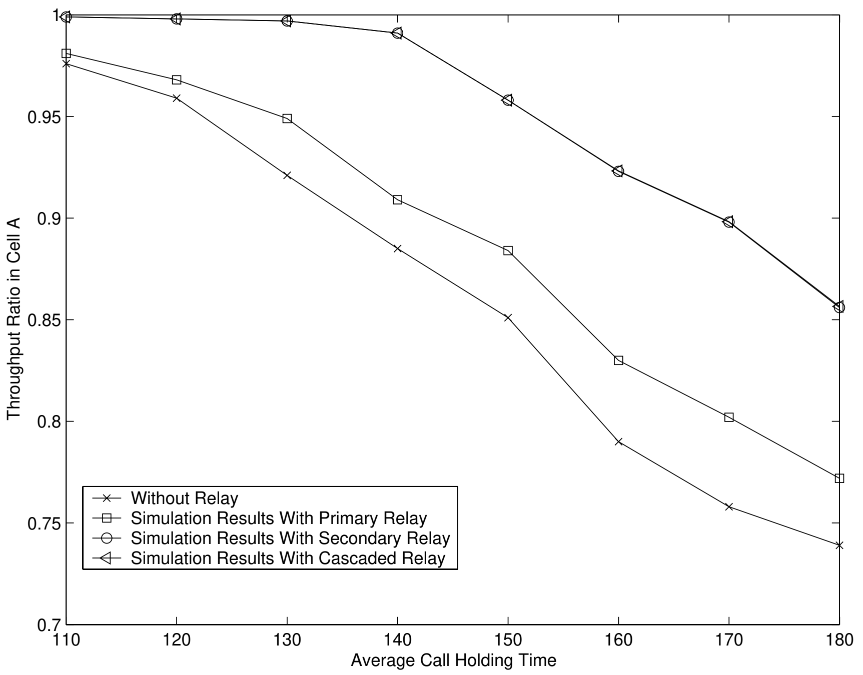


Figure 4.24: Throughput in cell A

Figure 4.25 shows the extra signaling overhead incurred at a BTS when the maximum MH moving speed is 1m/s. As can be seen from the figure, primary relaying do not cause much overhead. But when the traffic load in the system is very heavy, BTSs experience significantly high overhead while using secondary and cascaded relaying. This is because with increase in the traffic load, the probability that a call need to be relayed also increases. This results in a large number of requests for secondary relaying. For each request, the BTS will query MSC for DCH status information, send a broadcast message to all MHs (for secondary relaying) and process replies from the MHs. Our results also showed that the MHs suffer a higher overhead (as much as 2.5 times more than the case without relaying). This is because whenever a call tries secondary or cascaded relaying, all the active MHs using DCH in the cell is responsible to process and reply to the broadcast messages from BTS.

Notice that, the high overhead in BTSs and MHs is incurred only under very heavy traffic load (which may be unreasonably high because the blocking probability would be much more than 2%), and based on non-optimal signaling protocols. With a normal traffic intensity with average holding time equal or less than 120 seconds in this simulation, the extra overhead introduced by using all

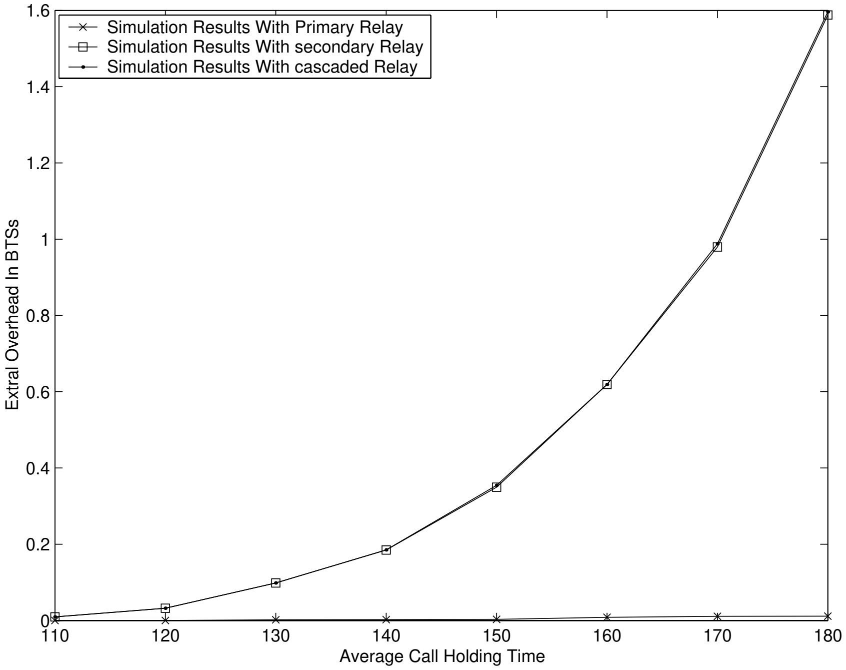


Figure 4.25: Extra overhead incurred by BTSs in the sub-system

three kinds of relaying at MSC, BTSs, and MHs are only about 1%, 3%, and 5% respectively, which is not significant. Nevertheless, further research is needed to improve the signaling protocols to reduce the overhead and to study the tradeoff introduced by dedicating one or more additional channels to carry control signaling information.

Finally, our simulation results also revealed that although with a higher MH moving speed, the MHs need to process more signaling messages because of the higher probability that a handoff call need relaying in order to avoid being dropped, mobility has little effect on the signaling overhead.

Chapter 5

ARS Placement

In this chapter, we discuss the important design issues in iCAR, i.e. the number and placement of ARSs. We will first discuss the number of ARSs needed to cover the entire system, and propose a seed growing approach for the case only a limited number of ARSs are available. Then, we will introduce the concept of the Quality of Coverage for the comparison of various ARS placement strategies, and propose three rules of thumb as the guideline to for ARS placement.

5.1 Full Coverage

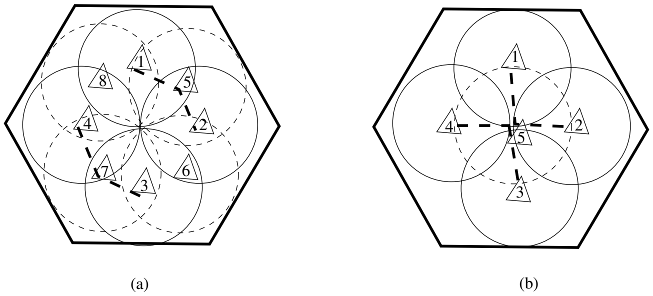
Let R be the radius of a cell and r the radius of the coverage area of an ARS. Typically, R is between 1 to 2 kilometers, and r is between 200 to 500 meters. Therefore, to cover the entire cell with ARSs, we need approximately $n = \lceil (\frac{R}{r})^2 \rceil$ ARSs. However, it is difficult to determine the number of ARSs needed to ensure that a relaying route can be established between any BTS and an MH located anywhere in any cell.

To establish a general guideline for the number of ARSs needed, note that in order to provide *overlapped* coverage of a cell with ARSs so that data can be relayed from one ARS to another, or in other words, in order to be able to establish a relaying route between any two locations in a cell (without involving a BTS), we would need at most $N_{max} = 2n$ ARSs (approximately). The table below shows some representative values of N_{max} ranging from 8 to 200.

A case for $N_{max} = 8$ (and $n = 4$) is illustrated in Figure 5.1 (a), where smaller circles denote

R r	1Km	1.5Km	2Km
200m	50	114	200
350m	18	38	66
500m	8	18	32

Table 5.1: The typical values of the number of ARSs needed for full coverage.



(a)

(b)

Figure 5.1: An example showing the number of ARSs needed in order to be able to establish a relaying route between any two locations in a cell. (a) the maximum is about 8, and (b) the minimum is about 5.

the coverage areas of the ARSs. As can be seen from Figure 5.1 (a), data from ARS 1 can be relayed to ARS 2 (each covering an area shown as a solid circle) via ARS 5 (whose coverage area is shown as a dashed circle).

Note that while placing $N = 8$ ARSs in a cell may not be too much of a burden (but placing 200 ARSs in a cell may), a network operator always wants the most cost-effective solution which uses a minimal number of ARSs to provide a satisfactory service. While a tight lower bound in the number of ARSs needed is difficult to obtain analytically, we observe that, because of the secondary relaying (see Figure 3.2), one can guarantee a successful CI-relaying with fewer ARSs. This is because, as mentioned earlier, the fact that a cell (say B) is congested means that there may be (up to) M MHs in cell B that can serve as MH Y, where M is the capacity of the cell and is typically large (e.g., a few hundreds). Hence, as long as one of these MHs can find a nearby ARS as a proxy from which a relaying route can be established, an MH (say X as in Figure 3.2) which just moved into cell B (or powered on) can be allocated a DCH without having to have any ARS nearby.

5.2 Seed Growing

When only a limited number of ARSs are deployed in a cell, a natural question is how many ARSs are reasonable and where to place them. Here, we propose a *Seed Growing* approach whereby one *seed ARS* is placed on each pair of "shared edges" (denoted by a bold line in Fig 5.2 (a)) along the border between two cells. Additional ARSs will *grow* from the seeds as to be discussed later (See Figure 5.2). Since a seed ARS is shared by two cells, and each cell has 6 edges, it is obvious that at most $3n$ seed ARSs are needed in a n-cell system. In fact, the total number of seed ARSs required will be less because it makes no sense to put any ARSs on a "non-shared edge" (denoted by a dotted line in Fig 5.2 (a)) of a "boundary cell". More formally, we have the following propositions stating the upper bound on the number of seed ARSs.

Proposition 1 *For a n-cell system where the cells are arranged in a near-circular shape (as in Fig 5.2 (b)), the number of seed ARSs needed is at most $3n - \lfloor 4\sqrt{n} - 4 \rfloor$.*

Proof : Suppose that the radius of each cell is R, then the total coverage of the n cells is about $n \times \pi R^2$. Since the cells are organized as a near-circle whose radius R' satisfies the equation $\pi R'^2 \approx n \times \pi R^2$, or $R' \approx \sqrt{n} \times R$, the estimated number of boundary cells is

$$\lfloor \frac{\pi R'^2 - \pi(R' - 2R)^2}{\pi R^2} \rfloor = \lfloor 4\sqrt{n} - 4 \rfloor \tag{5.1}$$

Since each boundary cell has at least two non-shared edges, the number of shared edges is at most $E = 6n - 2\lfloor 4\sqrt{n} - 4 \rfloor$. Hence, the maximum number of seed ARSs for a near-circular shaped n-cell system is $E/2$ or,

$$3n - \lfloor 4\sqrt{n} - 4 \rfloor \tag{5.2}$$

▮

Proposition 2 *For a n-cell system, the maximum number of seed ARSs needed is $3n - \lfloor 4\sqrt{n} - 4 \rfloor$.*

Proof : Obviously, fewer boundary cells imply fewer non-shared edges, or in other words, more shared edges and thus more required seed ARSs. According to the basic geometry theory, for a given coverage area equal to that covered by n cells, a circle will have the shortest perimeter. This

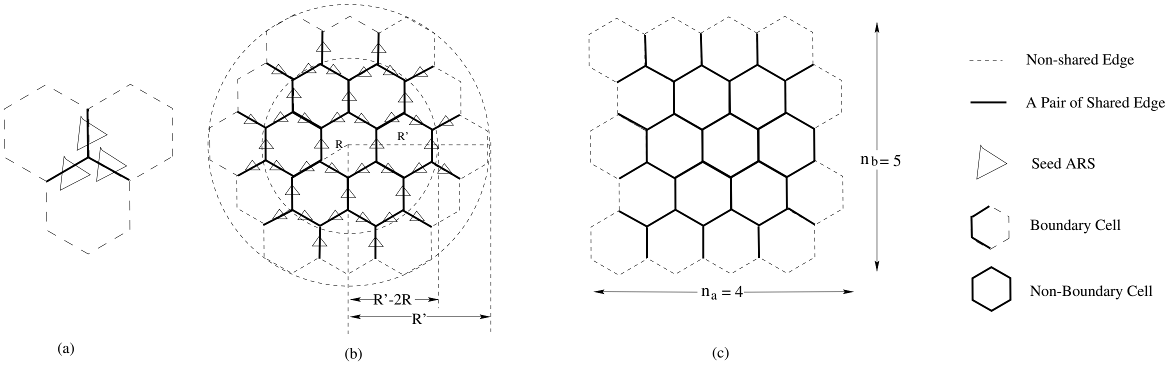


Figure 5.2: The number of seed ARSs with seed-growing approach. (a) One seed ARS for two shared edges. (b) a circular-shaped system with 19 cells. (c) a rectangular-shaped system with 20 cells.

means a circle will have the fewest boundary cells among all possible shapes. Thus, the upper bound on the number of seed ARSs needed in a circular-shaped system is also the upper bound on the number of seed ARSs needed in any shaped system. From Proposition 1, we know this upper bound is

$$S_u = 3n - \lfloor 4\sqrt{n} - 4 \rfloor$$

∎

Proposition 2 means that any other shaped system will require fewer seed ARSs. For example, Figure 5.2 (c) shows a rectangular shaped system with 20 cells, which requires only 43 seed ARSs, (which is the same as the circular- shaped system with 19-cells shown in Figure 5.2 (b), and certainly less than the number of ARSs required in a circular-shaped system with 20 cells). In fact, for any rectangular-shaped system with $n = n_a \times n_b$ cells, we can come up with the *exact* number of seed ARSs. More specifically, the total number of non-shared edges in such a system is $4n_a + 4n_b - 2$. So the number of seed ARSs needed is $S = 3n_a n_b - 2(n_a + n_b) + 1$. Since $n_a n_b = n$ and $n_a + n_b \geq 2\sqrt{n}$, we have $3n_a n_b - 2(n_a + n_b) + 1 \leq 3n - 4\sqrt{n} + 1$, or $S \leq S_u$.

5.3 Quality of Coverage (QoC) of ARSs

The ARS placement is a critical issue in iCAR as iCAR's performance improvement over a conventional cellular system is largely due to its ability of relaying traffic from one cell to another. Specifically, when only partial coverage by ARSs can be provided for an iCAR system, some locations are more important than others. For example, since an ARS provides a limited coverage (e.g., within a few hundreds of meters) compared to the size of a cell (e.g., a few kilometers in diameter), placing an ARS in the center of a cell without any nearby ARSs will be useless as it cannot relay any traffic between a MH in the cell and the BTS in another cell.

In this section aimed at quantitatively evaluating ARS placement strategies, we will introduce the concept of a new performance metric called quality of (ARS) coverage (QoC) and present the analytical results for the comparison between various ARS placement strategies, and propose three rules of thumb as guidelines for cost-effective ARS placement in iCAR. For simplicity, we assume that there is no bandwidth shortage along any relaying routes in the following discussion.

The major motivation of introducing the concept of QoC is to quantitatively evaluate the ARS placement via a simple and straightforward parameter, instead of the conventional performance metrics, such as the request blocking and dropping rate which may be obtained by either complex analysis or intensive simulations. We define the value (Q) of the quality of ARS coverage to be the *relayable traffic* in an iCAR system. Note that, the Q value is not always proportional to the ARS coverage. For example, the Q values of ARS 1 and ARS 2 in Figure 5.3 are zero because they cannot relay any traffic between cells although they do cover certain amount of traffic. Only when an ARS can directly or indirectly (i.e., via multi-hop relaying) relay traffic in one cell to a BTS in another cell, it may have a non-zero Q value which is the amount of traffic covered by the ARS minus the part which will be blocked by the reachable BTSs. More formally, we have the following equation for the value of QoC.

$$Q(t) = \iint_C f(x, y, t) \times (1 - b(\overline{x, y}, t))\, dx\, dy \tag{5.3}$$

where $f(x, y, t)$ is the location-dependent time varying traffic intensity at (x, y) at time t, C is the effective ARS coverage, and $b(\overline{x, y}, t)$ is the time varying blocking probability of the cells reachable

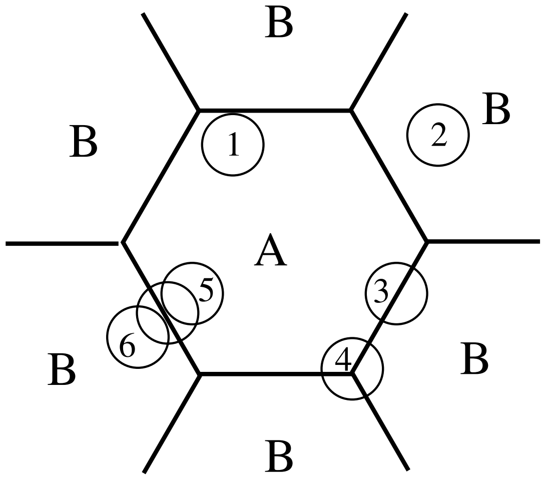


Figure 5.3: The ARSs placed in cell A and its six neighbors.

by the ARSs that cover the location (x, y), excluding the cell where (x, y) is located. When an ARS can reach more than one cells in addition to the cell where (x, y) is located, $b(\overline{x, y}, t)$ is the product of their blocking probabilities, which means the probability that all of them are congested. The higher the Q value, the better the ARS placement strategy is. As we discussed earlier, C equals to 0 if the area is covered by ARSs that can neither directly nor indirectly relay traffic in one cell to a BTS in another cell. Note that, if some area is covered by more than one ARSs, it is counted only once as the effective coverage. In other words, overlapping will not increase C.

As shown in Equation 5.3, the Q value depends on the traffic intensity distribution, and the number, placement and coverage of ARSs, etc. For any single system, one may place the ARSs at the optimized positions by searching for locations that result in the globally highest Q value. However, this optimization probe may be intractable and in addition, requires a known static traffic intensity function at every location. Thus, instead of searching for an optimized solution, we will provide several rules of thumb which can be used as guidelines for placing ARSs in an iCAR system.

5.4 Various ARS Placement Strategies

In this subsection, we will discuss various ARS placement strategies, evaluate them according to their Q values, and provide several rules of thumb as the guidelines for placing ARSs. We will consider a cell A in Figure 5.3, which has six neighboring cells with the same traffic intensity. For simplicity, we assume that the traffic is uniformly distributed in each cell, and only consider the time-averaged values. We denote T_A and T_B to be the traffic intensity in cell A and each of its neighboring cells, respectively, and denote T_a and T_b to be the traffic intensity per unit area in cell A and cell B, respectively. Clearly, $T_A = T_a \times cell\ coverage$ and $T_B = T_b \times cell\ coverage$. In addition, we denote S to be the coverage of one ARS, and M to be the total amount of available cellular bandwidth (CBW) at each BTS, given that one unit of bandwidth is required by one connection. If the MHs' moving speeds and directions are assumed to be uniformly distributed (which is a typical assumption in the mobile computing analysis), the average traffic intensity (i.e., T_a and T_b) will not vary with the mobility of MHs, and consequently, the Q values will not be affected either (which is also verified by our simulation results to be shown later). Hence, the MH mobility issue will be ignored in the following analysis.

5.4.1 Seed ARS Placement

We first discuss where to put the seed ARS to achieve the best performance. Considering cell A in Figure 5.3, there are two approaches to place the seed ARS. One is to put it at the shared border of two cells (see ARS 3 in Figure 5.3), and the other is to put it at a vertex (see ARS 4 in Figure 5.3). An ARS in the two approaches has the same effective coverage. However, their Q values may not be the same. On one hand, assuming cell A is the hot spot, the ARS placed at the border will cover more traffic than that covered by the ARS at a vertex. But on the other hand, the ARS at the vertex may relay traffic to two BTSs instead of one in the border approach, which in turn results in a lower blocking probability for the requests to be relayed. Applying Equation 5.3 in the above two approaches, we may have the following equations for their Q values.

$$Q_{Border} = \frac{S}{2} \cdot T_a \cdot (1 - b_B) + \frac{S}{2} \cdot T_b \cdot (1 - b_A) \tag{5.4}$$

$$Q_{Vertex} = \frac{S}{3} \cdot T_a \cdot (1 - b_B^2) + \frac{2S}{3} \cdot T_b \cdot (1 - b_A \cdot b_B) \tag{5.5}$$

in which b_A and b_B are the blocking probabilities without relaying in cell A and cell B. More specifically, $b_A = \frac{T_A^M/M!}{\sum_{i=0}^{M} T_A^i/i!}$ and $b_B = \frac{T_B^M/M!}{\sum_{i=0}^{M} T_B^i/i!}$, according to the Erlang-B model [97].

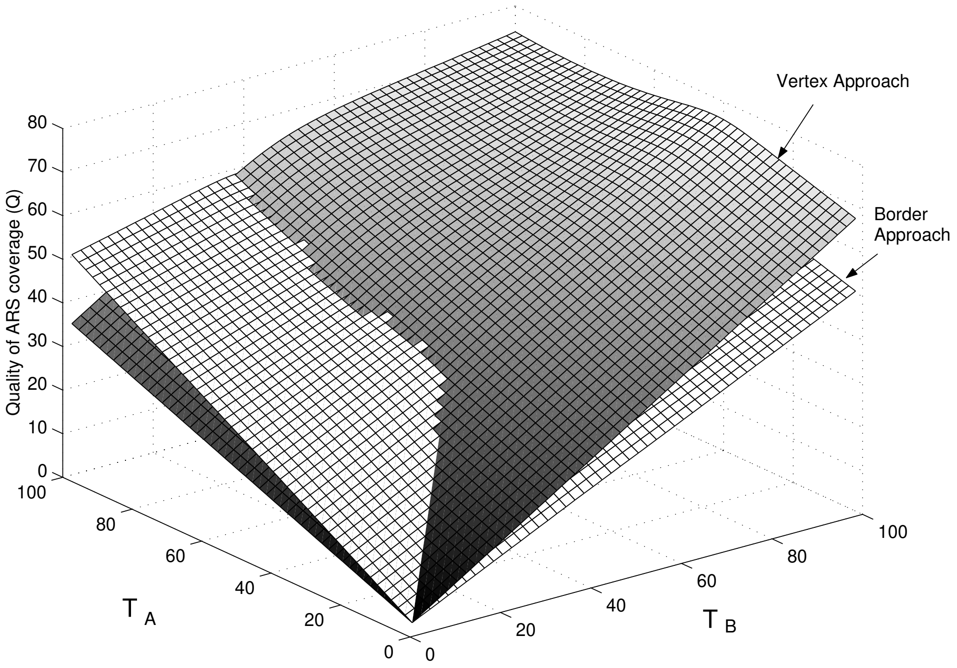


Figure 5.4: The quality of ARS coverage : vertex placement approach vs. border placement approach.

As shown in Equations 5.4 and 5.5, M will affect the Q values, because a different M may result in different b_A and b_B. However, although the results are not shown here, varying M within the normal range (e.g., choosing the values of M so that the blocking probability is around 2%), will not change the order of the Q values of different ARS placement approaches for any given T_A and T_B. Thus, in the following discussions, we will only consider the case where M is equal to the default value (50). In addition, although we have assumed that all neighboring cells of cell A have the same traffic intensity (T_B) and blocking probability (b_B), the analytical model can be easily extended to the case where the cell B's have different traffic intensities by replacing T_B and b_B with the corresponding values in Equations 5.4 and 5.5.

We have obtained Q_{Border} and Q_{Vertex} as shown in Figure 5.4 by varying T_A and T_B. The Q values of these two approaches depend on the traffic intensity in cell A and its neighboring cells. Approximately, when traffic intensity is high ($T_A, T_B > 50$ $Erlangs$) or $T_A < T_B$, Q_{Vertex} is higher than Q_{Border}. But, when $T_B < T_A < 50$ $Erlangs$, the border placement is better. Note that, 50 $Erlangs$ corresponds to about 10% blocking probability which is far beyond the normal

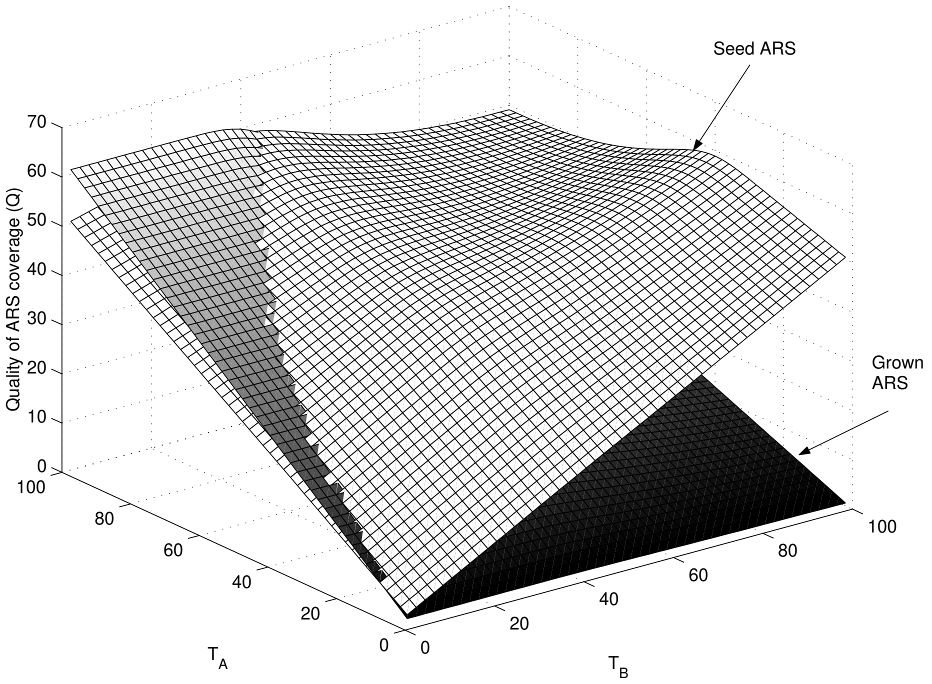


Figure 5.5: The quality of ARS coverage : seed ARS vs. grown ARS.

network operation range. If we assume cell A is the hot spot, i.e., $T_A > T_B$, then the border placement approach is usually a good choice for seed ARSs. So, we may have the first rule of thumb as follows.

Rule of Thumb 1: *Place the seed ARSs at cell borders.*

In addition, it has been shown in [11] that, for an n-cell system, the maximum number of seed ARSs needed for each shared border of two cells is $3n - \lfloor 4\sqrt{n} - 4 \rfloor$.

5.4.2 Seed ARS vs. Grown ARS

If additional ARSs are available, there are two approaches to place them. One is to place them as seeds according to what we discussed in Sec 5.4.1 without any overlap on the existing ARSs. This approach intends to maximize the total effective ARS coverage. The other way is to let them grow from the seeds which are already there (see ARS 5 shown in Figure 5.3, assuming border placement approach is adopted). The grown ARS is required to be within the coverage of at least one existing ARS so that they can relay traffic to each other. Thus, their coverage overlaps within some area, and

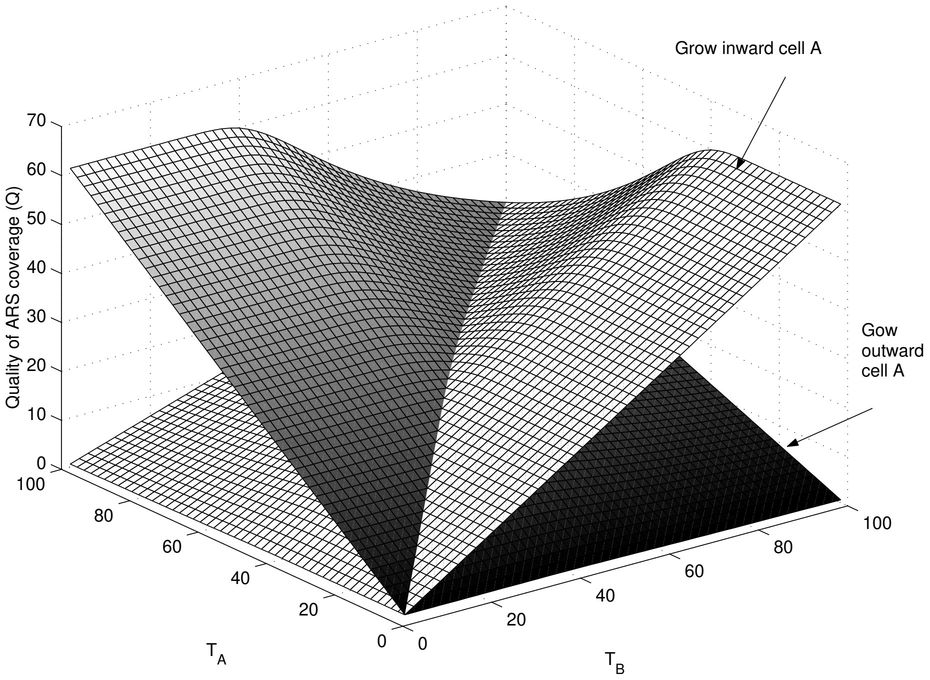


Figure 5.6: The quality of ARS coverage : the direction to grow.

not all of the area covered by the grown ARS will result in the increase of the system's Q value. To minimize the overlapped area and maximize the effective coverage of the grown ARS, we place it just within the transmission range of the existing seed ARS (i.e., let the distance between the two ARSs as far as possible while they can still communicate with each other). Thus, we may compute the additional coverage of the grown ARS (i.e., its coverage minus the overlapped area), which is about $\frac{\pi r^2 - 2(\frac{1}{6}\pi r^2 - \frac{\sqrt{3}}{4}r^2)}{\pi r^2} \approx 0.61S$, where r is the radius of an ARS, and the increased Q value to be

$$Q_{Grow\_in} \approx 0.61 \cdot S \cdot T_a \cdot (1 - b_B) \tag{5.6}$$

assuming it grows inward cell A. Comparing it with Q_{Border} in Figure 5.5, as we can see, only when T_B is very low and T_A is much higher than T_B, the grown ARS approach performs better than the seed ARS approach, and therefore we have the second rule of thumb.

Rule of Thumb 2: *Place an ARS as seed if it is possible.*

5.4.3 Direction of Growing

If the additional ARSs can not be placed as seeds because there is no free space at the shared boundaries of cells, we have to let them grow from some of the seeds. An ARS can grow inward cell A (see ARS 5 in Figure 5.3) or outward cell A (see ARS 6 in Figure 5.3). Both of them have the same ARS coverage (S). But since the ARSs cover different cells with different traffic intensities, they may result in different Q values. When an ARS grow inward cell A, its Q value (Q_{Grow_in}) has been shown in Equation 5.6. When an ARS grows outward cell A, its Q value is

$$Q_{Grow\_out} \approx 0.61 \cdot S \cdot T_b \cdot (1 - b_A) \tag{5.7}$$

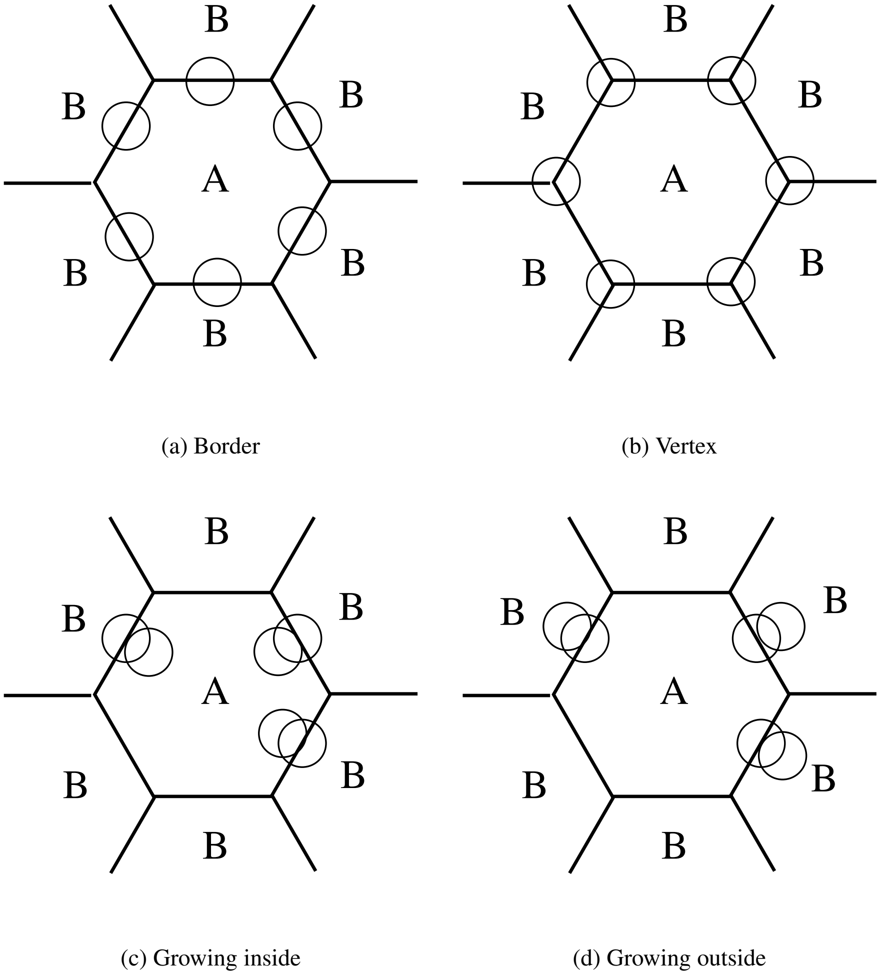
Figure 5.6 compares these two approaches. As we expected, it is better to let the ARS grow inward the cell with higher traffic intensity. Thus, we have the third rule of thumb.

Rule of Thumb 3: *Grow an ARS toward the cell with high traffic intensity.*

The three rules of thumb may serve as the guidelines for ARS placement. More specifically, to optimize the system performance, the operators may first place ARSs at the shared borders of the cells. If there are additional ARSs, they may let them grow in the cell with higher traffic load. However, depending on the size of the cells and the coverage of the ARSs, there may be some exceptions. For example, when a number of seed ARSs have been deployed in a system, placing another seed ARS later may result in some overlap with the existing ARSs, and therefore result in a lower Q value. In this case, growing the additional ARSs may result in a better performance. Similarly, when there are already many ARSs growing in the cells with high traffic intensity, placing an ARS in the neighboring cells may be more beneficial.

5.5 Simulation Results and Discussions

To evaluate the performance of various ARS placement approaches in terms of the system-wide (i.e., weighted average) request blocking and dropping probability, we have developed a simulation model using the GloMoSim simulator [94] and the PARSEC language [98]. The simulated system includes a cell A and six neighboring cells (see Figure 5.7), which are controlled by a PSC. The cells are modelled as hexagons with the center-to-vertex distance of 2 *Km*. We have assumed that



(a) Border

(b) Vertex

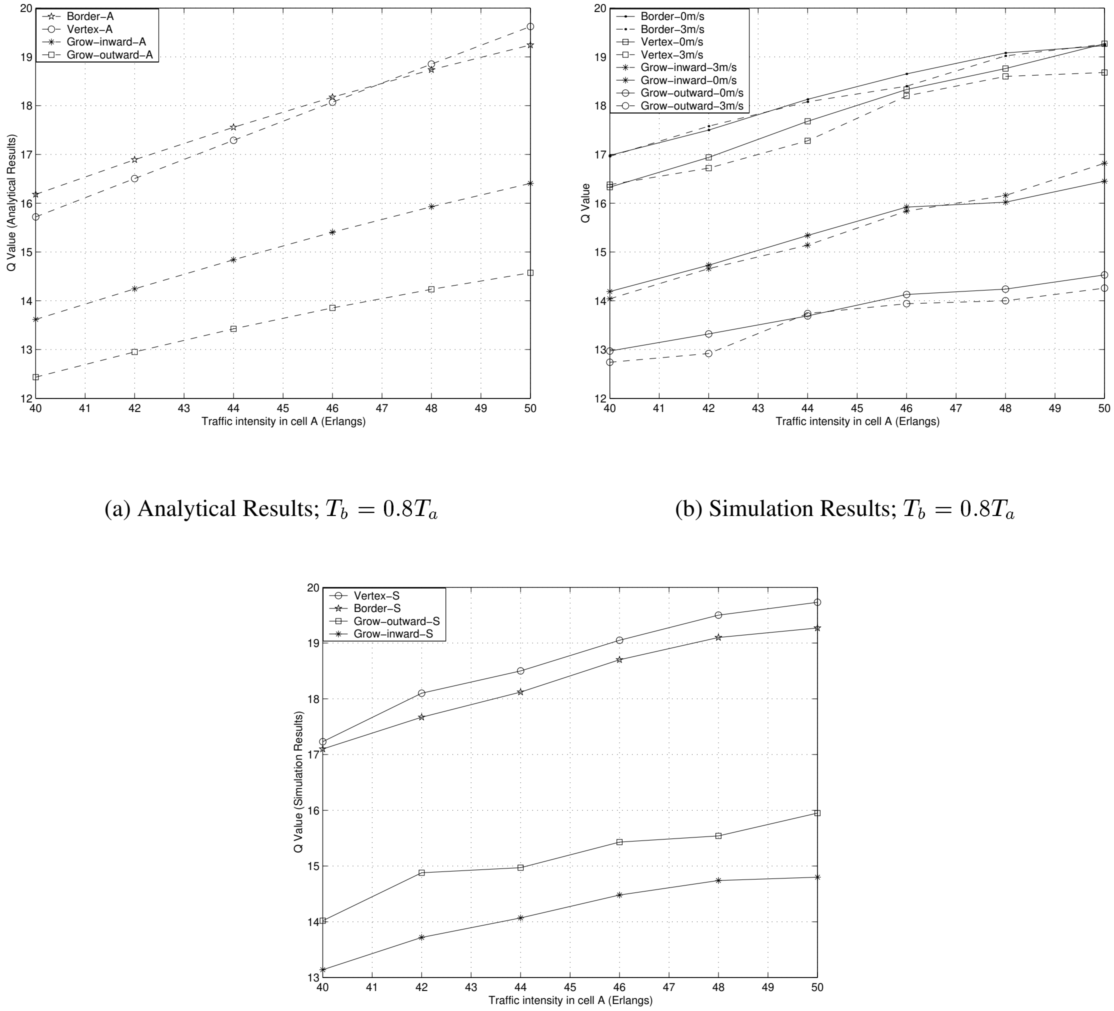
(c) Growing inside

(d) Growing outside

Figure 5.7: Four scenarios of ARS placement in the simulated system.

50 units of bandwidth are allocated for one cell, and for simplicity, each connection requires 1 unit bandwidth. In order to obtain converged statistical results, we have simulated $6,400$ MHs which are uniformly distributed in the system, and run the simulation for 100 hours for each traffic intensity before collecting the results. The traffic intensity is measured in Erlangs which is the product of the request arrival rate (Poisson distributed) and the holding time (exponentially distributed). In addition, we have used location dependent traffic pattern by default. More specifically, assuming cell A is the hot spot, the traffic intensity in cell B is about 80% of that in cell A. Six ARSs with 500 m transmission range have been simulated in four scenarios (see Figure 5.7 (a)-(d)), which implement the different ARS placement approaches described in Section 5. Figure 5.7(a) and (b) show six seed ARSs placed according to the border and the vertex approaches, respectively, while in Figure 5.7(c) and (d), there are 3 seed ARSs placed at the borders and 3 additional ARSs growing from the seeds inward and outward cell A, respectively.

We have obtained Q values of all six ARSs for different placement approaches from the simulation, and compared them with the analytical results in Figure 5.8. As we can see, the analytical results (in Figure 5.8(a)) and simulation results (in Figure 5.8(b)) show a very similar trend. The reason that the Q values obtained from the simulation are usually higher than those from the analysis is that the blocking probability without relaying is used in Equation 5.4 through 5.7, which is higher than the real blocking probability in iCAR (with relaying). In addition, as shown in Figure 5.8(b), the mobility of MHs has little affect on the Q values (although it does affect the connection dropping probability as to be shown later). In all cases within the normal operation range of an iCAR system (e.g., the traffic intensity of cell A is from 40 to 50 Erlangs), the grown ARSs yield lower Q values than that of the seed ARSs as we expected. However, the Q values of the border and the vertex approaches are very close, and when the traffic intensity is high, the vertex approach may result in higher Q values. This is because even though the ARSs in the border approach still cover more active connections in such a situation, a large fraction of the covered connections is *nonrelayable* because of the high blocking probability in the neighboring cells. On the other hand, the covered connections in the vertex approach may be relayed to either of the two neighboring cells, and therefore has higher Q values. As we discussed earlier, the real blocking probability of the cells



(a) Analytical Results; $T_b = 0.8T_a$

(b) Simulation Results; $T_b = 0.8T_a$

(c) Simulation Results; Speed=0m/s; $T_a = 0.8T_b$

Figure 5.8: Q values of different ARS placement approaches.

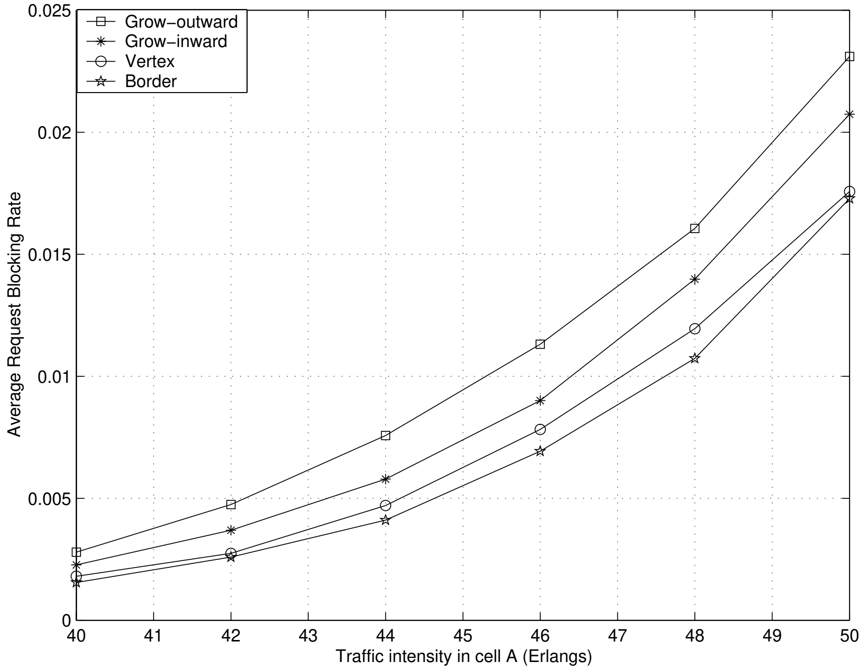


Figure 5.9: Call Blocking rates for different ARS placement approaches, MH Speed=0m/s.

in iCAR is lower than that we used in the analysis, thus the intersection point of the curves representing the Q values of the border and the vertex approaches in the simulation occurs at a higher traffic intensity than that in analysis (comparing Figure 5.8(a) and (b)). Finally, while Figures 5.8(a) and (b) are for the cases where cell A is the hot spot ($T_b = 0.8T_a$), Figure 5.8(c) shows the results when $T_a = 0.8T_b$. As one would expect, the vertex approach out-performs the border approach as an ARS at a vertex covers more area with a high traffic intensity than an ARS at a shared border of two cells. For the same reason, the ARS growing outward cell A has a higher Q value than the ARS growing inward cell A. The request blocking rates of MHs in the systems with different ARS placement approaches are shown in Figure 5.9. As we can see, an iCAR system with a higher Q value usually has a lower blocking rate. The results have also verified the usefulness of the three rules of thumb established in Section 5. More specifically, the border ARS placement has the lowest blocking rate among all of these approaches, which may be kept below 2% (the acceptable level) even when the traffic intensity is as high as 50 $Erlangs$. As a comparison (though the results are not shown), if six ARSs are randomly placed in the seven cells of the system (with the Q value being close to 0), the request blocking rate is from about 2% to above 10% when the traffic intensity of cell A increases from 40 to 50 Erlangs.

Although the MHs mobility may affect the dynamics in relaying capability of an iCAR system due to switch-over, our results indicate that the blocking rates in all ARS placement approaches increase very little with the MHs mobility. On the other hand, MH mobility affects the connection dropping probability significantly (see Figure 5.10[1]). More specifically, the dropping probability increases from 0 to the order of 10^{-3} and 10^{-2}, respectively, when the maximum MH moving speed increases from $0m/s$ to $1.5m/s$ and $15m/s$. Although the seed ARS placement approaches (i.e., the border and the vertex approaches) still perform better than the grown ARS placement approaches in terms of connection dropping rate, the difference among them is not as obvious as that in terms of the connection blocking rate. In addition, note that the vertex approach has a lower connection dropping rate than that of the border approach. This is because, when an active MH moves from one cell (i) to another cell (j), although there is the same probability that the MH is covered by ARSs at the moment crossing the shared border of the two cells, the ARS coverage is $\frac{2S}{3}$ in cell j in the vertex approach, which is larger than that in the border approach ($\frac{S}{2}$). The larger ARS coverage implies the longer time that the ARS can support the MH via relaying, and consequently results in a lower connection dropping rate.

We note that the proposed rules of thumb are based on the assumption that each ARS has an unlimited bandwidth at its R and C interface, that is, it can relay as many connections as needed to a BTS (provided that the BTS has free bandwidth). However, when there is only limited CBW, the grow-outward (cell A) approach may not be affected as much as the grow-inward approach, and hence the two approaches may perform equally well (or bad). This is because the amount of CBW determines the amount of traffic that can be relayed from cell A to its neighboring cells and consequently becomes the performance bottleneck, and placing an ARS outside cell A will increase the total amount of CBW (used to relay traffic from cell A to cell B) available to the ARS cluster, while placing inside cell A will not. Nevertheless, in a real system, only the connection requests that would be blocked without relaying, which is a small portion (e.g. about 5% of total requests

[1] For simplicity, we assume that there is no priority given to the handoff attempts over new connection attempts.

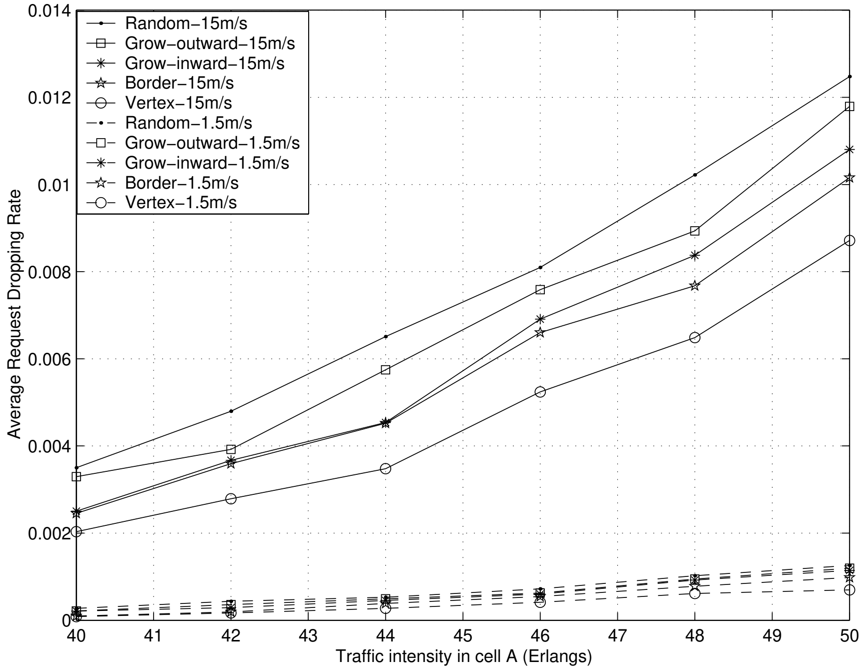


Figure 5.10: Call Dropping rates for different ARS placement approaches.

if the initial blocking rate is 5%), will be supported by relaying although the relayable traffic (i.e., the Q value) may be much higher than that, and thus the assumption of having enough relaying bandwidth is valid in most situations, and the presented rules of thumb will be good guidelines for ARS placement.

Chapter 6

Signaling and Routing Protocols

The explosive growth of Internet, and in particular, the introduction of IP version 6 (resulting a huge address space and a phenomenal increase in the number of mobile users and wireless nodes that all have their own globally unique IP addresses) has stimulated the interest in the development of packet switching data services in existing and future cellular systems. While General Packet Radio System (GPRS) and 3G system can support packet access in addition to conventional voice traffic, it is desirable to have a seamless converged next generation system that is based on IP techniques and connection oriented services at the same time. This is because IP is a connectionless protocol, and as such it is difficult for it to meet the quality of service (QoS) requirements of real time traffic.

To introduce IP [99] into wireless mobile networks, carriers and infrastructure providers face a major challenge in meeting the increased bandwidth demand of mobile Internet users, and the bursty and unbalanced IP traffic will exacerbate this problem of limited capacity in existing cellular systems. iCAR, with its ability to leverage both the cellular and ad hoc relaying techniques to increase system's capacity, is a promising evolution path for the cellular systems. Nevertheless, in order for iCAR to support real time IP-based applications in wireless mobile environment, efficient signaling and routing protocols are needed to set up a relaying path with reserved bandwidth, so as to guarantee the required QoS.

Although several connection oriented signaling protocols (e.g. RSVP, MPLS, etc) have been proposed for wired data networks, very little research has been done on QoS-capable connection oriented packet-switching in the wireless networks, especially for integrated networks with heterogeneous technologies such as cellular and ad hoc relaying. In this section, we describe the proposed signaling and routing protocols for iCAR to establish and release bandwidth guaranteed connections possibly involving ARS relaying. Such protocols aim to addressing the QoS need of IP based real time applications. Since iCAR integrates the cellular system [18, 86, 100, 101] and ad hoc networks [43, 48, 49, 102, 103, 104, 105], the signaling and routing protocols are a hybrid of those two systems, requiring novel design to search for relaying routes, for example, in order to achieve a high efficiency.

In this chapter, we describe the proposed signaling and routing protocols for iCAR to establish and release bandwidth guaranteed connections possibly involving ARS relaying. Since iCAR integrates the cellular system and ad hoc networks, the signaling and routing protocols have to take into consideration the characteristics of those two systems, and thus require novel designs in order to be able to establish relaying routes when needed. Further more, we will study the performance of the proposed protocols in terms of request blocking rates and signaling overhead via analysis and simulations.

The signaling and routing protocols of iCAR consist of the following three components.

- **Connection Request/Release Signaling**. We will focus on a establishing relaying path from a MH to a BTS (via one or more ARSs), instead of between MH_Y and MH_Y$'$ as in Figure 3.2 (b) for example. In order to support relaying, the signaling protocol in existing cellular systems has to be extended. In addition, the signaling protocol also needs to be tuned in concert with the other two components to be described below to optimize performance.

- **Routing (involving relaying via ARSs)**. The following two types of information are especially useful for QoS routing in iCAR, the topology (or connectivity) information on ARSs and BTSs, and the available bandwidth information including the *Relaying Bandwidth* (RBW) and *Cellular Bandwidth* (CBW) [1]. Since ARSs have low (or no) mobility, the topology is fairly stable and rarely needs to be updated. On the other hand, the bandwidth information may change frequently when the traffic load is high. Both topology and bandwidth infor-

[1] This is because the signal propagation delay will be largely determined by the number of hops.

mation may be maintained by either PSC or ARSs. When PSC maintains the bandwidth information, it treats the ARSs as if they were BTSs, and performs routing (i.e., determine the relaying path). On the other hand, if the bandwidth information is maintained by ARSs, it may or may not be up-to-date, depending on the frequency at which the updated information is exchanged among ARSs.

- **Bandwidth Reservation/Release**. The bandwidth along a selected relaying path can be reserved in two ways. One way is for PSC to multicast a reservation message including the entire path information to all ARSs, the MH (source) and the BTS (destination) along the path. The ARSs who are on the specified relaying path will reserve the requested bandwidth upon receiving this message. This method results in a fast bandwidth reservation process but the reservation message to ARSs and MHs consumes the scarce broadcasting bandwidth (e.g., the common control channel (CCH) in all cells that the relaying path traverses). The other way is for the source (or destination) to perform forward (or backward) hop-by-hop reservation, in which a reservation message is sent from the source (or destination) to the destination (or source) along the relaying path, reserving bandwidth at each hop. This method may result in a longer path setup delay and more signaling messages, but they consume the control channel bandwidth at the R-interface, which can be reused by ARSs (or MHs) that are far apart from each other (even though they may be in the same cell). Similarly, one can use these two approaches for bandwidth releasing.

Given the above primary choices for ARS routing and bandwidth reservation, we may devise the following protocols as shown in Table 6.1, where ”MCAST” and ”HOP” stand for reserving bandwidth via multicasting and hop-by-hop relaying, respectively. The symbol ”/” means that the global information is not maintained by either PSC or ARSs.

Briefly, In Protocol 1, PSC maintains both the topology and the bandwidth (especially RBW and CBW of ARSs) information, and selects a relaying route when needed. It is efficient in terms of signaling overhead and route optimization. The drawback of this protocol is that PSC becomes the single point of failure and performance bottleneck (due to processing overhead). On the other hand, in Protocols 2 and 3, the ARSs maintain the routing information and perform routing. The main

Protocols		1	2	3
Routing Information	Topology	PSC	ARS	ARS
	Bandwidth	PSC	ARS	/
Bandwidth Reservation/Release		MCAST	HOP	HOP

Table 6.1: The candidates of the protocols for iCAR

difference between Protocols 2 and 3 is as follows. In Protocol 2, the ARSs maintain the complete bandwidth (RBW/CBW) information of remote ARSs and use it to guide the routing, while in Protocol 3, the ARSs don't have the RBW/CBW information of remote ARSs and thus have to verify the availability of RBW/CBW along a possible relaying route. Note that, although the hop-by-hop bandwidth reservation/release is natural in Protocols 2 and 3, multicasting-based bandwidth reservation/release is also possible but will not be considered. Other choices (such as maintaining only the topology information at PSC, or only the RBW/CBW information at PSC/ARSs, or no information at all) will not be considered either.

In the rest of this chapter, we will discuss these three protocols in more detail. In particular, we will describe and evaluate these protocols in terms of their signaling/control overhead under the assumption that the signaling bandwidth is always enough even when BTSs or ARSs may have run out of bandwidth for data transmission. For simplicity, we assume that all BTSs, ARSs and MHs have unique addresses, and each MH sets up at most one connection.

6.1 Protocol 1: A PSC-Assisted Protocol

In this section, we introduce Protocol 1, also referred to as a PSC-assisted protocol, which takes advantage of the cellular infrastructure by letting PSC maintain both the topology and bandwidth (including RBW and CBW) information and perform routing. We will first present the signaling protocol for connection setup and release, and then discuss how routing and bandwidth reservation are done.

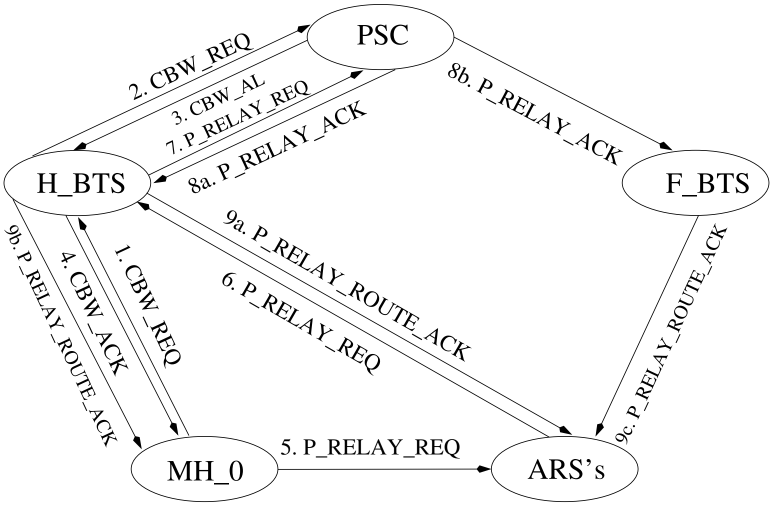


Figure 6.1: PSC-assisted Signaling protocol for connection request via primary relaying.

6.1.1 The Signaling Protocol For Connection Setup and Release

We first discuss the protocol for connection setup, which is illustrated in Figure 6.1. When a MH (e.g. MH_0) needs to set up a QoS guaranteed connection to the core network, it sends a CBW request message (CBW_REQ) to the BTS located in the same cell, (called home BTS) denoted by H_BTS (see step 1 in Figure 6.1). H_BTS will forward the message to PSC (see step 2 in Figure 6.1), which is responsible for admission control and bandwidth allocation. If there is enough CBW available, PSC responds with a CBW allocation message (CBW_AL) to H_BTS, which in turn generates a positive acknowledgement (CBW_ACK) to MH_0 (see steps 3 and 4 in Figure 6.1), and the connection request is satisfied. Otherwise MH_0 will receive CBW_NAK instead (although this is not shown in Figure 6.1). So far, this is pretty much the same as the process in a conventional cellular system. What is different in iCAR is that, instead of dropping the connection request, an attempt to establish a relaying path will be made. Note that, the PSC cannot initiate the relaying process at this point because it doesn't know whether MH_0 is covered by any ARSs, nor does it know which ARSs cover MH_0 (unless the PSC has the global *location* information of ARSs and MHs via the use of GPS, for example).

Connection Setup via Primary Relaying

To set up a primary relaying path, when there is not enough CBW available, H_BTS will start a timer T1[2] after sending NAK to MH_0. Upon receiving NAK, MH_0 will try primary relaying by broadcasting a P_RELAY_REQ message to nearby ARSs, which includes an unique sequence number (see step 5 in Figure 6.1). If MH_0 is not covered by any ARSs, H_BTS won't receive any response before T1 times out and thus primary relaying will fail. In this case, the request will be rejected unless secondary relaying is tried and succeeds (as to be discussed later). Otherwise, the ARSs, upon receiving P_RELAY_REQ, will forward the message to H_BTS (via MH_0 if the ARS is in a different cell other than that H_BTS locates in) which aggregates all requests with the same sequence number and sends it to PSC (see steps 6 and 7 in Figure 6.1). Of course, if the BTSs are just transceiver units, thus cannot perform any operations on the packets, they have to forward the request messages from ARSs one by one without aggregation.

Based on the system topology and RBW/CBW information, PSC will look for the shortest relaying path from one of the requesting ARSs to one of the non-congested BTSs, denoted by F_BTS (using the routing protocol to be discussed later in Sec 6.1.2). If there is a relaying route available, PSC will build a P_RELAY_ACK message including the full relaying route information (i.e., all nodes on the relaying route), and send it to relevant BTSs, i.e., H_BTS, F_BTS, and any other BTSs in the cells that the relaying path traverses (see step 8 in Figure 6.1). F_BTS will reserve CBW for the use by the gateway ARS, and all relevant BTSs will multicast the message to inform the ARSs in their cells to prepare for relaying (by reserving the RBW between adjacent ARSs as specified in the relaying path). In the meanwhile, H_BTS lets the requesting MH know that the relaying route is ready (see step 9 in Figure 6.1). Here (as well as in secondary relaying to be discussed below), we assume that the bandwidth reservation is done by multicasting. Nevertheless, the alternative approach (hop-by-hop reservation) can also be used with minor modifications, and we will compare their performance in Section 6.5.

Connection Setup via Secondary Relaying

If primary relaying fails, MH_0 will try secondary relaying by sending a S_RELAY_REQ message (also with a sequence number) to H_BTS (see step 1 in Figure 6.2), which will multicast the message to all active MHs (say MH_1, MH_2, ..., MH_n) that are using CBW in the same cell (see

[2] The timeout value of T1 should be limited by the maximum delay budget allowed for primary relaying.

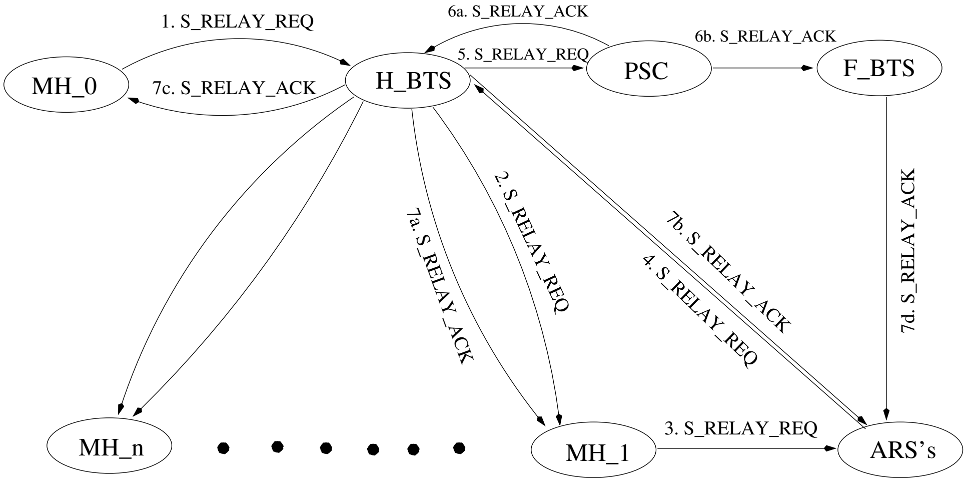


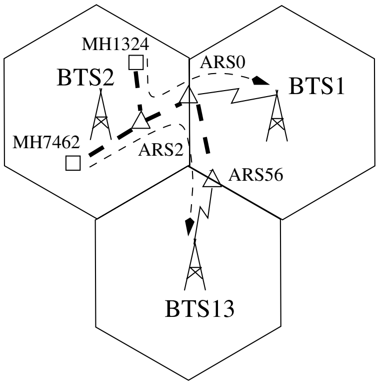
Figure 6.2: PSC-assisted Signaling protocol for connection setup via secondary relaying.

step 2 in Figure 6.2) and start a timer T2[3]. When MH_i ($1 \leq i \leq n$) receives a S_RELAY_REQ, it forwards the message to its nearby ARSs (see step 3 in Figure 6.2). The ARSs will take only the first request with the same sequence number and send it to H_BTS (possibly via a MH). If H_BTS receives no response before T2 times out, secondary relaying fails. Otherwise, similar to primary relaying, H_BTS aggregates the messages and send a single request to PSC (see steps 4 and 5 in Figure 6.2). PSC then selects a "best" MH (again using the routing protocol to be discussed later in Sec 6.1.2) for relaying, say MH_j, and multicasts the relaying path and an acknowledge to all relevant BTSs, ARSs and MHs (see Steps 6 and 7 in Figure 6.2). After receiving this message, F_BTS and ARSs on the relaying route will reserve requested bandwidth. Meanwhile, MH_j will switch over to the R-interface and release its CBW from H_BTS, which in turn assign it to MH_0. All other MH_i ($i \neq j$) will not be affected. If no relaying route is available, MH_0 will receive a S_RELAY_NAK, and its request is rejected.

Connection Release

The connection release in the PSC-assisted protocol is done by MH_0 sending a release request to PSC via BTSs. PSC will update the bandwidth information and multicast the release request to all relevant BTSs and ARSs (if the request was supported by relaying). When an ARS receives a release request, it will release the reserved bandwidth. Note that when H_BTS is no longer congested, PSC

[3] Similar to T1, the timeout value of T2 should be limited the maximum delay budget allowed for secondary relaying.



(a) Two relaying routes are established.

Conn_ID	Next Hop	Previous Hop
……	……	……
MH_7462	ARS_56	ARS_2
……	……	……
……	……	……
MH_1324	BTS_1	ARS_2
……	……	……

(b) A switching table maintained at ARS_0.

Figure 6.3: An example of ARS Routing.

may assign free CBW to a MH which is using a relaying path (e.g., MH_0 in primary relaying or MH_j in secondary relaying) and release the relaying path before the MH completes its data transmission. However, such a switch-over from the R-interface to the C-interface by a MH will be performed only when the free CBW in H_BTS exceeds some threshold in order to avoid thrashing (i.e., consecutive switching-over from the C-interface to the R-interface by one MH, and from the R-interface to the C-interface by another MH).

6.1.2 Routing and Bandwidth Reservation

In the PSC-assisted protocol, routing and bandwidth reservation is performed by PSC. More specifically, when ARSs power on, they discover neighbors including nearby ARSs and BTSs, and send the neighbor information to PSC via BTSs. PSC builds a network topology, maintains available bandwidth information, and makes bandwidth allocation and deallocation for BTSs and ARSs. On the other hand, the ARSs neither maintain the routing table nor exchange the routing information among themselves. They simply relay the requests from MHs to PSC and in turn, receive the relaying route from PSC and store the information in a switching table (as shown in Figure 6.3(b)), and forward data packets according to the switching table.

When PSC receives one or more primary or secondary relaying requests (forwarded by one or more ARSs), it will try to find the best relaying path from one of these ARSs to one of the BTSs. Since the notion of the best relaying route depends on the amount of requested bandwidth, the source ARSs and the destination BTSs, it does not make sense for PSC to maintain a routing table, instead, a routing algorithm is invoked for each individual relaying request. This implies that, we need an algorithm which can find a best path from multiple sources to multiple destinations. However, such an any-to-any problem can be easily transformed to the conventional problem of finding a shortest path from one source to one destination by adding a dummy source node S connected to every one of the multiple sources with a dummy edge of cost 0 (i.e., infinite bandwidth), and adding another dummy destination node D connected to every one of the multiple destinations with a dummy edge of cost 0. In addition, the cost of the links may be assigned by the following two approaches involving tradeoffs between bandwidth consumption and load balancing.

- **Minimum Distance-Bandwidth**. In this approach, when a link connects two ARSs, its cost will be 1 if both ARSs have available RBW, which is no less than the amount requested, denoted by RB, or ∞ otherwise. When a link connects an ARS and a BTS, its cost will be 1 if the available CBW between ARS and the BTS is no less than RB, or ∞ otherwise. A shortest path algorithm is applied to find a best relaying path. Accordingly, this approach will minimize the number of hops, the amount of consumed bandwidth and the delay (assuming each hop has the same queuing delay) along the relaying path.
- **Widest Path**. This approach assigns each link a cost of ∞ if its available bandwidth (either RBW or CBW) denoted by AB is less than RB, or $\frac{1}{AB}$ otherwise. When using the shortest path algorithm, this approach tends to balance the traffic load of the system.

Note that, as discussed, any resource shortage may result in the failure of relaying.

6.2 Protocol 2: A Link-State Based Distributed Protocol

While the PSC-assisted protocol can achieve efficient routing with a low signaling overhead, a PSC becomes the single point of failure and performance bottleneck. In a heavily congested system, the MHs may experience a long waiting time before they get responses from the system. These are two

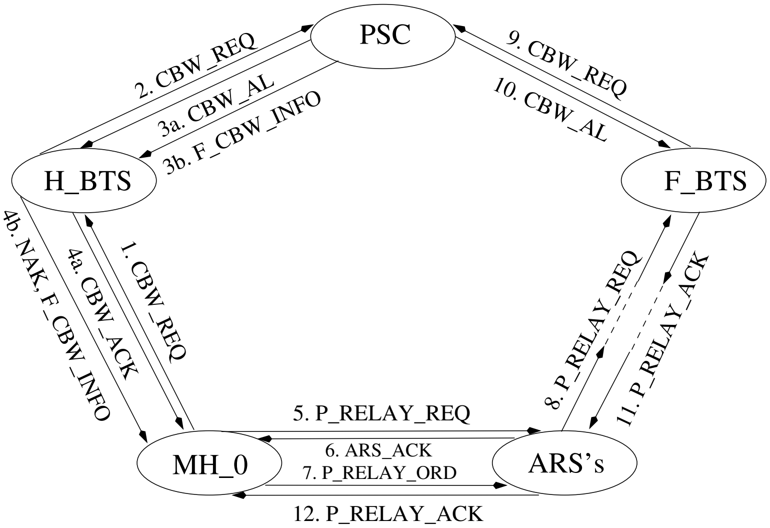


Figure 6.4: Link-state based distributed Signaling protocol for connection request via primary relaying.

major motivations to develop the distributed signaling and routing protocol for iCAR to be described in this section, in which each ARS exchanges link-state packets to maintain the topology as well as the bandwidth information. Again, we will first discuss the signaling protocol for connection setup and release, and then describe how routing and bandwidth reservation are done.

6.2.1 The Signaling Protocol For Connection Setup and Release

The signaling protocol is illustrated in Figure 6.4. Similar to the PSC-assisted protocol, MH_0 will receive a NAK if there is not enough bandwidth available when it requests a QoS guaranteed connection (see steps 1, 2, 3 and 4 in Figure 6.4). In addition, since PSC is the only single entity that has the CBW information of all BTSs, it is natural for PSC to send the up-to-date CBW information of BTSs to the requesting MH (see steps 3b and 4b in Figure 6.4), which can then initiate the relaying process.

Connection Setup via Primary Relaying

After receiving NAK and the CBW information of BTSs in the system, MH_0 will try primary relaying by switching to the R-interface. As in Protocol 1, it will contact the nearby ARSs by broadcasting a primary relaying request message (P_RELAY_REQ), but the difference is that

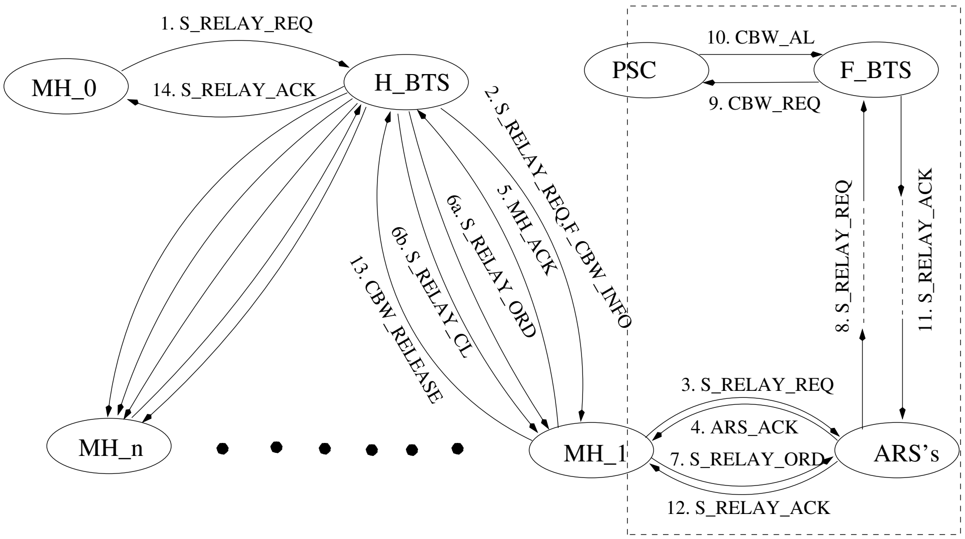


Figure 6.5: Link-state based distributed Signaling protocol for connection request via secondary relaying.

here, the message includes a set of nearby BTSs with available CBW, and in addition, MH_0 will start a timer (T1) at the same time (see step 5 in Figure 6.4). If the MH is not covered by any ARSs, it won't receive any acknowledge before T1 times out and thus primary relaying will fail. When a nearby ARS receives the relaying request, it computes the best relaying path (to be discussed later) and responds with an ARS_ACK message including the minimum cost of relaying to one of the desirable BTSs (see step 6 in Figure 6.4). When T1 times out, MH_0 will send an primary relaying order (P_RELAY_ORD) to the ARS that has responded with the lowest relaying cost (see step 7 in Figure 6.4). The ARS in turn tries to establish the relaying path in a normal (e.g., forward) hop-by-hop reservation fashion (see steps $8-11$), and send a P_RELAY_ACK to MH_0 (see step 12). Note that, since the ARSs may have computed the route based on out of date topology/bandwidth information, it is not guaranteed that the relaying path can be successfully established. For this reason, MH_0 will start another timer T2 which has a longer timeout interval than that of T1[4], and if the current attempt to establish a relaying path fails before T2 times out, MH_0 will try to establish an alternate relaying path until either T2 times out or a relaying path is established.

[4] Here, the sum of the timeout value of T1 and T2 should be no longer than the maximum delay budget for primary relaying

Connection Setup via Secondary Relaying

If primary relaying fails, MH_0 will send a second relaying request (S_RELAY_REQ) message to H_BTS, and try secondary relaying (see step 1 in Figure 6.5). After receiving S_RELAY_REQ, H_BTS will contact the active MHs by multicasting a S_RELAY_REQ message which includes a set of BTSs with enough free CBW (see step 2 in Figure 6.5), and starts a timer (T3). Whenever a MH (e.g. MH_1) receives the S_RELAY_REQ message, it will try to contact nearby ARSs in the same way as that used for primary relaying (see steps 3-4 in Figure 6.5) except that here each MH will present to BTS the best relaying path from itself to a F_BTS in a message MH_ACK (see step 5). If H_BTS receives no response from ARSs before T3 times out, secondary relaying fails. Otherwise (if H_BTS receives more than one MH_ACK messages), it will choose the best secondary relaying path, and send a S_RELAY_ORD (see step 6a in Figure 6.5) to the MH that responded with the best secondary relaying path and a S_RELAY_CL message to all other MHs to cancel their further relaying actions. After receiving a S_RELAY_ORD message, the MH (e.g. MH_1) will try to set up the relaying path (see steps 7-12 in Figure 6.5), and when succeeds, release its CBW (see step 13 in Figure 6.5) . If none of the active MHs can do a successful relaying, H_ BTS sends a NAK to MH_0 and its request is blocked.

Connection Release

When data transmission is completed, the MH sends a connection release message to either the BTS (if without relaying) or the proxy ARS (if with relaying). The ARSs on the relaying path release the reserved RBW/CBW, remove the corresponding entry in its switching table, and forward the release message to the BTS which provides CBW, and the latter will release the bandwidth and update the bandwidth information to the PSC. Similar to the connection release in Protocol 1, a relaying path may be released when H_BTS is no longer congested.

6.2.2 Routing and Bandwidth Reservation

In this subsection, we discuss multi-hop relaying among ARSs. The major difference between PSC-assisted protocol and the distributed protocols (Protocol 2, as well as Protocol 3 to be described later) is in routing. In the PSC-assisted approach, routing is done by PSC, while in Protocol 2, the ARSs need to maintain the topology and bandwidth information, and perform routing by themselves. Here, we propose a modified link state protocol for ARS routing.

When an ARS powers on, it discovers the reachable BTSs and neighboring ARSs. Then, the ARS builds and distributes the link state packets which include the addresses of its neighbors (ARSs and BTSs) as well as the bandwidth information. Based on these link state packets, each ARS can construct the cluster graph. The ARSs send the update link state packets once a while or only when needed. In the latter case, they are sent only when they have lost their relaying ability (because e.g., their CBW or RBW reduces to zero), or subsequently they become able to support new relaying requests.

Whenever an ARS receives a relaying request which includes the source MH address, the requested bandwidth, and a set of foreign BTS addresses, it computes a best relaying path in a way similar to that used by PSC in Protocol 1. More specifically, the ARS creates a dummy destination D and connects it to the set of foreign BTSs with the cost of 0 (i.e., infinite bandwidth). It then finds a best path from itself to the dummy destination. Let the foreign BTS which is along such a path be denoted by F_BTS. After creating an entry and stores the routing information into its switching table shown in Figure 6.3(b), it forwards the request to the next hop along the computed best relaying path, until the request reaches F_BTS.

6.3 Protocol 3: A Simple Route-Searching Protocol

In Protocol 2, both the topology and the available bandwidth information is maintained, and the routing function is performed in a distributed fashion by ARSs. However, this requires all ARSs to have a high computing power. In this section, we describe a simple route-searching protocol, which discovers the relaying routes using a depth-first search to eliminate the need for intensive computing and maintaining the RBW/CBW information of remote ARSs.

6.3.1 The Signaling Protocol For Connection Setup and Release

In the simple route-searching protocol, connection setup and release is very similar to that in Protocol 2. The only difference is as follows. In the link-state based protocol, since each ARS has

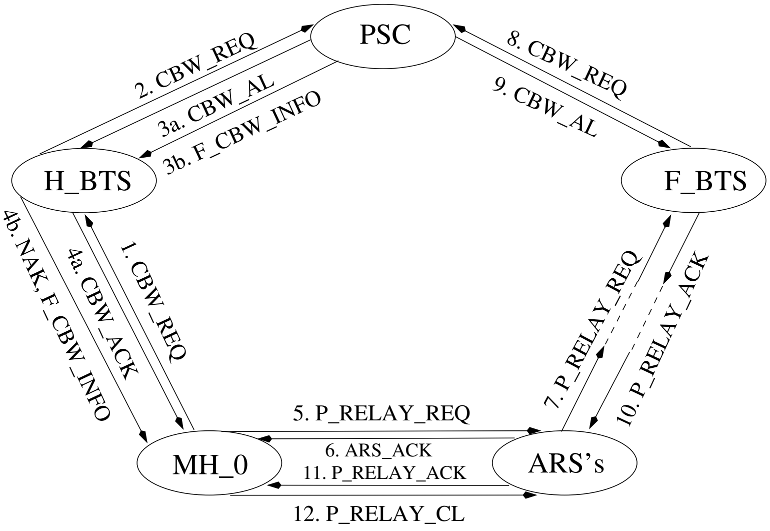


Figure 6.6: A simple route-searching protocol for connection request via primary relaying.

the global routing information, MH_0 can choose the best route according to the ARSs' responses it receives before T1 times out (see step 6 in Figure 6.4 and step 5 in Figure 6.5 for primary and secondary relaying, respectively), and asks only one ARS to set up the relaying route and reserve bandwidth. However, in the simple route-searching protocol, ARSs don't have the RBW/CBW information of remote ARSs, and therefore, it is necessary for MH_0 to ask multiple ARSs to *actually set up* the relaying routes simultaneously in order to achieve a high probability of finding a relaying path successfully. More specifically, when the nearby ARSs of MH_0 in Figure 6.6 in primary relaying (or MH_1 through MH_n in Figure 6.7 in secondary relaying) receive the relaying request, they will look up their routing table (to be discussed later in Section 6.3.2) and respond with a positive ARS_ACK if at least one of the desirable BTSs is reachable (topologically speaking only) as shown in step 6 in Figure 6.6 in primary relaying (or step 4 in Figure 6.7 for secondary relaying). After that, MH_0 sets another timer (T2) while the ARSs that have responded positively try to establish a relaying path as in Protocol 2. If the relaying request message is eventually relayed to a F_ BTS which has free CBW, the F_BTS will reserve the CBW and sends back a positive acknowledge (see steps 7-11 in Figure 6.6 or step 5-9 in Figure 6.7).

In primary relaying, MH_0 will start data transmission via the relaying route upon receiving the first acknowledge, and multicast a relaying cancel message P_REALY_CL to all other routes to

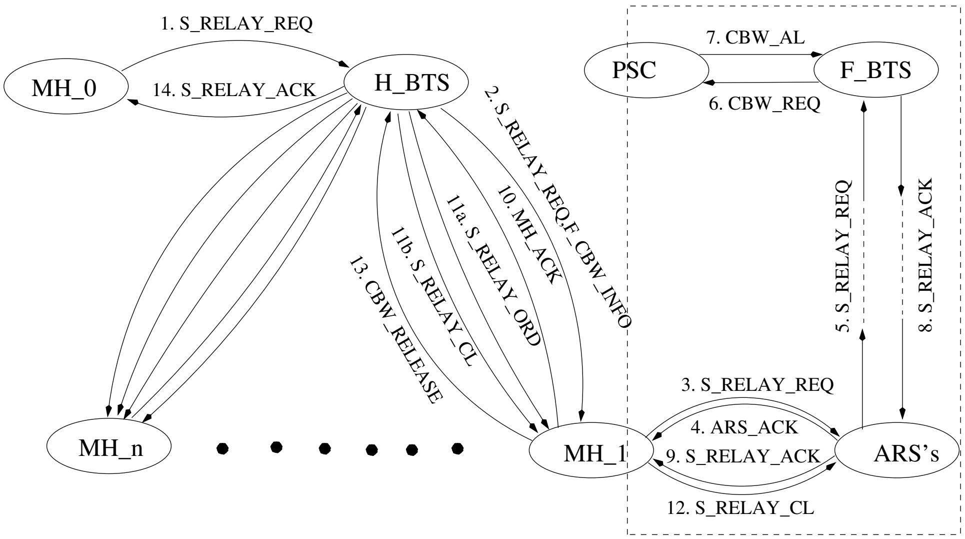


Figure 6.7: A simple route-searching protocol for connection request via secondary relaying.

release the reserved bandwidth as shown in step 12 in Figure 6.6. The P_RELAY_CL is also sent out when T2 times out.

In secondary relaying, MH_i ($1 \leq i \leq n$) will forward the first received acknowledge from ARSs to H_BTS, which in turn sends a secondary relaying order message S_RELAY_ORD to MH_i (i.e., the first MH which responded to the secondary relaying request) and a secondary relaying cancel message S_RELAY_CL to all other active MH_j ($j \neq i, 1 \leq j \leq n$). Upon receiving S_RELAY_ORD, MH_i can switch to R-interface and start data transmission via the selected relaying path, and multicasting a S_RELAY_CL message (see step 10 in Figure 6.7) to all other nearby ARSs to release the reserved bandwidth. The MH_j, who receives the S_RELAY_CL from H_BTS, will also multicast the cancel message to its nearby ARSs. Similar to the case for primary relaying, the S_RELAY_CL is also sent out when T2 times out.

The signaling protocol for connection release is very much the same as that in the link-state based protocol discussed in Sec 6.2 and is not discussed in further detail.

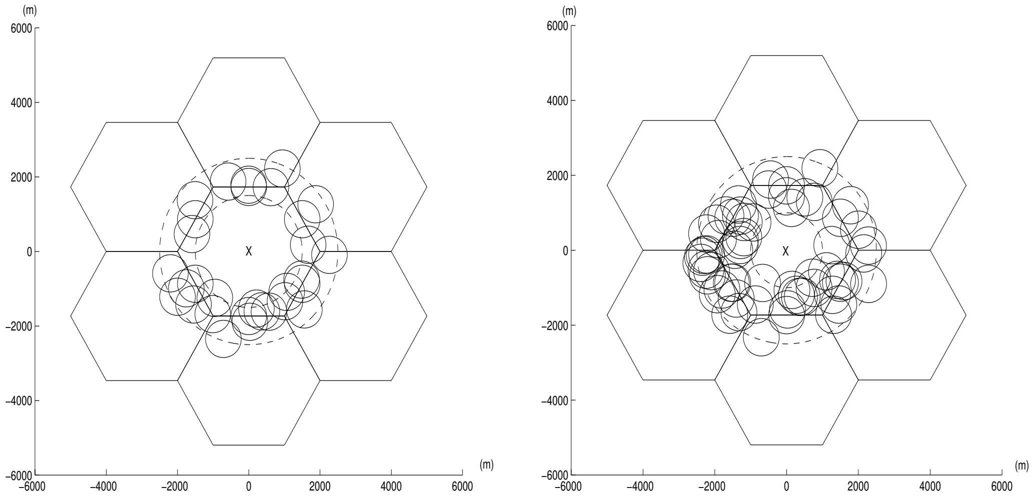
6.3.2 Routing and Bandwidth Reservation

In this approach, ARSs do not maintain the CBW/RBW information of other ARSs. Instead, this approach takes the advantage of the fact that the topology is fairly stable and thus each ARS can maintain a routing table based on the topology map. Unlike the switching table in Figure 6.3(b), each entry in such a routing table includes the address of a reachable BTS, the next hop to reach the BTS (the address of another ARS or the BTS itself) and the total number of hops to the BTS. Note that, since the number of BTSs, especially the reachable BTSs by an ARS via relaying in a system, is usually not large, it is feasible to include all reachable BTSs in a single routing table. In addition, since the size of each ARS cluster is small, more than one relaying routes instead of only the shortest path to a reachable BTS, each with a different next hop (ARS or BTS), can be stored in the routing table according to their distance. That is, there may be more than one entries for each reachable BTS. Even so, the maximum number of entries in its routing table, which is $N \times (M + 1)$ where N is the number of cells, and M is the number of neighboring ARSs, is still manageable.

Whenever an ARS receives a relaying request message which includes the source MH address and a set of foreign BTS addresses, it looks up the routing table to find all entries with matching destination BTSs. However, in order to limit the number of signaling messages due to further flooding, each ARS will attempt to establish one path at a time. More specifically, if only one foreign BTS is found in the routing table and free CBW (if the next hop is BTS) or RBW(if the next hop is ARS) is available, the ARS relays the message to the next hop. If there are more than one choices of next hop, one on the shorter path will be chosen first. If the destination BTS is reached, an ACK containing bandwidth information along the relaying path will be sent back to the source MH, and the relaying bandwidth will be reserved. If the request cannot be relayed further along the most preferred next hop, then the second choice next-hop will be tried. If no other choices are available, the ARS sends a negative acknowledgement (NAK) to the previous hop. Of course, such a root-search process initiated by a proxy ARS may be terminated earlier upon either a relaying path from another proxy ARS is found or T2 times out, as mentioned earlier.

6.4 Signaling Overhead Analysis

In this section, we analyze the signaling overhead of the proposed signaling and routing protocols, in terms of the average number of signaling messages (total received and transmitted) per satisfied connection request. We consider a cell X and its neighboring cells (see Figure 6.8), which are



(a) 30 ARSs

(b) 60 ARSs

Figure 6.8: The ARSs in an iCAR system.

controlled by a PSC. We assume that ARSs are randomly placed in the donut-shaped region of cell X, which is bounded by two dashed circles as shown in Figure 6.8, and for simplicity, there is no bandwidth shortage along the relaying path. Note that since the ARSs are randomly distributed, not all of them results in effective coverage [106]. In particular, some ARSs cannot relay traffic from one cell to another either directly or through other ARSs. We will first introduce the system parameters used in the analysis, and then discuss the signaling overhead for each protocol.

Table 6.2 lists the symbols to be used in the following discussion. The values of R, r, R_1, R_2, M and K are assumed to be given for a system. After the ARSs are placed in iCAR, the ARS coverage p may be estimated based on the system map (e.g., by evenly distributing a number of points and counting the fraction of them which are within the coverage of ARSs). Since the ARSs are randomly placed, and some of them are not able to relay traffic because they cover only one cell, the value of p is usually over-estimated. The request rejection rate B_o, B_p and B_s may be obtained either from a real system or by the analysis introduced in [15]. The average number of hops of an ARS relaying path ($\overline{H}$), the average number of active MHs covered by an ARS (N_M), and the average number of reachable ARSs for an active MH (N_A) may be obtained by the analysis shown

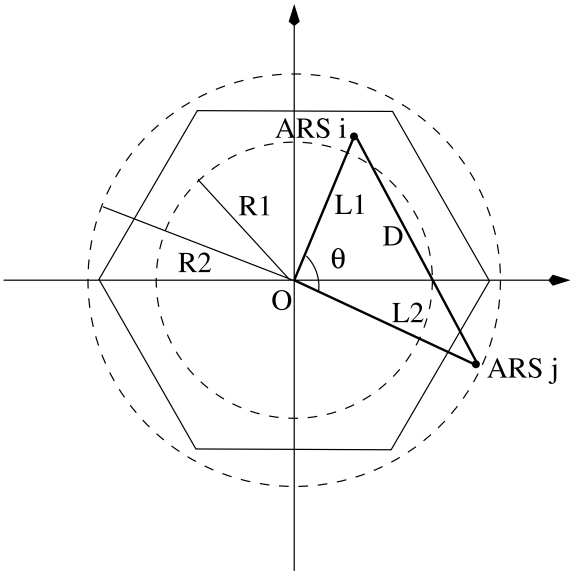


Figure 6.9: The distance between two ARSs.

as follows.

The average length of the ARS relaying path

We consider a system shown in Figure 6.8, in which the ARS are randomly placed in the donut-shaped area. We denote a random variable L to be the distance between an ARS to the origin (O), and a random variable θ to be the angle between two lines from two ARSs to the origin (see Figure 6.9). Both of them are assumed to be uniformly distributed with the density functions

$$f_{\mathbf{L}}(L) = \begin{cases} \frac{1}{R2-R1}, & R1 \leq L \leq R2 \\ 0, & otherwise \end{cases} \tag{6.1}$$

and

$$f_{\mathbf{\Theta}}(\theta) = \begin{cases} \frac{1}{\pi}, & 0 \leq \theta \leq \pi \\ 0, & otherwise \end{cases} \tag{6.2}$$

where $R1$ and $R2$ are the radius of the two bound circles.

According to the triangle equations, we can obtain the distance of two ARSs

$$D^2 = L_1^2 + L_2^2 - 2L_1L_2cos(\theta) \tag{6.3}$$

Since D is a function of three random variables L_1, L_2 and θ, we can derive the density function of D ($f_D(D)$) by defining two auxiliary variable $w_1 = L_1$ and $w_2 = L_2$. Accordingly, the Jacobian transformation is

$$J(L_1, L_2, \theta) = \begin{vmatrix} \frac{\partial D}{\partial L_1} & \frac{\partial D}{\partial L_2} & \frac{\partial D}{\partial \theta} \\ \\ \frac{\partial w_1}{\partial L_1} & \frac{\partial w_1}{\partial L_2} & \frac{\partial w_1}{\partial \theta} \\ \\ \frac{\partial w_2}{\partial L_1} & \frac{\partial w_2}{\partial L_2} & \frac{\partial w_2}{\partial \theta} \end{vmatrix} = \frac{D}{w_1 w_2 sin(arccos(\frac{w_1^2+w_2^2-D^2}{2w_1w_2}))}$$

and yields the jointed density function of D, w_1 and w_2

$$\begin{aligned} f_{Dw_1w_2}(d, w_1, w_2) &= \left|J^{-1}\right| f_{L_1L_2\theta}(w_1, w_2, arccos(\frac{D^2+w_1^2-w_2^2}{2w_1w_2})) \\ &= \frac{|J^{-1}|}{\pi(R2-R1)^2} \end{aligned} \quad (6.4)$$

Hence, the probability that D is smaller than the ARS transmission range r is

$$Pr\{D < r\} = \int_{R1}^{R2} \int_{R1}^{R2} \int_{|w_2-w_1|}^{min(w_2+w_1,r)} \frac{J^{-1}}{\pi(R2-R1)^2}\, dD\, dw_1\, dw_2 \quad (6.5)$$

Starting from an ARS, the probabilities that it may set up a relaying path including at least h hops is

$$\prod_{i=1}^{h}[1-(1-Pr\{D<r\})^{M-i}] \quad (6.6)$$

and thus the average number of hops of a relaying path is

$$\overline{H} = \sum_{j=1}^{M-1}\{\prod_{i=1}^{j}[1-(1-Pr\{D<r\})^{M-i}]\} \quad (6.7)$$

An estimation of N_A and N_M

We can estimate N_A and N_M which are used in the analysis as follows. According to the center-to-vertex distance of a cell (R) and the ARS transmission range (r), we can compute the cell size and ARS coverage to be $\frac{3\sqrt{3}}{2}R^2$ and πr^2, respectively. Assuming that there are K active MHs and M ARSs, then the MH density in a cell is $\frac{K}{\frac{3\sqrt{3}}{2}R^2}$, and thus $N_M = K \times \frac{\pi r^2}{\frac{3\sqrt{3}}{2}R^2}$.

For any given point inside the donut-shaped area defined by the two dashed circles whose size is $\pi(R_2^2 - R_1^2)$, in Figure 6.8, the degree of overlapped coverage of ARSs is $\frac{M\times\pi r^2}{\pi(R_2^2-R_1^2)}$. Hence, $N_A = p \times \frac{M\times\pi r^2}{\pi(R_2^2-R_1^2)}$.

The signaling overhead in the case without relaying is equal to the sum of connection requesting and releasing messages divided by the number of satisfied connection requests. Note that, only a satisfied connection request will result in a connection release. Therefore,

$$S_o = \frac{RQ_o + (1 - B_o) \times RL_o}{1 - B_o} \tag{6.8}$$

Similarly, the signaling overhead for primary and secondary relaying are

$$S_p = \frac{RQ_o + RQ_p \times B_o + (1 - B_o) \times RL_o + (B_o - B_p) \times RL_r}{1 - B_p} \tag{6.9}$$

and,

$$S_s = \frac{RQ_o + RQ_p \times B_o + RQ_s \times B_p + (1 - B_o) \times RL_o + (B_o - B_s) \times RL_r}{1 - B_s} \tag{6.10}$$

Without relaying, all three protocols have the same signaling overhead for each connection request (RQ_o). More specifically, PSC, BTSs and MHs will receive and send two (see steps 2 and 3 in Figures 6.1, 6.4 or 6.6), four (see steps 1, 2, 3 and 4 in Figures 6.1, 6.4 or 6.6), and two (see steps 1 and 4 in Figures 6.1, 6.4 or 6.6) messages for a connection request, respectively. In addition, PSC, BTSs and MHs will process two, four and two messages for each connection release, no matter the connection is via relaying or not (i.e., RL_o or RL_r). If a connection is not via relaying, ARSs will send and receive 0 message for releasing it. Otherwise, a number of messages (to be discussed later) will be processed by the ARSs on the relaying path. Also, the different protocols may result in different amount of signaling overhead for connection request when relaying is used, and we will analyze the values of RQ_p, RQ_s, and RL_r for the three protocols as follows.

6.4.1 Protocol 1

We first discuss the signaling overhead of Protocol 1. For the primary relaying, since MH_0 will send P_RELAY_REQ message (see step 5 in Figure 6.1), but receive an acknowledge only if it is covered by ARSs (with a probability of p), the total number of messages processed by MHs in primary relaying is $1 + p$. In the meanwhile, the nearby ARSs (i.e., the proxies) of MH_0 will send and receive $3 \cdot N_A$ messages in steps 5, 6 & 9a, and the ARSs on the relaying path will receive $\overline{H}$ messages if MH_0 is covered by ARSs (see steps $9a$ & $9c$). So, the total number of

R	the center-to-vertex distance of a cell
r	the transmission range of an ARS
R_1	the radius of the inner circle of the donut-shaped area
R_2	the radius of the outer circle of the donut-shaped area
K	the number of DCHs per BTS
p	the ARS coverage in terms of the percentage of a cell
$\overline{H}$	the average number of hops of an ARS relaying path
N_M	the average number of active MHs covered by an ARS
N_A	the average number of reachable ARSs of a active MH
B_o	the request rejection rate in a cell without relaying
B_p	the request rejection rate in a cell with primary relaying
B_s	the request rejection rate in a cell with secondary relaying
RQ_o	the number of signaling messages in each connection request without relaying
RQ_p	the number of signaling messages in each connection request with primary relaying
RQ_s	the number of signaling messages in each connection request with secondary relaying
RL_o	the number of signaling messages for releasing a connection without relaying
RL_r	the number of signaling messages for releasing a connection via relaying
S_o	the average signaling overhead per successful request without relaying
S_p	the average signaling overhead per successful request with primary relaying
S_s	the average signaling overhead per successful request with secondary relaying

Table 6.2: The symbols used in the analysis.

messages processed by ARSs in primary relaying is $3N_A + p\overline{H}$. In addition, as each MH can reach N_A ARSs on average, H_BTS will receive N_A messages in step 6 in Figure 6.1, and the BTSs (including H_BTS and F_BTS) may send and receive five messages (see steps 7, 8a, 8b, 9a-b & 9c in Figure 6.1) if MH_0 is covered by ARSs. Thus the total number of messages that will be received and sent by BTSs in primary relaying is $p \cdot 5 + N_A$. Finally, PSC will send and receive 2 messages (see steps 7 & 8a-b in Figure 6.1) if MH_0 is covered by ARSs.

To determine the number of signaling messages per connection request in secondary relaying, we note that PSC will receive a secondary relaying request unless none of the active MHs is covered by ARSs. Here, we assume there are K active MHs, and hence, according to steps 5 & 6a-b in Figure 6.2, the number of signaling messages processed by PSC in the secondary relaying is $2 \cdot [1 - (1-p)^K]$. Note that, since some CBW of the BTS may be used by the MHs in other cells via primary relaying, the actually number of active MHs may be smaller than K. So, this is the upper bound on the signaling overhead of PSC. For BTSs, the major overhead comes from the secondary relaying requests forwarded by ARSs (see step 4 in Figure 6.2). Since all K active MHs receive the secondary relaying request, and each MH can reach N_A ARSs on average, $K \cdot N_A$ requests in total are received by the ARSs. However, the ARSs will only respond to the first request of those with the same sequence number. In other words, if an ARS cover multiple active MHs, only one copy of the request will be forwarded to the BTS, and this is the reason why the number of secondary relaying requests received by BTSs is $K \cdot N_A$ divided by N_M. In addition, the BTSs will process 3 messages in steps 1, 2, & $7c$ in Figure 6.2, and another 5 messages in steps 5, $6a$, $6b$, $7b$ & $7d$ if at least one active MH is covered by ARSs. Thus, the total number of signaling messages processed by BTSs in secondary relaying is $K \cdot N_A/N_M + 3 + 5[1 - (1-p)^K]$. For MHs, if it is not covered by any ARSs, it won't receive the S_RELAY_ACK from BTSs. Thus, although each of the K MHs receives S_RELAY_REQ from H_BTS and forwards it to nearby ARSs (see steps 2 & 3 in Figure 6.2), resulting $2K$ signaling messages, only a fraction (p) of them will receive acknowledges (see step 7a in Figure 6.2), resulting in $p \cdot K$ messages. Two additional messages are processed by MH_0 in steps 1 & 7c in Figure 6.2. The nearby ARSs of the active MHs will receive $K \cdot N_A$ S_RELAY_REQ messages (see step 3 in Figure 6.2) , and send and receive $2K \cdot \frac{N_A}{N_M}$ messages

	RQ_o	RQ_p	RQ_s	RL_o	RL_r
PSC	2	$2 \cdot p$	$2 \cdot [1-(1-p)^K]$	2	2
BTS	4	$p \cdot 5 + N_A$	$K \cdot N_A/N_M + 3 + 5[1-(1-p)^K]$	4	4
MH	2	$1+p$	$2 \cdot K + p \cdot K + 2$	2	2
ARS	0	$3N_A + p\overline{H}$	$K \cdot N_A + 2K\frac{N_A}{N_M} + \overline{H}[1-(1-p)^K]$	0	$\overline{H}$

Table 6.3: The analytical results of signaling overhead for Protocol 1.

in steps 4 & 7b. In addition, the ARSs on the relaying path may receive $\overline{H}$ messages if at least one MH is covered by ARSs. Finally, to release a connection via relaying, each ARS on the relaying path will receive a message from the BTSs. A summary of the signaling overhead of Protocol 1 is shown in Table 6.3.

6.4.2 Protocol 2

The signaling overhead of Protocol 2 is shown in Table 6.4. When using primary relaying, MH_0 will send a P_RELAY_REQ message to nearby ARSs (see step 5 in Figure 6.4), but only when it is covered by ARSs, it will receive an ARS_ACK, send a P_RELAY_ORD and receive a P_RELAY_ACK message (see step 6,7 & 12 in Figure 6.4). Hence, the total number of messages sent and received by MHs in primary relaying is $1 + 3p$. Since a MH can reach N_A ARSs on average, the ARSs will receive N_A P_RELAY_REQ messages in step 5 and send N_A ARS_ACK messages in step 6. However, only one ARS will be selected to establish a relaying path if MH_0 is covered by ARSs, which results in $4p \times \overline{H}$ signaling messages (see steps 7, 8, 11 & 12 in Figure 6.4). So, $2N_A + 4p\overline{H}$ messages will be processed by ARSs in primary relaying. Similarly, if MH_0 is covered by ARSs, one relaying path may be established, and $4p$ and $2p$ messages will be processed by F_BTS and PSC, respectively.

In secondary relaying, MH_0 sends and receives one messages to and from H_BTS (see step 1 & 14 in Figure 6.5), respectively. In addition, each active MHs in the cell receives and sends one message in steps 2 & 3, which results in total $2K$ signaling messages. Since the ARSs only respond to the first received request with the same sequence number, $3K \cdot \frac{N_A}{N_M}$ messages will be sent and

	RQ_o	RQ_p	RQ_s	RL_o	RL_r
PSC	2	$2 \cdot p$	$2 \cdot [1-(1-p)^K]$	2	2
BTS	4	$4 \cdot p$	$2K \cdot N_A/N_M + 3 + 5[1-(1-p)^K]$	4	4
MH	2	$1+3p$	$2 + 2 \cdot K + 3K \cdot N_A/N_M + 3[1-(1-p)^K]$	2	2
ARS	0	$2N_A + 4p\overline{H}$	$K \cdot N_A + K \cdot N_A/N_M + 4 \cdot \overline{H}[1-(1-p)^K]$	0	$4\overline{H}$

Table 6.4: The analytical results of signaling overhead for Protocol 2.

received by MHs in steps 4, 5 & 6. If at least one active MH is covered by ARSs, one of them is selected by H_BTS to set up a relaying path, and thus will process 3 additional messages (see step 7, 12 & 13 in Figure 6.5). Hence, the total number of messages sent and received by MHs in secondary relaying is $2+2\cdot K+3K\cdot N_A/N_M+3[1-(1-p)^K]$. The ARSs receive $K \cdot N_A$ messages in step 3, and respond $K \cdot \frac{N_A}{N_M}$ messages in step 4. If at least one active MH is covered by ARSs, a proxy is selected to establish a relaying path, on which the ARSs need to process $4\overline{H}$ messages (see steps 7,8,11 & 12 in Figure 6.5). So, the total number of signaling messages processed by ARSs in secondary relaying is $K \cdot N_A + K \cdot N_A/N_M + 4 \cdot \overline{H}[1-(1-p)^K]$. Similar to that in the Protocol 1, the BTSs receive and send $2K \cdot \frac{N_A}{N_M}$ messages in steps 5 & 6, and process 3 messages in steps 1, 2, 14 and $5[1-(1-p)^K]$ messages in steps 8, 9, 10, 11, 13. In addition, PSC receives and sends $2 \cdot [1-(1-p)^K]$ messages. Note that, although the analytical results of the signaling overhead of PSC in secondary relaying in Protocol 2 is the same as that in Protocol 1, the actual overhead of Protocol 2 may be higher, because we have ignored the effect of multiple attempts to establish a relaying path which may happen in Protocol 2 when the bandwidth information used in routing is out of date. Finally, to release a connection via relaying, the ARSs will process $4\overline{H}$ messages.

6.4.3 Protocol 3

The signaling overhead of Protocol 3 is sown in Table 6.5. Similarly to that in Protocol 2, the total number of messages sent and received by MHs in primary relaying is $1+3p$. The nearby ARSs of MH_0 will receive N_A P_RELAY_REQ messages and send N_A ARS_ACK messages (see steps 5 & 6 in Figure 6.6), and may establish up to N_A relaying paths, which may result in $4N_A\overline{H}$

messages for the intermediate ARSs (see steps 7 & 10 in Figure 6.6) and $3N_A$ messages for the proxy ARSs (see steps 7, 10 & 11 in Figure 6.6). Note that, only one of established relaying paths will be maintained, and all others will be released by $2\overline{H}(N_A - 1)$ P_RELAY_CL messages. So the total number of messages processed by ARSs in primary relaying is $6N_A\overline{H} + 5N_A - 2\overline{H}$. In addition, the BTSs will receive and send $4N_A$ messages in steps 7, 8, 9 & 10, and process $2(N_A - 1)$ messages for releasing the established connections which are not to be used. Note that, if $N_A - 1 < 0$, which implies that there are no multiple paths established and no need to release the unused connections, we set the value of $N_A - 1$ to be 0. Similarly, PSC will receive and send $2N_A + N_A - 1$ messages.

In secondary relaying, MH_0 sends and receives one messages in steps 1 & 14 in Figure 6.7, respectively, and the active MHs in the cell receive and send $2K$ messages in steps 2 & 3 in Figure 6.7. Since the ARSs only respond for the first received request with the same sequence number, $5K\frac{N_A}{N_M}$ messages will be sent and received by MHs in steps 4, 9, 10, 11 & 12. In addition, one MH selected by H_BTS may send a CBW release message to H_BTS (see step 13 in Figure 6.7), if there is at least one active MH covered by ARSs. Thus, the total number of messages processed by MHs in secondary relaying is $2 \cdot K + 5K \cdot N_A/N_M + 3$. The nearby ARSs of the active MHs receive $K \cdot N_A$ messages in step 3, respond $K \cdot \frac{N_A}{N_M}$ messages in step 4, and try to establish $K \cdot \frac{N_A}{N_M}$ relaying paths which results in $K \cdot \frac{N_A}{N_M} \times 4\overline{H}$ messages for intermediate ARSs and $2K\frac{N_A}{N_M}$ messages for the proxy ARSs (see steps 5 & 8 in Figure 6.7). Similarly to that in primary relaying, only one relaying path will be maintained, and the S_RELAY_CL messages will be sent to release the other relaying paths, which results in $2\overline{H}(K\frac{N_A}{N_M} - 1)$ messages. So, the total number of messages processed by ARSs in secondary relaying is $6K\overline{H}\frac{N_A}{N_M} - K\frac{N_A}{N_M} + KN_A - 2\overline{H}$. The H_BTS will receive one message from MH_0, broadcast one S_RELAY_REQ message, receive $K\frac{N_A}{N_M}$ acknowledges from the active MHs, receive one $CBW\_RELEASE$ message, and finally send one S_RELAY_ACK to MH_0. The F_BTS may process $4K\frac{N_A}{N_M}$ messages in steps 5, 6, 7 & 8 for the relaying request and $2(K\frac{N_A}{N_M} - 1)$ messages for releasing the unused connections. So, the total number of messages processed by BTSs is $7K\frac{N_A}{N_M} + 2$. In addition, the PSC may process $2K\frac{N_A}{N_M}$ messages for relaying path requests and $K\frac{N_A}{N_M} - 1$ messages for releasing the unused connections. Finally, similar to that

	RQ_o	RQ_p	RQ_s	RL_o	RL_r
PSC	2	$3N_A - 1$	$3K\frac{N_A}{N_M} - 1$	2	2
BTS	4	$6N_A - 2$	$7K\frac{N_A}{N_M} + 2$	4	4
MH	2	$1 + 3p$	$2 \cdot K + 5K \cdot N_A/N_M + 3$	2	2
ARS	0	$6N_A\overline{H} + 5N_A - 2\overline{H}$	$6K\overline{H}\frac{N_A}{N_M} - K\frac{N_A}{N_M} + KN_A - 2\overline{H}$	0	$4\overline{H}$

Table 6.5: The analytical results of signaling overhead for Protocol 3.

in Protocol 2, the ARSs will process $4\overline{H}$ messages to release a connection via relaying.

Replacing RQ_o, RQ_p, RQ_s, RL_o and RL_r in Equations 6.8, 6.9 and 6.10 with the values in the Tables 6.3, 6.4, and 6.5, we can obtain the estimated signaling overhead of the three proposed protocols. Simulation results are also obtained for comparison and verification purpose, and more discussions will be presented in Sec 6.5.

6.5 Simulation Results and Discussion

To evaluate the system performance in terms of request rejection probability and signaling overhead, we have developed a simulation model using GloMoSim [94] and PARSEC language [98]. In this section, we introduce the simulation environment and present the results as well as some discussion.

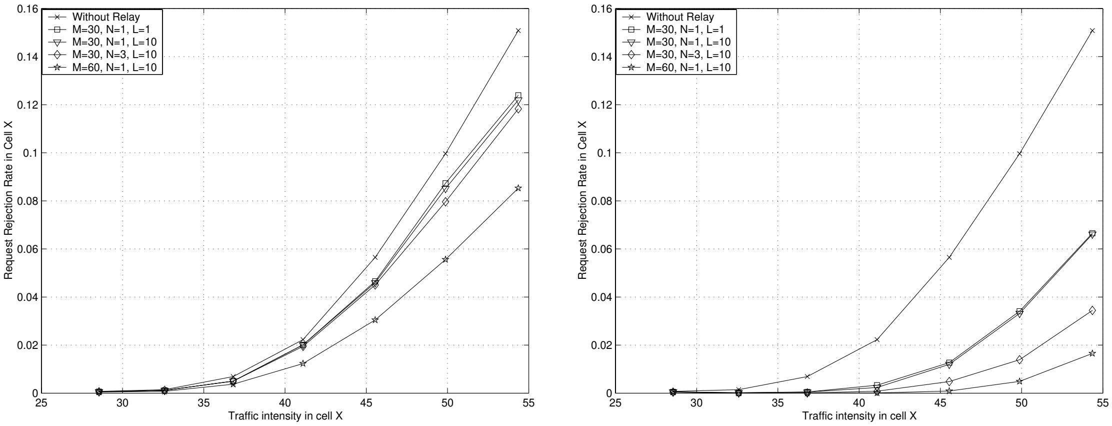
6.5.1 Simulation Model

As we have mentioned in Section 6.4, the simulated system includes a cell X and its neighboring cells (see Figure 6.8), which are controlled by a PSC. The cells are modelled as hexagons. We assume that the center-to-vertex distance of a cell is 2 *Km* and each cell has 50 units of bandwidth (CBW). For simplicity, we also assume that each connection requires 1 unit bandwidth. An ARS covers an area whose radius is 500 m, and we assume that ARSs are randomly placed in the donut-shaped region of cell X, which is bounded by two dashed circles as shown in Figure 6.8. Figure 6.8(a) shows the case where 30 ARSs are deployed in donut-shaped area between the two dashed circles whose radii are 2500 m and 1500 m respectively, while Figure 6.8(b) shows that there are 60 ARSs in the donut shaped area between the two dashed circles whose radii are 2500 m and 1000 m respectively.

We implement the signaling protocols discussed in Sec 6, with various parameters, such as the number of ARSs, the amount of RBW and CBW of each ARS, the mobility of MHs, and the traffic intensity in the system, to observe their effects on the system's performance. The traffic intensity is measured in Erlangs which is the product of the request arrival rate (in Poisson distribution) and the average holding time of a connection (in exponential distribution). In order to obtain converged statistical results, we simulate $1,600$ MHs which are randomly placed in the area within 2 *Km* from the BTS X, and run the simulation for 100 hours for each traffic intensity value before collecting the results.

6.5.2 Connection Request Rejection Rate

A request for establishing a connection will be rejected if at the time of the request, BTS X does not have any CBW left, and primary (or secondary) relaying fails. Figure 6.10 shows the request rejection rate of primary and secondary relaying in cell X with stationary MHs using Protocol 3. Although the PSC-assisted protocol can potentially select a better relaying path for a given request than the two distributed protocols, the three approaches result in almost the same performance in terms of the request rejection rate in this simulation, and hence, the results for the other two are not shown. As can be seen from the figure, the performance improvement of an iCAR system over a conventional cellular system (without relaying) in terms of the request rejection rate is due to the *relaying ability* of iCAR, which depends on the effective ARS coverage and the amount of relaying bandwidth. More specifically, with the increase in the number and coverage of ARSs (e.g. from 30 to 60), the rejection rates of both primary and secondary relaying are reduced significantly as we have expected. Increasing the RBW of ARSs (e.g., from 1 unit to 10 units) will also help reduce the request rejection rate, but not much because, in this simulation, the ARS clusters have a high connectivity, which results in many short, alternate relaying paths to a BTS, and thus making the RBW a non-critical resource. On the other hand, increasing the CBW of ARSs (e.g. from 1 unit to 3 units) may reduce the request rejection rate by about 50% in secondary relaying, although



(a) Primary Relaying

(b) Secondary Relaying

Figure 6.10: Request rejection rate in cell X. M is the number of ARSs; N is the amount of CBW of an gateway ARS; L is the amount of RBW of an ARS.

the reduction is insignificant in primary relaying. This is because primary relaying will fail if the requesting MH is not covered by any ARSs, and therefore its performance is mainly determined by the ARS coverage. However, in the secondary relaying, there is a high probability to find at least one active MHs covered by ARSs, and the probability to set up a successful relaying path is affected a lot by the CBW of ARSs.

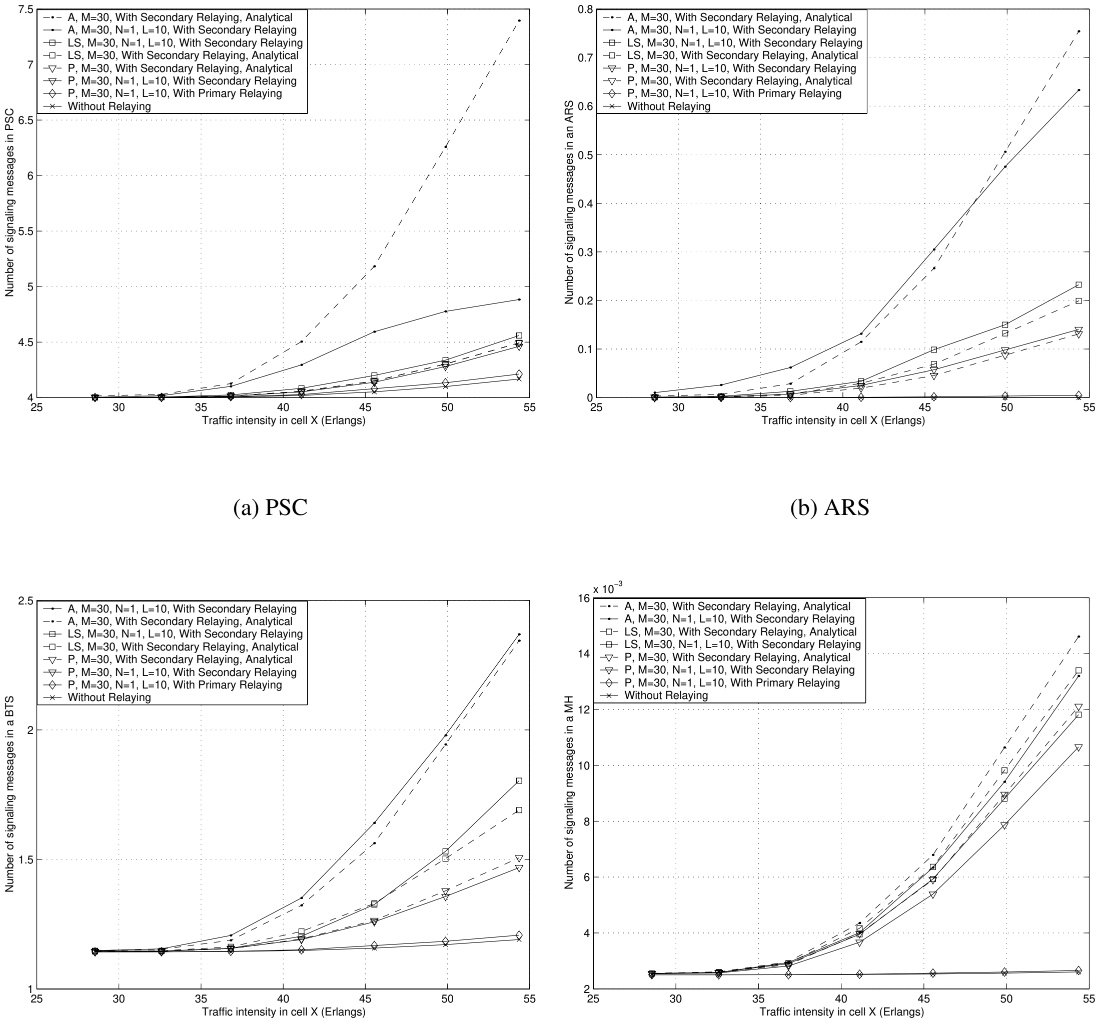
6.5.3 Signaling Overhead of Three Protocols

The signaling overheads incurred by PSC, ARSs, BTSs and MHs are shown in Figures 6.11(a) to 6.11(d) respectively, where 30 ARSs with 1 unit of CBW and 10 units of RBW are deployed in the system. The signaling overhead of different protocols due to primary relaying is almost identical, and hence only the results for the PSC-assisted protocol is shown in Figure 6.11. In addition to the simulation results, we also plot the analytical results in the figures for comparison. We have omitted the analytical results for the cases without relaying and with primary relaying since from the figures, because they are almost identical to the simulation results. As we can see, the analytical results are close to the simulation results except those for the PSC in Protocol 3, where we overestimate the signaling overhead a lot in the analysis because of the assumption that each proxy ARS may establish one relaying path. Note that, the major overhead of PSC in Protocol 3 comes from the secondary relaying requests sent by BTSs (see steps 6 & 7 in Figure 6.7) However, since some ARSs cannot relay traffic from one cell to another and the relaying bandwidth is limited in the simulation, not all of the proxy ARSs may establish the relaying paths and eventually send the CBW requests to PSC. Therefore, the actually number of messages processed by PSC may be much lower than what we have obtained in the analysis.

As we can see from these figures, the number of signaling messages increases with the traffic intensity since a higher traffic intensity results in more relaying requests, and therefore more signaling overhead. It is interesting to note that primary relaying increases the average overhead by a little but not much, because only the MHs associated with the requests initially blocked (i.e., without relaying) and their nearby ARSs will generate additional signaling messages due to the need for relaying. On the other hand, secondary relaying results in an exponential increase in the signaling overhead since many MHs and ARSs are involved in each secondary relaying request.

In particular, Protocol 3 results in a significantly higher signaling overhead than the other two protocols in secondary relaying, because in this protocol, all ARSs close to the active MHs will try to set up relaying path to a nearby BTS, and thus many paths are simultaneously being searched in many ARS clusters for each secondary relaying request. This creates a lot of hop by hop relaying requests and ACK's among the ARSs, BTSs and PSC, and results in a large number of signaling messages. In the link-state based protocol, on the other hand, although the ARSs will flood the bandwidth updating messages, the flooding is limited in only one ARS cluster. In addition, the relaying route is set up in one cluster only (i.e., Steps 8-12 in Figure 6.5 will only bother the ARSs in the cluster where the selected proxy ARS is located). Thus, Protocol 2 has a lower signaling overhead than the simple route-searching protocol. The PSC-assisted protocol has the lowest signaling overhead because it needs neither to flood updating messages nor to try multiple relaying paths[5]. But, as mentioned earlier, the drawback of the PSC-assisted protocol is that PSC becomes the single point of failure and performance bottleneck (in that one may experience a long delay due to the limited processing capability of PSC).

[5] In all three protocols, MHs and ARSs need to discover their neighbors via certain link management protocols.



(a) PSC

(b) ARS

(c) BTS

(d) MH

Figure 6.11: The average number of signaling messages per satisfied connection request. A, LS and P stand for the simple routing-searching protocol, the link-state based protocol and the PSC-assisted protocol, respectively.

6.5.4 Other Factors Affecting the Signaling Overhead

In addition to the traffic intensity and different signaling protocols, there are some other factors that affects the signaling overhead, such as the number or coverage of ARSs, the amount of ARS bandwidth (RBW/CBW), the MH moving speed, and so on.

Increasing the amount of CBW (or RBW) of ARSs has two opposite effects on the signaling overhead. It may increase the likelihood of being able to cope with the initially rejected request via primary relaying instead of secondary relaying, and find a shorter relaying path, which reduces the signaling overhead. On the other hand, it may also increase the probability to ”bother” the PSC, BTSs and ARSs via relaying (instead of resulting in a rejected request), and therefore increases the signaling overhead. Figures 6.12(a) shows an example of the signaling overhead in a BTS when CBW of ARSs increases from 1 unit to 3 units. Although the results shown in this subsection are for BTSs only, the results for PSC, ARSs and MHs have similar trend, and hence are omitted. As can be seen, in all of the three protocols, the BTS needs to process more signaling messages in the system with more CBW at ARSs. In addition, the overhead in Protocol 3 is affected by CBW a lot because more CBW will result in more probe messages in the simple route searching. On the other hand, the overhead in Protocol 1 increases only a little because routing of Protocol 1 is performed by PSC only.

Increasing the number of ARSs will increase the ARS coverage as well as the density (or connectivity) of the ad hoc network. As shown in Figure 6.12(b), the number of signaling messages increases with the number of ARSs for the PSC-assisted and the simple route-searching protocols, because the BTSs will receive more responses when performing secondary relaying in a system with more ARSs (see step 4 in Figure 6.2 and step 8 in Figure 6.7). But for the link-state based protocol, since the BTS won’t receive responses directly from ARSs, and the proxy ARS will choose the best relaying route and thus it is not necessary to try multiple paths, more ARSs won’t result in a significantly higher overhead as it does in the simple route-searching protocol. However, as more requests may be accommodated via primary relaying, the signaling overhead actually decreases with the increase in the number of ARSs.

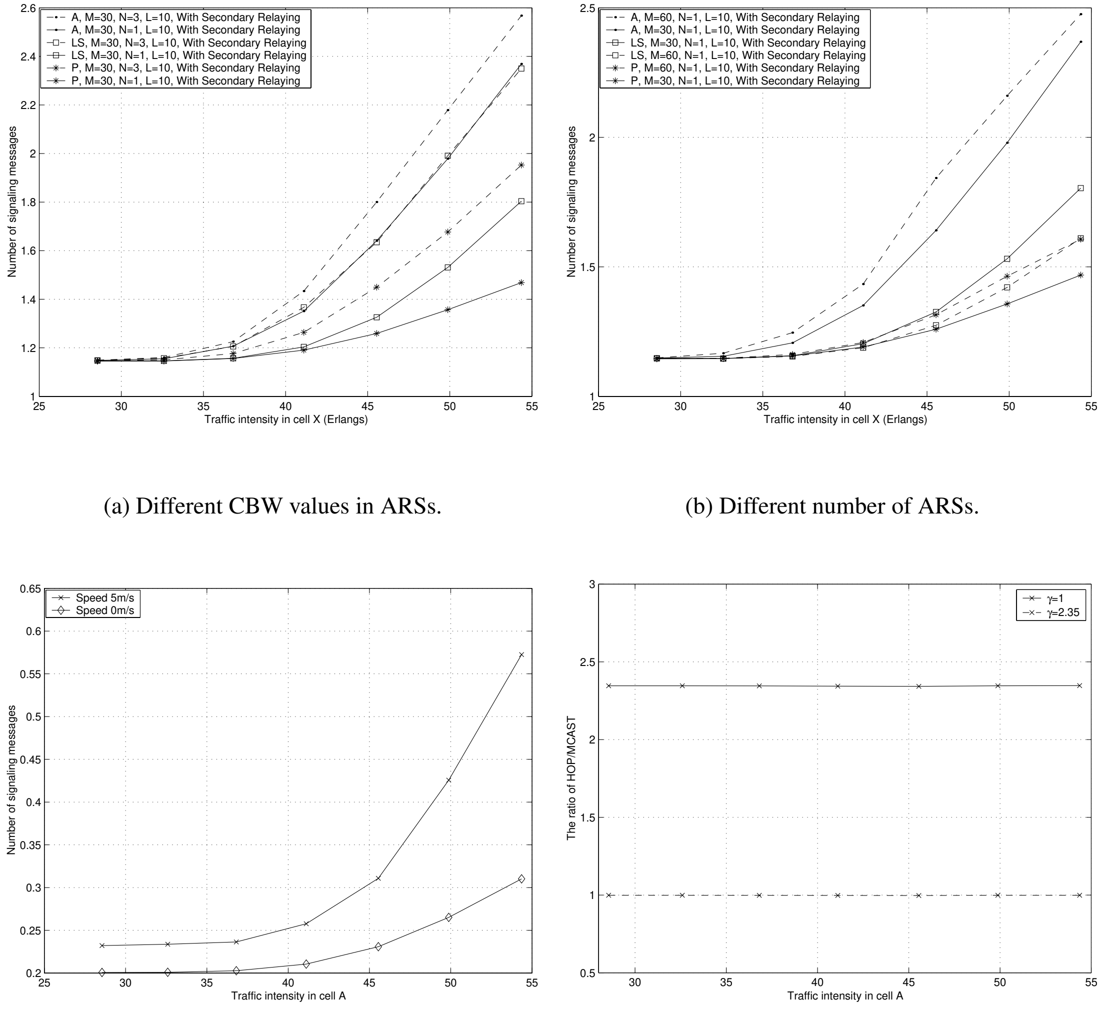
Figure 6.12(c) shows how MHs’ mobilities affect the signaling overhead in Protocol 2 (the same trend is exhibited for the other two protocols). With a higher moving speed of MHs, the number of signaling messages increases due to handoffs not only from one cell to another cell but also between ARSs and BTSs. In addition, since a higher moving speed will result in a higher connection blocking/dropping rate, the averaged overhead per satisfied connections will also increase.

As we discussed earlier in Section 6, there are two ways to do bandwidth reservation after a relaying route is selected. One is through multicasting by BTSs, and the other is via hop-by-hop relaying by ARSs. So far, since multicasting a message on a cellular channel prevents this channel from being used for any other purpose in the same cell at the same time while the same relaying channel may be used at the same time for different signaling messages on two different hops, we assign a weight ($\gamma \geq 1$) to each multicasted signaling message while assigning a weight of 1 to each hop-by-hop signaling messages. Figure 6.12(d) shows the ratio of the signaling overhead introduced by bandwidth reservation via multicasting and hop-by-hop relaying in a BTS when PSC-assisted protocol used (the same trend is exhibited for the other two protocols). As we can see, when $\gamma \approx 2.35$, the two approaches have a similar overhead. If $\gamma < 2.35$, multicasting has a better performance. Otherwise, unicasting shows its advantage.

Further Discussion

In Section 6, we have discussed the signaling protocols for new calls. When an active MH moves out of the coverage of a BTS or an ARS, it needs to perform handoff by sending a CBW request (including the addresses of the reachable BTSs) to PSC (via a BTS and possibly ARSs). This request will be processed by PSC similarly to a new connection request. If there is no CBW available in the reachable BTSs, either primary or secondary relaying may be performed. As a result of this request, a connection may be handed off from a BTS to another BTS, or from an ARS to another ARS, or between an ARS and a BTS. A connection will be dropped if all of the above attempts fail.

In addition, the scenario discussed in Section 6 assumes that a MH is the sender. If the MH is a receiver, and the cell in which it is located is congested, the BTS (e.g. BTS_Y) in the congested cell will request the MH to find a relaying path to another BTS which has free CBW as we discussed earlier. If either primary or secondary relay is successful, BTS_Y will inform PSC to re-route the



(a) Different CBW values in ARSs.

(b) Different number of ARSs.

(c) Different MH mobilities. Link-state based protocol is used.

(d) Multicasting v.s. Hop-by-hop reservation. PSC-assisted protocol is used.

Figure 6.12: Other factors affecting the signaling overhead (in a BTS). HOP and MCAST stand for by hop-by-hop and multicasting reservation respectively.

incoming data (either for this MH if primary relay is applied, or for another MH if secondary relay is applied) to the new BTS.

Although separate results for connection setup and release are not shown here, many of the extra signaling messages in iCAR are introduced by the search for relaying routes, especially in secondary relaying. Accordingly, the connection release requests result in only a small portion of total signaling overhead. In addition, note that, most of the signaling messages are short and therefore may be squeezed into existing signaling or data packets without significantly increasing additional bandwidth requirement for signaling.

In addition to the request rejection rate and the signaling overhead, there are other important performance metrics that have not been discussed in this paper, e.g. the connection setup latency, power consumption, implementation cost, etc. However, these metrics depend not only on the routing protocols but also very much on the MAC protocols (which are the subject of our future work). The selection of the protocol stack for iCAR should not based on one layer protocol or one single performance metric. For example, the PSC-assisted protocol shows the lowest signaling overhead in our simulation, but it is not necessary the best choice as PSC becomes the single point of failure. On the other hand, although Protocol 3 has the highest signaling overhead, it may reduce the cost of each ARS because of its simplicity, which in turn allow the operators to deploy more ARSs for a given budget.

Chapter 7

ARS Mobility Management

In this chapter, we address the ARS mobility management in iCAR. Intuitively, in iCAR, having more ARSs increases the relaying coverage which in turn means that more calls can be relayed from a congested cell to a non-congested cell, leading to a better grade of service (GoS) (i.e., lower call blocking probability). But on the other hand, more ARSs result in a higher system cost. Clearly, for a given number of ARSs, the effective ARS coverage can be increased by allowing ARSs to move so as to adapt to the dynamically changing locations of the MHs.

Note that, an ARS differs from an MH in that the former is deployed, used, and controlled by the system only, not by the end users. Accordingly, we will refer to the ARS mobility as managed mobility, to distinguish it from the MHs mobility which has been extensively studied in the context of MANET. To our knowledge, this is the first work that deals with such managed mobility.

The managed ARSs' mobility can be classified into two categories: macro-mobility and micro-mobility. With managed macro-mobility, an off-duty ARS (i.e., one that is not relaying any calls) can move a long distance (e.g., through several cells) to a location deemed more desirable by certain ARS placement strategies similar to those in [11, 15]. On the other hand, with managed micro-mobility, an active ARS (i.e., one that is relaying one or more on-going calls) can move only within a short range so as not to drop any on-going connections while still being able to relay a new or handoff call which otherwise would be blocked or dropped. In this paper, we will focus on the managed micro-mobility of ARSs.

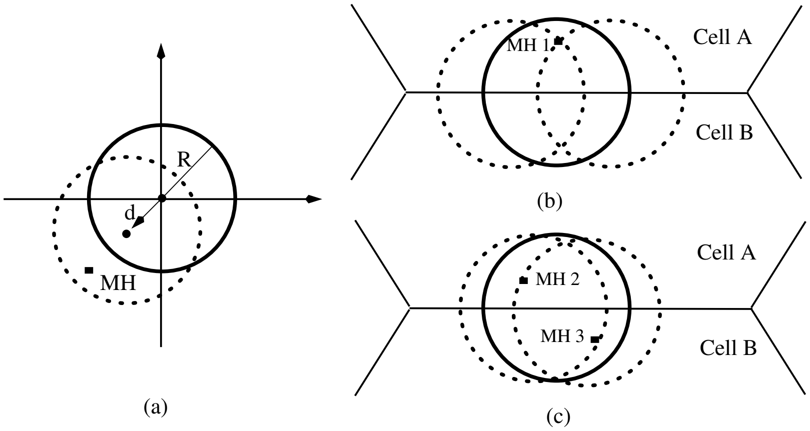


Figure 7.1: (a) Covering an MH with ARS mobility; (b) A seed ARSs mobility limited by one on-going relayed connection to/from MH 1; (c) A seed ARSs mobility limited by two on-going relayed connections to/from MH 2 and MH 3, respectively.

Introducing ARS mobility makes the iCAR system more like an ad hoc network. We anticipate that the proposed ARS micro-mobility model and the performance evaluation technique as well as performance results presented in this paper would also provide a new research direction for studying other ad hoc networks such as self-reconfigurable sensor networks, where an idle sensor node with a limited mobility may relocate to a more desirable location to aid communications among the neighboring sensor nodes [107, 108].

In the rest of this chapter, we first introduce the motivation and assumptions, and then describe the approaches to managing the ARS movement in iCAR.

7.1 Motivation and Assumptions

The motivation of allowing the ARSs to move is to increase their *effective coverage*, given their limited transmission range using the R-interface. In an iCAR system with stationary ARSs, the effective coverage of an ARS is limited to πR^2, where R is the ARSs transmission range (see the solid circle in Figure 7.1 (a)). So, if an MH is outside the circle, its call cannot be relayed by the ARS. However, if the ARS has a certain mobility, it may move close enough to the MH to provide relaying service (see Figure 7.1 (a)). Note that certain practical constraints (to be discussed later in this section) may limit the ARSs movement and consequently the increase in its effective coverage. Also, we cannot require an MH to move toward an ARS since MHs mobility cannot be controlled by any system. On the other hand, most (if not all) of the ARSs can move under the control of the iCAR system, and such ARSs will be called *Mobile ARS* or *MARS*. Hereafter, we will focus on MARS.

In the following discussion, we assume that the MARSs are initially placed at certain positions (according to some placement strategies). More specifically, these MARSs are grouped into clusters, and in each cluster, there is a *seed* MARS placed at the shared border of two cells , and additional MARSs may *grow* from the seed. Without loss of generality, we label the MARSs in a cluster located in one cell with a sequence of consecutive and increasing integers starting with the seed MARS.

For micro-mobility, an important practical constraint is that the movement of a MARS should not break any existing connections. For example, if the MARS is a seed, then it may still have to be a seed after moving (implying that it may only move along the shared border of two cells) as shown in Figure 7.1 (b) and (c). Otherwise, after the seed moves within cell A, the entire cluster will not be able to relay any traffic from cell A to cell B. If the MARS is not a seed, it can move in any directions as long as it is still connected to its upstream node (i.e., the neighbor MARS closer to the seed) after it moves. Of course, this may require its downstream MARSs to move accordingly. In other words, we want each MARS to remain in its cluster and to be relaying capable.

We assume that all of the mobile nodes in an iCAR system, including the MHs and the MARSs, are equipped with a Global Position System (GPS). The MARSs will periodically report their status including the location information to an ARS Mobility Controller (AMC) which can be co-located with the Base Station Controller (BSC). However, each MARS maintains the current positions of the MHs to which it is providing the relaying service as a proxy. In other words, it does not send such information to AMC so as not to create a bottleneck at AMC. Alternately, an AMC can maintain all the information about the MARSs and MHs that are receiving relaying service. But such a centralized control approach may not be scalable.

7.2 MARS Micro-Mobility Management

In this section, we discuss the micro-mobility management strategies for accommodating a relaying request, which is generated by (or on behalf of) an MH X in a congested cell after MH X fails to acquire a DCH in the cell for a new call. Such a relaying request may be satisfied by either primary relaying or secondary relaying without requiring any MARS movement. However, if both of them fail, AMC will try *primary movement* (in analogy to primary relaying) first, and then *secondary movement* if necessary, as to be described below.

7.2.1 Strategies For Primary Movement

The objective of the primary movement is to move a relaying capable MARS close enough to the MH requesting for the relaying service so as to provide primary relaying. We will first present the basic strategy for managing the primary movement, and then discuss possible extensions to improve the performance.

Basic Strategy

Using the basic strategy for primary movement, after receiving the $Move\_Req$ from the MH X (see step 1 in Figure 7.2) which includes the MHs location information, AMC will find the closest MARS (e.g., $MARS_i$) to MH X based on the locations of the MH and the nearby MARSs. AMC determines the destination to which $MARS_i$ will try to move (in order for it to become a proxy) by drawing a circle with the position of the MH to be the center and R to be the radius (Without loss of generality, here we assume that MHs transmission range using the R-interface is the same as that of a MARS). We will refer to the circle (shown as a dashed circle in Figure 7.3 (a) and (b)) as the destination circle, or D-circle. If $MARS_i$ is a seed along the shared border of two cells denoted by line AB (see Figure 7.3 (a)) and the D-circle intersects line AB at two points, the intersection point closer to $MARS_i$, denoted by H, is chosen as the destination (see Figure 7.3 (a)). In such a case, the destination is found. If $MARS_i$ is not a seed (see Figure 7.3 (b)), it can move within the circle centered at $MARS_{i-1}$ (with a radius of R), to be referred as the S-circle (so that it can still be connected to the seed after moving). Accordingly, AMC finds the intersection points of the

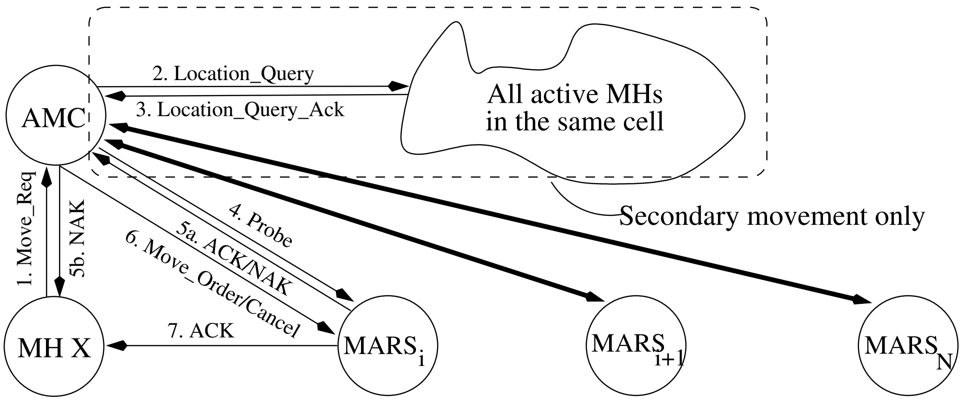


Figure 7.2: A signaling protocol for managed Micro-mobility of MARSs.

D-circle and the S-circle, and choose the intersection point closer to $MARS_i$ as the destination (see point H in Figure 7.3 (b)). In either case above, if there is no intersection points (or tangent point) between the D-circle and the line AB, or between the D-circle and the S-circle, no further actions will be taken, except that a NAK message will be sent to MH X (see step 5b in Figure 7.2) in the basic mobility management approach. (Nevertheless, in such as a case, the extended approach to be discussed later in this subsection may be employed).

After the moving destination is determined, AMC will compute the moving distance of $MARS_i$ to be $d_i = |O \leftrightarrow H|$ where O is the initial position of $MARS_i$ (see Figure 7.3 (a) and (b)), and the MARSs moving time T_i^m (e.g., based on d_i and the MARSs moving speed). If T_i^m is larger than the maximum delay budget t allowed for MARS movement, a NAK message will be again sent to MH X. Otherwise, AMC will compute the destination of the next hop $MARS_{i+1}$ by drawing a line connecting the new position of $MARS_i$ (point H) and the current position of $MARS_{i+1}$ (see Figure 7.3 (c)), and choose the intersection point of this line and the circle centered at H with a radius of R (e.g., point H' in Figure 7.3 (c)) to be the moving destination for $MARS_{i+1}$. Note that, the moving distance (thus the moving time) of $MARS_{i+1}$ will not be longer than that of $MARS_i$. More specifically, we have the following proposition.

Proposition 3 *If the moving distance of $MARS_i$ (which is not the last MARS in its cluster) is d_i, then the moving distance of $MARS_{i+1}$ (d_{i+1}) is not longer than d_i.*

Proof : Assume that the current location and the moving destination of $MARS_i$ are O and H respectively, where $d_i = |O \leftrightarrow H|$, and the current location and the moving destination of

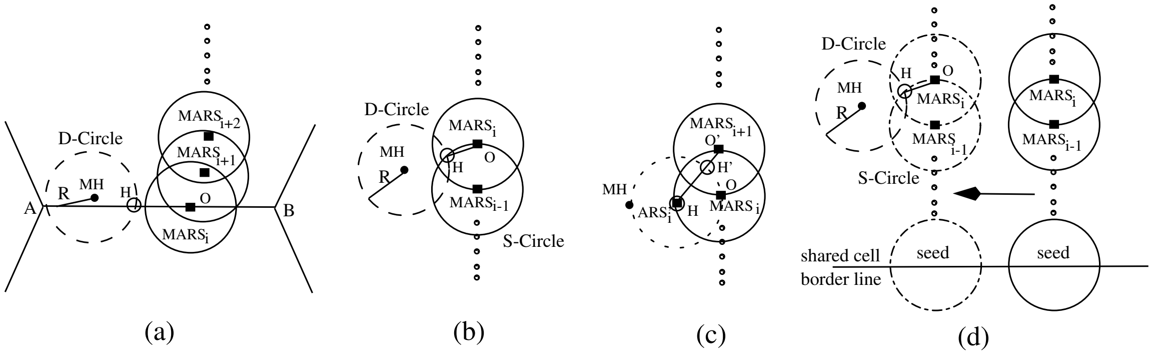


Figure 7.3: MARS Micro-mobility examples. (a) A seed $MARS_i$ is selected to move. (b) A grown $MARS_i$ is selected to move. (c) The movement of a downstream node ($MARS_{i+1}$). (d) The extended approach (cluster shifting).

$MARS_{i+1}$ are O' and H' respectively, where $d_{i+1} = |O' \leftrightarrow H'|$, (see Figure 7.3 (c)), then since $|O' \leftrightarrow O| \leq |H' \leftrightarrow H| = R$,

$$\begin{aligned} d_{i+1} &= |O' \leftrightarrow H'| = |O' \leftrightarrow H| - |H' \leftrightarrow H| \\ &= |O' \leftrightarrow H| - R \\ &< |O' \leftrightarrow O| + |O \leftrightarrow H| - R \\ &\leq d_i \end{aligned}$$

∎

Similarly, AMC will compute the destinations of other downstream nodes of $MARS_i$ (i.e., from $MARS_{i+2}$ to the last hop of this cluster $MARS_N$), and multicast a $Probe$ message containing the destination information to each of these MARSs (see step 4 in Figure 7.2).

After receiving the *Probe* message, each MARS will check if any on-going connections would be broken based on its destination and the current locations of MHs to which it provides relaying service (i.e., serves as a *proxy*). In case of potential drop of existing connection due to its movement, an NAK message will be sent to AMC. Otherwise, the MARS will send an ACK message to AMC (see step 5a in Figure 7.2). If AMC receives at least one NAK, it will send a $Move\_Cancel$ messages to the MARSs and no further actions will be taken. When AMC receives ACK messages from all these MARSs, it will send a $Move\_Order$ message to them (see step 6 in Figure 7.2).

After receiving the $Move\_Order$ message, the MARSs (including $MARS_i$ to $MARS_N$) can start moving. Upon arriving at its destination, $MARS_i$ will send an ACK message to MH X (see step 7 in Figure 7.2) and accordingly, MH X will perform a primary relaying.

In an alternative approach, instead of multicasting the $Probe$ message to all related MARSs, the AMC can send it to $MARS_i$ only, which will check the on-going relayed connections and forward the *Probe* message to the next hop ($MARS_{i+1}$) if none of the on-going connections would be dropped due to the movement, or send a $NACK$ to AMC otherwise. Similarly, other MARSs will forward the $Probe$ message and check their existing connections hop by hop. This approach may reduce the signaling load at AMC, but it will result in a longer delay.

Extended Approach

In the basic strategy discussed above, the primary movement attempt will fail if there are no intersection points between the D-circle and the line AB when $MARS_i$ is a seed (case 1), or between the D-circle and the S-circle centered at $MARS_{i-1}$ when $MARS_i$ is not a seed (case 2). In these cases, we can use the following extended approach. First, in case 1, if the circle centered at $MARS_i$ intersects (or is tangent) with the D-circle, then AMC will treat $MARS_{i+1}$ as "$MARS_i$" (i.e., will try to move $MARS_{i+1}$ and make it as a proxy) and proceed using the basic approach. Otherwise, it means no other MARSs' circle will intersect (be tangent) with the D-circle, and accordingly it is impossible to move any of the downstream nodes of $MARS_i$ to cover the MH. More clearly, we establish and prove the following proposition.

Proposition 4 *Assuming $MARS_i$ is the closest MARS to MH X, if the circle centered at $MARS_i$ with a radius of R does not intersect the D-circle of MH X, then none of the downstream nodes of $MARS_i$ can move to cover MH X.*

Proof : Denote the distance between a $MARS_j$ and MH X as d_j, $d_j \geq d_i$. Since the circle centered at $MARS_i$ with a radius of R does not intersect the D-circle, $d_j \geq d_i > 2R$. Thus, it is impossible to find a downstream nodes of $MARS_i$, whose S-circle (as shown in Figure 7.3 (b)) intersects the D-circle of MH X. Accordingly, none of the downstream nodes of $MARS_i$ can move to cover the MH without losing connection with other MARSs in the same cluster.

∎

Similarly, in case 2 above, AMC checks if the circle around $MARS_i$ intersects (or is tangent) with the D-circle. If so, AMC will try to move $MARS_{i+1}$ and make it as a proxy. Otherwise, there is no way to move the downstream nodes of $MARS_i$ to accommodate the MH. Of course, the MARSs still need to check if any of the on-going connections will be dropped or not.

If the above attempt (to move the downstream nodes of $MARS_i$ only) is failed, AMC will compute the moving direction, distance and time when the entire cluster of MARSs move as a single entity (without changing their relative positions) toward MH X while keeping the seed of this cluster on the cell border (i.e., still being a seed MARS), until the D-circle intersects the S-circle of at least one MARSs (see Figure 7.3 (d) for the cluster shifting approach). However, since the entire cluster of $MARS_i$ need to move, there is high probability that the movement may affect the on-going connections, and thus has to abort.

7.2.2 Strategies For Secondary Movement and Existing Relayed Connections

If primary movement is impossible, AMC will perform the secondary movement, whose objective is to move MARSs to facilitate secondary relaying. This can be accomplished by broadcasting a $Location\_Query$ message (see step 2 in Figure 7.2) to all active MHs in the cell where MH X is located. Upon receiving the $Location\_Query$ message, the MHs respond with a $Location\_Query\_Ack$ including their current GPS information to AMC (see step 3 in Figure 7.2). After AMC receives the locations of the MHs, it will find at least one pair of MH (which is an active MH using a cellular channel but not the MH requesting a relaying service) and MARS (e.g., $MARS_i$) with the shortest distance. Similar to the primary movement, AMC will compute the destinations of the related MARSs, and the MARSs will check to see if they can move or not based on the existing relayed connections. However, instead of primary relaying, a secondary relaying will be performed after a successful MARSs' movement.

We now briefly discuss the MARS mobility for keeping alive an existing relayed connection. More specifically, when an MH whose connection is being relayed is about to move out of the coverage of its proxy MARS, it can first try to handover to the BTS in the cell where the MH is located (note that although the BTS did not have a DCH then, it may have one now). However, the

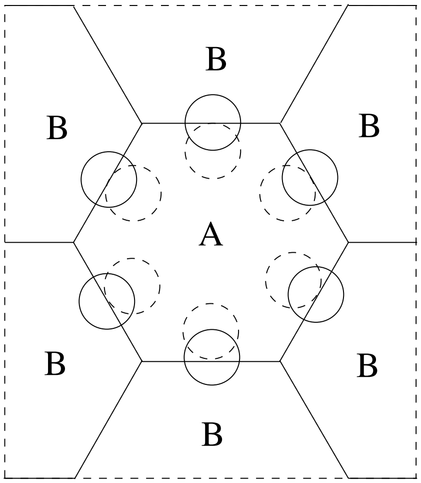


Figure 7.4: Simulation Environment. Seven cells with cell A to be the hot spot. The solid circles denote the seed MARSs. The dashed circles denote the grown MARSs.

handover may fail if the BTS (still) has no DCH available. In this case if the secondary relaying is also not possible, either the proxy MARS itself or another MARS may make primary movement to serve the MH. The proxy MARS or another MARS can also make secondary movement (i.e., to serve as a proxy for an active MH that is using its C-interface to communicate with the BTS in the same cell). Since the MARS mobility may be managed similarly as described earlier in this section by treating the handover request as a new relaying request, the detail is omitted.

7.3 Numerical Results and Discussion

In this section, we introduce the simulation model, and present the numerical results.

7.3.1 Simulation Model

To evaluate the performance improvement in an iCAR system due to the MARS mobility in terms of the request blocking probability, we have developed a simulation model using the PARSEC language [98] and the GloMoSim simulator [94]. The simulated system includes a cell A and six partial neighboring cells (see Figure 7.4), each modelled as a hexagon with the center-to-vertex distance of 2 *Km*. The traffic intensity is measured in Erlangs which is the product of the request arrival rate

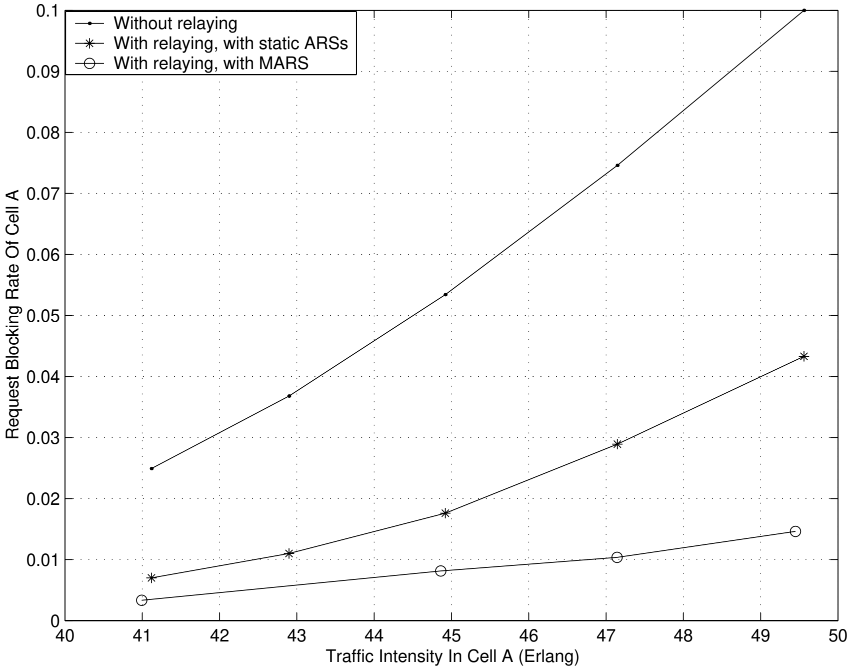


Figure 7.5: ARS mobility induced performance improvement. R=250m.

(Poisson distributed) and the holding time (which can be general-distributed). We assume that cell A is the hot spot with a varying traffic intensity from 40 to 55 Erlangs, while each of its neighboring cells has a fixed traffic intensity about $37.9\ Erlangs$. We also assume that 50 cellular channels are allocated to one cell's BTS, and for simplicity, each connection requires 1 channel. New calls arrive according to Poisson process. In order to obtain converging statistical results, we have simulated $6,400$ MHs whose locations are uniformly distributed in the system[1], and run each simulation for up to 100 hours for each set of parameter values. By default, we have assumed that each MARS has a moving speed of $15m/s$, a maximum moving distance of $2R$ and 10 relaying channels. With stationary ARSs, both primary and secondary relaying are used to relay the traffic. With MARS, both primary and secondary movement are used to provide relaying service for the MH.

7.3.2 Results

In this subsection, we present the simulation and analytical results.

[1] Since we do not examine the handoff performance in this study, there is no need to simulate MHs' mobility.

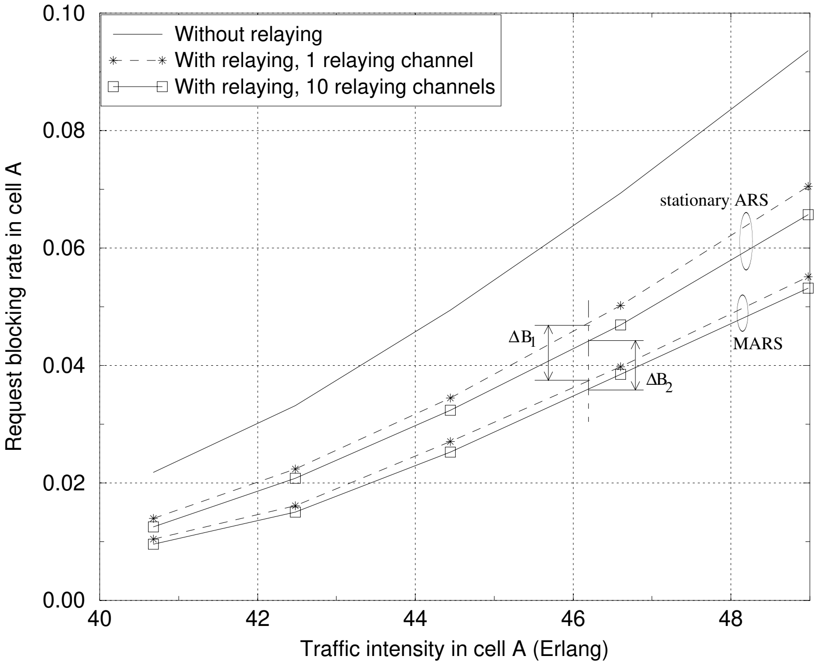


Figure 7.6: Performance with 18 relaying stations. R=120m.

Reduced request blocking rate

The request blocking probability defined to be the fraction of the total requests that are blocked is one of the most important performance criteria in the mobile wireless networks. We first consider the scenario where only 6 seed MARSs, each with $250m$ transmission range, are placed at the borders of cell A. With additional grown MARSs, for example, when there are 18 MARSs equally grouped into 6 clusters (of 3 MARSs each, with $250m$ transmission range), the blocking probability of cell A is reduced further due to the increased MARS coverage.

We have also simulated the scenario with smaller R values. Figure 7.6 shows the simulation results of the iCAR system which has 18 (stationary and mobile) relaying stations, each with $R = 120m$. As we can see, the iCAR system with MARSs always has a lower request blocking rate than that of the iCAR system with stationary ARSs.

Note that, compared to the case with stationary ARSs, the lower the relaying bandwidth, the higher the improvement due to the MARSs mobility (i.e., $\Delta B_1 > \Delta B_2$ in Figure 7.6). This is because with the increase in the relaying bandwidth, a MARS may serve as the proxy for more MHs, and as a result, it can only move within a smaller area in order to maintain the on-going

R	Stationary ARS	MARS
120m	1.75	3.25
250m	3.00	5.25
500m	14.75	21.75

Table 7.1: The improved capacity of cell A (Erlangs).

connections after moving. For example, the MARS which is the proxy of only one MH (e.g., MH 1 in Figure 7.1 (b)) can move a longer distance than the MARS which is the proxy of two MHs (e.g., MH 2 and MH 3 in Figure 7.1 (c)), and accordingly has a higher probability to move to the destination successfully.

Increased capacity

The capacity of a cell is defined to be the maximum traffic load that it can handle while keeping a certain GoS. Here we set the desirable GoS to be 0.02 (blocking probability) and consider a system with 6 seed relaying stations. In a conventional cellular system (without relaying), the capacity of a cell with 50 cellular channels is 40.25 Erlangs according to the Erlang-B formula. Table 7.1 shows the increased capacity of cell A in the iCAR systems with stationary ARSs and MARSs, respectively. As we can see, the bigger R results in a higher capacity improvement. The iCAR system with MARSs always has a significant higher capacity gain than that of the iCAR system with stationary ARSs. Specifically, when $R = 500m$, the iCAR system with MARSs can increase the capacity of cell A by 21.75 $Erlangs$, which means that, assuming the average connection holding time to be 120 seconds, cell A can accommodate over 600 more users per hour than the cell in a conventional cellular system can.

Decreased number of MARSs needed

The MARS mobility which increases the effective ARS coverage can result in a fewer number of MARSs than the iCAR system with stationary ARSs, and accordingly reducing the system's equipment cost while maintaining the same GoS.

Figure 7.7 shows the number of MARSs (with $R = 120m$) needed to achieve the required GoS

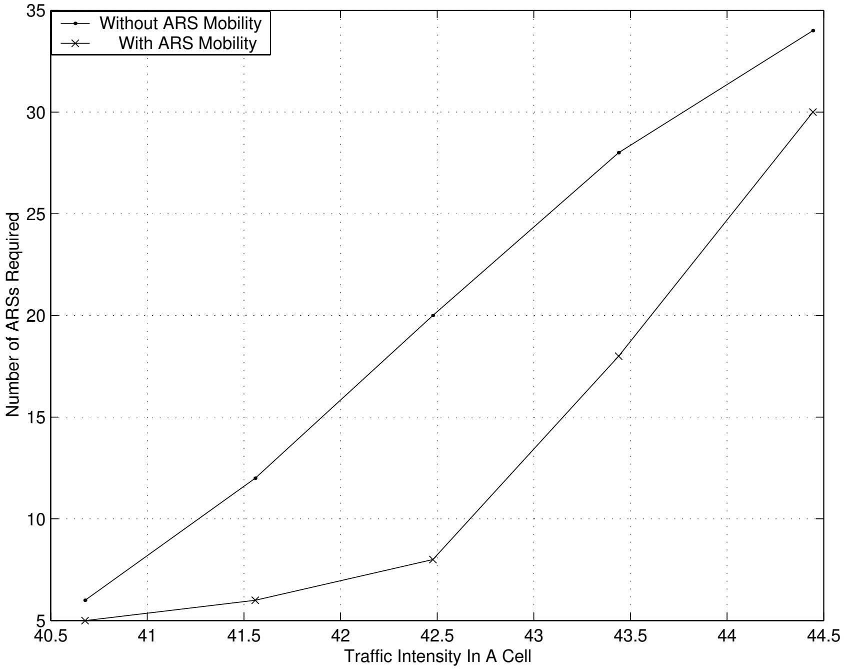


Figure 7.7: The number of MARSs needed in meeting a specified GoS (2%).

(2%). As we can see, the number of MARSs needed increases with the traffic intensity, but the system with MARS always needs fewer number of relaying stations than the system with stationary ARSs. For example, when the traffic intensity is about 42.5 $Erlangs$, only 8 MARSs are needed by the former, while the latter needs as many as 20 MARSs.

When the traffic intensity is very high, the number of MARSs needed in the two systems (i.e., the systems with and without ARS mobility) to achieve the GoS of 0.02 becomes close. This is because an MARSs mobility is limited by a significant number of relayed connections it serves under high traffic intensity.

Effect of t

The moving delay budget t is an important parameter in MARS mobility. A longer t allows the MARSs to move a longer distance, thus the MARS has a higher probability to finish a successful movement. Figure 7.8 shows the effect of t on the request blocking probability in an iCAR system with 6 seed MARSs with $R = 120m$. As we can see, longer t results in lower request blocking rate. But note that, t cannot be arbitrary large, it should be limited by the delay requirement of the

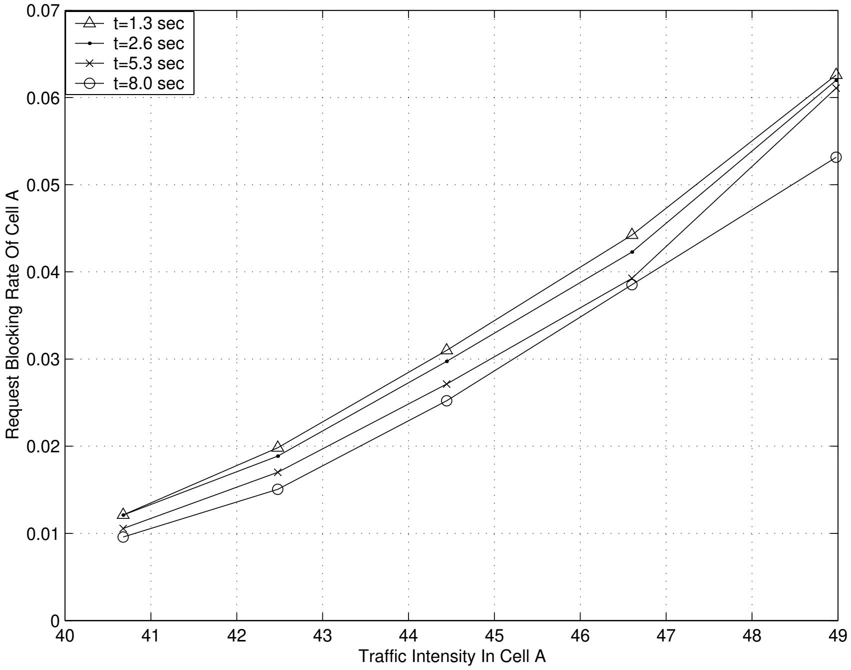


Figure 7.8: Effect of t. $R = 120m$.

requests.

Chapter 8

Summary

The objective of this work is to address the congestion problem due to the limited bandwidth in a cellular system, balance traffic among cells, increase system's capacity cost-effectively, and provide interoperability for heterogeneous networks. The major contributions of this dissertation are as follows.

1. We have proposed a new wireless system architecture based on the integration of cellular and modern ad hoc relaying technologies, called *iCAR*. It can efficiently balance traffic loads and share channel resource between cells by using *Ad hoc relaying stations* (ARS) to relay traffic from one cell to another dynamically. This not only increases the system's capacity cost-effectively, but also reduces transmission power for mobile hosts and extends system coverage.

2. We have analyzed the system performance in terms of the call blocking/dropping probability and queuing delay, and verified the analytical results via simulations. Our results have shown that with a limited number of ARSs and some increase in the signaling overhead (as well as hardware complexity), the call blocking/dropping probability in a congested cell as well as the overall system can be reduced.

3. we have discussed the number of placement of ARSs. In particular, we have proposed a seed-growing approach for ARS placement, and analyzed the upper bound on the number of seed ARSs needed in the system. We have also introduced a new performance met-

ric called quality of (ARS) coverage (QoC) for the comparison of various ARS placement strategies, and proposed three rules of thumb as guidelines for cost-effective ARS placement in iCAR.

4. We have also proposed the signaling and routing protocols for establishing QoS guaranteed connections for IP traffic in iCAR. In particular, we have discussed how a relaying route between an MH and a BTS in a nearby cell can be established via ARSs, and evaluate the performance of the protocols in terms of request rejection rate and signaling overhead.

5. Finally, we have introduced a novel concept called ”managed mobility” of ARSs, based on which we have proposed a signaling protocol and studied the strategies for the mobility management in iCAR.

This dissertation represents a first step in evolving to the next generation integrated wireless mobile networks. We have focused on and solved the problems in the network layer and the system level management. In our future work, we will address the issues in the Media Access Control (MAC) and physical layers to support the iCAR system. Specifically, iCAR needs a novel MAC protocol for the relaying. The existing wireless MAC protocols (such as IEEE802.11) may not be the best solutions for iCAR as the cellular infrastructure can help packet scheduling so as to avoid collisions. For the physical layer of the relaying interface, various approaches (e.g., the Ultra-Wide Band Radio, frequency hopping, etc.) will be studied and evaluated, and proper technology will be chosen for supporting iCAR. In addition, we will extend the concept of iCAR to a more general integrated system which takes the advantages of various technologies, such as the flexibility of Ad hoc and sensor networks, the coverage of the cellular and the satellite systems, and the wide bandwidth of the wired networks.

ACKNOWLEDGMENTS FOR FUNDING PROVIDERS

This research is in part supported by NSF under the contract ANIR-ITR 0082916 and Nokia Research Center.

Bibliography

[1] W. Lee, *Mobile Cellular Telecommunications Systems*. McGraw-Hill, 1990.

[2] D. M. Balston and R. C. V. Macario, *Cellular Radio Systems*. Artech House, Norwood, Massachussets, 1993.

[3] J. F. Whitehead, "Cellular system design: An emerging engineering discipline," *IEEE Communications Magazine*, vol. 3, no. 1, pp. 8–15, 1986.

[4] D. Bantz and F. Bauchot, "Wireless LAN design alternatives," *IEEE Network*, vol. 8, no. 2, pp. 45–53, 1994.

[5] V. Bharghavan, A. Demers, S. Shenker, and L. Zhang, "MACAW: A media access protocol for wireless LANs," in *Proceedings, 1994 SIGCOMM Conference*, (London, UK), pp. 212–225, 1994.

[6] J. Haartsen, M. Naghshineh, J. Inouye, O. Joeressen, and W. Allen, "Bluetooth: Vision, goals, and architecture," *Mobile Computing and Communications Review*, vol. 2, pp. 38–45, Oct 1998.

[7] http://www.bluetooth.com/.

[8] http://www.homerf.org/.

[9] K. Negus, A. Stephens, and L. Jim, "Homerf: Wireless networking for the connected home," *IEEE Personal Communications*, vol. 6, pp. 20–27, 2000.

[10] C. Qiao, H. Wu, and O. Tonguz, “Load balancing via relay in next generation wireless systems,” in *Proceeding of IEEE Mobile Ad Hoc Networking & Computing*, pp. 149–150, 2000.

[11] C. Qiao and H. Wu, “iCAR : an integrated cellular and ad-hoc relay system,” in *IEEE International Conference on Computer Communication and Network*, pp. 154–161, 2000.

[12] H. Wu, C. Qiao, and O. Tonguz, “A new generation wireless system with integrated cellular and mobile relaying technologies,” in *International Conference on Broadband Wireless Access Systems (WAS’2000)*, pp. 55–62, 2000.

[13] http://www.nwr.nokia.com/.

[14] H. Wu, C. Qiao, S. De, and O. Tonguz, “Performance analysis of iCAR (integrated cellular and ad-hoc relay system),” in *IEEE International Conference on Communications*, vol. 2, pp. 450–455, 2001.

[15] H. Wu, C. Qiao, S. De, and O. Tonguz, “Integrated cellular and ad-hoc relay systems: iCAR,” *IEEE Journal on Selected Areas in Communications special issue on Mobility and Resource Management in Next Generation Wireless System*, vol. 19, no. 10, Oct. 2001. Edited by Ian F. Akyildiz, David Goodman and Leonard Kleinrock.

[16] H. Wu and C. Qiao, “Modeling iCAR via Multi-dimensional Markov Chains,” *ACM Mobile Networking and Applications (MONET), Special Issue on Performance Evaluation of Qos Architectures in Mobile Networks*, 2002. To appear.

[17] http://www.att.com/.

[18] V. Garg and J. Wilkes, *Wireless and Personal Communications Systems*. Prentice Hall, 1996.

[19] V. MacDonald, “AMPS: The cellular concept,” *Bell Sys. Tech. Journal*, vol. 58, no. 1, 1979.

[20] I. Korn, “M-ary frequency shift keying with limiter discriminator integrator detector in satellite mobile channel with narrowband receiver filter,” *IEEE Trans. Commun.*, vol. 38, pp. 1771–1778, 1990.

[21] M. Mouly and M.-B. Pautet, *The GSM System for Mobile Communications*. Palaiseau, France: Cell & Sys, 1992.

[22] M. RAHNEMA, “An overview of the GSM system and protocol architecture,” *IEEE Communications Magazine*, vol. 31, 1993.

[23] www.etsi.org/.

[24] G. Gudmundson, J. Skld, and J. Ugland, “A comparison between CDMA and TDMA systems,” in *IEEE 42ndVeh Tech Conf VTC92*, 1992.

[25] K. Raith and J. Uddenfeldt, “Capacity of digital cellular TDMA systems,” *IEEE Transactions on Vehicular Technology*, vol. 40, no. 2, pp. 323–332, 1991.

[26] M. Honig, “Analysis of a TDMA network with voice and data traffic,” *AT&T Bell Laboratories Technical Journal*, vol. 63, no. 8, pp. 1537–1563, 1984.

[27] A. Viterbi, *CDMA: Principles of Spread Spectrum Communications*. Addison-Wesley, 1995.

[28] K. Gilhousen, I. Jacobs, R. Padovani, A. Viterbi, L. Jr, and C. On, “The capacity of a cellular CDMA system,” *IEEE Trans. Vehic. Technol.*, pp. 303–312, May 1991.

[29] C. Huitema, *IPv6 - The New Internet Protocol*. Prentice Hall PTR, 1996.

[30] S. Deerign and R. Hinden, *Internet Protocol, Version 6 (IPv6) Specification*, RFC 2460 ed., 1998.

[31] R. Kalden, I. Meirick, and M. Meyer, “Wireless internet access based on GPRS,” *IEEE Personal Comm.*, vol. 7, pp. 8–18, Apr. 2000.

[32] M. Meyer, “TCP performance over GPRS,” *IEEE Wireless Communications and Networks Conferance 1999 (WCNC’99)*, vol. 3, pp. 1248–1252, 1999.

[33] H. Matt, “3G wireless,” *Computerworld*, pp. 63–65, February 21, 2000.

[34] C. Comaniciu, N. Mandayam, D. Famolari, P. Agrawal, and G. for, “CDMA systems via admission and flow control,” in *IEEE Vehicular Technology Conference (VTC)*, September,2000.

[35] J. Huber, D. Weiler, and H. Brand, “The mobile multimedia vision for IMT2000: A focus on standardization,” *IEEE Comm. Magazine*, vol. 38, pp. 129–136, 2000.

[36] N. R. Prasa, “GSM evolution towards third generation UMTS/IMT2000,” in *IEEE International Conference on Personal Wireless Communications*, pp. 50–54, 1999.

[37] B. Kreller, “UMTS: A middleware architecture and mobile api approach,” *IEEE Personal Communications*, vol. 5, no. 2, 1998.

[38] G. Fleming, A. Hoiydi, J. de Vriendt, G. Nikolaidis, F. Piolini, and M. Maraki, “A flexible network architecture for umts,” *IEEE Personal Communications Magazine*, vol. 5, no. 2, pp. 8–15, 1998.

[39] T. Ojanpera and R. Prasad, “An overview of air interface multiple access for IMT-2000/UMTS,” *IEEE Communications Magazine*, vol. 36, no. 9, pp. 82–95, 2000.

[40] T. Ojanpera and R. Prasad, *Wideband CDMA for Third Generation Mobile Communications*. Artech House Publishers, 1998.

[41] E. Dahlman, P. Beming, J. Knutsson, F. Ovesjo, M. Persson, and C. Roobol, “Wcdma - the radio interface for future mobile multimedia communications,” *IEEE Trans. Vehicular Technology*, vol. 47, no. 4, pp. 1105–1118, 1998.

[42] H. Holma and A. Toskala, *WCDMA For UMTS*. John Wiley & Sons, 2000.

[43] S. Das, R. Castaneda, J. Yan, and R. Sengupta, “Comparative performance evaluation of routing protocols for mobile, ad hoc networks,” in *7th Int. Conf. on Computer Communications and Networks (IC3N)*, pp. 153–161, 1998.

[44] Y.-B. Ko and N.H.Vaidya, “Location-aided routing (LAR) in mobile ad hoc networks,” in *ACM/IEEE the 4th Annual Intl. Conference on Mobile Computing and Networking (MobiCom 98)*, 1998.

[45] C. Toh, *Wireless ATM and Ad-Hoc Networks: Protocols and Architectures*. Kluwer Academic Publishers, 1996.

[46] C. Perkins and P. Bhagwat, “Highly dynamic destination sequenced distance vector routing(dsdv) for mobile computers,” in *Proceedings of ACM SIGCOMM’94*, pp. 234–244, 1994.

[47] C. Perkins and E. Royer, “Ad-hoc on demand distance vector routing,” in *Proceedings of IEEE WMCSA’99*, pp. 90–100, 1999.

[48] V. Park and M. Corson, “A highly adaptive distributed routing algorithm for mobile wireless networks,” in *Proceedings of IEEE INFOCOM’97*, pp. 1405–1413, 1997.

[49] D. Johnson and D. Maltz, “Dynamic source routing in ad hoc wireless networks,” *Mobile Computing*, vol. 5, pp. 153–181, 1996.

[50] S. Murthy and J. Garcia-Luna-Aceves, “An efficient routing protocol for wireless networks,” *ACM/Baltzer Mobile Networks and Applications*, vol. 1, no. 2, pp. 183–197, 1996.

[51] A. Iwata, C.-C. Chiang, G. Pei, M. Gerla, and T.-W. Chen, “Scalable routing strategies for adhoc wireless networks,” *IEEE Journal on Selected Areas in Communications*, vol. 17, no. 8, pp. 1369–1379, 1999.

[52] S. Basagni, I. Chlamtac, V. Syrotiuk, and B. Woodward, “A distance routing effect algorithm for mobility (dream),” in *Proceedings of the ACM/IEEE MobiCom ’98 Conference*, pp. 76–84, 1998.

[53] S. Chen and K. Nahrstedt, “Distributed quality-of-service routing in ad hoc networks,” *IEEE Journal on Selected Areas in Communications*, vol. 17, no. 8, pp. 1488–1505, 1999.

[54] S.-B. Lee, G.-S. Ahm, X. Zhang, and A. T. Campbell, "Insignia: An ip-based quality of service framework for mobile ad hoc networks," *Journal of Parallel and Distributed Computing*, vol. 60, pp. 374–406, 2000.

[55] C.-K. Toh, "A novel distributed routing protocol to support ad-hoc mobile computing," in *15th IEEE Annual International Phoenix Conference on Computers and Communications*, 1996.

[56] J. Garcia-Luna-Aceves, "Source tree adaptive routing (star) protocol." Internet Draft, draft-ietf-manet-star-00.txt, http://www.ietf.org/internet-drafts/draft-ietf-manet-star-00.txt.

[57] H. Holma and A. Toskala, *WCDMA For UMTS*, p. 44. John Wiley & Sons, 2000.

[58] A. Gamst, "Study of radio network design stategies," in *IEEE Vehicular Technology Conference 36th*, pp. 319–328, 1986.

[59] J. Oetting, "Cellular mobile radio - an emerging technology," *IEEE Communications Magazine*, vol. 21, no. 8, pp. 10–15, 1983.

[60] V. Garg and J. Wilkes, *Wireless and Personal Communications Systems*, pp. 88–90. Prentice Hall, 1996.

[61] G. K. Chan, "Effects of sectorization on the spectrum efficiency of cellular radio systems," *IEEE Transactions on Vehicular Technology*, vol. 41, no. 3, pp. 217–225, 1992.

[62] G. Chan, "Effects of sectorization on the spectrum efficiency of cellular radio systems," *IEEE Trans. Veh. Technol.*, vol. 41, pp. 217–225, 1992.

[63] L. Wang, K. Chawla, and L. Greenstein, "Performance studies of narrow beam trisector cellular systems," *International Journal of Wireless Information Networks*, vol. 5, no. 2, pp. 89–102, 1998.

[64] V. Macdonald, "Advanced mobile phone service: the cellular concept," *Bell System Technical Journal*, vol. 58, pp. 15–41, 1979.

[65] S. Elnoubi, R. Singh, and S. Gupta, “A new frequency channel assignment in high capacity mobile communication systems,” *IEEE Trans. Veh. Technol.*, vol. 31, no. 3, 1982.

[66] M. Zhang and T. Yum, “Comparisons of channel-assignment strategies in cellular mobile telephone systems,” *IEEE Transactions on Vehicular Technology*, vol. 38, no. 4, pp. 211–215, November,1989.

[67] D. C. Cox and D. O. Reudink, “Increasing channel occupancy in large scale mobile radio systems: dynamic channel reassignment,” *IEEE trans. on commun.*, vol. COM-21, pp. 1302–1306, 1973.

[68] J. Tajima and K. Imamura, “A strategy for flexible channel assignment in mobile communication systems,” *IEEE Transactions on Vehicular Technology*, vol. 37, May 1988.

[69] J. Karlson and B. Eklundh, “A cellular mobile telephone system with load sharing and enhancement of directed retry,” *IEEE Trans Commun*, vol. 37, pp. 530–535, May 1989.

[70] H. Jiang and S. Rappaport, “Cbwl: A new channel assignment and sharing method for cellular communication systems,” *IEEE Transactions on Vehicle Technology*, vol. 43, pp. 313–322, May 1994.

[71] S. K. Das, S. K. Sen, and R. Jayaram, “A dynamic load balancing strategy for channel assignment using selective borrowing in cellular mobile environment,” in *Mobile Computing and Networking*, pp. 73–84, 1996.

[72] S. K. Das, S. K. Sen, and R. Jayaram, “A structured channel borrowing scheme for dynamic load balancing in cellular networks,” in *International Conference on Distributed Computing Systems*, pp. 0–, 1997.

[73] J. Qiu and J. Mark, “A dynamic load sharing algorithm through power control in cellular cdma,” in *PIRMC’98*, pp. 1280–1284, September 1998.

[74] http://wireless.iop.org/article/feature/2/3/4.

[75] http://www.metawave.com.

[76] http://www.3gpp.org/.

[77] X.-X. Wu, B. Mukerjee, and S.-H. G. Chan, "Maca – an efficient channel allocation scheme in cellular networks," in *IEEE Global Telecommunications Conference (Globecom'00)*, vol. 3, pp. 1385–1389, 2000.

[78] Y.D.Lin and Y.C.Hsu, "Multihop cellular: A new architecture for wireless communication," in *IEEE INFOCOM'2000*, pp. 1273–1282, 2000.

[79] I.F.Akyildiz, W. Yen, and B. Yener, "A new hierarchical routing protocol for dynamic multihop wireless networks," in *IEEE INFOCOM'97*, pp. 1422–1429, 1997.

[80] A. Campbell, J. Gomez, and A. Valko, "An overview of cellular ip," in *IEEE Wireless Communications and Networking Conference*, 1999.

[81] A. Valko, A. Campbell, and J. Gomez, "Cellular ip," 1998. Internet Draft, Work in Progress, draft-valko-cellularip-00.txt.

[82] A. Valko, "Cellular IP - a new approach to internet host mobility," *ACM Computer Communication Review*, 1999.

[83] C. E. Perkins and A. Myles, "Mobile IP," *Proceedings of International Telecommunications Symposium*, pp. 415–419, 1994.

[84] C. E. Perkins and D. B. Johnson, "Mobility support in ipv6," in *Mobile Computing and Networking*, pp. 27–37, 1996.

[85] C. E. Perkins, "Mobile networking through mobile IP," *IEEE Internet Computing*, vol. 2, no. 1, pp. 58–69, 1998.

[86] T. Rappaport, *Wireless Communications Principle and Practice*. Prentice Hall, 1996.

[87] Z. Haas and B. Liang, "Ad hoc mobility management with uniform quorum systems," *IEEE/ACM Transactions on Networking*, vol. 7, no. 2, pp. 228–240, 1999.

[88] L. Chen, H. Murata, S. Yoshida, and S. Hirose, “Wireless dynamic channel assignment performance under packet data traffic,” *IEEE Journal on Selected Areas in Communications*, vol. 17, no. 7, pp. 1257–1269, 1999.

[89] J.-L. Pan, S. Rappaport, and P. Djuric, “Multibeam cellular communication systems with dynamic channel assignment across multiple sectors,” *Wireless Networks*, vol. 5, no. 4, pp. 231–243, 1999.

[90] R. Syski, *Introduction to Congestion Theory in Telephone Systems*. Oliver & Boyd, 1960.

[91] Y. Fang, I. Chlamtac, and Y. Lin, “Call performance for a PCS network,” *IEEE Journal on Selected Areas in Communications*, vol. 15, no. 8, pp. 1568–1581, 1997.

[92] M. Meo and M. Marsan, “Approximate analytical models for dual-band GSM networks design and planning,” in *IEEE INFOCOM*, pp. 1263–1272, 2000.

[93] D. Hong and S. S. Rappaport, “Traffic model and performance analysis for cellular mobile radio telephone systems with prioritized and no-protection handoff procedure,” *IEEE Transactions on Vehicular Technology*, vol. 3, pp. 77–92, 1986.

[94] X. Zeng, R. Bagrodia, and M. Gerla, “GloMoSim: A library for parallel simulation of large-scale wireless networks,” in *Proc. Workshop on Parallel and Distributed Simulation*, pp. 154–161, 1998.

[95] H. Jung and O. K. Tonguz, “Random spacing channel assignment to reduce the nonlinear intermodulation distortion in cellular mobile communications,” *IEEE Transactions on Vehicular Technology*, vol. 48, no. 5, pp. 1666–1675, 1999.

[96] H. Ebersman and O. K. Tonguz, “Handoff ordering using signal prediction priority queueing in personal communication systems,” *IEEE Transactions on Vehicular Technology*, vol. 48, no. 1, pp. 20–35, 1999.

[97] R. L. Freeman, *Telecommunication System Engineering*. John Wiley & Sons Inc., 1996.

[98] R. Bagrodia, R. Meyer, M. Takai, Y. Chen, X. Zeng, J. Martin, B. Park, and H. Song, “Parsec: A parallel simulation environment for complex systems,” *Computer*, pp. 77–85, Oct. 1998.

[99] W. R. Stevens, *TCP/IP Illustrated, Volume 1.* Addison Wesley, 1994.

[100] U. Black, *Mobile and Wireless Networks.* Prentice Hall PTR, 1999.

[101] M. Gallagher and W. Webb, “UMTS: The next generation of mobile radio,” *IEE Review*, vol. 45, no. 2, pp. 59–63, 1999.

[102] S. Murthy and J. J. Garcia-Luna-Aceves, “Routing architecture for mobile integrated services networks,” *Mobile Networks and Applications*, vol. 3, no. 4, pp. 391–407, 1999.

[103] D. Maltz, J. Broch, J. Jetcheva, and D. Johnson, “Effects of on-demand behavior in routing protocols for multihop wireless ad hoc networks,” *IEEE Journal on Selected Areas in Communications*, vol. 17, no. 8, pp. 1439–1453, 1999.

[104] A. Iwata, C.-C. Chiang, G. Pei, M. Gerla, and T.-W. Chen, “Scalable routing strategies for ad hoc wireless networks,” *IEEE Journal on Selected Areas in Communications*, vol. 17, no. 8, pp. 1369–1379, 1999.

[105] D. Johnson and D. Maltz, “Protocols for adaptive wireless and mobile networking,” *IEEE Personal Communications*, vol. 3, no. 1, pp. 34–42, 1996.

[106] H. Wu, C. Qiao, S. De, E. Yanmaz, S. Mishra, O. Tonguz, and S. Dixit, “Quality of ARS (Ad hoc Relaying Station) Coverage in iCAR (Integrated Cellular and Ad hoc Relaying),” Submitted to a journal. A preliminary version appeared in ACM SIGCOMM 2001 conference student poster session and ACM Computer Communication Review (CCR), vol 32, no. 1, January 2002.

[107] C. C. Shen, C. Srisathapornphat, and C. Jaikaeo, “Sensor information networking architecture and applications,” *IEEE Personal Communications*, vol. 8, no. 4, pp. 52–59, August 2001.

[108] K. Sohrabi, J. Gao, V. Ailawadhi, and G. J. Pottie, “Protocols for self-organization of a wireless sensor network,” *IEEE Personal Communications*, vol. 7, no. 5, pp. 16–27, Oct 2001.