Routing Performance

In Ad Hoc Networks

Yuvaan Ramesh Gugrajah

Submitted in fulfilment of the academic requirements of the Master of Engineering Degree in the School of Electrical and Electronic Engineering, University of Natal, Durban, South Africa.

January 13, 2003.

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

Signed: ______________________

Name: Prof. Fambirai Takawira

Date: 13 January 2003

ABSTRACT

An ad hoc network is a multi-hop wireless network in which mobile nodes communicate over a shared wireless channel. The network is formed cooperatively without specific user administration or configuration and is characterised by a distributed network management system and the absence of a wired backbone. Military, law enforcement, and disaster relief operations are often carried out in situations with no pre-existing network infrastructure and can benefit from such networks because base stations, which are single points of failure, are undesirable from a reliability standpoint. The rising popularity of mobile computing has also created a potentially large commercial market for multimedia applications applied over wireless ad hoc networks.

This dissertation focuses on the routing aspects of ad hoc networking. The multi-hop routes between nodes constantly change as the mobile nodes migrate. Ad hoc network routing algorithms must therefore adapt to the dynamic and unpredictable topology changes, the random radio propagation conditions and portable power sources. Various routing protocols have been proposed in the literature for ad hoc networks. These protocols together with comparative simulations are discussed and a new protocol based on load balancing and signal quality determination is proposed and the simulation results are presented.

Currently the proposed routing protocols are compared using simulation packages which are often time consuming. This dissertation proposes a mathematical model for evaluating the routing protocols and the resultant end-to-end blocking probabilities. The mathematical model is based on a derivation of the reduced load approximation for analysing networks modelled as loss networks and the evaluation incorporates and adapts models that have been used for the analysis of cellular Code Division Multiple Access (CDMA) systems. While analytical methods of solving blocking probability can potentially generate results orders of magnitude faster than simulation, they are more importantly essential to network sensitivity analysis, design and optimisation.

PREFACE

The research work in this dissertation was performed by Mr. Yuvaan Gugrajah under the supervision of Prof. Fambirai Takawira in the School of Electrical and Electronic Engineering at the University of Natal, Durban, South Africa. This work was partially sponsored by Telkom South Africa and Alcatel Altech Telecoms as part of the Centre of Excellence Programme at the University of Natal.

The whole dissertation, unless specifically indicated to the contrary in the text, is the author's work, and has not been submitted in part, or in whole to any other university.

ACKNOWLEDGEMENTS

I would like to thank Prof. Takawira for his support, advice, encouragement and guidance during the past two years. His constructive criticism and patience are greatly appreciated.

I also wish to express my sincere appreciation to my father, mother and brother who have been a constant source of inspiration and support not only during my studies but also throughout my life. Special thanks are owed to my financè Reshna, for her encouragement and understanding during trying times and to the Mungar family for all their patience and assistance during the compilation of this dissertation.

I would like to thank Telkom SA for their valued financial support during my studies at university and Alcatel Altech Telecoms for the funding of this research project.

Acknowledgement also goes to my friends for making our time together enjoyable and for the hours of entertainment and thought provoking discussions.

CONTENTS

LIST OF FIGURES AND TABLES

Chapter 2

Chapter 3

Chapter 4

Chapter 5

LIST OF SYMBOLS

a_{js}	probability that a class s call will be admitted on link j
a_n	probability that a call will be admitted at node n
A_i	evaluation assignment to node I
B	bit rate
B_{avg}	average network blocking probability
B_j	blocking probability of link j
B_r	blocking probability of a call between node pair r
B_{rs}	blocking probability of a class s call between node pair r
C^f_j	free bandwidth on link j
C_{ik}	percentage of successful packet receptions and transmissions by node i, to and from node k
C_j	capacity of link j
d_f	Fraunhofer distance
$E[k]$	expected number of occupied bandwidth units on a link
$E_{b,k}$	bit energy of user k
e_l	probability of existence for link l
$E_{n,k}$	probability node n will accept a call given k calls in progress
G	processing gain
G_r	receiver gain
G_t	transmitter gain
h_{rm}	number of intermediate nodes on route (r,m)
I_0	maximum total acceptable interference density
$I_{u'}$	interference to reception of packet U
J	total number of links in the network
K_{rm}	most congested node on route (r,m)
λ	call arrival rate
$\ell_{(r,m)}$	most congested link on route (r,m)
M_r	total number of routes between node pair r
η	ratio of one-sided spectral noise density to total acceptable interference density
N	total number of nodes in a network
N_0	the one-sided spectral noise density
$p(k)$	probability that a system is in state k
P_{GT}	Packet Generation Time

$p_j(k)$	probability that k units of bandwidth are being used on link j
P_m	probability of an m-hop connection
$p_n(k)$	probability that k calls are in progress at node n
P_r	received power
P_t	transmit power
P_u	received power of packet U
Q_i	percentage of node i's transmit buffer that is empty
q_{rm}	probability that route (r, m) will be used for call request
q_{rms}	probability that route (r, m) will be used for call request (r, s)
(r,m)	mth route between node pair r
ρ	probability of activity of a call
R	total number of node pairs in a network
S_{ik}	average SIR of all packets received by node i from node k
S_k	system is in state k
τ	threshold for buffer occupation in DLAR scheme 3
$t_j(k)$	the probability of no more than k trunks being free on link j
T_m	monitoring time period
$1/\mu$	mean call duration
v_{js}	arrival rate of class s calls on link j
v_n	arrival rate of calls at node n
v_max	maximum velocity
w	weighting factor
W	system bandwidth
w_{ratio}	ratio of routing weighting factors
x_max	maximum x-dimension of network area
y_max	maximum y-dimension of network area
Z_k	noise plus interference in the CDMA system

LIST OF ACRONYMS

ABR	Associativity Based Routing
ACK	Acknowledgement
AMRIS	Ad Hoc Multicast Routing with Increasing ID numbers
AMROUTE	Ad Hoc Multicast Routing
AODV	Ad Hoc On Demand Distance Vector Routing
BQ	Broadcast Query
CAMP	Core-Assisted Mesh Protocol
CDMA	Code Division Multiple Access
CEDAR	Core Extraction Distributed Ad Hoc Routing
CGSR	Cluster-head Gateway Switch Routing
CLR	Route Clear Packet
CMU	Carnegie Mellon University
CSMA/CA	Carrier Sense Multiple Access with Collision Avoidance
CTS	Clear To Send
DAG	Directional Acyclic Graph
DBF	Distributed Bellman Ford
DCF	Distributed Coordination Function
DIFS	DCF Interframe Spacing
DLAR	Dynamic Load Aware Routing
DREAM	Distance Routing Effect Algorithm for Mobility
DSDV	Destination Sequenced Distance Vector
DSR	Dynamic Source Routing
DSSS	Direct-Sequence Spread Spectrum
EFPA	Erlang Fixed Point Approximation
FDMA	Frequency Division Multiple Access
FHSS	Frequency-Hopping Spread Spectrum
FSR	Fisheye State Routing
GPS	Global Positioning System
GSR	Global State Routing
HID	Hierarchical Identification number
HSR	Hierarchical State Routing
IARP	Intrazone Routing Protocol

IEEE	Institute of Electrical and Electronic Engineers
IERP	Interzone Routing Protocol
IETF	Internet Engineering Task Force
IrDA	Infrared Data Association
LAR	Location Aided Routing
LCC	Least Cluster-head Change algorithm
LP	Location Packet
LQ	Localised Query
LSP	Link State Packet
MAC	Medium Access Control
MANET	Mobile Ad Hoc Networking
MCDS	Minimum Connected Dominating Set
MCEDAR	Multicast Core Extraction Distributed Ad Hoc Routing
MPR	Multipoint Relay
MRL	Message Retransmission List
ns2	network simulator 2
NTDR	Near Term Digital Radio
ODMPR	On Demand Multicast Routing Protocol
OLSR	Optimized Link State Routing
PAN	Personal Area Network
PCF	Point Coordination Function
PCS	Personal Communication Service
PGT	Maximum Packet Generation Time
PRNet	Packet Radio Network
PSR	Partial Knowledge Spine Routing
QoS	Quality of Service
QRY	Route Query Packet
RD	Route Delete
RERR	Route Error
RRC	Route Reconstruction
RREP	Route Reply
RREQ	Route Request
RTS	Request To Send
SIR	Signal-to-Interference Ratio
SSR	Signal Stability Based Routing
SURAN	Survivable Adaptive Radio Network

SWAP	Shared Wireless Access Protocol
TDMA	Time Division Multiple Access
TORA	Temporally Ordered Routing Algorithm
UPD	Route Update Packet
WLAN	Wireless Local Area Network
WRP	Wireless Routing Protocol
ZHLS	Zone-based Hierarchical Link State
ZRP	Zone Routing Protocol

Chapter 1

INTRODUCTION

1.1. Ad Hoc Networking

The introduction of laptop and notebook computers, personal digital assistants, and palmtop devices has facilitated commercial mobile computing. The usefulness of these small and portable computers is augmented when the variety of information services provided by the Internet and local area networks are made accessible using wireless networking. However, the studies and the developments in wireless networking have primarily been driven by the success of the cellular architecture, where mobile computers communicate with a central controller, base station or access point. Many of these developments are still not directly applicable to satisfy the needs of wireless systems that require network architectures which do not necessarily follow the cellular model. When infrastructure is either not available, not trusted, or should not be relied on in times of emergency, infrastructure-less wireless networks can be used. They consist of mobile computers that communicate directly or via multiple hop paths with each other and are commonly referred to as ad hoc networks.

The lack of a central controller implies that each mobile node must behave both as a router and a host and multi-hop paths require the support of intermediate nodes to achieve connectivity. The mobile computers or nodes in an ad hoc network may be located in airplanes, cars, on people or any other small devices and are capable of being connected wirelessly and dynamically in an arbitrary manner. With the development of ad hoc networks truly ubiquitous computing and communication could become a reality.

There are a variety of application areas for ad hoc networks. Search and rescue missions in areas where existing infrastructure has been destroyed or is non-existent can benefit from these decentralized networks. Ad hoc networks could also be an advantage for data acquisition in inhospitable or unexplored terrain. Networks of wireless devices for voice and video communication, digital maps, and mobile sensors can provide valuable tactical information in military applications. The history of ad hoc networks can in fact be traced back to 1972 [Ram02] when the US Department of Defence introduced the Packet Radio Network (PRNet), which later became the Survivable Adaptive Radio Networks (SURAN) program. The Near-Term Digital Radio (NTDR) [Cotter02] is the latest incarnation of the US military's mobile packet radio networks and is the only non-prototype ad hoc network in use today.

The Institute of Electrical and Electronic Engineers (IEEE) replaced the designation "packet radio network" with "ad hoc network" when developing the 802.11 standard for Wireless Local Area Networks (WLAN) in the hope that it would indicate a new deployment scenario different from the associated multihop networks of large scale military operations. Commercial applications, where individual users need to share or exchange information without depending on a local network of access points, can employ ad hoc networks for collaborative wireless networking. Ad hoc networks can play a role in civilian forums such as classrooms, convention centres and office blocks. Quick deployment and configuration can be used in rural areas where cellular networks are unavailable. Personal Area Networks (PANs), which are small scale ad hoc networks, allow intercommunication between a person's various mobile devices, eliminating the need for cables and could also be used to extend the mobility provided by the fixed network.

Two of the technologies that are currently competing for commercial applications are the Bluetooth standard [Haartsen98], [Bluetooth] and the IEEE 802.11 standard [IEEE99].

Bluetooth is a universal radio interface that uses a frequency-hopping spread spectrum scheme in the unlicensed 2.4 GHz frequency band. Portable electronic devices can communicate wirelessly over short ranges with up to seven other devices in a small network referred to as a piconet. One device in the piconet becomes a master while the others are considered slaves. Two or more piconets can be interconnected to form a scatternet. However solutions for forming multi-hop ad hoc networks over Bluetooth scatternets still need to be introduced.

Bluetooth neither addresses routing, nor specifies which of the unicast and multicast routing protocols to use and the Bluetooth medium access control protocol does not specify how to cope with mobility. Initially, Bluetooth is being used as a replacement for point to point and point to multipoint cables. However, the small size of Bluetooth solutions (a chip is approximately 2 cm^2 centimetres in area), low cost (approximately $10 per chip) and low power requirements (1mW to 100mW for ranges of 0.1m to 100m), together with the involvement of many of the major wireless communication device manufacturers (such as Ericsson, Nokia, Toshiba, IBM, Lucent, Microsoft, 3Com and Intel) proves advantageous for the future of Bluetooth.

IEEE 802.11 specifies a wireless interface between a client and a base station or access point, as well as between clients, operating at between 1 and 2 Mbps. Two physical layer characteristics are defined: direct-sequence spread spectrum (DSSS), and frequency-hopping spread spectrum (FHSS), both of which operate in the 2.4 GHz range. The 802.11b "High Rate" amendment to the standard added two higher speeds at 5.5 Mbps and 11 Mbps to 802.11 and selected DSSS as the sole physical layer technique. The point coordination function (PCF) uses a centralized approach to allow an access point to control all traffic, while the distributed coordination function (DCF) allows direct communication between wireless clients. However, the IEEE 802.11 standard also does not specify a method for multihop ad hoc networking.

Other protocols that have potential for ad hoc networking include the HomeRF working group's Shared Wireless Access Protocol (SWAP) [CSDMag00] and the infrared communication protocols [IrDA] of the Infrared Data Association (IrDA).

The problems that face the development of ad hoc networks include limited power from portable power sources and the need for suitable power control algorithms, low bandwidth and high error rates from wireless channels, medium access control strategies, scheduling policies, security algorithms and routing protocols that can cope with the mobility and changes in topology. This dissertation focuses on routing in ad hoc networks.

1.2. Routing in Ad Hoc Networks

The Mobile Ad Hoc Networking (MANET) [Man02] working group was created within the Internet Engineering Task Force (IETF) in the mid 1990s when mobile computing became popular and viable communications equipment based on radio frequency and infrared were developed. The key focus of the working group is to develop and evolve ad hoc network routing specifications for standardization. Due to the constantly changing topology and lack of central controller the routing protocol implemented needs to be distributed and dynamic, efficiently utilizing available resources and achieving fast route convergence.

In current cellular networks, user mobility is handled by forwarding calls via the user's home network to the visited network. This forwarding principle also applies to Mobile IP [Perkins98]. A roaming user that enters a foreign network is associated with a visiting address provided by a foreign agent. The home agent establishes an Internet Protocol (IP) tunnel to the foreign agent using the provided visiting address. Any packet sent to the roaming user's home address is first sent to the home agent which redirects it to the foreign agent via the visiting address. This forwarding approach is only useful where only the nodes at the edges of fixed networks are mobile.

In ad hoc networks this is not the case. Nodes at the centre of the network can also move, with the whole network being based on the idea that mobile devices behave both as routers and as hosts. Therefore, in an ad hoc network, mobility is handled by the routing algorithm, which needs to take care of changes in topology.

The development of routing protocols for ad hoc networks first began with an attempt to modify routing protocols for wired networks [Bertsekas92]. This resulted in table-driven protocols that maintain consistent tables of network information [Perkins94], [Jaquet98], [Murthy95]. However, due to the constantly changing topologies in an ad hoc network, the table driven protocols were found to create too much congestion due to the updating of the tables [Broch98], [Johansson99]. This resulted in another type of routing strategy for ad hoc networks, referred to as on-demand routing or reactive routing [Johnson95], [Johnson96], [Park97], [Perkins99]. Reactive protocols only search for a route when a source requires one, and usually use some type of flooding mechanism to determine the state of the network topology.

Hybrid protocols and protocols with special adaptations have also been developed to overcome problems unique to ad hoc networking. These include location aided routing protocols [Basagni98], [Ko98], routing based on network stability [Dube97], [Toh99], routing with a virtual backbone [Siva98] and the multi-scope routing protocols which include clustering and hierarchical algorithms [Chiang97], [Gerla98], [Haas99], [Iwata99], [Joa99]. Attempts have also been made to develop protocols to maintain a certain degree of Quality of Service (QoS) [Chen99] since most of the routing protocols for ad hoc networks only consider the shortest path as the main routing metric.

1.3. Mathematical Analysis of Routing Protocols

The analysis of fixed networks has generally been achieved by employing a loss network model [Ross95]. In a loss network each call requires a fixed amount of bandwidth on every link on a route between the source and destination. The call is admitted to the network and holds the requested capacity for a certain amount of time if each link on the route has enough bandwidth to satisfy the requirements. The call is rejected otherwise, and the blocking probability is the probability that the call is rejected.

When there are multiple links and many alternate routes in a network between node pairs, the accurate evaluation of the blocking probability becomes practically unfeasible because of the number of possible states that the network can be in. One approach is to analyse loss networks using the reduced load approximation [Ross95]. The reduced load approximation is used to estimate the blocking probability of calls offered to a network and is based on two assumptions:

i) Link independence: blocking occurs independently from link to link.

ii) Poisson assumption: traffic flow to each individual link is Poisson

All of the comparisons of the routing protocols for ad hoc networks are performed using simulators, which are not only time consuming but also do not allow as rapid optimisation of routing protocols by a network designer as would analytical models. This therefore motivated the research of methods to evaluate routing protocols in ad hoc networks using mathematical models [Gugrajah02a], [Gugrajah02b].

In attempting to develop an analytical model for ad hoc networks, the wireless transmission environment needs to be considered. In wireless cellular networks employing Frequency Division Multiple Access (FDMA) and Time Division Multiple Access (TDMA), the number of channels available also fixes the capacity of the system. This hard capacity contrasts with the so called soft capacity of Code Division Multiple Access (CDMA) systems which use the same frequency for each transmitter but which are distinguished from one another through the use of distinct orthogonal codes. The spreading codes are however not always perfectly orthogonal and this results in multiple access interference on a link. A wireless link in a CDMA system therefore has a threshold signal-to-interference ratio at which the link should operate. The evaluation of blocking probability in CDMA cellular networks is frequently based on the Erlang Capacity of the networks [Viterbi93], [Narr99].

As far as the author is aware, there is no literature available on modelling ad hoc networks as loss networks. However, an ad hoc network can be modelled as a loss network if it is assumed that the routing protocol chooses its route from a list of possible routes based on the available resources. By combining the evaluation of blocking probability in CDMA networks with loss network modelling, the reduced load approximation can be adapted to evaluate the blocking probability in ad hoc networks. Analytical methods of solving blocking probability can potentially generate results orders of magnitude faster than simulation and can be used for network sensitivity analysis, design and optimisation.

1.4. Dissertation Outline

This dissertation focuses on the routing aspects of ad hoc networking. Chapter 1 has introduced the concept of ad hoc networking and described the application areas and technologies that have been associated with the expected deployment of ad hoc networks. The problems facing the development of ad hoc networks as well as the genealogy of the routing protocols for ad hoc networks have been discussed. The general methods for evaluating the blocking probability in wired and wireless networks were briefly mentioned.

In Chapter 2, the routing protocols that have been proposed in the literature are reviewed. The different types of proactive and reactive routing protocols are described together with the hybrid, multiscope and hierarchical routing protocols. Some of the comparative simulations that have been published are examined. The concept of multicasting is also discussed.

Chapter 3 is a description of a load balancing routing protocol and the modifications made in order to reduce traffic congestion and improve signal quality. The protocol is compared to other protocols using a custom built simulator. The physical layer model, medium access control model and mobility model used in the simulator are described. Results for packet delivery ratio and average end-to-end delay are presented. The results obtained show that the implementation of load balancing in the routing protocols for ad hoc networks is beneficial when the offered traffic is high.

Chapter 4 is a review of the methods used to evaluate blocking probability in wired networks and wireless networks employing CDMA. The Erlang Loss Formula is first discussed which introduces fundamental concepts in loss network modelling. Loss networks and the fixed-point approximation are then described. Chapter 4 concludes with a discussion on the Erlang capacity of cellular CDMA systems.

In Chapter 5 an analytical model to evaluate the blocking probability of ad hoc networks employing CDMA is presented. CDMA systems offer high spectrum efficiency, multipath resistance, inherent frequency diversity and interference rejection and the potential use of advanced antenna and receiver structures. The analytical model is based on a modification of the reduced load approximation proposed by Liu [Liu00] and combines it with the Erlang Capacity for cellular CDMA systems for adaptation to the ad hoc network scenario. Reasons for the difficulties in including mobility in the analysis are also mentioned. Analytical results for the blocking probability in an ad hoc network are presented.

Chapter 6 summarises the thesis and briefly discusses the work performed in each chapter. Some future directions for further work are also discussed.

1.5. Original Contributions

In Chapter 3 new extensions to enhance load balancing routing protocols for ad hoc networks are proposed. The Dynamic Load Aware Routing (DLAR) protocol [Gerla00] is modified with a fourth scheme which takes into account signal quality and load at nodes along the route. Simulations are performed comparing the three original DLAR schemes and the Ad Hoc On-Demand Distance Vector (AODV) routing protocol [Perkins99] to the newly proposed scheme.

A new method of evaluating the blocking probability of calls in an ad hoc network is proposed in Chapter 5. This method combines loss network modelling for fixed infrastructure networks with the Erlang capacity evaluation of wireless CDMA networks. The evaluation methods for cellular networks are modified for multi-hop ad hoc networks. While methods for analysis of cellular networks employing CDMA have been proposed in the literature, there is no available literature on the approach proposed in this dissertation to evaluate wireless multi-hop ad hoc networks.

The work has been published at the following conferences:

- SATNAC 2001, Wild Coast, South Africa [Gugrajah01]
- IEEE Africon 2002, George, South Africa [Gugrajah02a]
- SATNAC 2002, Drakensburg, South Africa [Gugrajah02b]

Chapter 2

ROUTING PROTOCOLS FOR AD HOC NETWORKS

2.1. Introduction

Routing is the function of the network that determines the path from a source to a destination for the establishment of viable communication. Ad hoc networking allows users to access the network independent of location and movement, but this mobility allows the topology of the network to change dynamically and unpredictably. The problem is amplified because routes often consist of multiple hops due to intentionally short transmission ranges that reduce interference and power consumption for nodes in the network. Several approaches have been proposed to provide routing for ad hoc networks by adapting techniques developed in wired networks. This chapter provides a survey of the existing schemes proposed for end-to-end routing in ad hoc networks.

Section 2.2 introduces unicast protocols. The unicast protocols include traditional distance vector and link state routing, as well as newer path-finding routing protocols and on-demand routing protocols. Unicast protocols can be divided into proactive or table-driven protocols, described in Section 2.2.1 and the reactive or on-demand routing protocols described in Section 2.2.2. Modifications and hybrid varieties of the two types of unicast protocols have also been proposed and are discussed in Sections 2.2.3 to 2.2.7.

Multicast routing protocols include variations on the shortest path tree algorithms, multicast trees and mesh-based protocols. In graph theory a tree is a path that does not contain any loops. Since this dissertation is focused on unicast routing protocols, multicasting is briefly reviewed in Section 2.3.

2.2. Unicast Routing Protocols

Unicast protocols for ad hoc networks are derived from the two fundamental types of routing strategies: distance vector routing and link-state routing. Link state algorithms flood routing information to all nodes in the network. Each router, however, only sends the portion of its routing table that describes the state of it's own links. This information is then used to construct an overview of the entire network topology and calculate routes to each node. Distance vector algorithms are based on every router sending its entire routing table to its neighbours. Nodes compute the shortest path and next hop towards a destination. Each node then compares the routes its neighbours has to its own routes, and updates its routing table if a more efficient route is found.

The unicast protocols are divided into proactive or table-drive routing protocols and reactive or on-demand routing protocols, determined by the manner in which routing information is obtained and maintained. Hybrid protocols and protocols with special adaptations have also been developed in order to take advantage of technological developments as well try to overcome problems unique to ad hoc networking. These include location aided routing protocols, routing based on network stability, routing with a virtual backbone and the multi-scope routing protocols which include clustering and hierarchical algorithms. Attempts have also been made to develop protocols to maintain a certain degree of Quality of Service (QoS) since most of the routing protocols for ad hoc networks only consider shortest path as the main routing metric.

2.2.1. Proactive Protocols

The proactive routing protocols are also referred to as table-driven protocols. This is because they attempt to maintain complete routes from each node to every other node in the network by maintaining one or more tables to record the route information. In order to maintain a consistent network view, updates of topological changes are continuously propagated throughout the network to all nodes. Information is therefore immediately available when a route is required. The earliest protocols that were proposed for routing in ad hoc networks were proactive distance vector protocols based on the Distributed Bellman Ford (DBF) algorithm [Bertsekas92]. However, the DBF algorithm suffers from poor convergence and high control traffic overhead. The DBF algorithm was therefore modified to overcome these problems which resulted in protocols such as the Destination Sequenced Distance Vector (DSDV) algorithm [Perkins94]. Link state protocols were also applied to address the convergence problem resulting in protocols such as the Optimized Link State Routing Protocol (OLSR) [Jaquet98]. The Wireless Routing Protocol (WRP) [Murthy95] is an example of a path finding algorithm with combines the features of the distance vector and link state approaches. The path finding algorithms attempt to reduce the amount of control traffic, reduce the possibility of temporary routing loops, and avoid the "counting-to-infinity" problem.

2.2.1.1. DSDV

In the Destination-Sequenced Distance-Vector (DSDV) routing protocol [Perkins94] each node maintains two tables. The routing table contains a list of all possible destinations in the network, the number of hops to each destination, the next hop and the route sequence number. Routing table updates are periodically transmitted throughout the network by each of the nodes. Nodes also transmit routing updates if they become aware of significant changes in the topology. Routing table updates are therefore both time-driven and event-driven. The route updates can be one of two types to reduce the potentially large amount of network traffic that can be generated. The "full dump" results in all available routing information from a node being transmitted to its neighbours and is performed infrequently when topological changes are occurring slowly. In an "incremental update", only information that has changed since the last full dump is transmitted. The second table that is maintained stores the data received in the incremental routing information packets.

DSDV improves the DBF algorithm by including sequence numbers in the routing updates that eliminate route looping by identifying stale routes. The sequence number for a route is assigned by the destination node and is increased by one for every new route found. The route with the highest sequence number is used. In the event that two routes have the same sequence number, the shorter route is used.

When an intermediate node along a path detects a broken route to a destination node it advertises its route to the destination with an infinite hop-count and the sequence number is increased by one. New route broadcasts contain the address of the destination, number of hops to reach the destination, sequence number of the routing update and a new identification number unique to the broadcast. Based on the past history, nodes also keep track of the weighted average time that routes to a destination fluctuate before the shortest route is received. This is called the settling time and is used to determine for how long the broadcast of a routing update should be delayed so that network traffic is reduced in case a better route is found very soon.

2.2.1.2. WRP

The Wireless Routing Protocol (WRP) is also a table based routing protocol and each node needs to maintain four tables [Murthy95]:

- Distance Table
- Routing Table
- Link Cost Table
- Message Retransmission List (MRL)

The Distance table is a matrix that contains the distance in hops of each destination node in the network via each of the neighbours of the node and also the predecessor node, which is the last node along the route before the destination.

The Routing table has an entry for each destination specifying the successor and predecessor of the shortest path chosen, the destination's identifier and distance (number of hops), and a tag to specify whether an entry is valid or not.

The Link Cost table lists the cost of relaying information through each neighbour of the node and the number of periodic update periods that have elapsed since any error-free messages have been received from the neighbour. The authors do not specify how links should be assigned costs [Murthy95].

The MRL contains information to let a node know which of its neighbours has not acknowledged its update message. After a broadcast, acknowledgments are expected from all neighbour nodes. If some acknowledgments are missing, the broadcast will be repeated, with the MRL specifying the subset of neighbours that need to respond. If a node is not sending updates, a “hello” message must be transmitted within a specified time interval to ensure connectivity.

An update message is only sent between neighbouring nodes. Nodes only send updates after processing updates from neighbours or detecting a change in a link to a neighbour. On receiving an update message the node modifies its distance table and looks for better paths using new information. Any new path so found is relayed back to the original nodes so that they can update their tables. A unique feature of this algorithm is that it checks the consistency of the predecessor nodes for each destination with all of its neighbours every time it detects a change in link of any of its neighbours. Consistency checking in this way eliminates looping situations and provides faster route convergence when a link failure event occurs.

2.2.1.3. OLSR

Optimized Link State Routing (OLSR) [Jacquet98] is a link-state protocol where the link information is disseminated through an efficient flooding technique. The key concept in OSLR is the multipoint relay (MPR). The MPR set of a node is a subset of its neighbours whose combined radio range covers all nodes within a distance of two hops from the node. MPRs minimize the overhead of flooding messages in the network by reducing duplicate retransmissions in a network region. Figure 2-1 demonstrates the concept of MPRs for an arbitrary node A in an ad hoc network. The neighbours of a node that are not in its MPR set receive and process broadcast messages but do not retransmit the broadcast messages. Only those neighbours in the node's MPR set participate in forwarding the message. Each node obtains the two-hop topology through its neighbours' periodic broadcasting of “hello” packets containing the neighbours' lists of neighbours. The smaller the MPR set is, the more optimal is the routing protocol [Jacquet98]. Furthermore, a node only originates link updates concerning those links between itself and the nodes in its MPR set. Routes are therefore computed using a node's partial view of the network topology.

Each node also maintains information about another set of its neighbours. This is the set of neighbours, called the MPR selector set, which have selected the node as an MPR. A node also obtains this information from the periodic "hello" messages received from the neighbours. Besides being used to reduce broadcast transmissions, OLSR relies only on MPR nodes to perform as intermediate nodes in the path between a source and a destination. All routes therefore consist of a sequence of hops through the MPRs from source to the destination.

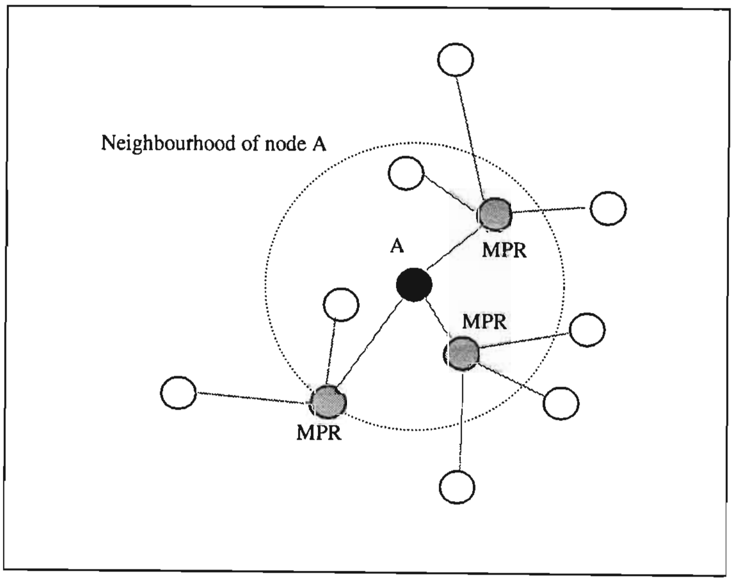

Figure 2-1. Multipoint relays for an arbitrary node A

2.2.1.4. Discussion on Proactive Protocols

The major problem with all the proactive protocols is the increased network traffic overhead due to constant updating of the tables. While DSDV does allow incremental updates, the periodic updates regardless of the number of changes in the network topology still results in poor efficiency. Besides increases in congestion, power consumption is increased because most of the mobile nodes use batteries and other portable power sources. Conservation of power in "sleep" or "standby" mode is also prevented. The "full dump" in DSDV is required when topological changes are frequent, which means that DSDV will perform poorly when there is high mobility. OLSR has less traffic overhead than DSDV because of the use of the MPR set. However, there is also the requirement of periodic updates with OLSR. WRP and OLSR have lower time complexity than DSDV in terms of convergence during link failures since only neighbouring nodes are informed about link status changes. Another problem is the required storage capacity at the node for the tables that are maintained especially with WRP that needs to maintain four tables. With an increase in network size there will be a substantial increase in the memory required for the storage of the tables.

2.2.2. Reactive Protocols

In 1995, Johnson [Johnson95] proposed that in ad hoc networks, instead of having periodic routing updates transmitted throughout the network, a node requiring a route to a destination should use some type of route discovery mechanism to find a suitable route when required. As long as conditions remain unchanged the route should continue to work for as long as it is needed. If conditions change, the node needs to reinitiate a route discovery or modify the route. This effectively divides the routing procedure into a route discovery phase and a route maintenance phase. Routing protocols of this type are referred to as reactive protocols. The procedure to find a route to a destination is often some type of flooding mechanism when a request is made for a route. Reactive protocols are therefore also referred to as on-demand routing protocols.

Examples of reactive protocols include the Dynamic Source Routing (DSR) protocol [Johnson96], Ad hoc On Demand Distance Vector (AODV) routing protocol [Perkins99] and Temporally Ordered Routing Algorithm (TORA) [Park97]. Other reactive routing protocols will be discussed in later sections where their special modifications are emphasised.

2.2.2.1. DSR

DSR uses source routing. Source routing is a technique in which the sender of a packet determines the complete sequence of nodes through which to forward the packet [Johnson96]. The route is explicitly listed in the packet's header identifying each hop to the destination. Each mobile node maintains a route cache of routes that it has used or is aware of. When a source node requires a route to a destination node it first consults the cache for any routes that have not yet expired. If there are no previous routes, the source node broadcasts a route request packet to all its neighbours. The route request contains the addresses of the source and destination and a unique identification for the request. The identification number is used by neighbouring nodes receiving the packet to decide whether to retransmit the packet or to discard it if they have received the route request previously. Each node that receives the route request checks its own cache to determine whether it has any valid routes to the required destination node. If it cannot provide a valid route, it appends its address to the route request and rebroadcasts it, provided it has not received the same route request before.

A route reply is generated when the route request is received by the intended destination node or when an intermediate node replies to the source node due to it being aware of a valid route to the destination. The route request contains a route record yielding the sequence of hops taken. If the destination node is transmitting the route reply, it places the route record from the route request packet in the route reply. If an intermediate node is transmitting the route reply it appends its cached route to the route record in the route request. The return path is the route that has been recorded in the route request or a route that the destination node or intermediate node has cached to get back to the source node. If there is no cached route and bi-directional links are not supported, the destination node or intermediate node can initiate a new route request to determine a route back to the source.

Route error packets are generated when it is found by a lack of acknowledgements or when transmission problems are encountered, that a link has changed due to node movement. When a node receives a route error packet, the hop in error is removed from the node's route cache and all routes containing the hop are truncated at that point.

DSR has proposed modifications for optimisation such as a "promiscuous mode", in which a node listens to route requests, route replies and route errors not intended for itself and updates its route cache correspondingly. Another DSR feature is the expanding ring search procedure, in which the route request packets are first sent with a limited hop count, which can be increased if the destination is not found within the hop count limit.

2.2.2.2. AODV

Ad Hoc On-Demand Distance Vector (AODV) routing [Perkins99] incorporates the destination sequence number technique of DSDV into an on-demand protocol while minimizing the number of required broadcasts by creating routes on a demand basis. Each node keeps a next-hop routing table containing the destinations to which it currently has a route. A route expires if it is not used or reactivated for a threshold amount of time. Like DSR, AODV initiates a route discovery process by broadcasting a route request (RREQ) to neighbours when a route is required. The neighbours forward the RREQ to their neighbours until the destination receives the RREQ or an intermediate node with a valid route receives the RREQ. Destination sequence numbers are utilized to ensure all routes are loop free and contain the most recent route information. RREQs contain the source's own sequence number and unique RREQ broadcast identification number together with the latest known sequence number for the destination. A RREQ is therefore uniquely identified. If the same RREQ is received again, the node discards it. An intermediate node can only reply to the route request if the intermediate node's recorded sequence number for the destination is greater than or equal to that contained in the RREQ. Figure 2-2 demonstrates the route discovery procedure consisting of the RREQ and route reply (RREP) phases.

While the RREQ is being forwarded through the network, each IN records in its routing table the address of the neighbour from which the RREQ was first received and the source address. A next hop reverse path is thereby formed and the intermediate node knows which neighbour to transmit to in order to get back to the source. The RREP from the destination or an intermediate node that has a route to the destination follows the reverse path back to the source. Upon receiving the RREP packet, each intermediate node updates its next-hop table entries with respect to the destination node, dropping the redundant RREP packets and those RREP packets with a lower destination sequence number than one previously seen. The forward path is thus formed where each intermediate node receiving the RREP knows the next hop to get to the destination. When an intermediate node discovers a broken link in an active route, it broadcasts a route error (RERR) packet to its neighbours, which in turn propagate the RERR packet along the reverse path towards all nodes that have an active route using the broken link. The affected source node can then re-initiate route discovery if the route is still needed.

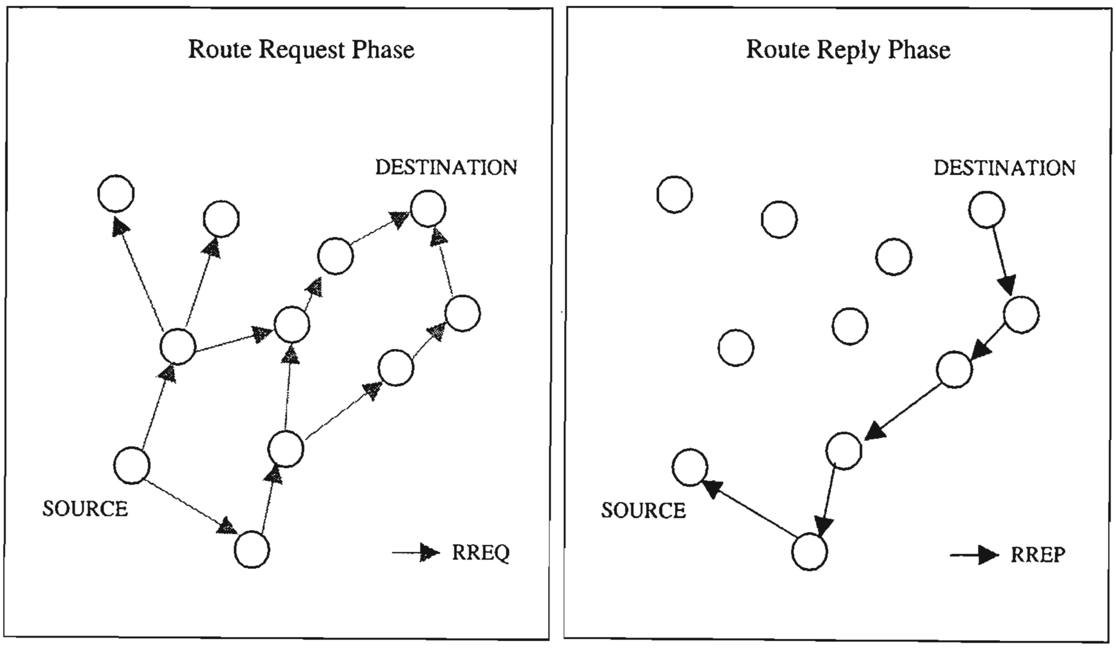

Figure 2-2. The route discovery procedure in AODV.

2.2.2.3. TORA

In the Temporally Ordered Routing Algorithm (TORA) [Park97] routes to a destination are defined by a Directional Acyclic Graph (DAG) rooted at the destination. The DAG is formed during the route creation phase using a "height" metric. Each link in the network is assumed to be bi-directional, but in order to form the DAG with respect to a destination a logical direction of the link is defined based on the relative height metric of neighbouring nodes. The destination is assigned a height metric of zero. If a source has no route to a destination, it broadcasts a route query packet (QRY), which is propagated outwards by its neighbours. After receiving the QRY, a node that has a route to the destination broadcasts a route update packet (UPD) containing its own height relative to the destination. Each node receiving the UPD sets its own height metric for the required destination one higher than the node from which it received the UPD. This results in a series of directed links from the source to the destination in order of decreasing height. As a consequence of this multiple routes are often present for a given destination, but none of them are necessarily the shortest route. The analogy can be made between the path taken from source to destination and a stream flowing downhill.

When a node discovers that a link is broken, it sets its height metric higher than that of its neighbours, and issues a UPD to that effect reversing the direction of the link between them. Links are reversed to reflect the change in adapting to the new reference level. If a node finds that it has no downstream neighbours, the destination is presumed lost and a clear (CLR) packet is issued to remove invalid links from the rest of the network.

Timing is an important factor for TORA because the "height" metric is dependent on the logical time of a link failure and TORA therefore assumes that all nodes have synchronized clocks. Figure 2-3 shows the height metrics for the destination after a QRY has been broadcast by the source and the destination has replied with a UPD. Figure 2-4 shows the re-establishment of a route after failure of a link. The assignment (x, y) in Figure 2-3 and Figure 2-4 to each node gives the sequence number x (based on time) for the height metric and the height metric y for the required destination.

When multiple sets of coordinating nodes are concurrently detecting partitions, erasing routes, and building new routes, there is a potential for oscillations to occur in TORA. This instability is similar to the "count-to-infinity" problem in distance vector routing protocols except that such oscillations are temporary and route convergence will ultimately occur. This can however lead to potentially lengthy delays while waiting for the new routes to be determined. TORA's reliance on synchronized clocks is also a drawback because besides being more costly to implement hardware such as global positioning systems (GPS), it introduces a single point of failure if the time source becomes unavailable.

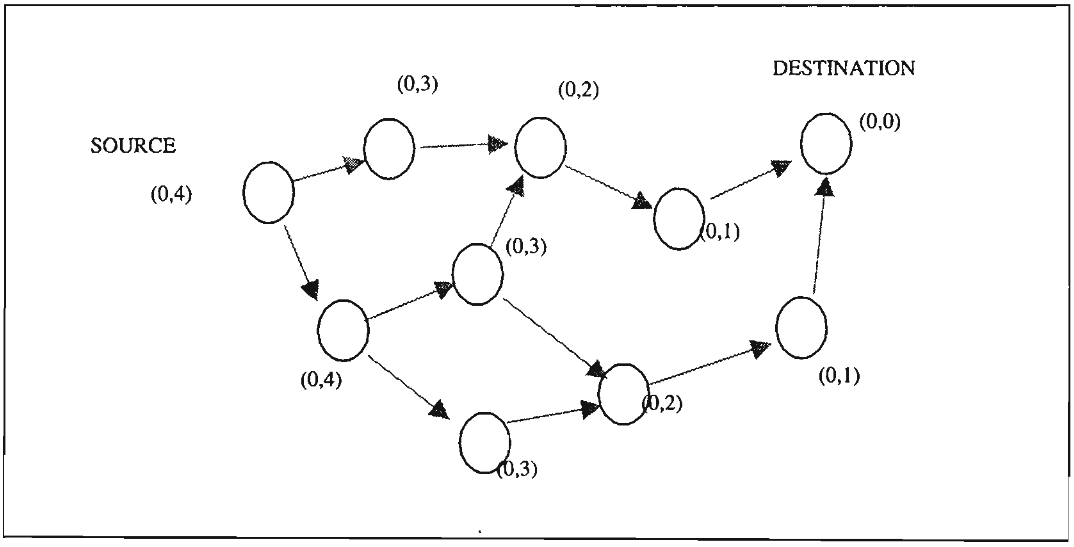

Figure 2-3. TORA height metrics for destination shown after UPD transmitted.

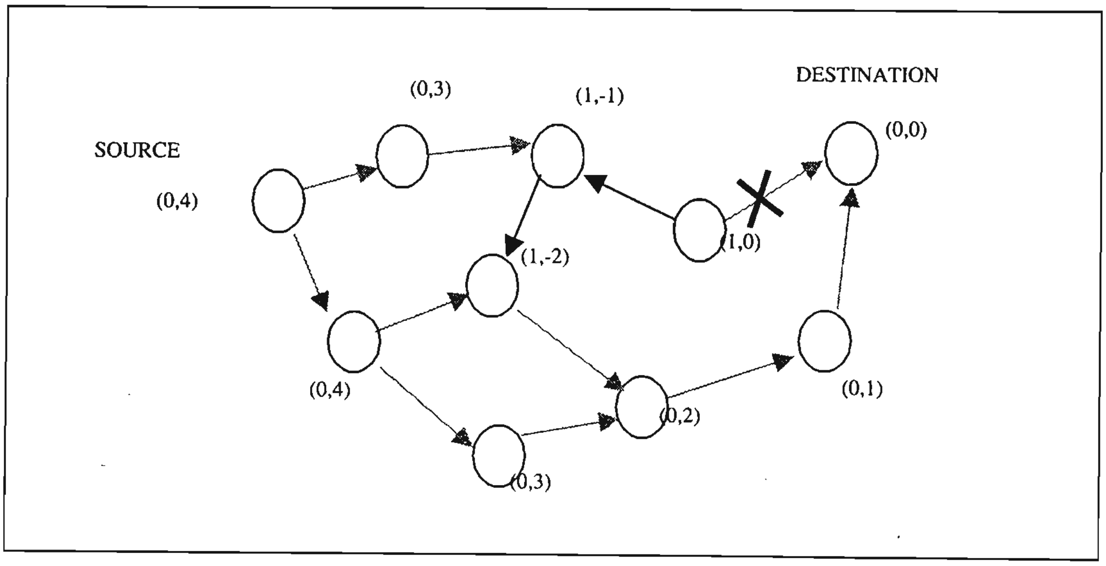

Figure 2-4. Network reaction to link failure in TORA.

2.2.2.4. Discussion on Reactive Protocols

A reactive routing protocol attempts to discover a route to a destination only when it is presented with a packet for forwarding to that destination. This increases the latency of delivering the packet since discovery must be completed before the packet can be sent. Also, without additional information, a protocol using on-demand routing must search the entire network for a node to which it must send packets. Discovering a new route therefore remains a costly operation. The use of some sort of route cache can be used to avoid the need to re-discover each routing decision for each individual packet. However, the cache itself may contain out-of-date information indicating that links exist between nodes that are no longer within wireless transmission range of each other. This stale data represents a liability that may degrade performance rather than improve it.

Various comparative simulations of AODV and DSR have been published in the literature. Broch et al [Broch98] at Carnegie Melon University (CMU) developed ad hoc networking extensions for the ns2 network simulator [Ns02] to compare DSDV, TORA, DSR and AODV. Another simulation was scenario-based [Johansson99] which simulated DSDV, AODV and DSR in a conference setting, event coverage situation and a disaster area. Das et al [Das01] compared AODV to DSR.

[Das01] concluded that in terms of throughput and delay DSR outperforms AODV when average node speed and load is low while AODV outperforms DSR with widening performance gaps with high load and higher node speeds. However, it was found that DSR consistently generates less routing overhead than AODV. This is because AODV made more frequent route requests. The aggressive caching used in DSR and the inability to expire stale routes were recognized as DSR's weaknesses. [Johannson99] also found that DSR and AODV perform relatively similarly with similar conclusions to [Das01]. [Broch98] found that while DSR and AODV performed relatively similarly, at higher node speeds the routing overhead of AODV is more expensive. Both [Broch98] and [Johannson99] found that DSDV had difficulties in maintaining valid routes when the average node speed was increased due to DSDV being a proactive protocol. [Johannson99] concluded that the reactive protocols (AODV and DSR) were superior to the proactive protocol (DSDV) in all event scenarios simulated because of the failure of DSDV to convergence when mobility was introduced.

2.2.3. Routing Based on Network Stability

Associativity Based Routing (ABR) [Toh96], [Toh99] and Signal Stability Based Adaptive Routing (SSR) [Dube97] are two reactive protocols that have unique metrics for determining which route is optimal based on notions of network stability.

2.2.3.1. ABR

ABR is an on-demand routing protocol that uses a unique metric known as the "degree of association stability". Each node maintains an associativity table that contains "ticks" for every other node in the network. The nodes periodically generate beacons which when received by a neighbouring node causes the neighbouring node to update its associativity table with respect to the beaconing node by incrementing the number of ticks associated with the beaconing node. A node is considered more stable if there are many ticks associated with it. Nodes are therefore more likely to use routes through neighbours that have displayed a high degree of association stability. A low degree of association stability indicates a high state of node mobility while a high degree of association stability indicates a low state of node mobility. The associativity ticks for a neighbouring node are reset when the neighbouring node moves out of proximity. The objective of ABR is to utilise stable routes.

ABR consists of three phases: route discovery, route reconstruction (RRC) and route maintenance. When a node requires a route to a destination, a broadcast query (BQ) is broadcast to its neighbours. When a node receives a BQ, it appends to the BQ its ticks associated with the neighbour it has just received the BQ from, together with its address and current load information. In this way, each BQ arriving at the destination will contain associativity ticks of the nodes along the route to the destination on which the BQ has travelled. The destination then selects the best route by examining each of the BQs. When multiple paths have the same degree of association stability, the route with the minimum number of hops is selected. A REPLY is then transmitted back to the source along the best route selected. Each node along the route that receives the REPLY marks their routes as valid.

RRC consists of a localized search for a valid route when a link along the chosen route fails. If the source node moves, a broadcast query and reply (BQ-REPLY) cycle is reinitiated. If the destination node moves the predecessor on the route attempts a localized query (LQ). The LQ determines whether the destination is still reachable via a minimum number of hops. The hop count is incremented if the destination is still not found and the process backtracks to the next upstream neighbour, who again initiates a LQ. If the destination is reached the destination transmits a REPLY to the source using the new partial route and the remainder of the original route. If the destination node is not reached, the process continues backtracking upstream until it reaches the node that was originally halfway along the route. This node then informs the source of the error and the source reinitiates a BQ-REPLY cycle. The movement of intermediate nodes invokes the same backtracking process.

A route delete (RD) is propagated by a full broadcast when the source node no longer requires a route that has been discovered. A direct unicast is not used because the source may not be aware of any changes that have been made during RRC.

Although the route selected is not necessarily the shortest, it will tend to be longer lived resulting in fewer RRCs. Also, since only the selected route is marked as valid, ABR is free of packet duplicates being transmitted through the network. However, periodic beaconing is a potential problem and may result in additional power consumption. In [Gerla99], ABR is compared to the Distributed Bellman-Ford (DBF) algorithm and DSR. Both ABR and DSR had significantly lower control message overhead (up to 76.56% less control message overhead) than DBF, which is a proactive protocol. It was found that when the node speed was low, ABR had more control message overhead than DSR because of the beaconing but as speed was increased, ABR became more efficient due to ABR's local route recovery feature. DSR propagates the route error message all the way back to the source. It was also found that ABR has higher throughput than DSR because when the average node speed in the network was increased DSR uses shortest hop routes, while ABR uses routes that are more likely to last longer. DBF performs poorly because more event-triggered updates are generated when mobility increases. The delay results showed that ABR performed better in this regard as well. This can be attributed to the fact that beaconing provides a method of faster convergence when finding suitable routes. It is admitted [Gerla99] that power consumption is a problem with ABR, especially since nodes use portable power supplies. Another possible problem is the interception of beacons by unintended receivers, such as the enemy in a battlefield scenario, which would reveal a nodes location.

2.2.3.2. SSR

Signal Stability Based Adaptive Routing [Dube97] is another on-demand routing protocol that selects routes based on network stability metrics. In SSR a route is selected based on the signal strength between nodes and a node's location stability. Periodic beacons are broadcast which allow neighbouring nodes to determine signal quality, which is recorded as either "weak" or "strong" in a Signal Stability Table (SST). A route search process is initiated if a valid route is not contained in the Routing Table (RT). Route requests are broadcast throughout the network but are only forwarded to the next hop if they are received over strong channels and have not been processed before. Once the destination receives a route request it ignores later copies that have taken different routes because it is assumed that the first route request arriving has arrived over the shortest path with the least congestion. The destination then replies using the same route along which the first route request arrived because it is assumed to be the route with the best signal quality. Route requests are dropped if the channel was recorded as weak. If the source does not receive a route reply within a specified timeout period, the source specifies in the header of the next route request that weaker links may be used since they may be the only links available.

Erase messages are sent by the source node once it is informed of failed links by an intermediate node along a route. The erase message is sent to notify all nodes of the broken link. The source then reinitiates a route-search process.

As in ABR, the routes selected by the SSR algorithm are more stable and longer-lived resulting in fewer route reconstructions, while not necessarily being the shortest. However no attempt is made to make partial route recovery as in ABR with the result that the source has to reinitiate the route search. Conversely, the "backtracking" method of localized route reconstruction in ABR may not reduce delay as much as is expected, compared to the source immediately initiating a route request when a broken link is found. Longer delays may also result in the route search process because unlike AODV and DSR, intermediate nodes are not allowed to respond with a route reply.

2.2.4. Location Based Routing Protocols

Mobile nodes can be designed to obtain physical location information using Global Positioning System (GPS) receivers. While these receivers may be relatively expensive, they will become cheaper with advancements in technology. The gains in using them for location information with routing in ad hoc networks could be significant due their being able to potentially reduce routing related overheads [Ko97]. The selective paging scheme in Personal Communication Service (PCS) networks has a purpose similar to utilizing location information with ad hoc networks. When a host needs to be located using selective paging, the network pages only a selected subset of the cells close to the last reported location. Tracking cost is thereby decreased.

This subsection describes two routing protocols that use location information in the routing process. Location Aided Routing (LAR) [Ko98] is an on-demand routing protocol. In LAR a source node estimates the range of a destination node's location based on the destination node's last reported velocity, and broadcasts a route request only to nodes within a geographically defined request zone. In Distance Routing Effect Algorithm for Mobility (DREAM) [Basagni98] a node periodically transmits its location coordinates to other nodes in the network. DREAM is therefore considered a proactive protocol. The period of location transmission depends on the node's velocity and the geographic distance to nodes for which the location information is intended.

2.2.4.1. LAR

The basic idea behind Location Aided Routing (LAR) [Ko98] is that route discovery cost and routing related overhead can be decreased by using a limited search space for the destination. The limited search space is called a request zone and only nodes in the request zone are allowed to forward a route request message to their neighbours.

Part of the request zone is the "expected zone". The expected zone is a circular area in which the destination node is expected to be, determined by the last position and expected average velocity of the destination node that the source node is aware of. If the source node does not have any previous information about the destination node then the route request procedure becomes a flooding process as in the previously mentioned on demand routing protocols. There are two different approaches to LAR that are proposed: LAR scheme 1 and LAR scheme 2, referred to here as LAR1 and LAR2 respectively. The two schemes differ in the way in which the request zone is defined.

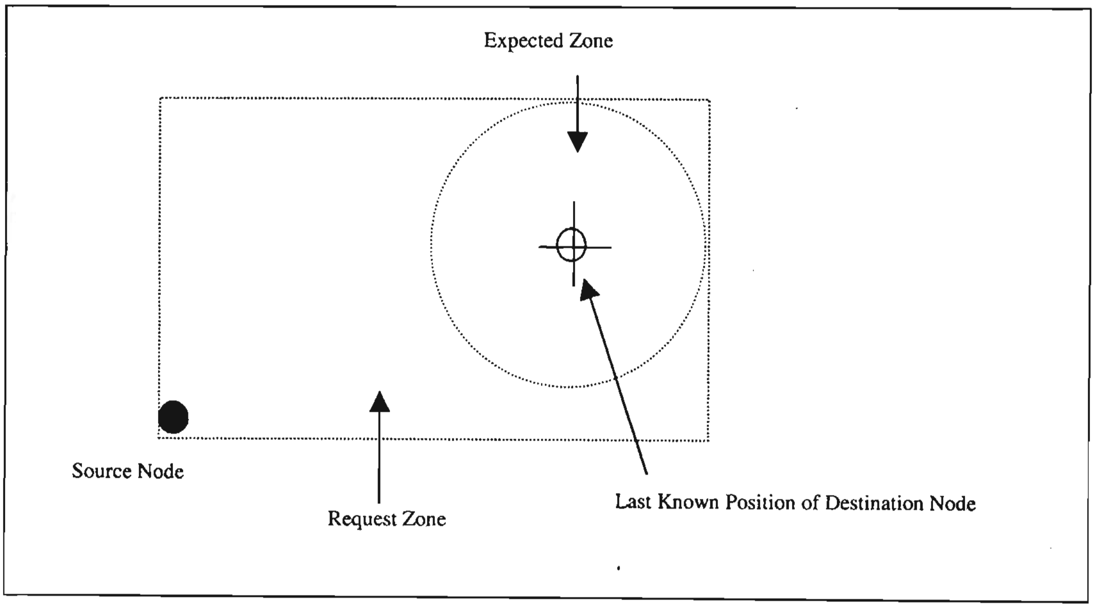

Figure 2-5. Request Zone and Expected Zone for LAR Scheme 1

LAR1 uses a rectangular request zone that is the smallest rectangle containing both the source and the expected zone, as shown in Figure 2-5. The source node determines the coordinates of the four corners of the rectangle and transmits this with the route request. A node which is not in the request zone that receives the route request discards the route request. When the destination node receives the route request it replies using the path along which the route request first arrived and includes its location information for future use by the source node.

In LAR2 the source node uses its own location information to calculate the distance D_s between itself and the last known location of the destination node, (x_d, y_d). The calculated distance D_s and (x_d, y_d) are included in the route request. When a neighbouring node receives the route request, it uses the information in the route request to determine if it is closer to (x_d, y_d) than the source node is. If it is closer then it forwards the route request to its neighbours but replaces D_s with its own distance from (x_d, y_d). If it is not closer than the source node to (x_d, y_d) then it discards the route request. In this way the route request is only forwarded by nodes which are closer to (x_d, y_d) than the neighbouring node from which the route request was received.

The problem with LAR is that it has to have prior location information, otherwise its route request procedure simplifies to the original on-demand protocols. But LAR relies on the fact that although nodes may not initially be aware of the locations of other nodes, as time progresses, each node can get location information by its own route discovery or by "eavesdropping" on packets that are being forwarded. There is also the suggestion [Ko98] that location information may be propagated by piggybacking it on any message. LAR also allows for localized queries after link breakages so that the node that determines a link failure is allowed to initiate a local search using its own request zone. Another optimisation that is recommended [Ko98] is to allow the request zone to be dynamically modified by nodes with more recent location information about the destination node when they receive a route request.

2.2.4.2. DREAM

The Distance Routing Effect Algorithm for Mobility (DREAM) [Basagni98] is a proactive routing protocol. Each mobile node transmits a location packet (LP) to all other nodes in the network. The transmission period of the LPs is determined by the mobility of a node. The faster a node moves, the more often it must communicate its location. Dream also uses "distance effect" which is the fact that the greater the distance separating two nodes, the slower they appear to be moving with respect to each other. Nodes that are far apart from each other therefore need to update each other's locations less frequently than nodes closer together. Each LP has a certain "lifetime" in terms of geographical distance which determines how far an LP travels before being discarded. Faster moving nodes transmit shorter lived LPs more frequently while slow moving nodes transmit longer lived LPs less frequently. By differentiating between nearby and faraway nodes, DREAM attempts to limit the overhead of LPs.

LPs are used to update the location tables of the receiving nodes and contain the source identification, coordinates of the source, and the time the LP originated. When a source node needs to transmit data to a destination node, it calculates a circle around the most recent location information for the destination nodes, much like the expected zone in LAR. The source then defines the forwarding zone to be a cone whose vertex is at the source node and whose sides are tangent to the circle calculated for the expected location of the destination node. The authors of DREAM [Basagni98] use a minimum cone angle of 30° in their simulation. The source node then forwards the data packet to the neighbours in the forwarding zone. This is unlike LAR and other on-demand routing protocols where a route request is forwarded first. The neighbours receiving the data packet then compute their own forwarding zones based on their own location tables and forward the data packet accordingly. When the destination receives the data packet an acknowledgement (ACK) is returned to the source along the reverse of the path that was used for the data. If the source does not receive the ACK within a timeout period, the source resorts to the DREAM recovery procedure, which involves flooding the data packet through the network. A destination node receiving a flooded data packet does not return an ACK. DREAM also defines a timeout value on location information. A source node resorts to flooding if the location information is older than the specified limit.

2.2.4.3. Comparison of LAR and DREAM

[Camp02] presents a simulation comparison of DSR, LAR, DREAM and a pure flooding algorithm. With zero average node speed the data packet delivery ratio of the DSR and LAR protocols is 100 % while that of DREAM and flooding is approximately 68% since there is much congestion in the network with the flooding algorithm and DREAM. When average speed is increased, LAR performs better than DSR. This is because LAR's use of location information to find a new route is more efficient than DSR's route discovery method. The promiscuous mode of DSR was found to significantly aid nodes in finding routes. DREAM had the highest average end-to-end delay of all protocols simulated. This is because the DREAM recovery procedure is used approximately 40 % of the time at low mobility and almost all the time at high speeds since the ACK is not received within the timeout. [Camp02] concluded that the added location capability of DREAM did not provide benefits over simple flooding algorithms. The location information used in the LAR protocols however was found to be significantly more beneficial than using the route request procedure of DSR. The LAR protocols were found to be considerably more efficient than DREAM in terms of packet delivery ratio and throughput.

2.2.5. Routing with a Backbone

Sivakumar et al [Siva98] describe a self organizing network structure called a spine which functions as a virtual backbone to facilitate routing in ad hoc networks. The spine is chosen to be a small and relatively stable sub-network of the ad hoc network whose primary role is to compute and maintain routes as opposed to carrying data packets. Every node in the network is either in the spine or is a neighbour of a node in the spine. The spine nodes maintain local copies of the global topology of the network and collectively compute routes between any pair of nodes in the network. They are thus able to minimize the access overhead for routing information. Besides computing routes and tracking topology changes, the spine nodes also provide temporary backup routes for fault tolerance. The spine of an ad hoc network is depicted in Figure 2-6.

The spine is obtained by using an approximation to the minimum connected dominating set (MCDS) [Siva97] of the ad hoc network topology. The MCDS of an ad hoc network is depicted in Figure 2-6. A dominating set in a network is the subset of nodes such that every node in the network is either in the dominating set or is a neighbour of one of the nodes in the dominating set. A connected dominating set is a dominating set consisting of nodes which are able to communicate with each other without needing to use nodes that are not elements of the dominating set.

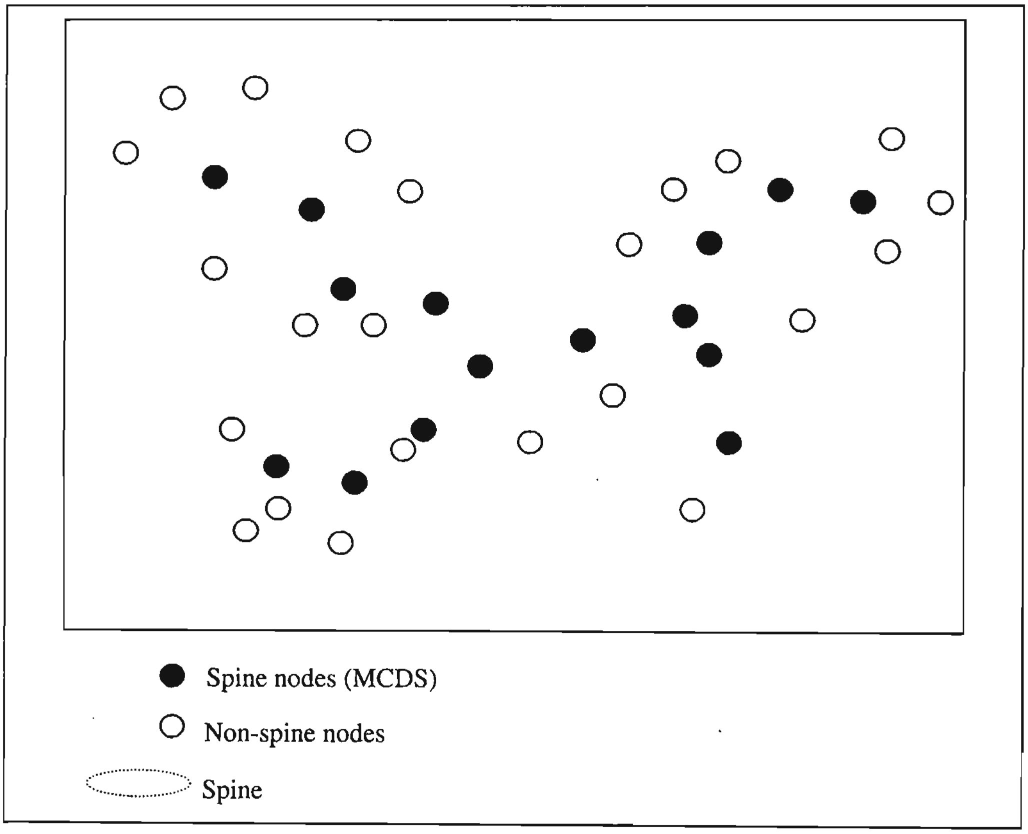

Figure 2-6. Depiction of the MCDS and spine in spine routing

The MCDS approximation consists of an initialisation phase where nodes flood messages to their neighbours informing their neighbours of the number of neighbours that they have. It is assumed that nodes initially are aware of their own neighbours. The nodes are then able to determine who has the most neighbours and which node should form part of the dominating set by "marking" nodes with the most neighbours. For example, if two neighbouring nodes A and B find that node A has more neighbours than B then A also becomes the dominating node of node B. The procedure is repeated until all nodes are either in the dominating set or are at least connected to one of the nodes in the dominating set. The second phase involves another flooding process which connects the nodes of the dominating set and this finally results in a single spine that runs through the network.

Once the spine has been formed, spine nodes gather topology information from nodes that they dominate and the spine nodes transmit this link information to all other nodes on the spine. Using this global knowledge of the topology, the spine nodes determine the shortest paths for all node pairs. When a source node requires a route to a destination, the source node requests a route from its dominator in the spine. Route maintenance by the spine nodes are event-driven where the spine nodes update routes when nodes move and time-driven with topology updates being periodic.

Core-Extraction Distributed Ad Hoc Routing (CEDAR) algorithm [Siva99] was introduced in an attempt to provide Quality of Service (QoS) routing in ad hoc networks. CEDAR is based on the original spine routing algorithm using MCDS but bandwidth information is also included in the information stored at each dominator. When a source seeks a destination with a required bandwidth, a core-path is first established from the dominator node of the source to the dominator node of the destination. The core-path is then used to provide the direction in which to iteratively find a set of partial routes using only local information each of which satisfies the bandwidth requirement. Together the partial routes are expected to form a single QoS admissible route.

The main problem with the spine routing is that with node movement the structure of the spine needs to change often. This means that it will be common for a new node to be added to the spine or for a spine node to move into a position where it finds a dominator and no longer needs to be a spine node. Changes in the spine structure mean that the topological information needs to be updated on all nodes along the spine. To overcome this problem, Partial Knowledge Spine Routing (PSR) [Siva98] was introduced, which only allows spine nodes to maintain local information of the nodes for which they are the dominator. Routing is then achieved by using on-demand methods. However, the spine routing infrastructure may still be unable to cope with highly mobile networks. Another problem is that nodes forming part of the spine will need to utilize their resources (processing time and power) more than other nodes in the network.

2.2.6. Multi-scope and Hybrid Routing Protocols

The multi-scope routing protocols distinguish nodes by their relative positions. More resources are devoted to maintaining the topology information of nearby nodes rather than maintaining a global view of the network. Scalability is therefore the main advantage of the multi-scope routing protocols. The multi-scope routing protocols can be divided into flat and hierarchical routing protocols. The flat multi-scope routing protocols include DREAM (discussed in section 2.2.4.2), the Zone Routing Protocol (ZRP) [Haas99] and Fisheye State Routing (FSR) [Iwata99]. The hierarchical routing protocols include Cluster-Head Gateway Switch Routing (CGSR) [Chiang97], Hierarchical State Routing (HSR) [Iwata99] and Zone-based Hierarchical Link State (ZHLS) [Joa99] routing.

2.2.6.1. ZRP

The Zone Routing Protocol (ZRP) [Haas99] provides a hybrid routing framework that is locally proactive and globally reactive. Each node proactively advertises its link state only to its routing zone, which includes all nodes within a certain number of hops away from the node. The zone radius is adjustable and provides an optimisation parameter. The local advertisements give each node an updated view of its own routing zone. The Intrazone Routing Protocol (IARP) is responsible for maintaining routes within each node's zone through periodic routing table updates. [Haas99] states that the type of proactive routing used in the IARP (distance vector or link state) has very little effect on the performance of ZRP. Figure 2-7 illustrates the concept of a zone in ZRP with zone radius of 2 hops.

The neighbouring nodes that are exactly the number of hops away from the node as the zone radius dictates, are referred to as peripheral nodes. The peripheral nodes represent the boundary of the routing zone and play an important role in zone based route discovery. The Interzone Routing Protocol (IERP) uses on-demand routing to communicate between zones. ZRP uses knowledge of routing zone connectivity to guide its global route discovery. Rather than blindly broadcasting route queries from a node to all its neighbours, ZRP employs a service called bordercasting, which directs the route request from a node to its peripheral nodes. Peripheral nodes that have been covered by the route query and that belong to the routing zone of a node that has already bordercast the query ignore the bordercast thus preventing further unnecessary bordercasts. This encourages the query to propagate outward, away from its source.

Multiple hop paths within the routing zone can bypass link failures. Similarly, sub optimal route segments can be identified and traffic can be re-routed along shorter paths. The difficulty in implementing ZRP is determining the optimum zone radius.

The zone radius affects the overhead traffic since if IARP is used more frequently then the negative effects of proactive protocols are experienced, while if IERP is used more frequently then the negative effects of reactive protocols need to be endured. The general rule of thumb [Haas99] is that a sparse network favours a large routing zone while a dense network favours a small routing zone. Figure 2-8 illustrates the ZRP zone routing radius optimisation.

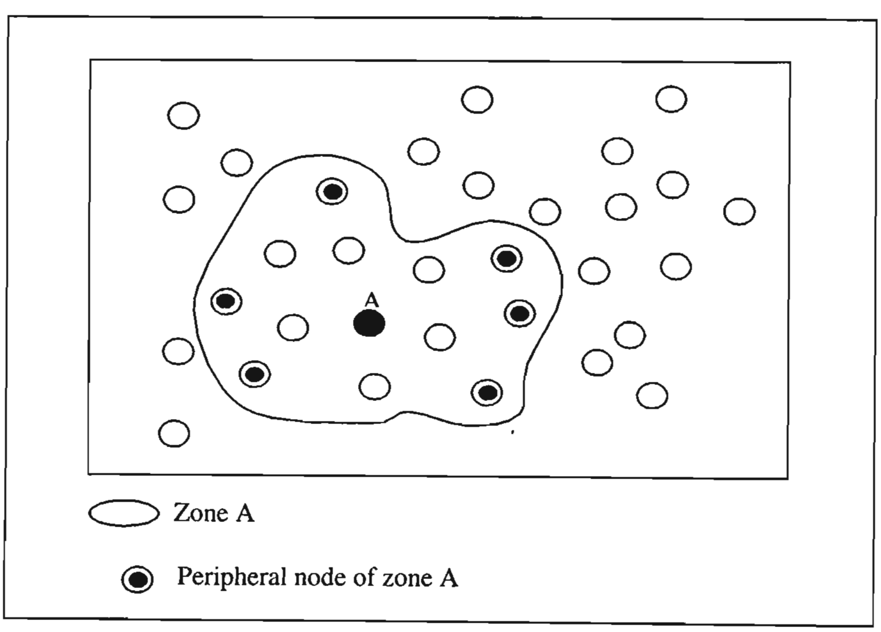

Figure 2-7. Concept of a zone in ZRP. (zone radius = 2 hops)

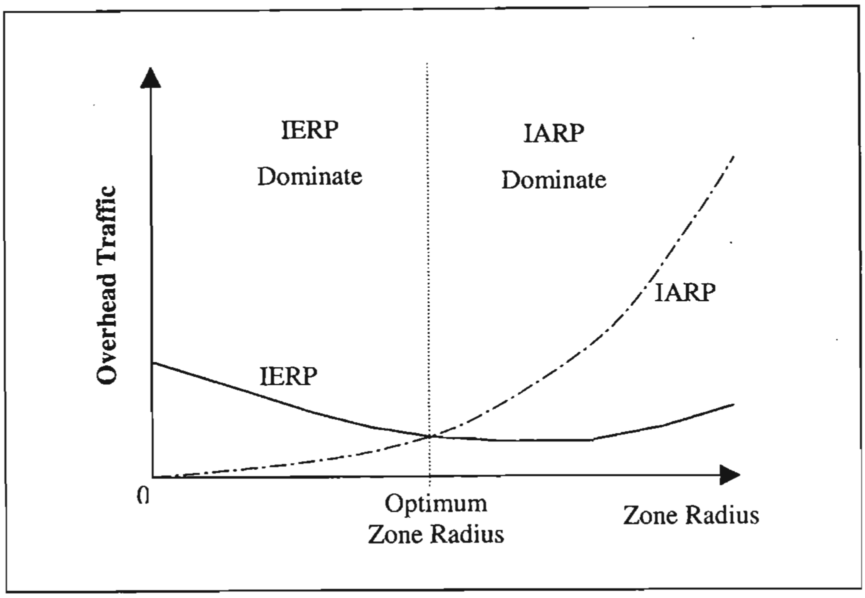

Figure 2-8. The optimum region for the zone radius resides between the IERP and IARP dominated regions [Haas99].

2.2.6.2. FSR

The Fisheye State Routing (FSR) protocol [Iwata99] models the routing methodology on the way in which the eye of a fish functions. The eye of a fish captures with high detail the pixels near the focal point and the detail decreases as the distance from the focal point increases. Figure 2-9 illustrates the concept of Fisheye State Routing.

The aim of FSR is to reduce routing update overhead in large networks. Nodes maintain a link state table based on the up-to-date information received from neighbouring nodes and periodically exchange it with their local neighbours only, which prevents flooding. Table entries with larger sequence numbers replace the ones with smaller sequence numbers. The circles with different shades of grey in Figure 2-9 define the fisheye scopes with respect to the centre node. Three scopes are shown for 1 hop, 2 hops and hops greater than 2, respectively. Different exchange periods are used for different entries in the routing table. Entries corresponding to nodes within the smallest scope are propagated to the neighbours most frequently.

The imprecise knowledge of the best path to a distance destination is compensated by the fact the route becomes progressively more accurate as the packet gets closer to the destination.

FSR is based on Global State Routing (GSR) [Gerla98]. GSR can be considered a special case of FSR where there is only one fisheye scope level. Although information is still only exchanged between direct neighbours, the overhead is high in GSR because the entire topology table is exchanged among neighbours. Unlike FSR, GSR does not distinguish between nodes based on their relative distance away from the node in question. FSR was designed to be able to scale to large networks, but avoid on-demand techniques while keeping overhead low and still provide adequate routes which become more accurate closer to the destination.

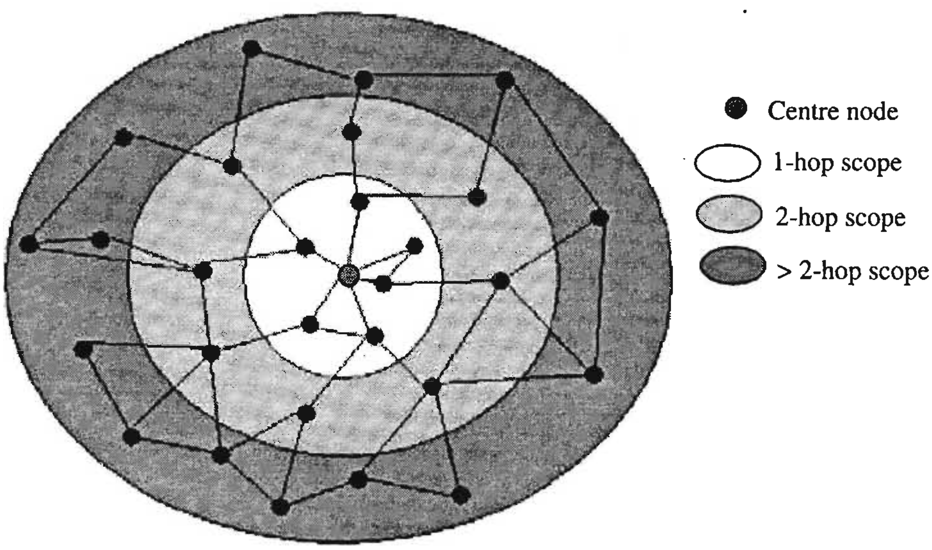

Figure 2-9. Fisheye scope in FSR.

2.2.6.3. CGSR

Cluster-head Gateway Switch Routing (CGSR) [Chiang97] aggregates nodes into clusters. Each cluster is controlled by a cluster-head and adjacent clusters communicate via gateway nodes, which are members of two or more clusters. A node is always within transmission range of the cluster-head of its cluster. The cluster-head selection process is determined by the Least-Cluster-head Change

(LCC) algorithm [Chiang97]. In LCC only two conditions allow a cluster-head to change : two cluster-heads come into contact with each other or a node moves out of range of all cluster-heads and becomes a cluster-head itself. When two cluster-heads come into contact with each other, the node that continues being the cluster-head is the node with the lower ID or the node with the greater number of internal nodes in its cluster, dependent on the choice of the network designer.

In CGSR, the DSDV protocol is modified so that it can take an advantage of the clustering architecture. Each node maintains two tables: a cluster member table which maps the destination node address to its cluster-head address, and a routing table which shows the next hop to reach the destination cluster. Both tables contain sequence numbers to purge stale routes and prevent looping. In CGSR, packets are routed alternatively through cluster-heads and gateways. In other words, the typical route looks like $C_1G_1C_2G_2$...CiGi, where Ci is a cluster-head and Gi is a gateway node.

The advantage of CGSR is that only the routes to the cluster-heads are maintained due to the hierarchical routing. Also, power control, code scheduling, and medium access control (MAC) can be more easily implemented. However, there is overhead associated with maintaining clusters. Each node needs to periodically broadcast its cluster member table and update its table based on the received updates. The cluster-heads and gateway nodes usually do more work than the ordinary nodes and so not all nodes have the potential to be cluster-heads or gateway nodes.

2.2.6.4. HSR

The Hierarchical State Routing (HSR) protocol [Iwata99] uses logical partitioning and physical clustering to maintain a hierarchical topology. At the physical level, nodes form clusters and a cluster head is selected. The elected cluster heads of the lower levels become members of the next higher level cluster, as shown in Figure 2-10. The authors [Iwata99] do not specify which clustering algorithm to use for the formation of clusters and the election of cluster heads.

There are three kinds of nodes in a cluster: cluster heads, gateways and internal nodes. The cluster head maintains topological information of the nodes in the cluster and acts as a local coordinator of transmissions within the cluster. The cluster head also summarizes link state information within its cluster and propagates it to the neighbouring cluster heads via the gateway nodes. The knowledge of connectivity between neighbour cluster heads leads to the formation of the next higher cluster level. A virtual link is formed between two cluster heads using the internal nodes and gateways nodes between the cluster heads. Nodes within a cluster exchange virtual link state information as well as summarized lower level cluster information.

The Hierarchical Identification (HID) is used to identify each node and consists of a sequence of MAC addresses of the nodes from the top hierarchy to the node itself. For example, in Figure 2-10, the HID of node 10 is <3, 3, 10>. For the delivery of a packet from node 5, with HID(5) = <1, 5, 5> to node 10 with HID(10) = <3, 3, 10>, the packet is first forward to node 1 (top hierarchy) which uses the virtual link (1,6,2,8,3) to deliver the packet to node 3. Node 3 then delivers the packet to node 10 along the downwards hierarchical path.

The drawback of HSR is that nodes have to maintain longer hierarchical addresses and have to continuously update the cluster hierarchy and hierarchical address as nodes move. Since a continuously changing hierarchical address makes it difficult to locate and keep track of nodes, logical partitioning is used in HSR. Logical partitioning involves using subnets of the network, each with its own home agent to manage membership. The subnets correspond to a particular user group, for example, students from the same class or tanks in the same battalion. Each member of a logical subnet knows the HID of its home agent from the routing table and registers its own HID with the home agent. The registration with the home agent is both event driven and time driven. It is assumed [Iwata99] that since members of the same subnet would move as a group, registration overhead is modest because the members will tend to reside in neighbouring clusters. When a source node requires a route to a destination it uses the logical address of the destination to send the packet to the home agent of the destination. The home agent redirects the packet to the destination, which can then continue communication using the newly acquired and current HID of the source. The source also then becomes aware of the current HID of the destination.

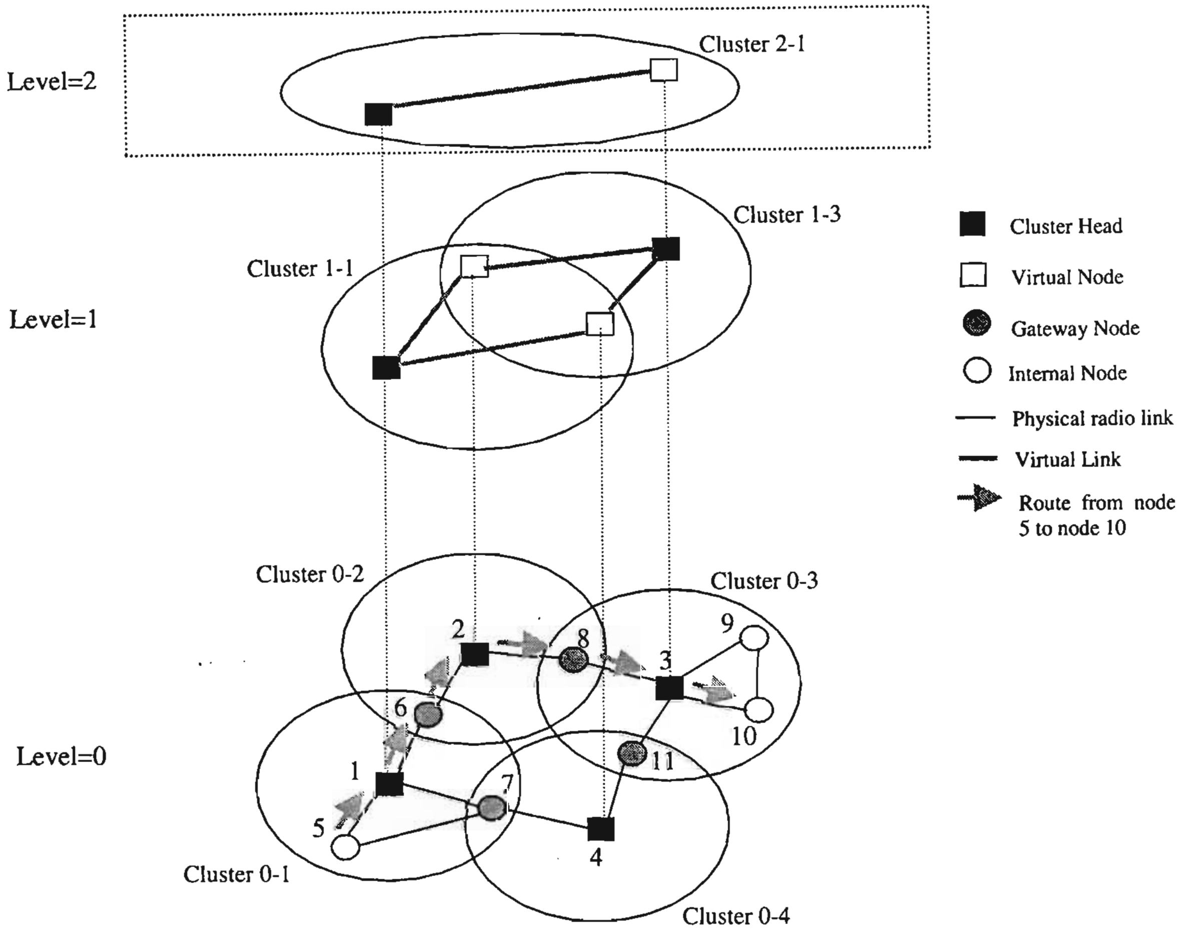

Figure 2-10. Clustering and hierarchical structuring in HSR [Iwata99].

Although the authors [Iwata99] suggest the logical partitioning as a means of avoiding overhead caused by the hierarchical clustering, the process of members of a subnet communicating continuously with the home agent and the process of a packet being sent all the way to the home agent to be redirected to the destination still appears to consume more resources than would be necessary in a flat routing protocol, especially in a highly mobile environment. Note that the home agent subnet approach is similar to that used in Mobile IP [Perkins98], except that in Mobile IP the home agent is not mobile.

2.2.6.5. ZHLS

The Zone-based Hierarchical Link State (ZHLS) routing protocol [Joa99] divides the network into non-overlapping zones. A node determines the zone it is in using GPS and by mapping its physical location onto a predefined zone map. Unlike other hierarchical and clustering routing algorithms, ZHLS does not implement a cluster-head for each zone thus overcoming the problem of higher computation and communication burden that cluster-heads experience in HSR and CGSR.

Each node in a zone broadcasts a link request to its neighbours to determine the node level topology within the zone and to determine which of its neighbouring nodes are in other zones. The link state routing is performed on two levels: node level and zone level. This results in two types of link state packets (LSP) that are propagated. The node LSP of a particular node contains a list of its connected neighbours and is propagated locally within its zone. The zone LSP contains a list of a zone's neighbouring zones and is propagated globally throughout the network by gateway nodes. Gateway nodes are those nodes that can communicate with nodes in other zones due to being within transmission range. Each node therefore knows the path to each node in its own zone and the path to each of the other zones.

Figure 2-11 shows an example of the node level and zone topology obtained by the nodes in zone A. When a source requires a route to a destination not in its zone, a request is sent to each of the other zones. The gateway node for the zone that the destination is in, replies to the request. The source can then initiate communications using only the node ID and the zone ID.

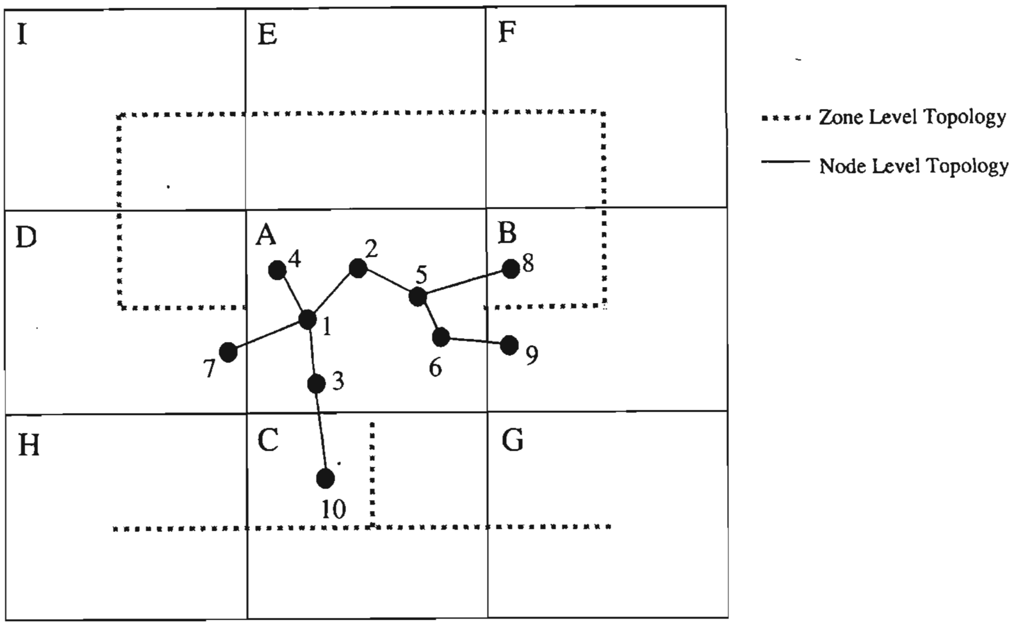

Figure 2-11. Network topology obtained by nodes in zone A using ZHLS.

2.2.7. QoS Routing Protocols

Quality of Service (QoS) provision is commonly accepted to be the ability of the network to guarantee pre-specified service attributes to the user in terms of delay, delay variance, availability of bandwidth and probability of packet loss, among others. QoS routing for ad hoc networks is complicated because the network cannot guarantee that mobile nodes will remain in the same position for the duration of the communication. But in a wireless network, mobility is not the only reason QoS guarantees are difficult. The nature of the wireless medium also makes maintaining the precise link information very difficult. The research into QoS routing for ad hoc networks has therefore moved towards an adaptive soft QoS approach instead of the hard QoS guarantee of wired networks. Soft QoS implies that there may be certain periods of time when breaks in links or network partitions result in the required QoS not being guaranteed.

CEDAR has already been described in section 2.2.5 as a proposed QoS routing algorithm which uses a spine or core of nodes in the network to establish and maintain routes. Another QoS routing strategy for ad hoc networks is ticket-based probing [Chen99]. Nodes maintain local state information about the delay (channel delay, queuing delay, processing delay) of each outgoing link, residual bandwidth on each link, and the cost of the link, which is lower for links that have been in existence for longer. The basic idea behind ticket-based probing [Chen99] is to limit the number of candidate paths and avoid flooding by using probe messages with a limited number of so called tickets. Each ticket corresponds to one path searching so the maximum number of searched paths is bounded by the tickets issued from the source. When an intermediate node receives a probe message it decides how to split the tickets and where to forward the probe depending on its local state information. When the destination node receives a probe message, a possible path from the source to the destination is found.

More tickets are issued for the connections requiring higher requirements and the intermediate nodes assign more tickets to the links with larger residual bandwidth that satisfy the QoS requirements. Probes carrying “yellow” tickets prefer paths with smaller delays and more residual bandwidth in order to satisfy the given requirements. Probes with “green” tickets attempt to maximize the probability of finding a low-cost path, which may or may not have larger delays and the required bandwidth. Algorithms are presented [Chen99] to determine the optimum number of each type of ticket to assign to a probe depending on the QoS requirements. In general, the lower the constraints, the lower the number of tickets required.

The dynamic nature of ad hoc networks makes it difficult to provide QoS. The unpredictable way in which the network changes coupled with the fact that the state information is usually imprecise results in high overheads in attempting to provide QoS and QoS provision can be impossible [Chen99] if the ad hoc network changes too fast. The proposed QoS routing protocols for ad hoc networks are therefore based on the assumption that the nodes are very slow moving and topology changes occur infrequently enough for the routing protocol to be successful.

2.3. Multicast Routing Protocols

Multicasting can be beneficial in many of the applications for ad hoc networks such as group video conferencing and file distribution. Multicasting allows a single stream of data to be distributed to a large number of recipients without network congestion since the data is transmitted once and duplicated only when necessary. Multiple unicast transmissions are unable to achieve the same results in such situations because the same data must be repetitively sent to each destination node independently.

In order to multicast, multicast groups are first defined. A node must first join a multicast group in order for it to receive a multicast message. There are two key approaches in designing multicast routing for ad hoc networks: mesh-based protocols and tree-based protocols. The tree-based protocols are so named because a tree in graph theory is a path that has no cycles and the tree-based protocols use this structure. The majority of multicast routing protocols in the Internet are based on shortest path trees, because of their ease of implementation. Also, they provide minimum delay from sender to receiver, which is desirable for most real-life multicast applications. However, tree-based multicasting in ad hoc networks is fragile and requires frequent control to maintain the tree. Multicasting in ad hoc networks is also more challenging than in the Internet because of the need to optimise the use of several resources simultaneously.

Mesh-based protocols use a set of nodes that are in charge of forwarding multicast packets. The mesh provides richer connectivity among multicast members compared to tree-based protocols but this results in high control traffic because the number of nodes to forward multicast packets is larger than tree-based multicasting.

The multicast routing protocols for ad hoc networks include the On Demand Multicast Routing Protocol (ODMPR) [Chiang99], Ad Hoc Multicast Routing with Increasing ID-numberS (AMRIS) [Wu99], Core-Assisted Mesh Protocol (CAMP) [Madruga99], Ad Hoc Multicast Routing (AMRoute) [Talpade98], Multicast AODV

[Royer99], Neighbor Supporting Ad Hoc Multicast Routing Protocol [Lee00], and Multicast Core Extraction Distributed Ad Hoc Routing (MCEDAR) [Sinha99].

Geocasting is a variation of multicasting and is the problem of sending a message to all nodes located within a particular region. The multicast group is implicitly defined as the set of nodes within a specified area. Geocast routing protocols which use flooding and adaptations of the GPS based unicast routing protocols are described in [Navas97], [Ko99] and [Basagni99].

Multicasting in ad hoc networks is difficult because after sufficient information about the topology of the network is obtained and a tree or mesh is computed, there may be very little time before the calculated mesh or tree becomes inadequate.

2.4. Summary

One of the areas of active research in ad hoc networks is in the design of routing protocols. The mobility of the nodes in the network results in dynamic and unpredictable topologies. The routing protocols for ad hoc networks need to determine viable routes from the source node to the destination node without compromising the data, but at the same time need to be able to adjust to changes in the topology.

The unicast protocols can be divided into proactive routing protocols and reactive routing protocols. The proactive protocols are modifications of the routing protocols used for wired networks and maintain updated routing tables. The reactive protocols search for a route only when one is needed by a source, and usually implement some type of flooding mechanism in order to find a current route to the destination. Comparative simulations of the routing protocols have shown that the reactive protocols are better equipped to contend with the constantly changing ad hoc network environment. The proactive routing protocols have been found to add to the congestion of the network due to the constant updates of the routing tables.

Modifications have been made to the originally proposed routing protocols for ad hoc networks to incorporate various enhancements such as using signal and location stability of nodes as metrics for routing. Other enhancements include incorporation of the Global Positioning System (GPS) to allow nodes to determine the location of the other nodes in the network using both reactive and proactive mechanisms. The use of virtual backbones has also been attempted to mimic the fixed infrastructure of cellular networks. Hybrid protocols that use proactive routing locally and reactive routing globally as well as clustering and zone-based protocols were also discussed. Figure 2-12 depicts some of the unicast routing protocols proposed in the literature for ad hoc networks. The concept of multicast routing is also briefly described.

Each routing protocol has its advantages and disadvantages. The implementation of the enhancements to the basic routing protocols depends on a trade-off of cost vs. performance. For example, the inclusion of a GPS unit would need to be justified, depending on whether or not the GPS-based routing protocol would significantly improve performance in the application under consideration. The zone-based, hierarchical and clustering protocols have their disadvantages in that while reducing overhead for a majority of the nodes, certain nodes are overburdened and consume more resources than other nodes in order to maintain and organize the network. It therefore becomes difficult to justify the hierarchical and clustering structures when different users are using devices all of the same type, especially where service costs are involved, unless some type of billing system is implemented.

Most of the routing protocols proposed for ad hoc networks use the shortest path as the main routing criterion. In Chapter 3 load balancing is discussed and extensions to the reactive routing protocols are proposed for a routing protocol that uses routing metrics which incorporate both signal quality determination and load balancing.

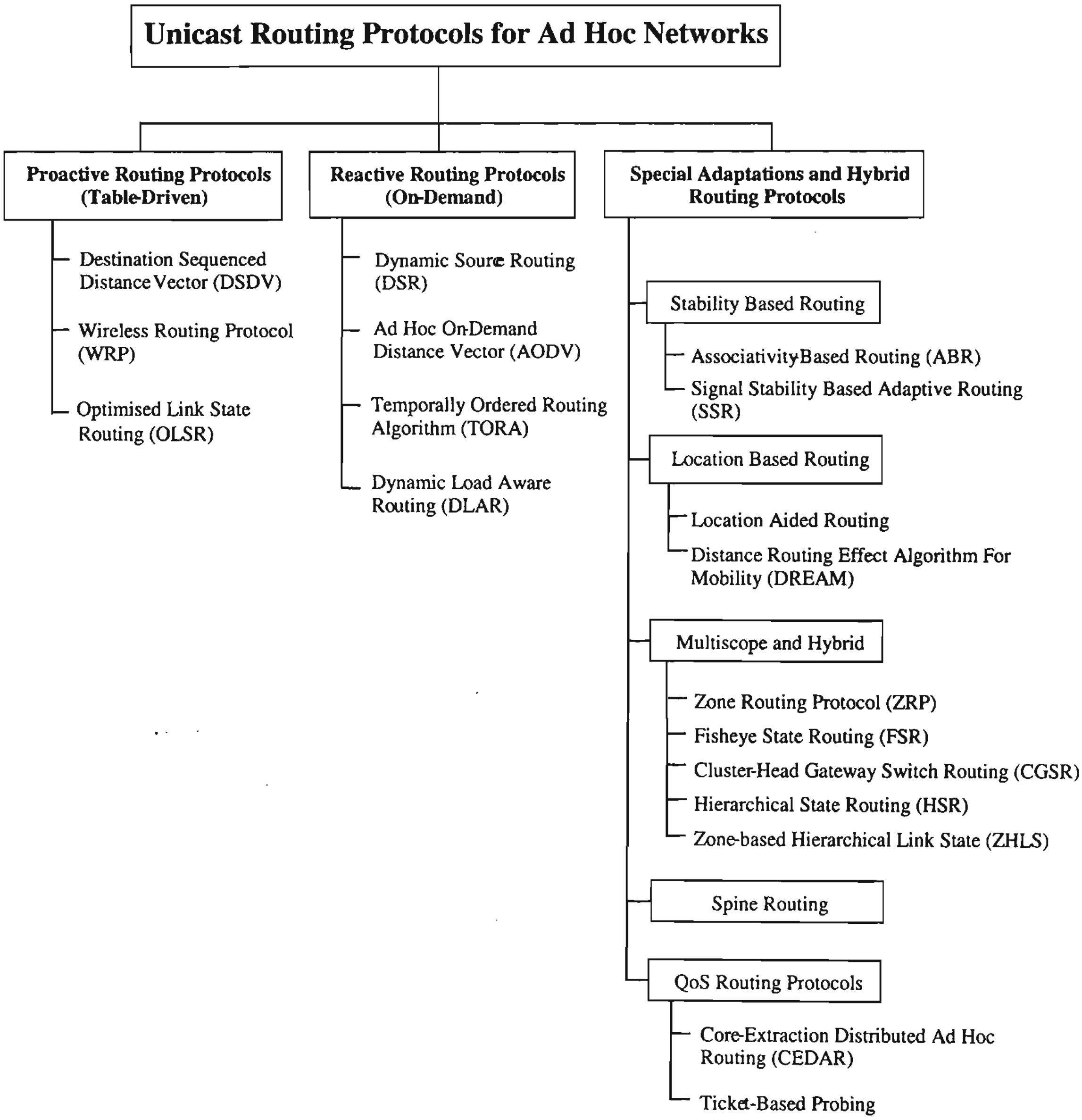

Figure 2-12. Unicast routing protocols for ad hoc networks

Chapter 3

SIMULATION OF A LOAD BALANCING ROUTING PROTOCOL

3.1. Introduction

The results of simulations comparing on-demand routing protocols for ad hoc networks to the proactive table-driven routing protocols [Broch98], [Johannson99], [Camp02] show that the table-driven protocols struggle to maintain valid global state information in ad hoc networks with high mobility. The resulting overhead causes network congestion and increased delay. The on-demand protocols discussed in Chapter 2 also do not account for load balancing to avoid congestion and often use the shortest path or the first path found by a route request procedure. While the shortest path is likely to result in decreased delay, using the shortest path may actually increase delay in situations where traffic load is high. Certain nodes may become overburdened due to their respective locations in the network and as a result the packet buffers overflow, throughput performance deteriorates and delay increases.

With reference to a simple example in Figure 3-1, node 6 becomes overburdened because of its position in the network. The shortest path algorithms would frequently employ node 6 as a node along the shortest route to some other node on the opposite side of node 6. However, congestion at node 6 would eventually result in longer delays. By avoiding node 6 and routing traffic along the outer ring when node 6 is congested, network throughput and delay will be improved. The configuration in Figure 3-1 is purely for demonstrating the problem of load balancing; however mobility and the resulting topologies can always result in similar topologies and even more complicated topologies where this unfavourable condition could arise.

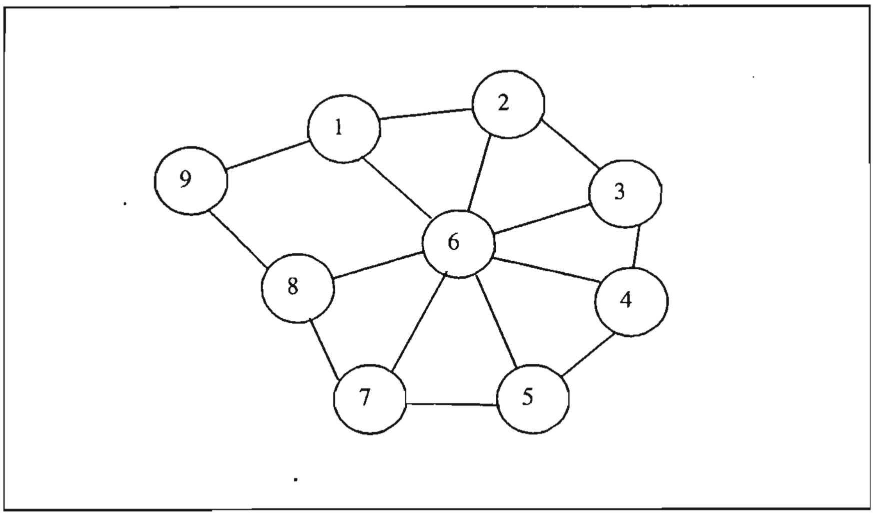

Figure 3-1. Example topology resulting in network congestion due to shortest path routing.

An aim of the research described herein was therefore to evaluate a routing protocol that would perform load balancing in order to ensure that congestion in the network would be avoided. Table-driven routing protocols were not considered suitable because the mobility of the ad hoc network results in unnecessary overhead being created due to table updates. Hierarchical, zone and cluster based schemes were avoided because the aim was to treat all nodes equally and in these types of routing protocols selected nodes tend to use more resources than other nodes when maintaining local information. The aim was also to avoid periodic beaconing to conserve power.

Dynamic Load Aware Routing (DLAR) [Gerla00] is an on-demand routing protocol that takes load balancing into account by determining the state of the buffers of nodes along possible routes. DLAR has three original schemes, each differing in the manner in which the buffer contents are evaluated for route selection. This chapter discusses the modifications made to DLAR to include a fourth scheme that also takes into consideration expected signal quality and the signal to interference ratio (SIR) along the route. Although SSR (described in Section 2.2.3.2) also takes into consideration signal quality, the DLAR implementation proposed herein avoids the periodic beacons of SSR and allows adaptive route selection from multiple possible routes. Dynamic reselection and maintenance of routes occurs during communication so that the route selected minimizes congestion in the network.

Section 3.2 discusses the modifications to the DLAR protocol and the sections that follow describe the simulations performed using a custom built simulator written in C++ to evaluate the DLAR schemes and compare them to AODV (described in Section 2.2.2.2). AODV is a prominent ad hoc network routing protocol that has performed well in previous comparative simulations especially under high traffic loads and high mobility [Broch98], [Johannson99], [Das01]. The simulations are performed in a network area of 1000m x 1000m for 50 mobile nodes. The environment consists of flat terrain with no obstacles to movement or radio propagation, except for the boundaries of the network area. Varying load and mobility conditions are used to evaluate the performance of the protocols in terms of delay and packet delivery ratio. The simulator implements physical layer modelling, the medium access control (MAC) protocol of the IEEE 802.11 Distributed Coordination Function (DCF), and node mobility. The results show that implementing load balancing in routing protocols for ad hoc networks has advantages over the shortest distance method.

3.2. Routing with Load Balancing

3.2.1. Overview of Dynamic Load Aware Routing

DLAR [Gerla00] is an on-demand routing protocol similar to AODV and DSR, both of which are described in Chapter 2. The major difference with DLAR is the use of load information by the destination node from intermediate nodes. The load information is used to determine the route to be used instead of using only the shortest path. A route request procedure is initiated when a source node requires a route to a destination. A sequence number, and the addresses of the source and destination nodes, uniquely identify the route request. Intermediate nodes propagate route requests received by them for the first time and ignore route requests that have been received previously. The intermediate nodes attach their load information and other metrics that can be used to determine optimal routes to the route request packet.

The type of information appended depends on the DLAR scheme that is implemented. The destination accepts multiple route requests within an appropriate time frame, *reply_delay*, after receiving the first route request. This allows the destination to determine which routes are available and the quality of the available routes. The destination chooses the most suitable route and transmits a route reply packet using this route to the source. Once the source receives the route reply, it can begin transmitting the data on the selected route.

While the route is being used, intermediate nodes piggyback load information on the data packets. The destination node uses this information to monitor the status of the route to determine if the route is becoming too congested. The destination can decide to find a new route before the current route fails and broadcasts a route request in the same way that the source initiates a route request procedure. The route request is eventually propagated to the source, which can analyse the load information attached to the route request by intermediate nodes and send the next packet on the most suitable route. In this way routes are dynamically selected during a session in order to perform load balancing and reduce congestion in the network.

Unlike AODV and DSR, DLAR does not allow intermediate nodes to respond to route requests. Only the destination node is allowed to respond to route requests, which prevents stale route information from the caches of intermediate node being used. Preventing intermediate nodes from generating route replies also prevents a flood of route replies from multiple intermediate nodes with cached information and reduces the congestion caused by the reply storm.

Figure 3-2 demonstrates the congestion that is created when intermediate nodes are allowed to respond to route requests with route replies, hence neglecting the potential congestion that can result. Assume that node 2 initiates a route request for a route to node 7 and obtains the route {2, 4, 5, 6, 7}. When node 3 requests a route to node 7, node 4 responds with a cached route resulting in node 3 using the route {3, 4, 5, 6, 7}. This new route overlaps the previous route and the same process occurs when node 1 requests a route to node 6. Node 1 receives the route reply from node 4 indicating that {1, 8, 4, 5, 6} is a suitable route. It can be seen that although allowing intermediate nodes to respond to route requests prevents the further propagation of the route request, congestion can result in the network.

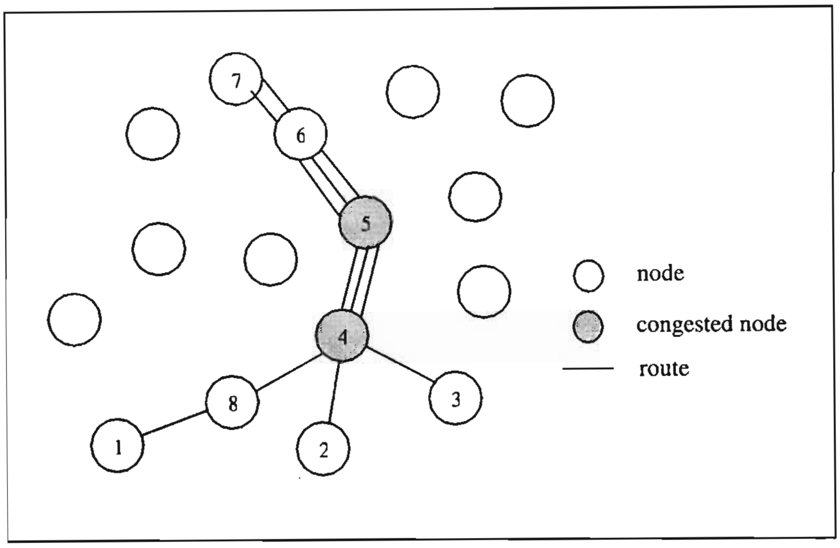

Figure 3-2. Network congestion due to intermediate routes replying to route requests

3.2.2. Original DLAR Schemes

The authors [Gerla00] of DLAR originally suggested 3 schemes that could be used. The three schemes vary in the manner in which the buffers at each intermediate node are evaluated. The three DLAR schemes operate as follows:

a) DLAR scheme 1: The first scheme adds the routing load of each intermediate node on the route and selects the route with the least sum of data packets in the buffers. If two or more routes have the same sum of data packets, then the shorter route is selected. If the shortest routes are the same length then the route whose route request arrived first is the route that is selected as the most appropriate.

b) DLAR scheme 2: The second scheme uses the average number of packets buffered at each intermediate node. The route with the lowest average number of buffered packets is selected as the best route to use. The tiebreaker used is also the length of the route and then the time of arrival of the route request.

c) DLAR scheme 3: For the third scheme, a threshold value τ is used in order to determine when an intermediate node can be considered congested. The route selected is the route with the least number of intermediate nodes with load higher than the threshold value. Similar tiebreakers as that in scheme 1 and scheme 2 are used.

3.2.3. SIR-based DLAR Scheme

The scheme proposed in this dissertation, referred to as DLAR scheme 4, not only takes into account the buffer contents but also the interference experienced by the intermediate nodes due to their relative locations in the network. This new DLAR scheme takes into consideration three factors as measured within a monitoring time frame of T_m seconds. Assuming that node i has just received a route request from node k then the following factors will be considered:

a) S_{ik} is the average SIR of all packets received by node i from node k within the last T_m seconds, including the current route request. This measurement determines the expected strength of the signal on the route from source to destination between intermediate nodes k and i. Assuming U packets have been received from node k within the last T_m seconds, with packet u being received with power P_u and interference I_u,

$$S_{ik} = \frac{\sum_{u=1}^{U}(P_u / I_u)}{U} \tag{3-1}$$

b) C_{ik} is the percentage of successful packet receptions and transmissions by node i, to and from node k within the last T_m seconds without collisions occurring. C_{ik} includes all routing information packets including the current route request.

c) Q_i is the percentage of node i's transmit buffer that is empty.

Node i uses the three metrics to obtain an evaluation value, A_i to attach to the route request before rebroadcasting it. A_i identifies the current capabilities of node i in terms of being a suitable intermediate node of the route and is calculated as follows:

$$A_i = w_{A1}S_{ik} + w_{A2}C_{ik} + w_{A3}Q_i \tag{3-2}$$

where w_{A1} , w_{A2} and w_{A3} are weighting factors. The route request is then rebroadcast with the evaluation information and the process is repeated at each intermediate node.

If there are N nodes in the network then there are R=N(N-1)/2 possible node pairs in the network, each indexed r. Each node pair will have a possible M_r routes between the pairs terminal nodes, listed from best to worst, with the mth route being indexed r_m. The destination node receives multiple route requests representing different routes and sums the intermediate node evaluation values contained in each route request to obtain an evaluation value A_{rm} for the entire route r_m. The route selected is the route that maximizes the cost function

$$w_{m1}A_{rm} + \frac{w_{m2}}{h_{rm}} \tag{3-3}$$

where h_{rm} is the number of intermediate nodes on route r_m and w_{m1} and w_{m2} are weighting factors.

DLAR scheme 4 combines the use of load information with the evaluation of the signal quality being experienced at each node and the length of the route. By monitoring the signal quality and using the routes that have lighter loads, delay is decreased and congestion is alleviated. The reason nodes which have had more collisions in the recent past are avoided is that the queues would generally be longer at those nodes due to the problem of having to retransmit packets.

3.3. Medium Access Control

In order to model the contention of nodes for the wireless medium, the Distributed Coordination Function (DCF) of the IEEE 802.11 Medium Access Control (MAC) [IEEE99] was implemented. Both physical carrier sensing and virtual carrier sensing are used by the DCF. Physical carrier sensing uses Carrier Sense Multiple Access / Collision Avoidance (CSMA/CA). Nodes wanting to transmit first check the channel to ensure that the channel is idle. Once the node has determined that the medium has been idle for a minimum time period, known as the DCF Inter-Frame Spacing (DIFS), it determines a random back-off period by setting an internal timer. When the timer reaches zero, the node may begin transmission. However, if the channel is seized by another node before the timer reaches zero, the timer setting is retained at the decremented value for subsequent transmission.

Instead of using physical carrier sensing only, virtual carrier sensing was implemented in the simulator to reduce the probability of collisions due to hidden terminals. The hidden terminal problem is demonstrated in Figure 3-3. If node A is transmitting to node B, node C may not be in range of the transmission and therefore will not be aware of the transmission from node A. Node C will therefore assume that the medium is free and will begin transmitting to node B. This will result in a collision at node B. Virtual carrier sensing therefore allows nodes to reserve the medium for a specified period of time through the use of Request to Send (RTS) / Clear to Send (CTS) transmissions to avoid the hidden terminal problem. Referring again to Figure 3-3, when node A wants to transmit to node B, it first sends an RTS packet. The RTS packet includes the receiver address and duration required to be reserved. Physical carrier sensing is used before transmitting the RTS. Once the packet is received by node B, node B replies with a CTS that also includes the duration of the reservation by node A contained in the RTS. This allows node C to be informed of the ensuing transmission even though it is not within range of node A.

All correctly received unicast packets are followed by the transmission of an ACK. Broadcast packets however are not preceded by an RTS/CTS combination and are not acknowledged by their recipients but they are only sent when physical carrier sensing indicates that the medium is clear.

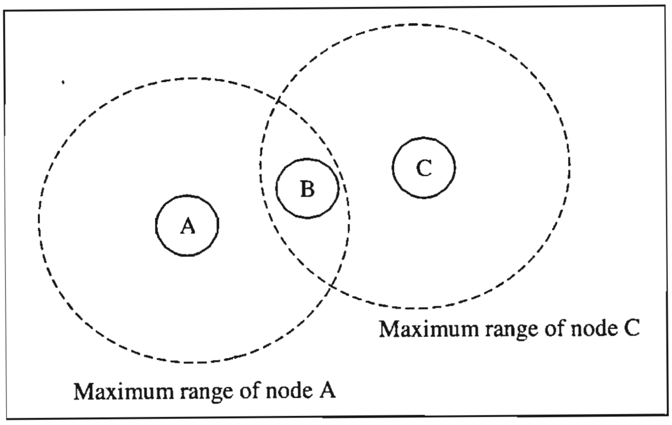

Figure 3-3. RTS/CTS transmissions avoid hidden terminal problem

3.4. Physical Layer Model

In contrast to wired networks, the properties of wireless channels are highly unpredictable and time varying. The propagation of the radio signal is strongly influenced by such factors as the distance between the transmitter and the receiver, the geographical obstructions and terrain over and through which the signal traverses resulting in fading, multi-path distortion due to reflection, refraction and shadowing, and interference from adjacent signals operating in the same frequency band.

An important consideration in radio networks is the distribution of received signal energies as a result of the fact that terminals are spatially separated with different and varying distances. An important phenomenon related to the aforementioned fact is the near-far effect, a process that favours signal reception from transmitting terminals that are closer to a receiving terminal.

Radio propagation is a complex topic and implementing a complete physical layer model is beyond the scope of this work. The physical layer model is therefore limited to supporting propagation delay and attenuation and includes a free space model with a two-ray ground reflection model.

To determine radio propagation, assuming the antennas used are isotropic (radiates equally in all directions), the ratio of the power radiated by the transmit antenna to the power available at the receive antenna is known as the path loss. The minimum loss on any given path occurs between two antennas when there are no intervening obstructions and ground losses. In such a case, when the receive and transmit antennas are isotropic, the path loss is known as free space path loss. The power received (P_r) by an antenna with gain G_r from a transmitter transmitting with power P_t and having a gain G_t, when the antennas are separated at a distance d apart is given by the Friis transmission equation [Jan01]

$$P_r(d) = P_t G_t G_r \left(\frac{\lambda}{4\pi d} \right)^2 \tag{3-4}$$

where λ is the wavelength with the same units as d. Equation 3-1 can be simplified to

$$P_r(d) = K_1 \frac{P_t}{d^2} \tag{3-5}$$

where K_1 is a constant.

The Friis model is valid providing the distance d exceeds the far-field distance or Fraunhofer distance d_f where $d_f = (2D^2)/\lambda$ and D is the largest physical dimension of the antenna. Since the Friis equation does not hold for $d = 0$, a received power reference point, d_{ref} is chosen such that $d > d_{ref} > d_f$. The reference distance for practical systems using low-gain antennas in the 1-2 GHz region is usually chosen to be 1m in an indoor environment and between 100m or 1km in outdoor environments. With reference to Figure 3-4, when the distance d between the receive and transmit antennas is much greater than the product of the height above ground of the receive (h_r) and transmit (h_r) antennas, the low angle of incidence of the radio wave allows the earth to act as a reflector. The reflected signal could be out of phase and the ground ray destructively interferes with the line-of-sight path. The two-ray ground reflection model considers both the direct path and the reflected path, resulting in a higher path loss exponent for the relationship of the received power to the distance between antennas;

$$P_r(d) = P_t G_t G_r \frac{h_t^2 h_r^2}{d^4} \tag{3-6}$$

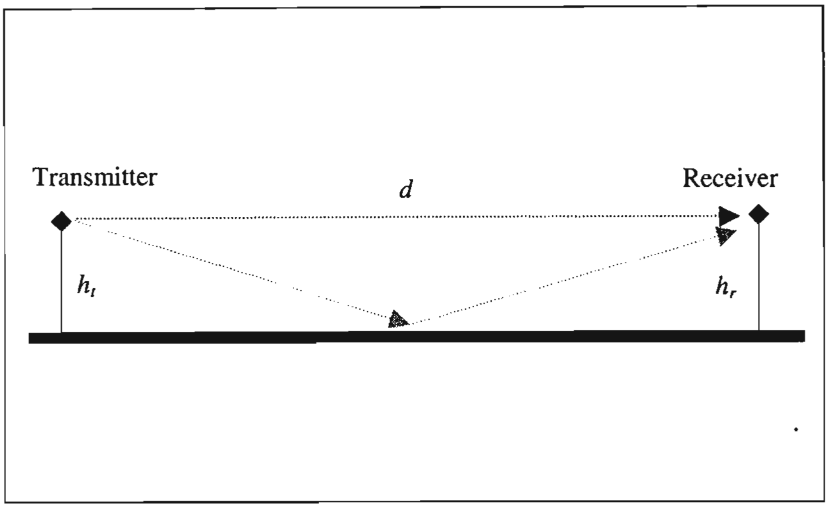

Figure 3-4. Two-ray ground reflection model

Both theoretical and measurement-based propagation models indicate that average large-scale path loss for an arbitrary transmitter-receiver separation is expressed as a function of distance by using a path loss exponent [Jan01]. In order to take these considerations into account, an attenuation model similar to that used in [Broch98] is implemented, where the attenuation is proportional to $1/d^2$ for distances less than a cross-over distance and proportional to $1/d^4$ for distances greater than the cross-over distance. The cross-over distance used in the simulations was 200m, based on antennas transmitting at 2.5GHz at a height of 1m above the ground. All nodes transmit at constant power. Power control was not implemented.

The simulator schedules a packet reception event when the physical channel object receives a packet to transmit. The propagation delays are calculated so that the packet reception event notifies the appropriate network interface for each node when a packet is being received. The power levels are then calculated based on the distance between the receiver and transmitter. If the received power is above the receive threshold, the packet is passed up to the MAC layer. If the MAC layer is idle a "packet reception complete" event is scheduled after computing the transmission time of the packet. If the MAC layer is not idle the new packet is ignored provided the power of the original packet being received is 10dB greater than the new packet. If this is not the case then a collision occurs and both packets are dropped.

3.5. Mobility Model

One of the difficulties in simulating and analysing ad hoc networks to evaluate a new protocol is accurately taking into account the mobility of nodes in the network. The mobility pattern is dependent on the application to which the ad hoc network is applied. Currently network simulations utilize two types of mobility models: traces and synthetic models. Traces are mobility patterns that are derived from real life observations. Traces have been used to derive traffic and mobility prediction models in the study of various problems in cellular systems, such as handoff, location management, paging, registration, calling time and traffic load. However new network scenarios such as those of ad hoc networks are not easily modelled if traces have not yet been created. Synthetic models attempt to represent the behaviour of mobile nodes without the use of traces.

While realistic mobility models are often sought, random models provide sufficiently robust engineering approximations and provide insight into the more general problems. Random models have been used in network design and performance analysis in almost all areas of telecommunication research for achieving analytically tractable results and accurately representing the aggregate behaviour of the convergence of a large number of independent sources. Random mobility with uniformly distributed trajectory is frequently assumed in performance analysis of resource control and mobility management algorithms for cellular or PCS networks [McDonald99]. The majority of mobility models proposed are continuous-time stochastic processes, which characterise the movement of nodes in a two-dimensional space [Sanchez98], [Shukla01], [Zonoozi97]. It is assumed that if there is a large enough population of nodes and high frequency of events occurring, the result is an aggregate effective movement that can be modelled by a suitable random process.

The mobility model implemented in the simulator is the random waypoint model [Broch98]. The random waypoint model [Broch98] has been used by the MANET workgroup of the IETF and was first introduced by the CMU group for the ns2 network simulator.

In the Random Waypoint Model the nodes are initially assigned random positions and then select a randomly distributed destination and a random speed between predefined limits, and travel to the destination at that speed. On arrival the node waits for a "pause time" before repeating the procedure to move to a new destination. [Davies00] found that the network is more stable in terms of link breakages over time in simulations with fast mobile nodes and long pause times as opposed to slow mobile nodes and short pause times. Figure 3-5 is a flowchart of the implementation of the Random Waypoint Model for a mobile node in the simulator. For the simulations the values for *x_max* and *y_max* were both 1000 m and *v_max* was 20 m/s. The *pause_time* used was 0, 25s, 50s and 150s for different simulation runs. By varying the pause times, different mobility scenarios were simulated.

[Royer01] modified the Random Waypoint Mobility model after it was found that density waves were created in the average number of neighbours during a simulation. The average number of neighbours seen at a given node periodically increases and decreases as the simulation progresses, where the frequency of the change is relative to the speed of the nodes. The density waves are as a result of nodes repeatedly moving towards the centre or passing through the centre of the designated network area since statistically the nodes choose a new position from where there are most positions to choose.

The modification resulted in the Random Direction Model to alleviate this behaviour and promote a semi-constant number of neighbours during the simulation. Nodes are requested to choose a direction and speed, instead of destination, until a boundary is reached. The nodes then wait for a certain "pause time" before choosing another direction and speed. But since nodes move towards the boundaries and pause at the boundaries, the average hop count between node pairs is higher than the average hop count in simulations using other mobility models. Network partitions also occur more frequently [Davies00].

The simulation results discussed in this chapter are based on simulations using the original Random Waypoint Mobility Model, since it was not necessary to ensure that nodes had a constant number of neighbours and the periodic increase and decrease in the average number of neighbours could actually help to demonstrate the robustness of the routing protocol.

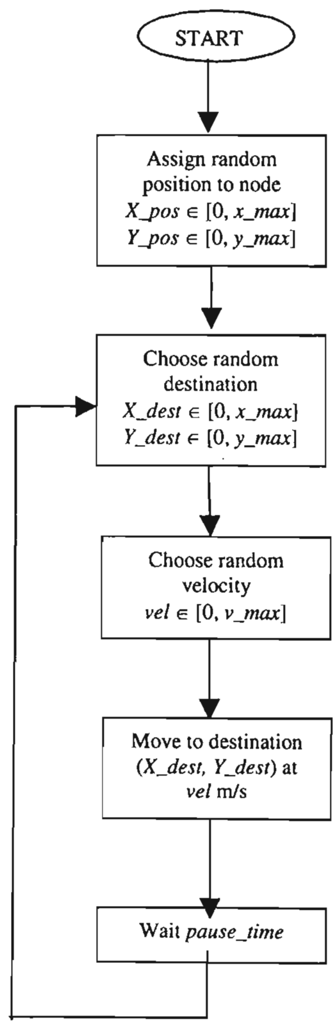

Figure 3-5. Flowchart for Random Waypoint Mobility model for a single node

3.6. Traffic Model

Two packet buffers are maintained at each node, each being *buffer_length* packets long. One buffer contains data packets and the other routing information packets (route requests, route replies and route error packets). Routing information packets have the higher priority. Packets are dropped if they are in a queue for longer than *pkt_timeout* seconds.

[Broch98], [Johannson99] and [Das01] used constant bit rate (CBR) sources and varied the number of sources for different simulation runs in order to determine the effects of traffic load. The source nodes selected to be CBR sources were the only nodes throughout the simulation that generated traffic and the associated destination nodes for each source node also never changed. A different approach was implemented for traffic generation for the simulations discussed herein in an attempt to create stochastic behaviour.

Each node is assigned a random Packet Generation Time (P_{GT}), normally distributed between 0 and a maximum packet generation time PGT. After P_{GT} seconds, the node selects a destination node randomly and generates *session_length* packets each of *pkt_length* bytes to be transmitted to the selected node. The packets are added to the appropriate buffer and the node assigns itself a new P_{GT}.

3.7. Simulation Procedure

The custom built simulator was used to compare AODV to the four DLAR schemes in terms of packet delivery ratio and delay. A total of 24 simulations were run for each of the four routing protocols for 3600s of simulation time. By varying the pause time from a low pause time to a high pause time, the network topology was varied from highly dynamic topologies with constant node movement to slow changing topologies with longer pause times. There were four different pause times used: 0, 25s, 50s and 150s. For each pause time the load was varied for six different values of PGT: 0.25s, 0.5s, 1s, 2s, 5s and 10s.

To evaluate the packet delivery ratio and delay, all data packets generated at all nodes during the 3600s of simulation time were considered. Routing information packets were not considered. Packet delivery ratio is defined as the ratio of the number of packets received successfully by the intended destination to the number of packets that were generated. Delay was measured from the time the packet was generated to the time it was successfully received by the intended destination and includes the route request and reply phases, queuing at intermediate nodes and MAC interactions. Packets that did not arrive successfully at the intended destination were not considered for delay calculations. Simulation parameters are summarised in Table 3-1.

Table 3-1. Simulation parameters

x_max	1000m
y_max	1000m
v_max	20 m/s
nodes	50 nodes
pkt_length	64 bytes
session_length	10 packets
buffer_length	100 packets
pkt_timeout	2s
PGT	1s, 2s, 4s, 10s, 20s, 40s
pause_time	0, 25s, 50s, 150s
reply_delay	50 ms
T_m	20 s
τ	70 packets
w_{A1}	0.1
w_{A2}	3
w_{A3}	2
w_{m1}	1
w_{m2}	10

3.8. Results

Figures 3-6 to 3-9 show the packet delivery ratio results obtained for each of the four protocols with different mobility pause times and varying PGT values. Since both DLAR and AODV use on-demand route selection and similar mechanisms for acquiring routes, the general trend for all the protocols with varying load is similar. However the effects of load balancing can clearly be seen when the offered load is high and the pause time is low, which is the most demanding scenario. In Figure 3-6, where there is no pause time and hence continuous node mobility, the packet delivery ratio for AODV with PGT = 1s is 39.8 % while DLAR 4 has a packet delivery ratio of 54.2 % for the same scenario. DLAR 4 therefore has a 14.4 % larger packet delivery ratio than AODV for the most severe situation. AODV also performs worse than the other DLAR protocols, with DLAR 4 performing the best. The differences in packet delivery ratio decrease as the load is decreased. With a PGT of 40s, which is the lowest offered load, AODV has a 3.8 % difference in packet delivery ratio as compared to DLAR 4. The other DLAR protocols appear to perform similarly.

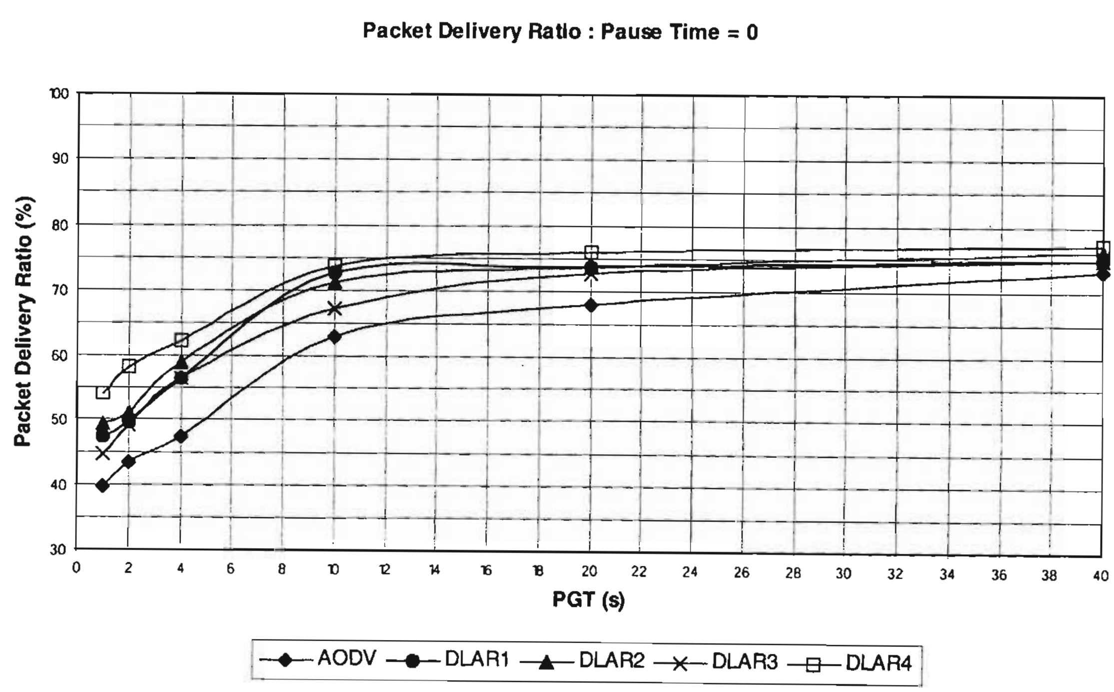

Figure 3-6. Packet delivery ratio with varying load and no pause in node movement

For a pause time of 25s as shown in Figure 3-7, DLAR 4 also performs the best with AODV again having the lowest packet delivery for the highest load when PGT = 1s of 52.5%, which is 9.3 % lower than DLAR 4. For the slowest changing mobility scenario, the results in Figure 3-9 show that all protocols deliver approximately 97 % of the packets for the lowest offered load, while for the highest offered load, there is still a difference of 9.2 % between DLAR 4 and AODV.

The general trend is that the difference in packet delivery ratio between AODV and DLAR 4 decreases as the load is decreased for all mobility scenarios. This can be seen when comparing the plots in Figure 3-10 and Figure 3-11, which display the results obtained to show the effect of pause time on packet delivery ratio. In Figure 3-10 the packet delivery ratio is plotted against mobility pause time, with a PGT of 40s. With this PGT the difference between the better performing DLAR 4 and AODV varies from 3.8 % for continuous movement to less than 0.2 % for pause time of 150s. However with a PGT of 1s, as shown in Figure 3-11, the difference between AODV and DLAR 4 is 14.4 % for continuous mobility and 9.2 % for pause time of 150s.

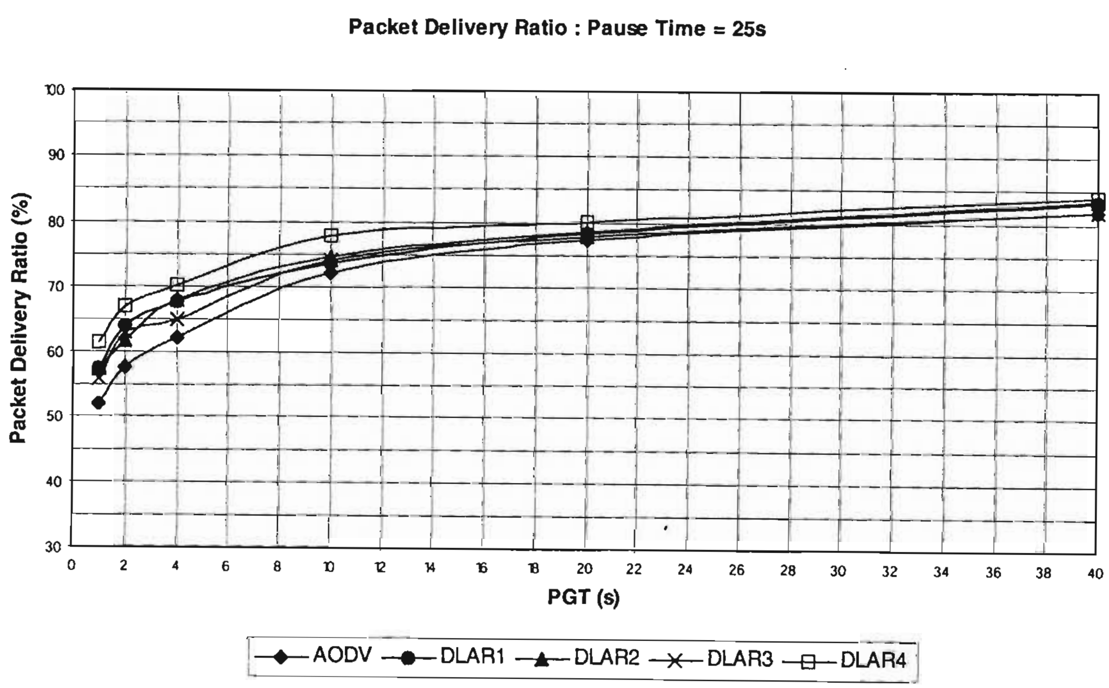

Figure 3-7. Packet delivery ratio with varying load and Pause Time of 25s

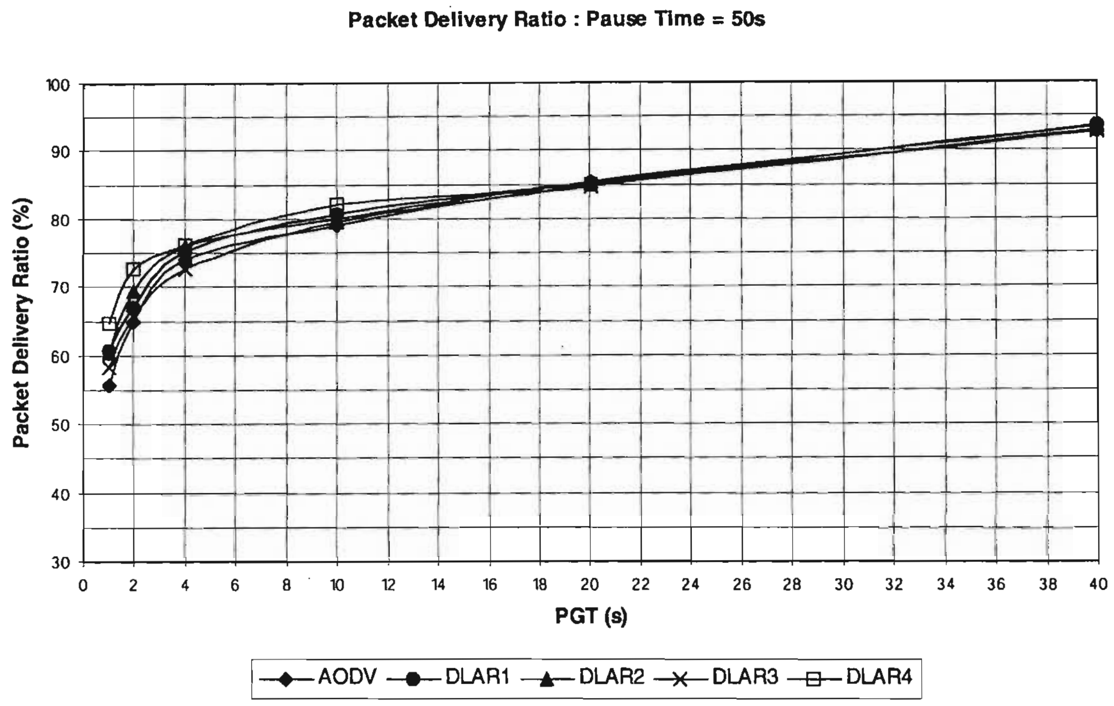

Figure 3-8. Packet delivery ratio with varying load and Pause Time of 50s

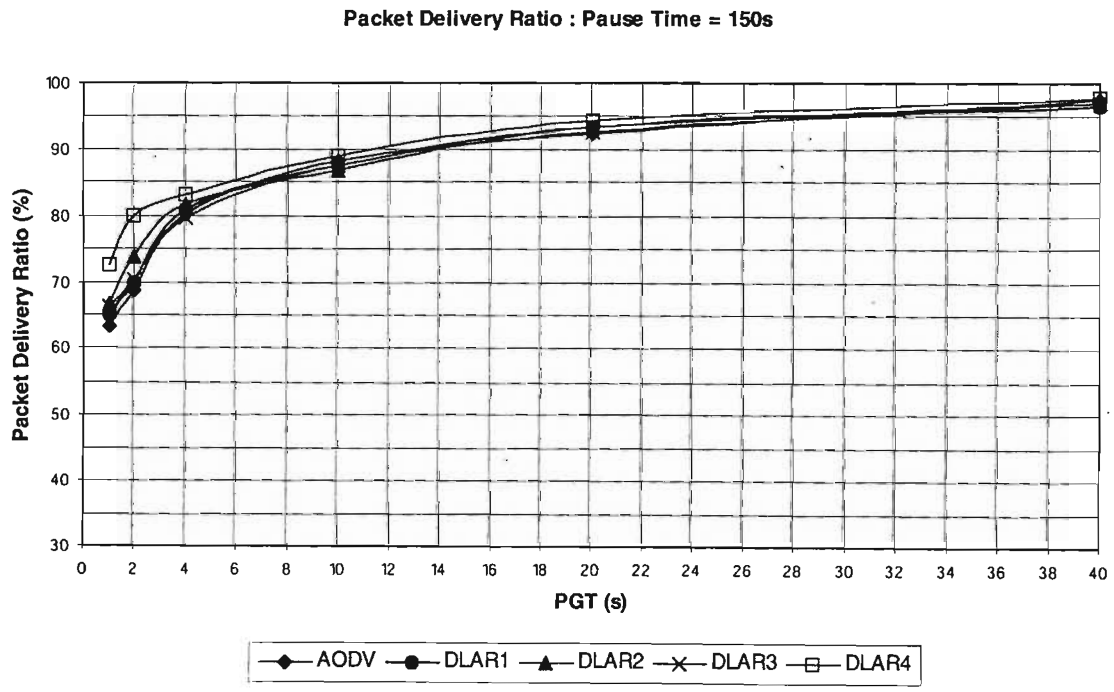

Figure 3-9. Packet delivery ratio with varying load and Pause Time of 150s

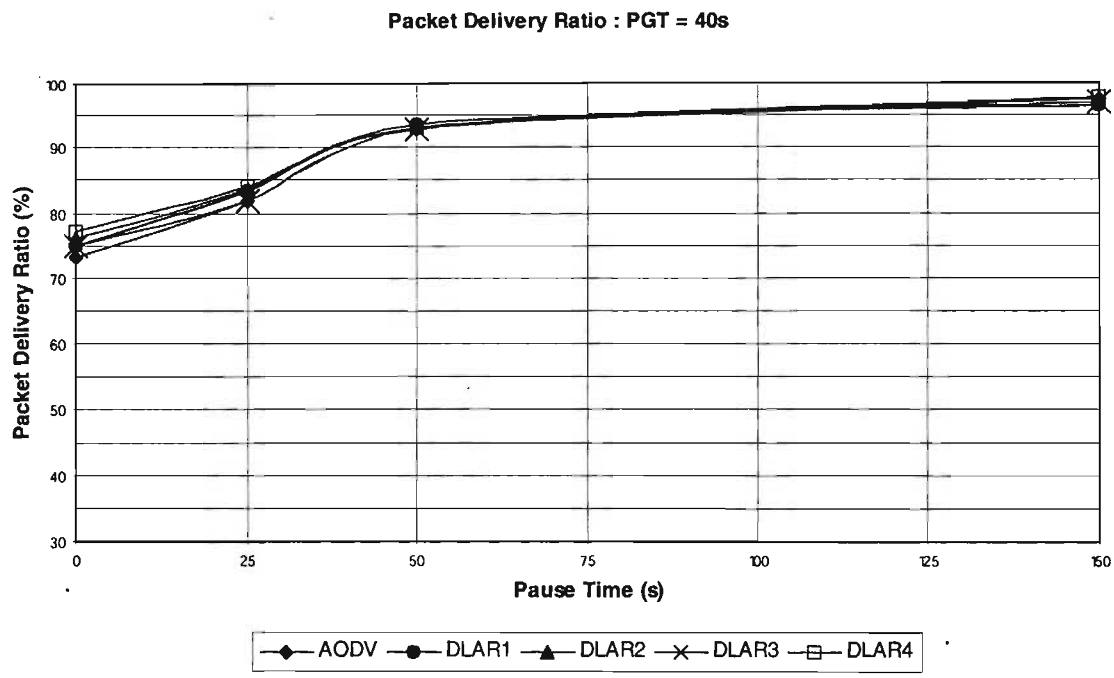

Figure 3-10. Packet delivery ratio with varying Pause Time and PGT of 40s

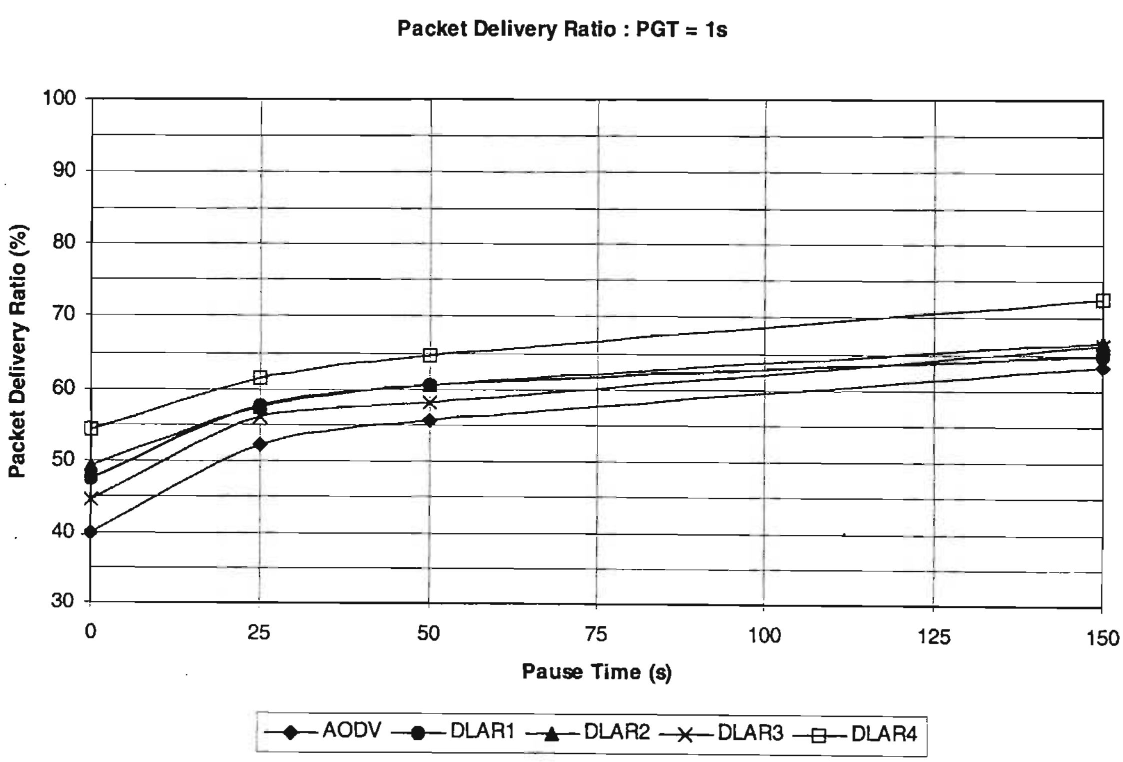

Figure 3-11. Packet delivery ratio with varying Pause Time and PGT of 1s

The results can be explained as follows. In both AODV and DLAR packets are dropped once limited retransmission of a packet by an intermediate node fails. The probability of retransmission is reduced when congested nodes are avoided. Also, packets were dropped if not repeatedly retransmitted. The DLAR protocols outperform AODV for high offered traffic loads due to the ability to reduce congestion and avoid congested nodes. AODV uses the shortest path regardless of congestion, resulting in more packets being dropped.

Figures 3-12 to 3-15 show the average end-to-end results obtained for each of the four protocols with different mobility pause times and varying PGT values. The DLAR protocols again outperform AODV for the high traffic situations in all mobility scenarios. The congestion experienced with AODV results in longer buffering times at intermediate nodes and hence longer end-to-end delays.

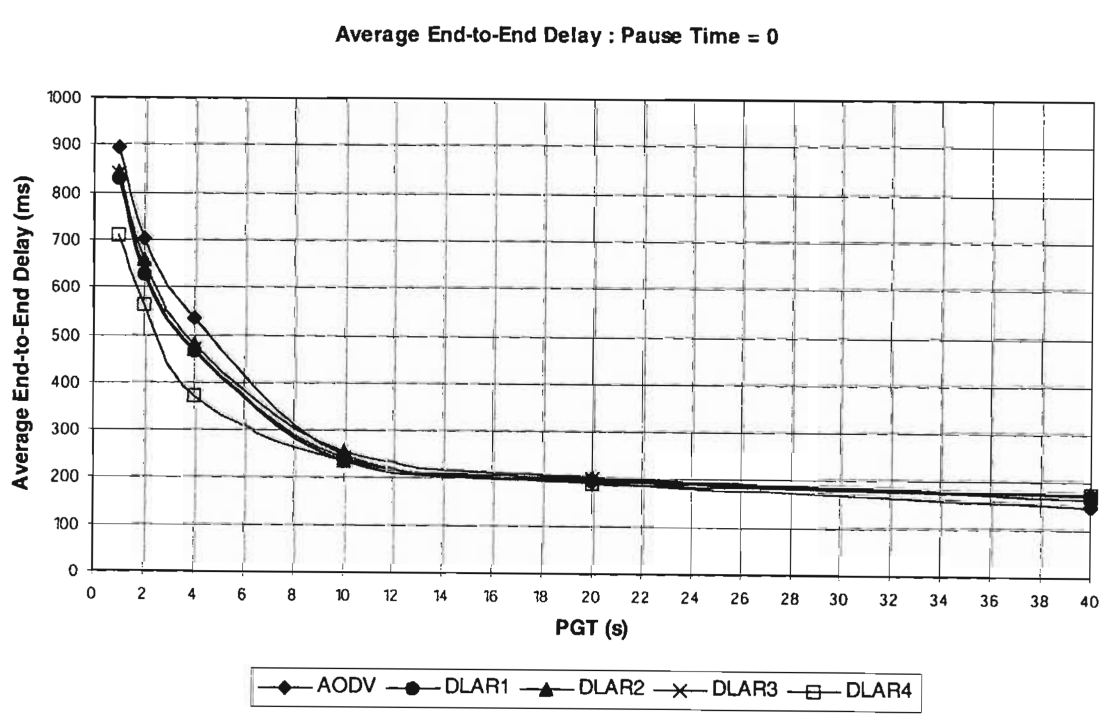

Figure 3-12. Average end-to-end delay with varying load and no pause in node movement

In Figure 3-12, the average end-to-end delay of DLAR 4 for a PGT of 1s and no pause time is 709 ms while it is 894 ms for AODV in the same scenario and approximately 840 ms for the other DLAR protocols. The performance with regards to end-to-end delay does however improve for AODV as the load is decreased. For PGT = 10s in the zero pause time scenario, there is only a 16 ms difference between DLAR 4 with 234 ms delay and AODV with 250 ms delay. The worst performing protocol for PGT =10s and zero pause time is DLAR 2 with an average end-to-end delay of 255 ms.

As shown in Figure 3-16, for all mobility scenarios AODV outperforms the DLAR protocols when the offered load is lowest at PGT = 40s, while Figure 3-17 shows that DLAR 4 has the best average end-to-end delay performance for the highest load with PGT = 1s across varying pause times. The end-to-end delay behaviour can possibly be attributed to the reply delay that was implemented for the DLAR protocols, which actually works against them when offered traffic is low. However the advantages of the load balancing capabilities of the DLAR protocols for average end-to-end delay is clearly evident when offered traffic is high.

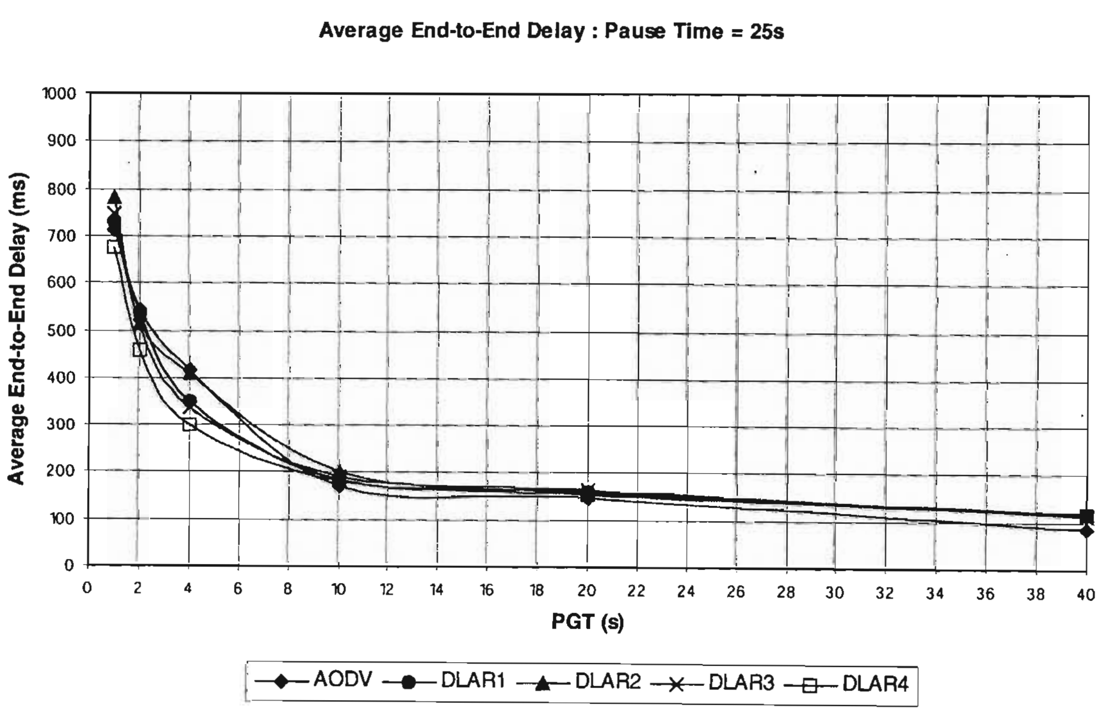

Figure 3-13. Average end-to-end delay with varying load and Pause Time of 25s

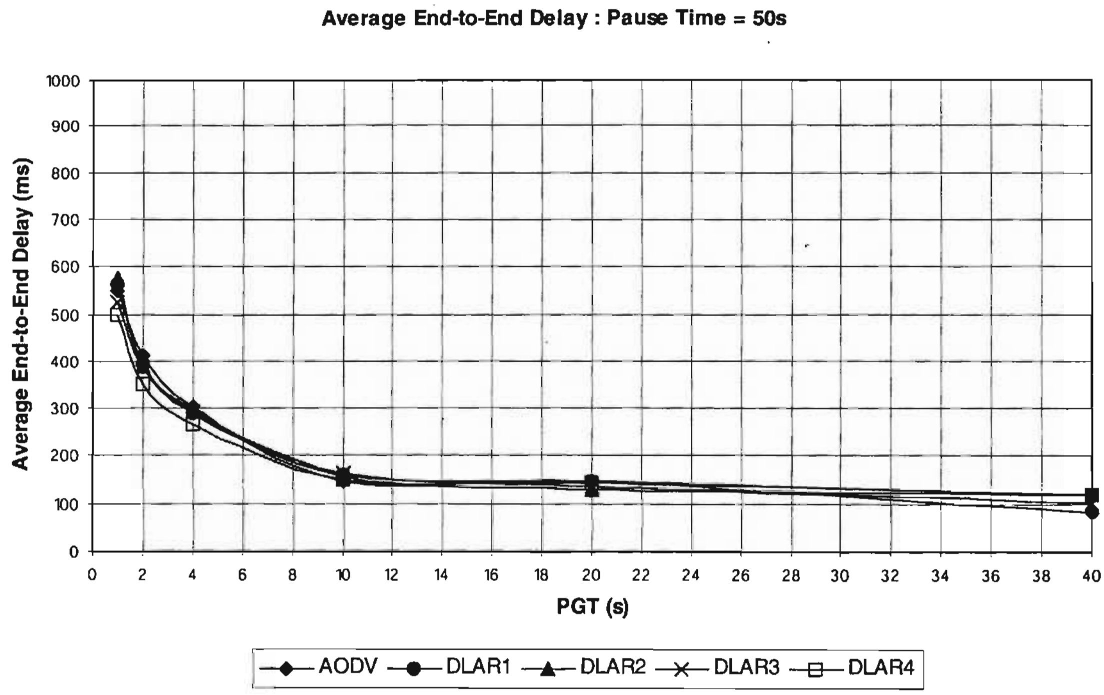

Figure 3-14. Average end-to-end delay with varying load and Pause Time of 50s

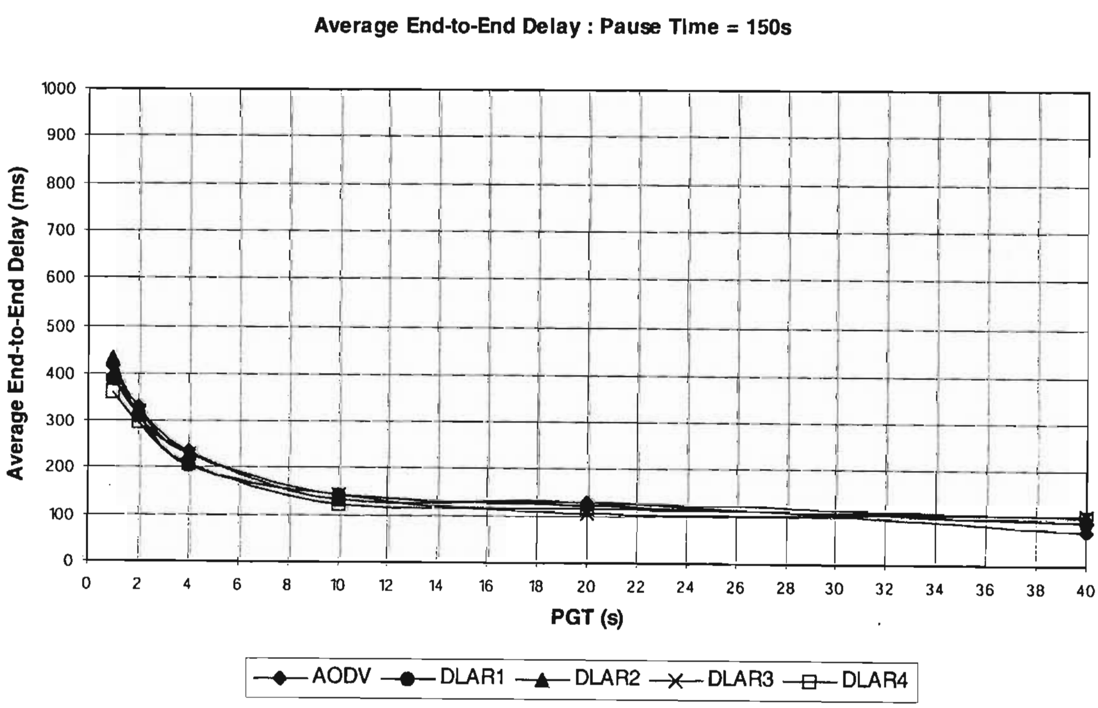

Figure 3-15. Average end-to-end delay with varying load and Pause Time of 150s

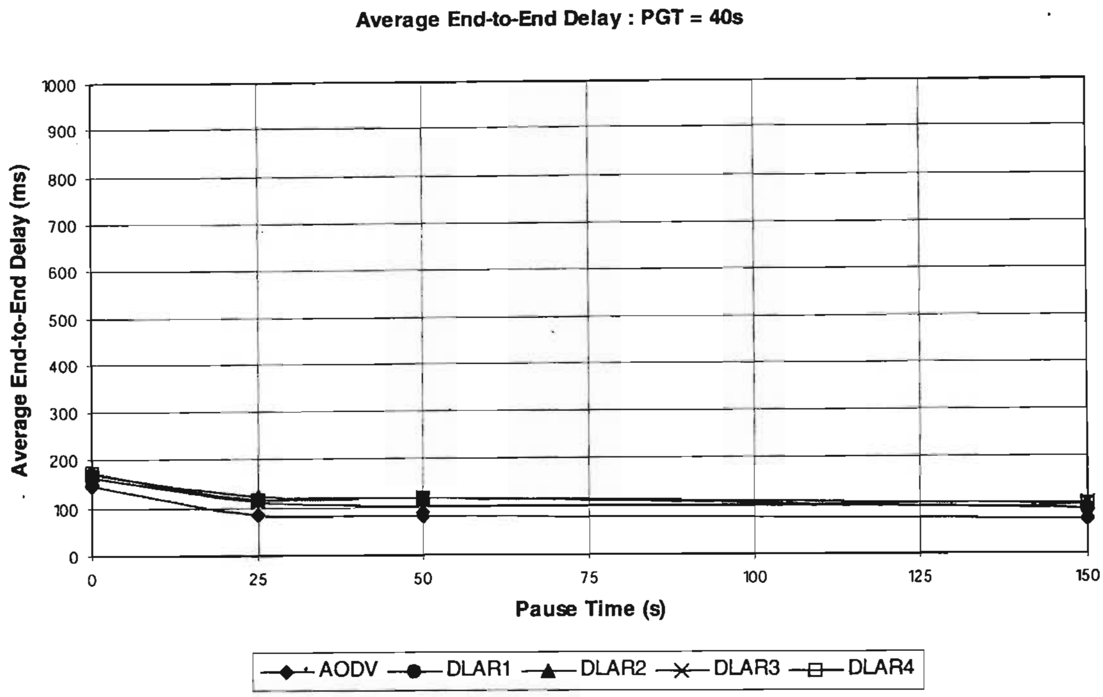

Figure 3-16. Average end-to-end delay with varying Pause Time and PGT of 40s

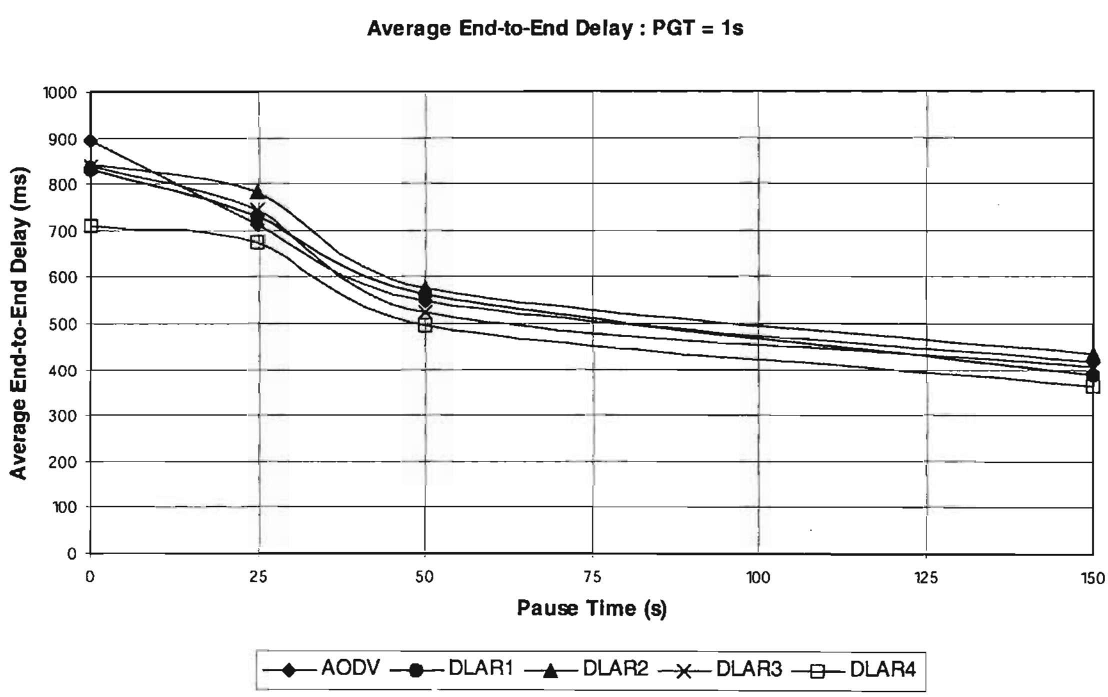

Figure 3-17. Average end-to-end delay with varying Pause Time and PGT of 1s

3.9. Summary

The routing protocols that have been proposed for ad hoc networks generally consider the shortest path as the main routing criteria. However, simple examples can be used to show that congestion may be caused when the shortest path is always used. In this chapter the Dynamic Load Aware Routing protocol [Gerla00] is described, with 3 schemes that each employ load balancing by determining the usage of the buffers at each node. A new fourth DLAR scheme is proposed in this chapter that takes into account signal quality and network load in order to alleviate congestion in the network and reduce delay.

Simulations were performed to compare the new DLAR protocol to the other DLAR protocols and to AODV, which is a prominent routing protocol for ad hoc networks. The two quantities of interest that were measured was average packet delivery ratio and average end-to-end delay. The simulator implements physical layer modelling, medium access control and a mobility model, all of which have been described. The mobility of the network was characterized by using a pause time for which the nodes do not move. A high pause time implies a slow changing topology, while a low pause time implies a rapidly changing topology. The load offered to the nodes was determined by a packet generation time assigned repeatedly to each node. When the packet generation time expires, the node begins transmitting packets to a randomly selected destination. The node is then reassigned a random packet generation time which is no greater than the maximum allowed for the simulation run.

The results obtained using the simulator are discussed. It was found that the DLAR 4 routing scheme has the best performance in comparison to the other protocols in terms of average packet delivery ratio for all mobility scenarios and offered traffic, while AODV has the worst performance due to the selection of routes based purely on their length. It was also found that as the offered load was reduced, the difference between AODV and the DLAR 4 protocols reduced. This demonstrates that the implementation of load balancing does increase packet delivery ratio for high offered load.

In terms of average end-to-end delay, when offered load is high, the DLAR protocols again outperform AODV, with DLAR 4 producing the least average end-to end delay for all mobility scenarios and highest load. However, the AODV protocol begins outperforming the DLAR protocols when the load is decreased. This was attributed to the reply delay implemented in the DLAR protocols that allow the destination nodes to wait for multiple route requests before transmitting the route reply so that the destination can determine which is the best route to use. This delay therefore disadvantages DLAR when traffic is low.

The results obtained show that the DLAR protocols, in particular DLAR 4, reduce congestion in the network and therefore provide higher average packet delivery ratio and lower average end-to-end delay for high offered traffic as compared to AODV, which uses the shortest path. The decision to implement a load balancing protocol as opposed to a conventional routing protocol that uses the shortest path needs to be weighed against the added computational complexity inherent in determining the optimal path as topology changes. Determination of signal quality and load in the buffers implies combining information from the physical layer, data link layer, and network layer, all of which employs valuable processor time. A network designer needs to determine whether the improvements in performance justify the incorporation of load balancing techniques especially with regard to determining whether or not the load offered to the network will become high enough for the advantages of the load balancing protocols to become apparent.

Chapter 4

EVALUATION OF NETWORK BLOCKING PROBABILITY

4.1. Introduction

Due to the frequently changing and unpredictable topology of ad hoc networks, an efficient, distributed and dynamic routing protocol is required to ensure economical usage of the bandwidth available. The routing protocols that have been proposed are usually evaluated using discrete event simulators. One of the aims of the research described in this dissertation is to develop a mathematical model for ad hoc networks that can be used to evaluate the routing protocols and their resultant end to end blocking probabilities. The blocking probability of a network is the probability that a call offered to the network will be blocked due to insufficient resources. Analytical methods for evaluating blocking probability can potentially generate estimates orders of magnitude faster than simulation, and once refined can be used to determine optimum bandwidth, power requirements, network user density and other factors relevant to the development of viable ad hoc networks.

Throughout the 20^{th} century telecommunication engineers have employed loss networks to model the performance of telephone systems [Ross95]. In a loss network each call requires a fixed amount of bandwidth on every link on a route between the source and destination. The call is admitted to the network and holds the requested capacity for a certain amount of time if each link on the route has enough bandwidth to satisfy the requirements. The call is rejected otherwise, and the blocking probability is the probability that the call is rejected.

In 1917 the Danish mathematician A. K. Erlang published his loss formula for the loss probability of a telephone system [Kelly95]. Erlang's formula refers to a single link with calls at a single rate. Although a loss network can be used to model an entire network of links, it needs to be modelled as a multidimensional Markov process when there are multiple links, multiple call rates with different bandwidth requirements and a fixed route associated with each source-destination pair. Since the number of calls of each class on each feasible route uniquely defines a state of the network, the dimension of the state space of the process is the product of the number of the service bandwidth requirements and the number of routes allowed in the network.

The difficulty arises when alternate routes are present in addition to fixed routes. The Markov process no longer has a product form, and the whole set of detailed balance equations need to be written out and solved to obtain the equilibrium state probabilities. Since the computational complexity of the detailed balance equations is both exponential in the number of routes and exponential in the number of service classes, this approach is not practical when there are many routes. To provide accurate estimates of blocking probabilities in reasonable time frames, other computational techniques are required.

One approach is to model wired networks as loss networks and analyse them using the reduced load approximation. The reduced load approximation is based on two assumptions:

i) Link independence: blocking occurs independently from link to link.

ii) Poisson assumption: traffic flow to each individual link is Poisson

[Liu00] proposes a fixed-point approximation that is one type of reduced load approximation for use with multihop, multirate, loss networks with state-dependent routing. An iterative process is used to solve simultaneously for fixed-point variables that are then used to evaluate the blocking probability.

In trying to develop an analytical model for ad hoc networks, the wireless transmission environment needs to be considered. In wireless cellular networks employing Frequency Division Multiple Access (FDMA) and Time Division Multiple Access (TDMA), the number of channels available also fixes the capacity of the system. This hard capacity contrasts with the so called soft capacity of Code Division Multiple Access (CDMA) systems which use the same frequency for each transmitter but which are distinguished from one another through the use of distinct orthogonal codes. The spreading codes are however not always perfectly orthogonal and this results in multiple access interference on a link. A wireless link in a CDMA system therefore has a threshold signal-to-interference ratio at which the link should operate. The evaluation of blocking probability in CDMA cellular networks is frequently based on the Erlang Capacity of the networks [Viterbi93], [Narr99].

CDMA systems offer high spectrum efficiency, multipath resistance, inherent frequency diversity and interference rejection and the potential use of advanced antenna and receiver structures [Viterbi95]. These advantages of CDMA over FDMA and TDMA systems are the reasons for the incorporation of CDMA into the blocking model proposed in Chapter 5.

This chapter is a review of the reduced load method as applied to wired networks, modelled as loss networks and the use of Erlang Capacity for wireless CDMA networks to assess the blocking probability under varying load conditions. The Erlang Loss Formula is presented in Section 4.2 in order to illustrate the fundamentals behind loss network modelling. The problems associated with the large state space of a loss network are described in Section 4.3. In Section 4.4 one type of reduced load approximation, the Erlang Fixed Point Approximation (EFPA) is described. The fixed-point approximation proposed in [Liu00], which is the basis for the model proposed in Chapter 5, is detailed in Section 4.5. In Section 4.6 the development of the Erlang Capacity for CDMA cellular systems is described based on the approach proposed by [Gilhousen91] and [Viterbi93].

An ad hoc network can be modelled as a loss network if it is assumed that the routing protocol chooses its route from a list of possible routes based on the available resources. Chapter 5 describes the manner in which the reduced load approximation proposed in [Liu00] is modified to take into account the effects of wireless communication using CDMA in ad hoc networks.

4.2. Erlang Loss Formula

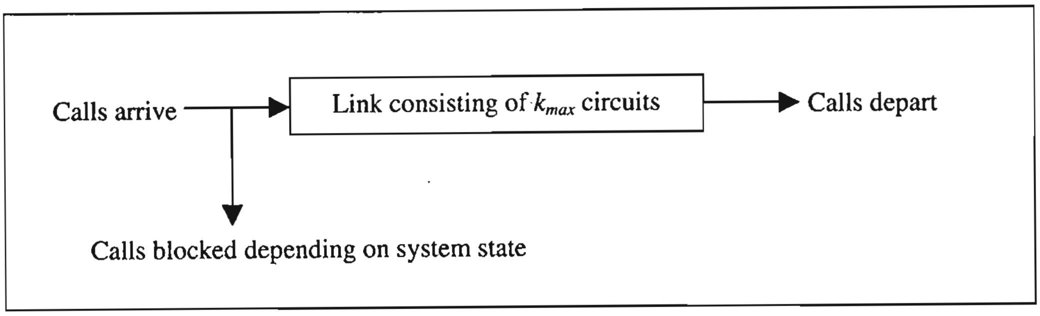

Figure 4-1. Erlang loss system

The Erlang Loss Formula is a core subroutine for the algorithmic analysis of more complex models [Ross95]. Consider a system as shown in figure 4-1, consisting of a single link of k_{max} circuits, with calls arriving at a rate of λ calls per unit time. The arrival times of the calls are assumed to be independent, which leads to a Poisson distribution for the number of calls generated in a given time interval, T:

$$\Pr\{k \text{ calls generated in } T\} = \frac{(\lambda T)^k}{k!} e^{-\lambda T}, n \geq 0 \tag{4-1}$$

The probability density function for the time between call arrivals is exponential:

$$\mathrm{p_a}(t) = \lambda e^{-\lambda t}, t \geq 0 \tag{4-2}$$

The Poisson process has been employed in analysis of telephone systems [Kelly91], [Ross95] and most analyses of cellular mobile networks [Narr99] and is a reasonable assumption since offered traffic is independent of network statistics.

It is a well established rule that a call holding time in telephone networks is exponentially distributed and this assumption has been made repeatedly in the literature. Assuming that the mean call duration is $1/\mu$, the probability density function of the service time for a given channel is exponential:

$$\mathrm{p_s}(t) = \mu e^{-\mu t}, t \geq 0 \tag{4-3}$$

When the number of customers in the system is k and the maximum number of circuits available is k_{max}, the system is said to be in a state S_k, (k = 0, 1, ... k_{max}). Let $p(k)$ be the proportion of time that k circuits are being used, in other words, $p(k)$ is the probability that the system is in state S_k. Assuming that calls arrive with overall rate λ, and since the proportion of time the system spends in state S_k is $p(k)$, the rate at which the transition $S_k \rightarrow S_{k+1}$ occurs is $\lambda p(k)$. When k=k_{max}, the state S_{kmax+1} is physically impossible, therefore the transition $S_{kmax} \rightarrow S_{kmax+1}$ is 0.

To obtain the downward transition $S_{k+1} \rightarrow S_k$, (k = 0, 1, ... k_{max}-1), recall the mean holding time of a circuit is $1/\mu$. This means that the termination rate for a single call is μ. Similarly, if two calls are in progress simultaneously and the average duration of a call is $1/\mu$, the average number of calls terminating during an elapsed time $1/\mu$ is 2 and the termination rate for two simultaneous calls is therefore 2μ. By this reasoning the termination rate for k+1 simultaneous calls is $\mu(k+1)$. Since the system is in state S_{k+1} with probability $p(k+1)$, the downward transition $S_{k+1} \rightarrow S_k$ occurs at a rate $\mu(k+1)\, p(k+1)$ transitions per unit time, (k = 0,1,... k_{max}-1).

If the system is to be in statistical equilibrium, that is, if the relative proportion of time that the system spends in each state is to be a stable quantity, then the upward transition $S_k \rightarrow S_{k+1}$ must occur with the same rate as the downward transition $S_{k+1} \rightarrow S_k$. Therefore,

$$\lambda p(k) = (k+1)\mu \quad , k = 0, 1, \ldots, k_{max}\text{-}1 \tag{4-4}$$

Equation (4-4) can be solved recurrently obtaining a result that expresses each $p(k)$ in terms of the value $p(0)$,

$$p(k) = \frac{\left(\frac{\lambda}{\mu}\right)^k}{k!} p(0) \tag{4-5}$$

together with the normalization requirement,

$$\sum_{k=0}^{k\max} p(k) = 1 \tag{4-6}$$

Using the normalization equation (4-6) together with equation (4-5), $p(0)$ can be determined,

$$p(0) = \frac{1}{\sum_{k=0}^{k\max} \frac{(\lambda/\mu)^k}{k!}} \tag{4-7}$$

Substituting equation (4-7) into (4-5), the probability that k circuits are being used is given by,

$$p(k) = \frac{(\lambda/\mu)^k / k!}{\sum_{c=0}^{k\max} \frac{(\lambda/\mu)^c}{c!}} \tag{4-8}$$

The probability of the system being in a particular state depends on the arrival rate λ and the mean holding time $1/\mu$ only through the product λ/μ. This product is a measure of the demand made on the system and it is often called the offered load, $A=\lambda/\mu$ expressed in units called erlangs (erl) after A. K. Erlang. When $k = k_{max}$ in equation (4-8) it becomes the Erlang loss formula:

$$E(A, k_{\max}) = \frac{A^{k\max}}{k_{\max}!}\left(\sum_{k=0}^{k\max} \frac{A^k}{k!}\right)^{-1} \tag{4-9}$$

Erlang's loss formula, also known as the Erlang-B formula, gives the proportion of calls that are lost to the system. If call holding times are exponentially distributed, the formula gives the steady state probability that all circuits are busy. The formula was obtained by Erlang from his development of the concept of statistical equilibrium and he demonstrated the concept's ability to deliver exact formulae for many of the critical problems of telephony [Kelly91].

4.3. Loss Networks

Erlang's formula refers to a single link with calls at a single rate. When there are many links and different classes of calls, a generalisation of Erlang's model allows an analysis of a network of links, where the number of circuits used by the call depends on the type of call being offered, as demonstrated by [Kelly95]. The circuit switched network of the kind depicted in figure 4-2 consists of J links, indexed j, each with a capacity C_j. Let M be the set of all possible routes and assume that calls requesting route m arrive as a Poisson stream of rate λ_m with a holding time of μ_m. The offered load to a route is therefore $\rho_m = \lambda_m / \mu_m$. A call on route m uses U_{jm} circuits from link j.. M therefore indexes independent Poisson processes. A call requesting route m is blocked and lost if on any link j there are less than U_{jm} circuits free. The constraint

$$\sum_{m=1}^{M} U_{jm} k_m \leq C_j \tag{4-10}$$

where k_m is the number of calls in progress on route m, expresses the condition that the number of circuits required from link j by calls in progress cannot exceed the capacity of link j. The call is otherwise admitted and simultaneously holds U_{jm} circuits from link j for the holding period of the call. Let $\mathbf{k} = (k_m, m \in M)$, $\mathbf{C} = (C_1, C_2, \ldots, C_J)$ and $\mathbf{U} = (U_{jm}, m \in M, j = 1, \ldots, J)$. The usual model for a circuit-switched network [Kelly91] is a continuous-time Markov chain $(\mathbf{k}(t), t \geq 0)$ taking values in

$$S = S(\mathbf{C}) = \{\mathbf{k} \in Z_+^M : \mathbf{Uk} \leq \mathbf{C}\} \tag{4-11}$$

where Z_+ is the set of non-negative integers, and its unique stationary distribution is given by

$$p(k) = \mathbf{\Phi}^{-1} \prod_{m \in M} \frac{\rho_m^{k_m}}{k_m!}, \mathrm{k} \in S \tag{4-12}$$

where

$$\mathbf{\Phi} = \Phi(\mathbf{C}) = \sum_{\mathbf{k} \in S(\mathbf{C})} \prod_{m \in M} \frac{\rho_m^{k_m}}{k_m!} \tag{4-13}$$

The stationary probability that a route m call is blocked is then given by

$$L_m = 1 - \frac{\Phi(\mathbf{C} - \mathbf{Ue}_m)}{\Phi(\mathbf{C})} \tag{4-14}$$

where $\mathbf{e}_m$ is the unit vector from S(C) describing just one call in progress on route m.

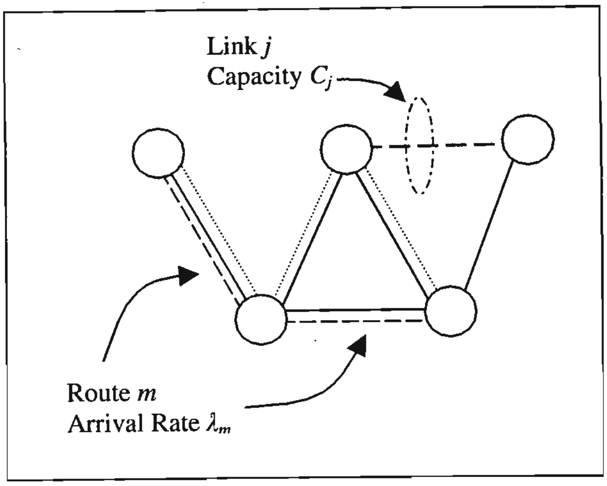

Figure 4-2. Typical circuit-switched network with 5 nodes, 6 links and 5 routes.

Although the expression in (4-14) is explicit, (4-13) cannot be calculated in polynomial time [Kelly91], [Kelly94], [Ross95], [Liu00]. As a simple example, if **U** = **I**, which implies that $U_{jm} \in \{0;1\}$ and the network is fully connected with all possible one-link routes such that $|M| = J$, then the size of the state space becomes

$$|S(\mathbf{C})| = \prod_{j=1}^{J} C_j = C^J \tag{4-15}$$

[Kelly94] shows that the exact determination of $\Phi(\mathbf{C})$ is NP-complete in the number of distinct routes. A fast and exact sequential algorithm therefore probably does not exist. Therefore for networks of even moderate capacity, alternative methods are required.

4.4. Erlang Fixed Point Approximation

The reduced load approximation is one such alternate method and there is a fairly broad class of approximation techniques referred to as reduced load approximations. The general approach is to reduce the arrival rates of offered traffic to a subnetwork by a factor equal to the probability that a new call on that route would not be blocked on the other links of its path. This leads to a set of fixed point equations for which there exists a solution, which is not necessarily unique. The ideas behind all reduced load approximation methods use the following two assumptions:

i) Link Independence: The blocking occurs independently from link to link, so the probability that a call is admitted on a given route is the product of the probabilities that the call is admitted on each individual link of that route.

ii) Poisson Arrival: The offered load to a link is a Poisson process with rate reduced by blocking on other links. In cases of multi-rate traffic, the offered load of each class of traffic onto a link is a Poisson process with its rate reduced by blocking on other links.

The Erlang Fixed Point Approximation (EFPA) is a member of the reduced load class, one that analyses each link as a separate subnetwork. The EFPA performs well asymptotically and [Kelly91] proved that the estimates for a network with fixed routing tends towards the exact probabilities when,

i) the link capacities and arrival rates are increased simultaneously keeping the network topology fixed, and

ii) when the number of links and routes are increased while the link loads are held constant.

The EFPA is a solution to the set of fixed point equations

$$B_j = E(\rho_j, C_j) \tag{4-16}$$

$$\rho_j = \sum_{m \in M} U_{jm} \lambda_r \prod_{\substack{i \in m \\ i \neq j}} (1 - B_i) \tag{4-17}$$

B_j is the probability that link j is full given its offered traffic load is ρ_j, and uses the Erlang Loss Formula from equation (4-9). The offered traffic load is an approximation obtained by considering the sum of the contributions made by each route m to j's carried load. Applying the independent blocking assumption results in

$$\begin{aligned} L_r &= 1 - \Pr(\text{call is accepted on each link } i \in r) \\ &= 1 - \prod_{i \in m} \Pr(\text{call is accepted on each link } i) \\ &= 1 - \prod_{i \in m} (1 - B_i) \end{aligned} \tag{4-18}$$

[Kelly86] proved that there is a unique vector $(B_1, \ldots, B_J) \in [0,1]^J$ satisfying (4-16) and (4-17).

Other approximation techniques include the Knapsack Approximation [Kaufman81], Pascal Approximation [Ross93] and the Markov Random Field Method [Zachary99]. None of the abovementioned approximations take into account multihop networks together with state dependent routing. The networks tend to be modelled as fully connected, so even when adaptive routing is used, it is assumed that the one-hop route is always first attempted and overflows to a selected two-hop route, based on expected free capacity. The call is lost if it cannot be completed on the two-hop route.

[Liu00] proposes a scheme in order to tackle the following characteristics of currently evolving integrated service networks:

i) Sparse topologies, where the assumption of a direct route between source and destination nodes does not always hold.
ii) Routes comprise a large number of hops with many possible routes between the source and destination nodes.
iii) Different traffic classes characterized by widely varying bandwidth requirements and different mean holding times.

The following Section describes the Fixed Point Approximation (FPA) proposed by [Liu00]. In Chapter 5 the adaptation of this FPA to evaluate the call blocking probability of an ad hoc network employing CDMA is described.

4.5. Blocking Probability in Multirate Multihop Loss Networks

In [Liu00] a Fixed Point Approximation to evaluate the end-to-end blocking probability in multirate multihop loss networks is proposed. The approximation method involves solving simultaneously for four fixed point variables using three mappings. Once solved iteratively, the fixed point variables are used to calculate the blocking probability of each class of call between each node pair. The terms "units of circuits", "trunks" and "units of bandwidth" are used interchangeably in the following discussion.

4.5.1. Notation

The network consists of N nodes, indexed n and has a total of $R = N(N\text{-}1)/2$ node pairs, each indexed r. Between each node pair r there are a total of M_r routes listed in order from least cost (best route) to highest cost (worst route). The mth route between node pair r is referred to as (r, m). The network consists of J links with link j having a capacity of C_j. The network is offered S classes of traffic and b_s is the bandwidth required for a class s call. A call request (r, s) between node pair r of class s arrives with a Poisson arrival rate of λ_{rs} and has an exponential mean holding time of μ_{rs}. Each link is considered to be bi-directional and duplex. The offered traffic between a node pair consisting of nodes a and b is the sum of the traffic from a to b and of traffic from b to a. k_{js} defines the number of class s calls on link j.

4.5.2. Fixed Point Method

There are four fixed point variables that are used to evaluate the end-to-end blocking probability B_{rs} of a class s call between node pair r. v_{js} is the arrival rate of class s calls on link j. a_{js} is the probability that a class s call will be admitted on link j. The probability that k units of bandwidth are being used on link j, $k \in \{0, 1, \ldots, C_j\}$, is given by $p_j(k)$. The fourth fixed-point variable is q_{rms} which is the probability that route (r, m) will be used for call request (r, s) given that route (r, m) can admit the call. Three mappings are used to solve simultaneously for the four fixed-point variables as follows:

i) a_{js}, $q_{rms} \rightarrow v_{js}$: the admissibility probabilities and route selection probabilities are held fixed to obtain the arrival rate on each link.

ii) $v_{js} \rightarrow a_{js}$, $p_j(k)$: the arrival rates are held fixed to calculate the admissibility probability and the occupancy probability of the links.

iii) $p_j(k) \rightarrow q_{rms}$: the occupancy probabilities are held fixed to obtain the route selection probability of each route between each node pair for each class of traffic.

4.5.3. Routing Strategy

A simple adaptive alternate routing strategy is implemented where the route with the maximum free bandwidth on the most congested link is selected. The source destination node pair r evaluates its list of possible routes M_r to first pick the link with the minimum free bandwidth for each route. The route (r, m) is in a state of admitting the call (r, s) if and only if

$$C_j^f \geq b_{js}, \forall j \in (r,m) \tag{4-19}$$

where C^f_j is the free bandwidth on link j.

When a route (r, m) can admit a call request (r, s), the most congested link $\ell_{(r,m)}$ on the route is defined as the link with the fewest free circuits available,

$$\ell_{(r,m)} = \arg\min_{j \in (r,m)} C_j^f \tag{4-20}$$

4.5.4. Mapping 1

For the first mapping, which holds the values of the admittance probabilities and route acceptance probabilities fixed to determine the link arrival rates, define v_{jrms} as the arrival rate on link j from a call request (r, s) on route (r, m) given that link j will admit the call. Therefore the arrival rate of call requests on link j from all routes is given as

$$v_{js} = \sum_{(r,m)} v_{jrms} \tag{4-21}$$

where

$$v_{jrms} = \lambda_{rs} q_{rms} I[j \in (r,m)] \prod_{i \in (r,m), i \neq j} a_{is} \tag{4-22}$$

and $I\,[j \in (r, m)]$ is the indicator function,

$$I[j \in (r,m)] = \begin{cases} 1, & j \in (r,m) \\ 0, & \text{otherwise} \end{cases} \tag{4-23}$$

The indicator function is used in equation (4-22) because only routes that include the link being evaluated make a contribution to the arrival rate on the link. Equation (4-22) represents the first mapping and takes into consideration the arrival rate of each call request, the probability that the route will be used by the call request and the probability that the call request will be accepted on each of the other links making up the route.

4.5.5. Mapping 2

The occupancy probability of each link and the probability that the link can admit a call request can be calculated by holding the value of the link arrival rate fixed, that is, $v_{js} \rightarrow a_{js}$, $p_j(k)$. [Kaufman81] gives a one dimensional recursion for calculating the link occupancy distribution probabilities:

$$kp_j(k) = \sum_s b_{js} \frac{v_{js}}{\mu_{rs}} p_j(k - b_{js}), k = 1,..., C_j \tag{4-24}$$

where

$$k = \sum_s b_{js} k_{js} \tag{4-25}$$

is the total bandwidth being used on link j.

Additionally, $p_j(k) = 0$ if $k < 0$ and the normalisation requirement is

$$\sum_{k=0}^{C_j} p_j(k) = 1 \tag{4-26}$$

The probability that a class s call is admitted to link j is given by

$$a_{js} = 1 - \sum_{k=C_j-b_{js}+1}^{C_j} p_j(k) = \sum_{k=0}^{C_j-b_{js}} p_j(k) \tag{4-27}$$

4.5.5. Mapping 3

The third mapping, $p_j(k) \rightarrow q_{rms}$ derives the probability that a route will be used, given that it can admit the call based on the status of the links. Since the routing scheme employs the route with the most available bandwidth on its most congested link, the stochastically most congested link on the route needs to be found. Let $E[k]$ be the expected number of occupied bandwidth units on a link which is given by

$$E[k] = \sum_{s=1}^{S} \frac{v_s b_s}{\mu_s} a_s \tag{4-28}$$

The proof of equation (4-28) can be found in [Liu00] and is based on Kaufman's recursion [Kaufman81]. Using equation (4-28) and equation (4-20) the stochastically most congested link on route (r, m) becomes:

$$\ell_{(r,m)} = \arg\min_{j \in (r,m)} (C_j - \sum_{s=1}^{S} \frac{v_{js} b_s}{\mu_s} a_s) \tag{4-29}$$

Define for link j the probability of no more than k trunks being free as:

$$t_j(k) = \sum_{i=0}^{k} p_j(C_j - i) \tag{4-30}$$

A call will be admitted on route (r, m), which is one of the M_r routes between node pair r, if all the routes listed before (r, m) have fewer free trunks on their most congested link and all routes listed after (r, m) have at most the same number of free trunks on their most congested link. The probability that route (r, m) will be used can be expressed as:

$$q_{rms} = \sum_{k=1}^{C_{\ell(r,m)}} \left[p_{\ell(r,m)}(C_{\ell(r,m)} - k) \prod_{i=1}^{m-1} t_{\ell(r,m)}(k-1) \prod_{i=m+1}^{M} t_{\ell(r,m)}(k) \right] \tag{4-31}$$

4.5.6. Blocking Probability

Once the four fixed-point variables have been solved, the blocking probability for a call request (r, s) is given by

$$B_{rs} = 1 - \sum_{m=1}^{M_r} q_{rms} \prod_{j \in (r,m)} a_{js} \tag{4-32}$$

The equilibrium fixed-point is obtained using repeated substitution and the fixed point is used to calculate the end-to-end blocking probabilities using equation (4-32).

4.6. Erlang Capacity of CDMA Cellular Systems

The Erlang-B formula, can be used to evaluate conventional multiple access systems, such as those employing Frequency Division Multiple Access (FDMA) and Time Division Multiple Access (TDMA). In networks using FDMA and TDMA, traffic channels are allocated to users as long as there are channels available. When all channels are being used, arriving traffic is blocked until a channel becomes free at the end of a current call. In a Code Division Multiple Access (CDMA) system, however, the number of channels available to mobile users on the reverse link is not fixed. The reverse link (or uplink) is the link from the mobile user to the base station and instead of access being restricted to the number of channels, its restriction is based on the amount of interference and noise present at the base station [Viterbi93], [Gilhousen91].

[Viterbi93] defines blocking to occur when the users within the cell of interest and in other cells introduce an interference density so great that it exceeds the background noise level by an amount that is taken to be 10dB. The total interference at a base station receiver is the sum of the intracell interference, intercell interference, and the thermal noise. The intracell interference includes interference from all mobiles within the cell of interest while the intercell interference includes interference from mobiles in other cells. The thermal noise is the product of the spread bandwidth W and the one-sided spectral noise density N_0. For a call to be admitted, the total interference must be less than the acceptable interference, I_0W, where I_0 is the maximum total acceptable interference density,

$$\text{Intracell int.} + \text{Intercell int.} + N_0 W \leq I_0 W \qquad (4\text{-}33)$$

with the following assumptions:

(i) The number of active calls is a Poisson random variable with mean λ/μ where λ is the call arrival rate and $1/\mu$ is the mean call holding time.

(ii) Each mobile involved in a call actively transmits with probability γ.

(iii) The required bit energy-to-interference ratio $E_{b,k}/I_0$ of each user k is varied according to propagation conditions to achieve the desired frame error rate.

Let the bit rate be B. The intracell interference power from each of the K active mobiles within the cell is $BE_{b,k}$ and similarly for each of the L active mobiles outside of the cell of interest, the intercell interference power is $BE_{b,l}$. Equation (4-33) then can be written as,

$$\sum_{k=1}^{K} BE_{b,k}v_k + \sum_{l=1}^{L} BE_{b,l}v_l \leq W(I_0 - N_0) \qquad (4\text{-}34)$$

where v is a binary random variable that takes on values between 0 and 1 and represents activity factor with probability $\Pr\{v = 1\} = \gamma$. Dividing equation (4-34) by I_0B and defining $\varepsilon_k = E_{b,k}/I_0$ and $\varepsilon_l = E_{b,l}/I_0$ results in

$$\sum_{k=1}^{K} \varepsilon_k v_k + \sum_{l=1}^{L} \varepsilon_l v_l \leq G(1-\eta) \qquad (4\text{-}35)$$

where $G = W/B$ is the spread spectrum processing gain and $\eta = N_0/I_0 = 0.1$ (nominally) [Viterbi93]. Let the total interference on the left hand side of equation (4-35) be Z_k. The blocking probability for the system then becomes:

$$\Pr\{\text{call blocked}\} = \Pr\{Z_k > G(1-\eta_{\text{crit}})\} \qquad (4\text{-}36)$$

where η_{crit} is the critical value for the loading experienced by the base station. When the condition in equation (4-36) is exceeded call quality suffers, so this probability needs to be kept sufficiently low to ensure high availability of good quality service. The call blocking in CDMA systems is referred to as "soft blocking" because the condition can be relaxed by allowing $1-\eta_{\text{crit}}$ to increase.

4.7. Summary

Traditionally, wired networks have been modelled as loss networks. In a loss network each call requires a fixed amount of bandwidth on every link on a route between the source and destination. The call is admitted if the resources available are adequate and holds the requested resources for a certain time. The blocking probability is the probability that the call will be rejected due to the unavailability of satisfactory resources.

The Erlang Formula is a fundamental routine upon which more complicated models for loss networks are built. A brief description of the development of the Erlang formula is described followed by a discussion of loss networks and the complications that arise due the large state space that needs to be evaluated in order to accurately determine the blocking probability. These complications lead to the development of a group of estimation algorithms referred to collectively as reduced load approximations. The arrival rates of offered traffic to a link are reduced by a factor equal to the probability that a new call on that route would not be blocked on the other links of the route. This leads to a set of fixed-point equations that need to be solved simultaneously in order to obtain a stationary distribution of variables for the calculation of the blocking probability. The reduced load approximations are based on two assumptions that are described.

The Erlang Fixed-Point Approximation is described in Section 4.4 as one type of reduced load approximation. Other reduced load approximations include the Knapsack Approximation [Kaufman81], Pascal Approximation [Ross93] and Markov Random Field Method [Zachary99]. However since none of these approximations take into account multihop networks together with state-dependent routing, the fixed point approximation proposed by Mingyan Liu [Liu00] is described, which attempts to deal with the characteristics of current and future integrated service networks.

Ad hoc networks communicate wirelessly, therefore the multiple access method for a wireless system needs to be considered. A CDMA system is chosen for this analysis since it has advantages in terms of spectrum efficiency and interference rejection over TDMA and FDMA systems. In Section 4.5 the Erlang capacity of a cellular CDMA system is described.

One on the aims of the research work described herein is to develop an analytical model to evaluate the blocking probability of the routing protocols in ad hoc networks. In Chapter 5 it is shown how the fixed-point approximation is combined with methods used to evaluate the Erlang Capacity in cellular CDMA networks to develop a model that can eventually be used for the assessment of blocking probability in wireless ad hoc networks.

Chapter 5

ADAPTATION OF THE FIXED-POINT APPROXIMATION TO AD HOC NETWORKS

5.1. Introduction

In this chapter an analytical model is proposed to evaluate the call blocking in a wireless ad hoc network employing CDMA. In the literature [Kaufman81], [Kelly94], [Ross95], the reduced load method, in particular the fixed-point approximation, has been used to evaluate blocking probability in wired networks modelled as loss networks. The call blocking probability in wireless CDMA systems has been evaluated using the Erlang capacity of the CDMA system and is based on the total interference being experienced by the base station. There is nothing in the literature, as far as the author is aware, that attempts to analytically evaluate the blocking probability in an ad hoc network due to the routing protocol that is implemented. The analytical model proposed is a combination and modification of the blocking probability methods used for wired networks proposed in [Liu00] and the blocking probability evaluation of wireless cellular CDMA systems as proposed in [Gilhousen91], [Viterbi93], [Miller95], [Viterbi95] and [Narr99]. The blocking probability evaluation of wired networks and wireless cellular networks are discussed in Chapter 4.

Since the ad hoc network is a wireless environment, it is assumed that congestion is due to the number of simultaneous calls being handled by a particular node, instead of using the capacity of links as the limiting factor. Nodes in an ad hoc network receiving transmissions from other peer nodes in the network are analogous to a base station in a cellular network receiving transmissions from mobiles within its cell. Since the only transmissions in an ad hoc network intended for a target node are transmitted from nodes within a certain range of the target node, the nodes in the ad hoc network also have an area that surrounds them that can be considered their "cell" within which other nodes can transmit to them. The power control algorithm used in the ad hoc network is assumed to be optimum, so that each node transmitting to its target node is able to provide the signal power as required by the target node. To simplify the analysis the medium access control is neglected and it is assumed that a sufficient number of receiver processors are provided to each node such that the probability of a new arrival finding all receiver processors busy is negligible. This in effect implies that blocking is only interference-limited, since the greater the number of calls in progress, the higher the interference is and the worse the resulting signal quality will be.

A problem not yet overcome is the ability to include topology changes in the analytical model. Since the fixed-point approximation is based on knowledge of the routes that are available in an equilibrium state it is difficult to incorporate topology changes into the analytical model for ad hoc networks. This problem is described further in section 5.2. A fixed topology wireless multihop network is therefore assumed. The analytical model is described in section 5.3 where the mappings that are evaluated iteratively to find the fixed-point variables are described. Numerical results obtained using the analytical model to evaluate a given fixed topology wireless network are presented and discussed in section 5.4.

5.2. Problems with Analysing Topological Changes

The prediction of node position in an ad hoc network is significantly more complicated than that of the cellular network due to it involving more degrees of freedom in mobility and there being no fixed point of reference. In order to take into account all of the possible routes that could be available in an ad hoc network as the topology varies over time, a fully connected network could be considered, consisting of the number of nodes of the network of interest. Then each link has a certain probability of being in existence. Consider for example Figure 5-1 where the random topological structure of an ad hoc network consisting of 5 nodes is equivalent to a fully connected network with some of the links not available. There are 5 active links in the random topology network due to nodes being within range of each other. Each link in the fully connected network is active if the equivalent link in the random topology network is in existence.

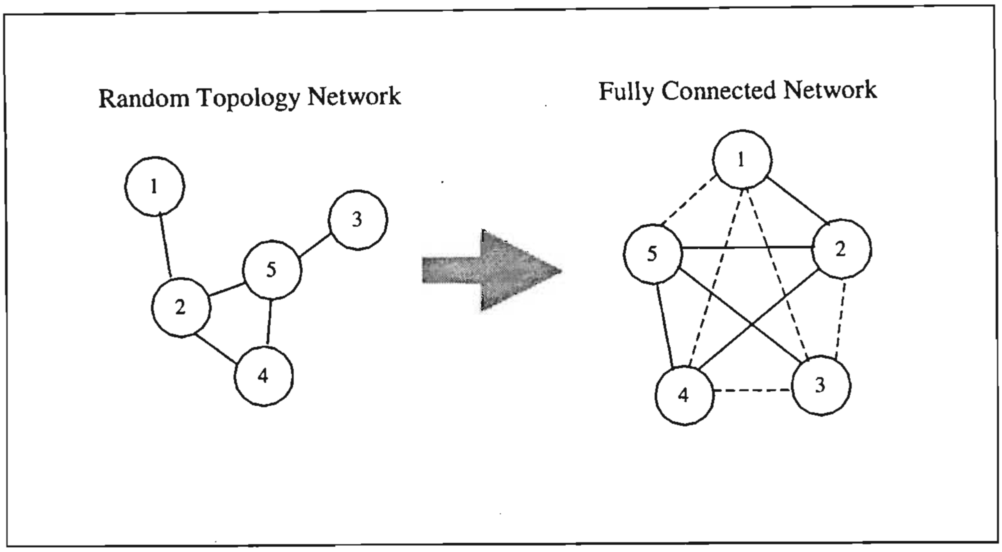

Figure 5-1. Random topology network and equivalent fully connected network with some links inoperable

A probability of existence e_l, for a link l, needs to be found for the modelling of the random topology network as a fully connected network. [McDonald99] develops a mobility model that is used to predict the evolution of an ad-hoc network's topology. Expressions are derived for the probability of the link being active as a function of time based on different initial conditions for the nodes of interest. The distribution of the mobility of one node with respect to the other is first be determined in order to characterize the availability of a link over a period of time. It is therefore first necessary to derive the mobility distribution of a single node in isolation and then extend the distribution to derive the joint mobility distribution that accounts for one node with respect to the other. Assuming that links fail independently, once the link availability metric is known for each link along a path, [McDonald99] determines the path availability as the product of the individual link availability metrics. This approach is however not applicable to the analytical model for blocking probability since path availability is determined as a function of time, whereas the fixed-point approximation is based on an equilibrium state. Also the assumption that links fail independently is not valid since it involves the movement of two nodes that have other links to other nodes.

[Miller01] provides a probability for the event that any two nodes are connected by a two-hop path and are not directly connected. The motivation for this study is that all node pairs that are connected by a single relay node (i.e. are two hops away) have the potential for creating a hidden terminal problem, which is described in section 3.3. [Miller01] also obtains an upper bound for the probability of an h-hop connection where $h \leq N-1$ in an N node network. The resulting expression is then applied to the calculation of an upper bound for the average hop distance in the network. The nodes are assumed to be randomly distributed with the x and y coordinates having independent zero-mean Gaussian distributions. After deriving the one-hop and two-hop connection probabilities, the asymptotic probability of an h-hop connection, is shown to be [Miller01]:

$$\mathrm{P}_h = e^{-(h-1)^2 D^2/4\sigma^2} - e^{-h^2 D^2/4\sigma^2} \tag{5-1}$$

where D is the transmission range of each node, and σ is the standard deviation of the x and y coordinates.

A possible procedure for incorporating the changes in topology which result in changes in the routes would be to consider all possible routes consisting of up to h-hops and then assigning a probability of existence for each link between all possible node pairs in the network. Since any possible topological configuration is possible if it is assumed that nodes move with equivalent mobility patterns, all $N(N-1)/2$ links would have the same probability of being in existence. However this would lead to an unnecessarily high blocking probability because each route acceptance probability needs to be multiplied by the product of the link existence probabilities of each link along the route.

Besides obtaining an unrealistically high blocking probability, there are too many possible routes between each node pair for every possible topological configuration for this method to be a practical solution. The number of routes between each node pair in an N node network, where the maximum length of a route is H intermediate hops, is given by

$$M = 1 + \sum_{h=2}^{H} \frac{(N-2)!}{(N-h-1)!} \quad ; H \leq N-1 \tag{5-2}$$

Equation (5-2) is proved in the Appendix. If the number of hops in a 50 node network is limited to 10 hops, then the number of possible routes between each node pair is 6.24×10^{14} routes, if all possible topologies are considered. Even in a moderate sized network of 10 nodes with a maximum route length of 5 hops, the number of possible routes between each of the 45 node pairs is 2081 routes.

Due to the above-mentioned problems the analytical model is limited to the evaluation of a fixed topology multihop wireless ad hoc network. The fixed topology allows only routes that exist to be evaluated.

5.3. Analytical Model

The fixed-point approximation described here is based on that proposed by [Liu00]. The approximation method involves solving simultaneously for four fixed-point variables using three mappings. Once solved iteratively, the fixed-point variables are used to calculate the blocking probability for each node pair.

5.3.1. Notation

The network consists of N nodes, indexed n and has a total of $R = N(N-1)/2$ node pairs, each indexed r. Between each node pair r there are a total of M_r routes listed in order from least cost (best route) to highest cost (worst route). The evaluation of the routes depends on the particular routing protocol that is under evaluation. The mth route between node pair r is referred to as (r, m) and consists of h_{rm} intermediate nodes. Calls arrive to a node pair with a Poisson arrival rate of λ_r and have an exponential mean holding time of $1/\mu$. The offered traffic between a node pair consisting of nodes a and b is the sum of the traffic from a to b and of traffic from b to a.

5.3.2. Fixed Point Method

There are four fixed point variables that are solved for simultaneously to evaluate the end-to-end blocking probability B_r of a node pair r. The arrival rate of calls at node n is v_n. The probability of admitting a call to node n is a_n. The steady state probability that node n is receiving k calls simultaneously is $p_n(k)$, where $0 < k < \infty$. The fourth unknown is q_{rm}, which is the probability that route (r, m) will be selected for a call between nodes of node pair r. Three mappings are used to solve simultaneously for the four fixed-point variables as follows:

i) a_n, $q_{rm} \rightarrow v_n$: the admissibility probabilities and route selection probabilities are held fixed to obtain the arrival rate at each node.

ii) $v_n \rightarrow a_n$, $p_n(k)$: the arrival rates are held fixed to calculate the admissibility probability and the occupancy probability of the nodes.

iii) $p_n(k) \rightarrow q_{rm}$: the occupancy probabilities are held fixed to obtain the route selection probability of each route between each node pair.

5.3.3. Mapping 1

The arrival rate of calls to be relayed at a node is also as a result of multiple routes between various node pairs that would be using the node as an intermediate node. Therefore the arrival rate at node n for calls to be relayed is:

$$v_n = \sum_{(r,m)} \lambda_r q_{rm} I[n \in (r,m)] \prod_{i \in (r,m), i \neq n} a_i \tag{5-3}$$

The indicator function is used in equation (5-3) because only routes that include the node being evaluated make a contribution to the arrival rate at the node. Equation (5-3) represents the first mapping and takes into consideration the arrival rate of each call request, the probability that the route will be used by the call request and the probability that the call request will be accepted at each of the other nodes making up the route.

5.3.4. Mapping 2

In section 4.6 the probability that a base station blocks a call is shown to be

$$\Pr\{\text{call blocked}\} = \Pr\{Z > G(1-\eta)\} \tag{5-4}$$

where there are k mobiles communicating with the base station under consideration and L mobiles communicating with other base stations. G is the processing gain of the CDMA system and $\eta = N_0/I_0$ is the ratio of the one-sided spectral noise density to the total acceptable interference density. Z_k is the sum of the total interference experienced by the base station due to the mobile users communicating with the base station (intracellular interference) and mobile users who are communicating with other base stations (intercellular interference). If the analogy is taken between a base station receiving transmissions from mobiles and a node receiving transmissions from other nodes, let

$$\begin{aligned} E_{n,k} &= \Pr\{\text{call request is accepted given } k \text{ calls in} \\ &\qquad \text{progress at node } n\} \\ &= \Pr\{Z_k \leq G(1-\eta) \end{aligned} \tag{5-5}$$

where Z_k is the total interference experienced by a node when it is processing k calls. The state of node n can be defined as $S_k^{(n)}$, $0 \leq k < \infty$.Two possible state transitions can occur as shown in Figure 5-2:

i) $S_k^{(n)} \cdot \quad S_{k+1}^{(n)}$ at a rate $k\mu$

ii) $S_k^{(n)} \cdot \quad S_{k-1}^{(n)}$ at a rate $v_n E_{n,k}$

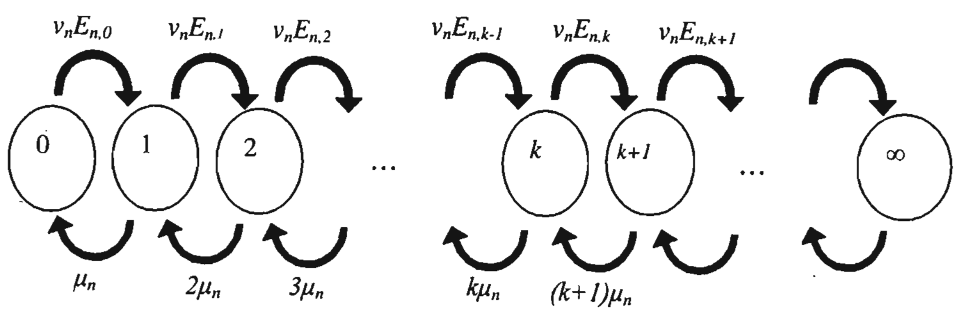

Figure 5-2. State transition model for number of calls in progress at a node

In order to evaluate $E_{n,k}$ the distribution of Z_k needs to be found. The E_b/I_0 ratio of a single user depends on the power control scheme that is used to equalize the performance of all users. Experimentation and field trials for cellular CDMA systems show that the E_b/I_0 ratio can be approximated by a log-normal distribution [Viterbi93], [Miller95]. Figure 5-3 [Viterbi93] was obtained from tests conducted with all cells fully loaded and E_b/I_0 being varied to maintain a frame error rate of less than 1%. A log-normal probability density with the mean of the normal exponent being 7dB and a standard deviation of 2.4 dB is used to approximate the histogram.

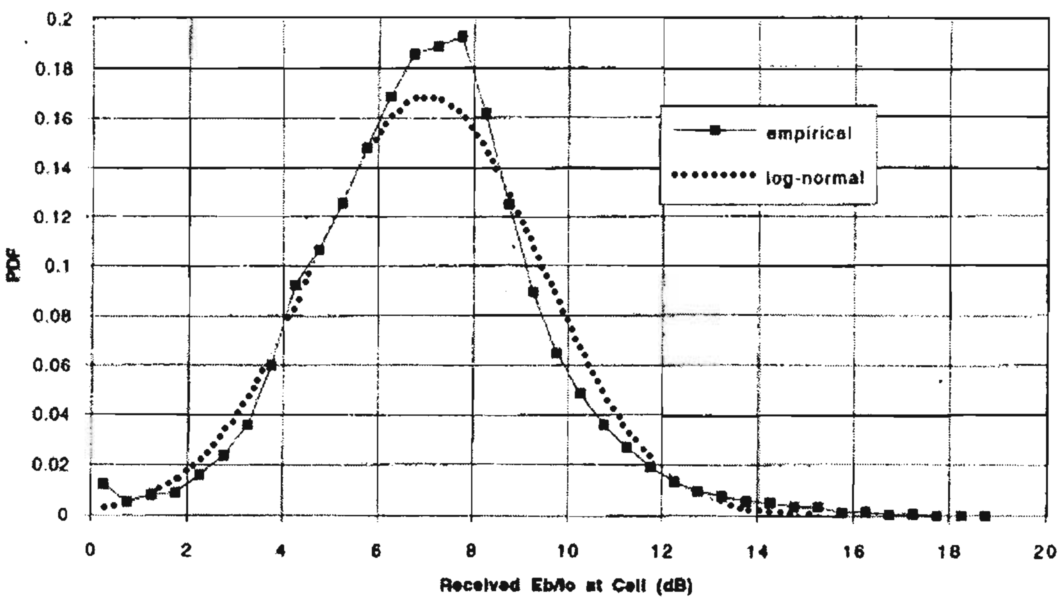

Figure 5-3. Probability density of E_b/I_o obtained empirically and its log-normal approximation with mean = 7 dB and standard deviation of 2.4 dB.[Viterbi93]

Since there are no field trials available for ad hoc networks the assumption is that the approximation will also follow a log-normal distribution in the case of ad hoc networks. Since Z_k is a sum of random variables, the Gaussian distribution can be used to estimate the value of $E_{n,k}$ by invoking the central limit theorem which states [Math02] that "the mean of any set of variates with any distribution having a finite mean and variance tends to the Gaussian distribution." This method of estimation has also been employed in [Miller95], [Viterbi95], [Narr99]. Therefore,

$$E_{n,k} = \int_{0}^{G(1-\eta)} \frac{1}{\sigma_{Zk}\sqrt{2\pi}} e^{\frac{(Z_k-\bar{Z}_k)^2}{2\sigma_{Zk}^2}} \tag{5-6}$$

where $\overline{Z_k}$ and $\sigma^2{}_{Zk}$ are the mean and variance of Z_k respectively. Equation (5-6) can be simplified using the error function *erf*(*x*) so that

$$E_{n,k} = \frac{1}{2} + \frac{1}{2} erf\left(\frac{G(1-\eta) - \overline{Z}_k}{\sigma_{Zk}\sqrt{2}}\right) \tag{5-7}$$

The error function *erf*(*x*) is given by :

$$erf(x) = \frac{2}{\sqrt{\pi}} \int_{0}^{x} e^{-t^2} dt \tag{5-8}$$

Since it has been assumed that Z_k is Gaussian due to power control inaccuracies being approximated by a log-normal distribution, [Viterbi93] shows that the mean and variance of Z_k can be simplified to

$$\overline{Z}_k = \mathrm{E}_n[k]\,\rho\, e^{\frac{(\beta\sigma)^2}{2}} \tag{5-9}$$

$$\sigma_{Zk} = \mathrm{E}_n[k]\,\rho\, e^{2(\beta\sigma)^2} \tag{5-10}$$

where $\mathrm{E}_n[k]$ is the expected number of calls being processed by node n, ρ is the probability of activity of a call, $\beta = (\ln 10)/10$ and $\sigma \approx 2.5$ dB is the standard deviation of the log-normal approximation of E_b/I_0. For the ad hoc network case, the expected number of calls being processed by node n is given by

$$\mathrm{E}_n[k] = \frac{v_n}{\mu} a_n \tag{5-11}$$

Once the value of $E_{n,k}$ is obtained, with reference to section 4.2 and equations 4-4 to 4-7, the birth-death model in Figure 5-2 is solved by

$$p_n(k) = A_n(k) p_n(0) \quad , 1 \le k < \infty \tag{5-12}$$

$$p_n(0) = \frac{A_n(0)}{\sum_{k=0}^{\infty} A_n(k)} \tag{5-13}$$

where the normalisation requirement is

$$\sum_{k=0}^{\infty} p_n(k) = 1 \tag{5-14}$$

and $A_n(k)$ and $A_n(0)$ are defined as

$$A_n(k) = \frac{(v_n E_{n,0})(v_n E_{n,1})(v_n E_{n,2})...(v_n E_{n,k-1})}{(\mu_n)(2\mu_n)(3\mu_n)...(k\mu_n)} \quad ,k > 0 \tag{5-15}$$

$$A_n(0) = 1 \tag{5-16}$$

Finally, the probability that a call will be admitted at node n is given by

$$a_n = \sum_{k=0}^{\infty} p_n(k)\left[E_{n,k}\right] \tag{5-17}$$

The second mapping therefore obtains the probability distribution of the calls in progress at node n, $p_n(k)$ and the probability of a call being admitted at a node n, a_n.

5.3.5. Mapping 3

The determination of the route acceptance probability q_{rm} depends on the routing protocol being implemented. A simple routing protocol was selected to focus on the development of the analytical model. To simplify the route determination procedure it is assumed that the available routes and their suitability are determined by traffic included in the parameters for offered traffic.

Once routes have been determined, the routes between each node pair are listed from shortest to longest. Both the length of the route and the congestion level at each intermediate node are taken into account and the route selected is the shortest route that has the least congestion on its most congested intermediate node.

Define the most congested node on route (r, m) as K_{rm}. The route selected is then the route that minimizes the cost function

$$w_1\varphi(K_{rm}) + w_2 h_{rm} \tag{5-18}$$

where w_1 and w_2 are weighting factors and $\varphi(K_{rm})$ is the number of calls being received by the most congested node on route (r, m).

The most congested node on a route (r, m) can be found from the expected number of calls being routed on the node as follows:

$$K_{rm} = \arg\max_{n\in(r,m)}(\mathrm{E}_n[k]) \tag{5-19}$$

Define the probability that there are at most k calls in progress at node n as

$$t_n(k) = \sum_{i=0}^{k} p_n(i) \tag{5-20}$$

The probability that a route (r, m) will be selected from M_r routes between node pair r due it being able to admit the call and minimise the cost function is given by

$$q_{rm} = \sum_{k=1}^{\infty} p_{Krm}(k)(E_{Krm,k})\prod_{\substack{i=1 \\ i\neq m}}^{Mr} t_{Krm}(\frac{w_2}{w_1}(h_{ri} - h_{rm}) + k) \tag{5-21}$$

5.3.6. Blocking Probability

Once the four fixed-point variables have been solved iteratively, the blocking probability for a call between a node pair r with M_r routes between them is given by

$$B_r = 1 - \sum_{m=1}^{Mr} q_{rm} \prod_{n\in(r,m)} a_n \tag{5-22}$$

5.4. Analytical Results

The analytical model described in section 5.3 was used to calculate the blocking probability for calls between the 45 node pairs in an ad hoc network containing 10 nodes. The topology with no node mobility is shown in Figure 5-4. The links shown are those that are assumed to be operational in a power controlled environment where a node communicates to only those nodes within a specified range. The area shown in Figure 5-4 is 1000m x 1000m with a maximum node transmission range of 400m. For all the results that follow the following constants were used: G = 32; η = 0.1; $1/\mu$ = 10s for all calls between all node pairs. Furthermore, define $w_{ratio} = w_2/w_1$, where w_2 and w_1 are the routing protocol weighting factors from equation 5-18.

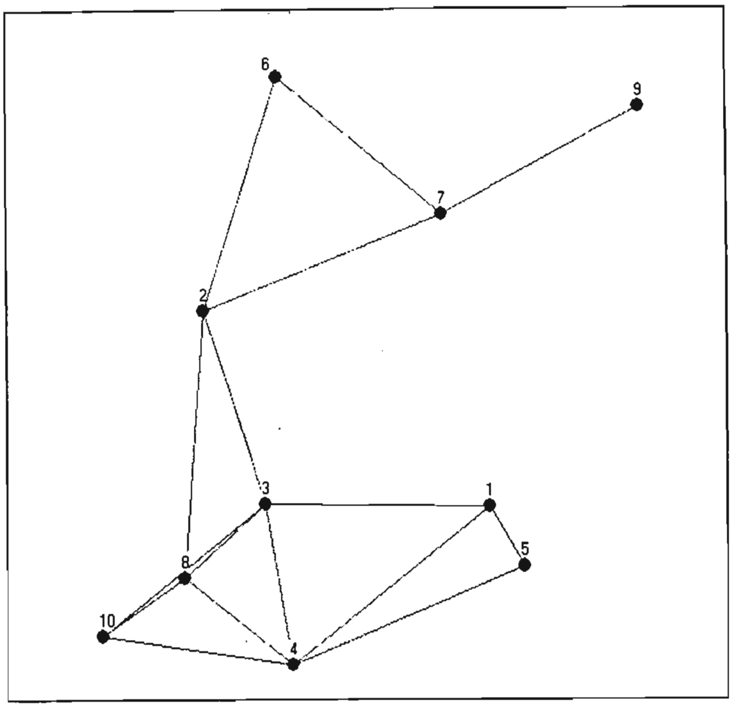

Figure 5-4. Network topology used to obtain results

The average blocking probability for the network for varying mean arrival rates per node pair is shown in Figure 5-5. The mean arrival rate of calls for each node pair is identical and the average blocking probability for the network is calculated as

$$B_{avg} = \frac{\sum_{r=1}^{R} B_r}{R} \qquad (5\text{-}23)$$

In Figure 5-5 the call activity probability ρ, labelled as "rho" in the graphs, was varied to investigate the effect of call activity probability on the blocking probability. The ratio of the weighting factors, w_{ratio} is 2. As expected, the larger the probability of activity of a call, the larger the blocking probability for the network, however the large difference in blocking probability for different ρ values is worthy of further consideration. In the analysis of a communications system, when a node is assumed to be transmitting with a certain probability while engaged in a call, it is important to able to adequately characterise the value of ρ based on experimental data. Due to the lack of experimental data for ad hoc networks, ρ = 0.45 is used for the results that follow. Another striking feature of the graph in Figure 5-5 are the oscillations that occur for a high probability of call activity. The analytical model becomes unstable for high blocking probabilities and it was found that this also occurs for other values of ρ when the mean call arrival rate is increased to obtain blocking probabilities in the 90 % region. This occurs because of the iterative process in the fixed-point approximation approximating the fixed-point variables near local minima instead of finding the absolute minima. However since the blocking probability of a network should not be greater than 2% for reasonable performance [Bertsekas92] the instability at high blocking probabilities it not a major concern. The 2% blocking for $\rho = 0.45$ occurs for a mean arrival rate of approximately 0.192 calls/s per node pair.

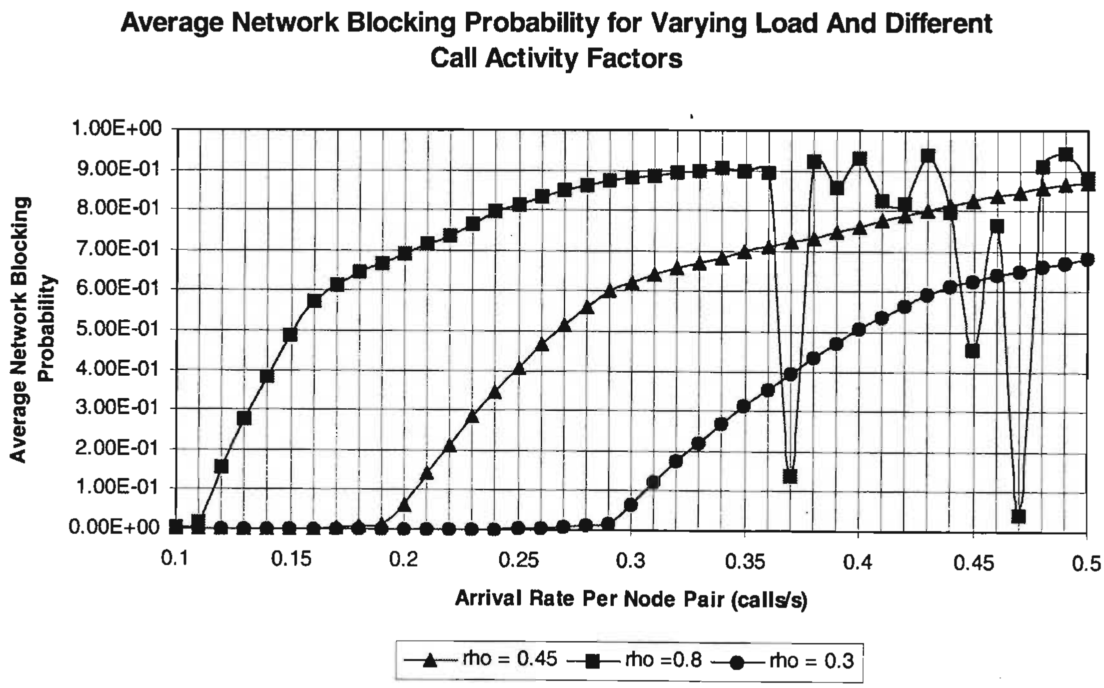

Figure 5-5. Average network blocking probability for varying load and different call activity factors

In Figure 5-6 the route weighting factor ratio, w_{ratio}, is varied to determine the optimum ratio for implementation in the simple routing protocol proposed in section 5.3.5. The mean arrival rate per node pair is fixed at 0.2 calls/s per node pair. From Figure 5-6, the optimal w_{ratio} is found to be 10, with a w_{ratio} of 3 satisfying the 2% blocking criteria for 0.2 calls/s per node pair. There is an order of magnitude difference in blocking probability between w_{ratio} values of 1 and 10.

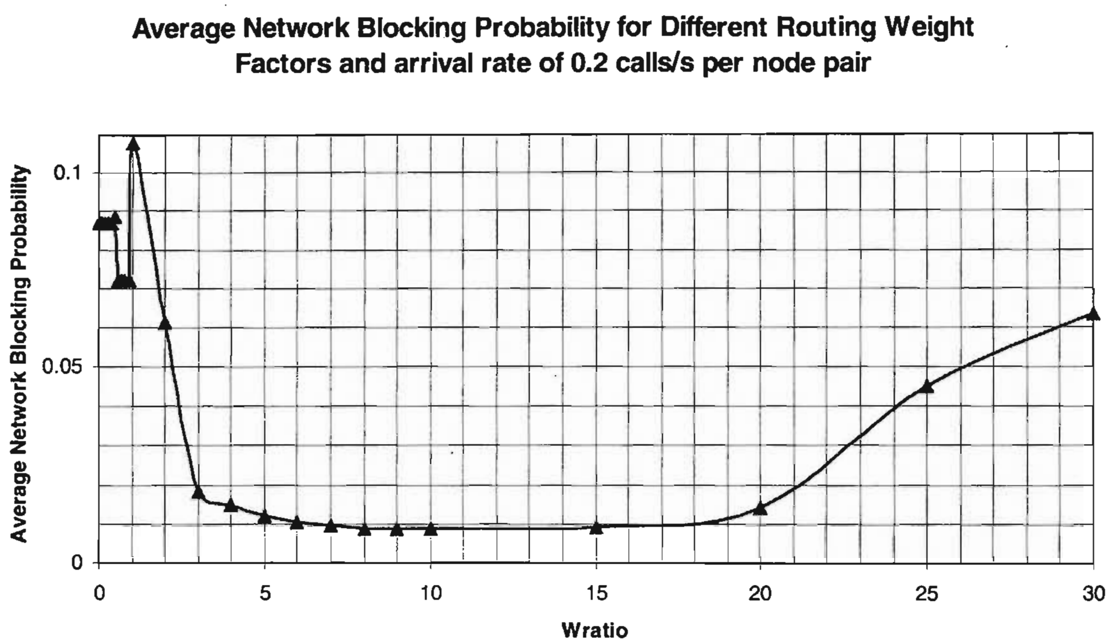

Figure 5-6. Average network blocking probability for varying routing weighting factors and a call arrival rate of 0.2 calls/s per node pair

Figure 5-7 shows the results obtained for w_{ratio} of 2 and for the optimal w_{ratio} of 10. The 2 % blocking criteria is achieved for mean arrival rates of less than 0.192 calls/s per node pair with w_{ratio} of 2 and less than 0.218 calls/s per node pair for w_{ratio} of 10.

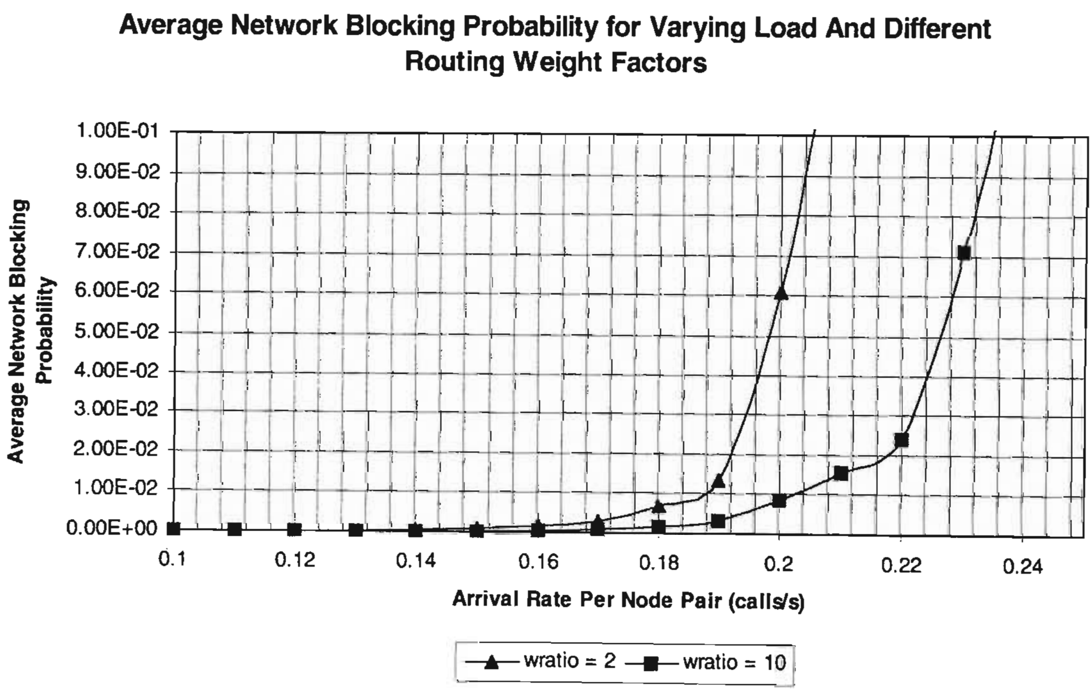

Figure 5-7. Average network blocking probability for varying arrival rates and different routing weight factors

Figure 5-8 is a comparison of the results obtained with the analytical model and results obtained used a modified version of the simulator described in Chapter 3. The mobility model was removed and each node was assigned a fixed location as in Figure 5-4. The MAC model was also removed to allow nodes to transmit simultaneously and it was assumed that each node was able to receive multiple calls simultaneously. The traffic model was adjusted to incorporate the circuit switched model with calls arriving at arrival rate of λ calls/s per node pair with a mean call duration of $1/\mu$ = 10s. Calls are blocked if at least one node along each route available is experiencing interference greater than the critical limit. The routing protocol described in Section 5.3.5 was implemented and the following constants were used for the comparison: $G = 32$; $\eta = 0.1$; $w_{ratio} = 10$; $\rho = 0.45$.

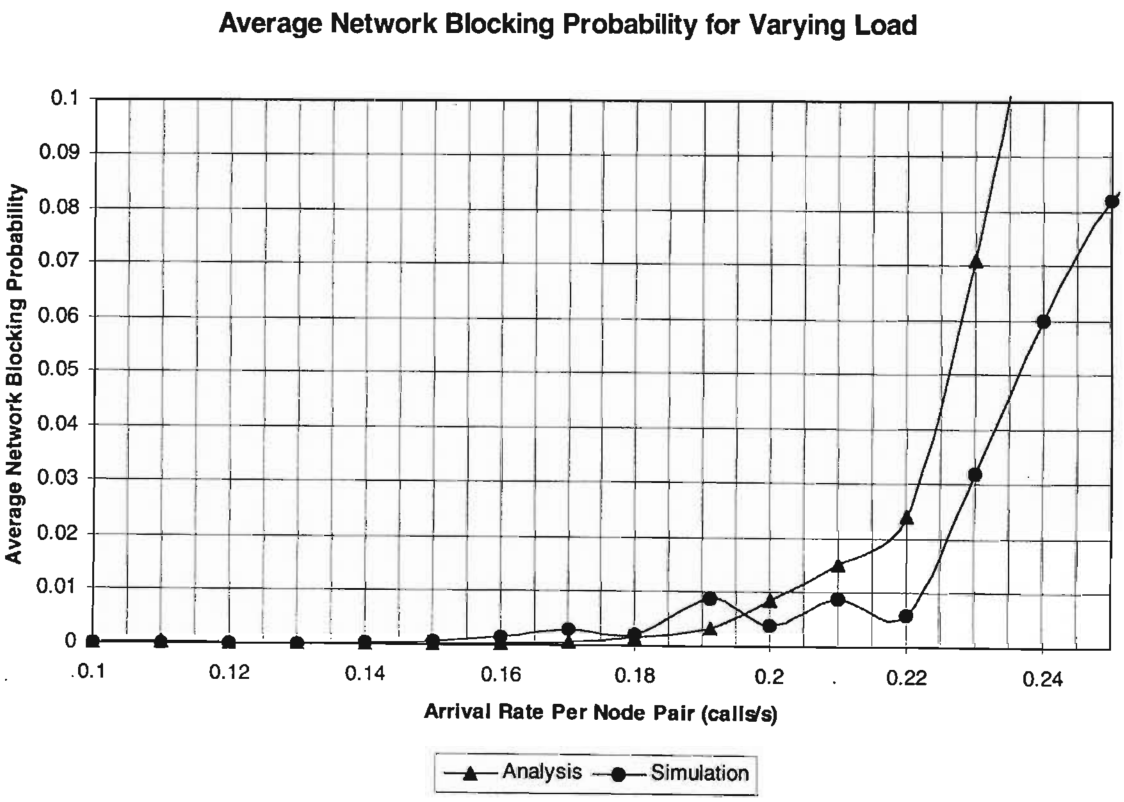

Figure 5-8. Comparison of analytical results with simulation results

The 2% blocking criteria is achieved with an arrival rate of less than 0.226 calls/s per node pair for the simulation and less than 0.217 calls/s per node pair for the analytical model, which is an error of 3.9%. For blocking below the 2% level the simulation results appear to oscillate between 0.18 calls/s per node and 0.22 call/s per node. While it is difficult to explain this phenomenon, the errors obtained when comparing the analytical results to simulations can most likely be attributed to the assumptions made with regard to the signal propagation and the use of a Gaussian distribution for the estimation of the interference distribution.

Figure 5-9 and Figure 5-10 show the stationary distribution of calls in progress at nodes 3 and 9 for mean arrival rates of 0.1 and 0.9 calls/s per node pair respectively, which are obtained by plotting $p_n(k)$ for $n \in \{3,9\}$ and $0<k<100$. The mean of the stationary distribution for node 9 increases with increased mean arrival rate due to there being more calls in progress. However due to the location of node 3, the mean of the stationary distribution for node 3 decreases due to higher interference from neighbouring nodes. Note that node 9 does not experience increased interference with increased traffic due to its location. This implies that as the number of calls in progress in the network increases, the capacity of nodes experiencing high traffic will decrease. The stationary distribution plots can therefore be used as an indication of maximum capacity depending on the network arrival rates.

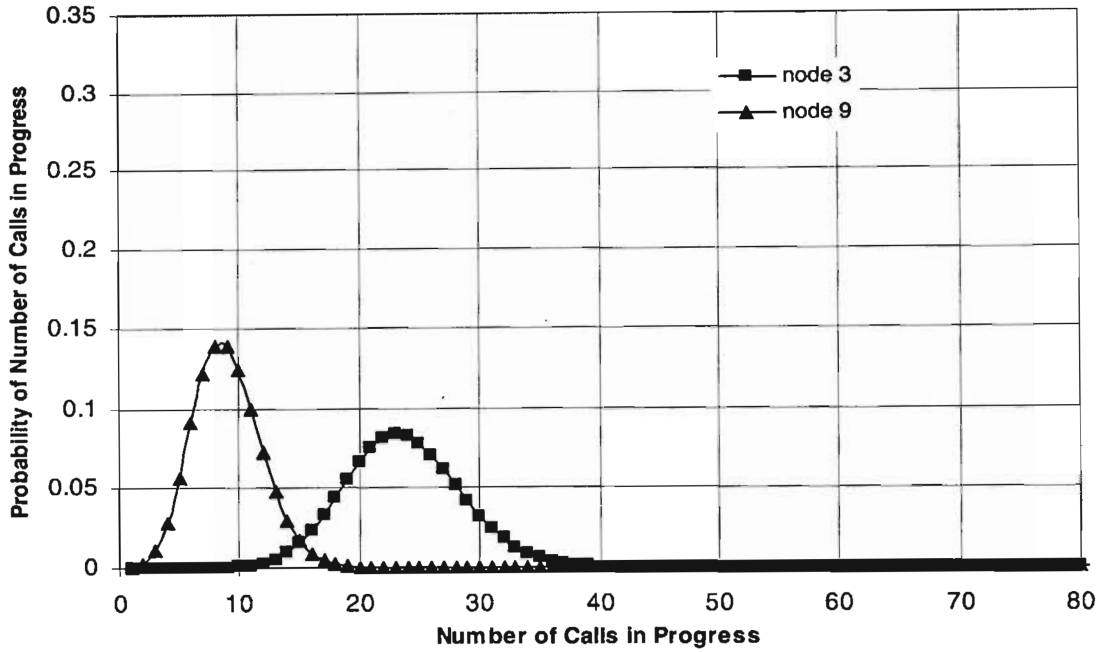

Figure 5-9. Stationary distribution of calls in progress at nodes 3 and 9 for arrival rate of 0.1 calls/s per node pair

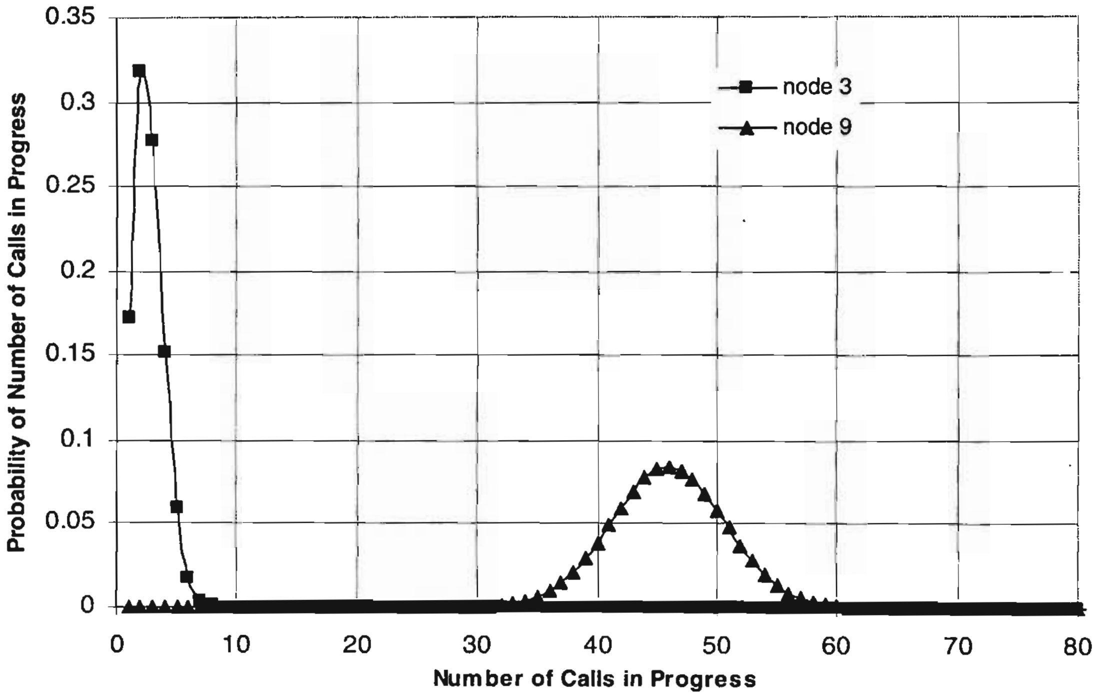

Figure 5-10. Stationary distribution of calls in progress at nodes 3 and 9 for arrival rate of 0.9 calls/s per node pair

5.5. Summary

This chapter describes an analytical model for the evaluation of a routing protocol's resultant blocking probability in a CDMA ad hoc network. The ad hoc network is modelled as a loss network and a fixed-point approximation proposed by Liu[00] is modified and combined with models for evaluating the Erlang capacity of cellular CDMA systems. As far as the author is aware, there is no literature available on the analysis of ad hoc networks using these methods.

The problem of continual topology changes in an ad hoc network has not been overcome in the development of the analytical model. Section 5.2 discusses the reasons for the difficulty in taking into consideration changes in topology and describes the reasons for the impracticality of considering all possible topological configurations.

The analytical model is detailed in Section 5.3 and involves using three mappings to iteratively solve for 4 fixed-point variables in order to find the blocking probability between each node pair. A simple routing protocol is proposed for the development of the analytical model in Section 5.3.5. The analytical model was used to evaluate a multihop fixed topology ad hoc network and the results are presented in Section 5.4. The results obtained show that the call activity probability is an important modelling parameter. The results are also used to find the optimum route weighting factor ratio so that the performance of the routing protocol is optimised.

Results obtained by comparing a simulation of the network to the analytical results are discussed. The results compare favourably with an error at the 2% blocking criterion of 3.9%. Finally, the call probability distributions for selected nodes are plotted and it is shown that these plots can be used to determine the steady state capacity of nodes in the network.

Chapter 6

CONCLUSIONS AND FUTURE WORK

6.1. Dissertation Summary

Ad hoc networks consist of mobile nodes that communicate wirelessly and achieve peer-level connectivity by means of multihop paths. These networks are therefore distributed and infrastructureless. The main focus of this dissertation has been on the performance and evaluation of routing protocols for wireless ad hoc networks. This work has been prompted by the potential for ad hoc networks to provide ubiquitous computing and communication. The applications for ad hoc networks vary from implementation in large-scale military scenarios to small personal area networks for commercial mobile computing and Internet access. Among the issues that need to be solved is the provision of adequate distributed and robust routing protocols that are able to cope with rapidly changing topologies, high error rates and limited power supplies.

The routing protocols are classified according to the method in which routes are obtained. Routing protocols that maintain constantly updated tables are referred to proactive routing protocols, while those that attempt to find routes only when routes are required are referred to as reactive protocols. Comparative simulations conducted by various researchers have shown that the reactive protocols are better suited to the highly mobile environment of ad hoc networks. The proactive protocols are not able to cope with constantly updating the tables and actually trigger more congestion in the network. The proactive protocols do have their merits in small networks, but as the size of the network grows, it becomes unfeasible to maintain global topological views.

While the reactive protocols are able to cope better with mobility, there is a greater delay in providing routes due to the route request procedure that is only initiated when a route is required. Another disadvantage of reactive routing is that the entire network needs to be searched before a route is found, which makes the route discovery procedure very costly. The inclusion of route caches for nodes using reactive protocols could be useful, but the disadvantage is that stale routes represent a liability that may degrade performance rather than improve it.

Various modifications have been made to the basic reactive and proactive routing protocols, which include special adaptations that are unique to ad hoc networking. By using physical layer properties such as signal quality and spatial properties such as mobility prediction, routes can be chosen such that they are more reliable, with the cost of the route request procedure being decreased.

Hybrid varieties of the proactive and reactive routing protocols employ the table-driven approach locally and the on demand approach globally, thereby attempting to exploit the advantages of both classifications of routing. The hierarchical routing protocols attempt to mimic the fixed infrastructure of cellular networks, with the aim of simplifying medium access control, code assignment, and routing functions. However selected nodes become overburdened and resources are consumed unfairly in the network

All the routing protocols have advantages and disadvantages, and the particular application for the ad hoc network would determine what type of routing protocol to implement and whether or not higher cost and greater computational complexity can be justified.

This dissertation proposes a new routing protocol based on DLAR [Gerla00], which is a load balancing routing protocol. The new routing protocol, referred to as DLAR 4 is unique because it combines load balancing with signal quality determination, which allows the routes selected to be of better quality, hence providing higher throughput. Simulations are used to compare the protocol to the original DLAR schemes and to another prominent reactive routing protocol, AODV [Perkins99].

Results for packet delivery ratio and average end-to-end delay are obtained and show the new protocol performs better than the other protocols compared in all scenarios where there is high mobility and high offered load. This shows that load balancing in ad hoc networks is beneficial under certain circumstances. However, once again the computational cost of implementing more complex routing protocols need to be weighed against the benefits. Under low offered loads, the less complex AODV routing protocol performed almost as well as the DLAR protocols.

This dissertation also proposes an analytical model for the evaluation of the blocking probability of a call offered to an ad hoc network that results from a particular routing protocol being implemented. All of the comparisons of the routing protocols are performed using simulators, which are not only time consuming but also do not allow as rapid optimisation of routing protocols by a network designer as would analytical models. Techniques for modelling fixed networks as loss networks are combined with those for the evaluation of cellular CDMA networks. CDMA was chosen because CDMA systems offer high spectrum efficiency and the potential use of advanced antenna and receiver structures. No previous work has been published on the type of full-scale analytical evaluation of ad hoc networks that is consider in this dissertation. Numerical results obtained provide a 3.9 % error as compared to simulation results of a similar network, with the analytical model providing a more conservative estimate.

While the results of the analytical model compare favourably to simulation results, a major shortcoming of the analytical model proposed is the inability to account for mobility. Due to the potential number of topological configurations possible, all routes cannot be accounted for. The analytical model therefore is only capable of evaluating fixed wireless multihop ad hoc networks. Networks of this sort include wireless local loop extensions to the fixed wired telephone network in rural areas. Methods for evaluating and optimising routing protocols using the analytical model are demonstrated.

6.2. Future Work

6.2.1. Load Balancing and Cross-Layer Adaptations

The simulations results in Chapter 3 show that load balancing and signal monitoring are advantageous to the performance of the routing protocol. Other methods of monitoring the environment by allowing the interaction of various layers of the protocol stack can be investigated. For example, the medium access control protocol has already been used in some simulations discussed in the literature to provide the routing protocol with connectivity information. Route selection can benefit from information gained due to these interactions.

6.2.2. Simulator Mobility Models

Software simulators are used in an attempt to simulate real life statistics. However, the assumptions that are made, such as arrival statistics, service statistics, propagation models and mobility models, are questionable. A potential area of investigation is in the determination of realistic mobility models that would apply to particular application scenarios. While the random movement models do provide certain insights in the behaviour of the protocols, the next step is to obtain actual user mobility traces so that the routing protocols can be evaluated in a more realistic scenario. Experiments could also provide the required physical layer propagation models that would allow the simulator to consider more realistic wireless environments. While too much detail results in slow and cumbersome simulators, simulations that lack necessary detail can result in misleading or incorrect results.

6.2.3. Analytical Model Assumptions

The Poisson arrival is the fundamental assumption underlying the Erlang Formula. The reduced load approximation and many other strategies for evaluating networks rely on this assumption. The importance of this assumption is that the distribution of the call duration only shows through its mean value in the blocking probability. It is foreseeable that with long holding times in data networks (which implies infinite variance), it could take infinitely long for a system to reach steady state. The traditional steady state assumption is therefore worth investigating.

The adaptation of this assumption to ad hoc networks also needs to be investigated further. Accounting for topology changes has not been included in the analytical model because it is currently difficult to determine the steady state of an ad hoc network. Further work would involve overcoming this hurdle and incorporating mobility in the analytical model. Routing protocols would then be easily evaluated and optimised.

The interference distribution has been modelled as Gaussian in keeping with the results obtained from field trials with cellular CDMA networks. In order to determine whether these estimations and assumptions are correct, experiments need to be conducted to determine the physical layer attributes of wireless multihop networks.

REFERENCES

[Basagni98] Basagni, S, Chlamtac, I, Syrotiuk, V, Woodward, B, "A Distance Routing Effect Algorithm for Mobility (DREAM)", *Proceedings of the Fourth Annual ACM/IEEE International Conference on Mobile Computing and Networking, MobiCom'98*, Oct. 1998, pp. 76-84.

[Basagni99] Basagni, S, Chlamtac, I, Syrotiuk, V, "Geographic Messaging in Wireless Ad Hoc Networks", *Proceedings of the 49th IEEE International Vehicular Technology Conference, VTC'99, May 1999,* Vol. 3, pp. 1957-1961.

[Bertsekas92] Bertsekas, D, Gallager, R, "Data Networks", *2nd Edition, Prentice Hall Inc.*, 1992.

[Bluetooth] Bluetooth Special Interest Group (SIG) [Online] http://www.bluetooth.com [Last Accessed: November 2002]

[Broch98] Broch, J, Maltz, D, Johnson, D, Hu, Y-C, Jetcheva, J, "A Performance Comparison of Multihop Wireless Ad Hoc Network Routing Protocols", *ACM/IEEE International Conference Mobile Computing & Networking (MobiCom98)*, 1998, pp. 85-97.

[Camp02] Camp, T, et al, "Performance Comparison of Two Location Based Routing Protocols for Ad Hoc Networks", *Proceedings of IEEE INFOCOM*, 2002.

[Chen99] Chen, S, Nahrstedt, K, "Distributed Quality-of-Service Routing in Ad Hoc Networks," *IEEE Journal on Special Areas in Communications, Special Issue on Wireless Ad Hoc Networks*, vol.17, no.8, Aug. 1999, pp.1488-1505.

[Chiang97] Chiang, C, Wu, H, Liu, W, Gerla, M, "Routing in Clustered Multihop, Mobile Wireless Networks with Fading Channel", *Proceedings of IEEE Singapore International Conference on Networking (SICON '97),* Apr. 1997, pp.197-211.

[Chiang98] Chiang, C, "Wireless Network Multicasting", PhD Thesis, Department of Computer Science, University of California, Los Angeles, 1998.

[Chiang99] Chiang, C, Lee, S, Gerla, M, "On-Demand Multicast Routing Protocol", *Proceedings of IEEE WCNC'99*, Oct.1999, pp.1298-1304.

[Colbourn87] Colbourn, C, "The Combinatorics of Network Reliability", *Oxford University Press*, London, 1987.

[Cotter02] Cotter, P, "The Role of Retrans in the IBCT", *Military Review May-June 2002*, June 2002, pp 50. [Online] http://www.cgsc.army.mil/milrev/english/MayJun02/cotter.asp (Last Accessed: August 2002)

[CSDMag00] CSDMag, "HomeRF : Wireless Networking for the Connected Home", [Online] http://www.csdmag.com/main/2000/02/0002home.htm, (Last Accessed: November 2002)

[Das01] Das, S, Perkins, C, Royer, E, Marina, M, "Performance Comparison of Two On-Demand Routing Protocols", *IEEE Personal Communications,* Feb. 2001, pp. 16-28.

[Davies00] Davies, V, "Evaluating Mobility Models Within an Ad Hoc Network", M.Sc. Thesis, Department of Mathematical and Computer Sciences, Colorado School of Mines, Colorado, 2000.

[Dube97] Dube, R, Rais, C, Wang, K-Y, Tripathi S, "Signal Stability Based Adaptive Routing For Ad Hoc Mobile Networks", *IEEE Personal Communications*, Feb. 1997, pp. 36-45.

[Esta84] Estahanian, A, Hakimi, S, "On Computing the Connectivity of Graphs and Digraphs", *Networks*, vol. 14, 1984, pp 355-366.

[Gerla98] Gerla, M, Chen, T-W, "Global State Routing: A New Routing Scheme for Ad hoc Wireless Networks," *IEEE International Conference on Communications (ICC 98)*, Jun. 1998, pp171-175,

[Gerla99] Gerla, M, Lee, S, Toh C-K, "A Simulation Study of Table-Driven and On-Demand Routing Protocols for Ad Hoc Networks", *IEEE Network,* vol. 13, no. 4, Aug. 1999, pp. 48-54.

[Gerla00] Gerla, M, Lee, S, "Dynamic Load Aware Routing in Ad Hoc Networks," *Technical Report, Computer Science Department, University of California, Los*

Angeles, Aug. 2000.

[Gilhousen91] Gilhousen, K, Jacobs, M, Radovani, R, Viterbi, A, Weaver, L, Wheatly, C, "On the Capacity of a Cellular CDMA System", *IEEE Transactions on Vehicular Technology,* vol. 40, May 1991.

[Gugrajah01] Gugrajah, Y, Takawira, F, "Performance of Wireless Ad-Hoc Network Routing Protocols Using Load and Capacity Metrics", *Proceedings of the South African Telecommunications Networks and Applications Conference (SATNAC 2001),* Sep. 2001.

[Gugrajah02a] Gugrajah, Y, Takawira, F, "Analytical Model for Evaluating Blocking Probability in Wireless Ad Hoc Networks", *Proceedings of IEEE Africon 2002 Conference,* Volume 1, Oct. 2002, pp. 277-282.

[Gugrajah02b] Gugrajah, Y, Takawira, F, "Blocking Probability Evaluation in Multihop CDMA Networks", *Proceedings of the South African Telecommunications Networks and Applications Conference (SATNAC 2001),* Sep. 2002.

[Haas97] Haas, Z, "A New Routing Protocol For Reconfigurable Wireless Networks", *Proc. IEEE International Conference on Universal Personal Communications (ICUPC),* Oct. 1997, pp. 562-565.

[Haas99] Haas,J, Pearlman, M, "Determining the Optimal Configuration for the Zone Routing Protocol," *IEEE Journal on Selected Areas in Communications, Special Issue on Wireless Ad Hoc Networks,* vol.17, no.8, Aug. 1999, pp.1395-1414.

[Haartsen98] Haartsen, J, "Bluetooth – The Universal Radio Interface for Ad Hoc Wireless Connectivity", *Ericsson Review,* vol. 75, no. 3, 1998, pp. 110-117.

[Hong99] Hong, X, Gerla, M, Pei, G, Chiang, C, "A Group Mobility Model for Ad Hoc Wireless Networks", *Proceedings of the ACM International Workshop on Modeling and Simulation of Wireless and Mobile Systems,* Aug. 1999.

[IEEE99] IEEE, "Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications", *IEEE 802.11-1999,* 1999.

[IrDA] Infrared Data Association, "Technical Summary of IrDA DATA and IrDA

CONTROL", [Online] http://www.irda.org/standards/standards.asp (Last Accessed: November 2002)

[Iwata99] Iwata, A, et al, "Scalable Routing Strategies for Ad Hoc Wireless Networks," *IEEE Journal on Selected Areas in Communications, Special Issue on Wireless Ad Hoc Networks*, vol.17, no.8, Aug. 1999, pp.1369-1379.

[Jacquet98] Jacquet, P, Muhlethaler, P, Qayyum, A, "Optimized Link State Routing Protocol," *IETF MANET*, Internet Draft, draft-ietf-manet-olsr-98.txt, Nov. 1998.

[Jan01] Janaswamy, R, "Radiowave Propagation and Smart Antennas for Wireless Communications", *Kluwer Academic Publishers,* 2001, pp. 22.

[Joa99] Joa-Ng, M, Lu, I-T, "A Peer-to-Peer Zone-Based Two-Level Link State Routing for Mobile Ad Hoc Networks," *IEEE Journal on Selected Areas in Communications, Special Issue on Wireless Ad Hoc Networks*, vol.17, no.8, Aug. 1999, pp.1415-1425.

[Johannson99] Johannson, P, Larsson, T, Hedman, N, "Scenario-based Performance Analysis of Routing Protocols for Mobile Ad Hoc Networks", *Proceedings of the IEEE/ACM International Conference on Mobile Computing and Networking, MOBICOM'99*, Aug. 1999.

[Johnson95] Johnson, D, "Routing in Ad Hoc Networks of Mobile Hosts", *Proceedings of the IEEE Workshop on Mobile Computing Systems and Applications*, Dec. 1994

[Johnson96] Johnson, D, Maltz, D, "Dynamic Source Routing in Ad Hoc Wireless Networks", Mobile Computing, *Kluwer Academic Publishers,* 1996, pp. 153-181.

[Kaufman81] Kaufman, J, "Blocking in a Shared Resource Environment", *IEEE Transactions on Communications,* vol. 29, Oct. 1981. pp. 1474-1481.

[Kelly86] Kelly, F. P, "Blocking in Large Circuit-Switched Networks", *Advances in Applied Probability*, vol. 18, 1986. pp. 473-505.

[Kelly91] Kelly, F. P, "Loss Networks", *Annals of Applied Probability*, vol. 1, no. 3, 1991, pp. 319-378.

[Kelly94] Kelly, F. P, Louth, G, Mitzenmacher, M, "Computational Complexity of Loss Networks", *Theoretical Computer Science*, 1994. pp. 45-59.

[Kelly95] Kelly, F. P, "Mathematical Models of Multiservice Networks", *Complex Stochastic Systems and Engineering, Oxford University Press*, 1995. pp. 221-234.

[Ko97] Ko, Y-B, Vaidya, N, "Using Location Information to Improve Routing in Ad Hoc Networks", *Technical Report 97-013, Department of Computer Science, Texas A&M University,* Dec. 97.

[Ko98] Ko, Y-B, Vaidya, N, "Location-Aided Routing (LAR) In Mobile Ad Hoc Networks", *ACM/IEEE MobiCom 98,* Oct. 1998. pp. 66-75.

[Ko99] Ko, Y-B, Vaidya, N, "Geocasting in Mobile Ad Hoc Networks: Location Based Multicast Algorithms", *Proceedings of WMCSA*, 1999.

[Lee00] Lee, S, Kim, C, "Neighbor Supporting Ad Hoc Multicast Routing Protocol", *First Annual Workshop on Mobile and Ad Hoc Networking and Computing*, 2000, pp. 37-44.

[Lin99] Lin, R, Liu, J, "QoS Routing in Ad Hoc Wireless Networks," *IEEE Journal on Selected Areas in Communications, Special Issue on Wireless Ad Hoc Networks*, vol.17, no.8, Aug. 1999, pp.1426-1438.

[Liu00] Liu, M, "Network Performance Modelling, Design and Dimensioning Methodologies", *Ph.D. Thesis, Faculty of the Graduate School of the University of Maryland, College Park,* 2000.

[Madruga99] Madruga, E, Garcia-Luna-Aceves, J, "A Multicast Routing Protocol for Ad Hoc Networks", *Proc. Joint Conference of the IEEE Computer and Communication Societies (INFOCOM),* 1999, pp. 784-792.

[Man02] IETF Mobile Ad Hoc Network Charter, [Online] http://www.ietf.org/html.charters/manet-charter.html (Last Accessed: November 2002)

[Math02] "Gaussian Distribution from Mathworld", [Online] http://mathworld.wolfram.com/GaussianDistribution.html (Last Accessed: November 2002)

[McDonald99] McDonald, A. B, Znati, T, "Link Availability Models for Mobile Ad

Hoc Networks", *Technical Report TR99-07, Department of Computer Science, University of Pittsburgh,* 1999.

[Miller95] Miller, L, Lee, J, "On the Erlang Capacity of CDMA Cellular Systems", *IEEE GlobeCom '95,* 1995.

[Miller01] Miller, L, "Probability of a Two-Hop Connection in a Random Mobile Network", *35th Conference on Information Sciences and Systems*, Mar. 2001

[Murthy95] Murthy, S, Garcia-Luna-Aceves, J, "A Routing Protocol for Packet Radio Networks," *Proceedings of ACM Mobile Computing and Networking Conference, MOBICOM'95,* Nov. 1995.

[Navas97] Navas, J, Imielinski, T, "GeoCast-Geographic Addressing and Routing", *Proceedings of ACM Mobile Computing and Networking Conference, MOBICOM'97,*1997, pp. 66-76.

[Narr99] Narrainen, R, "Call Admission Control in CDMA Cellular Networks," *Ph.D. Thesis, University of Natal,* 1999.

[Ns02] ns2 website [Online] http://www.isi.edu/nsnam/ns/ (Last Accessed : November 2002)

[Park97] Park, V, Corson, M, "A Highly Adaptive Distributed Routing Algorithm for Mobile Wireless Networks," *IEEE INFOCOM '97*, 1997

[Perkins94] Perkins, C, Bhagwat, P, "Highly Dynamic Destination-Sequenced Distance Vector Routing (DSDV) for Mobile Computers", *ACM SIGCOMM: Computer Communications Review, vol. 24, no.4,* Oct. 1994, pp. 234-244.

[Perkins98] Perkins, C, "Mobile IP, Design Principles and Practices", *Addison Wesley Longman*, 1998.

[Perkins99] Perkins, C, Royer, E, "Ad-hoc On-Demand Distance Vector Routing," *Second IEEE Workshop on Mobile Computing Systems and Applications*, Feb. 1999, pp.90-100.

[Ram02] Ramanathan, R, Redi, J, "A Brief Overview of Ad Hoc Networks:

Challenges and Directions", *IEEE Communications Magazine,* May 2002, pp 20-22.

[Ross93] Ross, K.W, "Reduced Load Approximations for Multirate Loss Networks", *IEEE Transactions on Communications,* vol. 41, no. 8, Aug. 1993, pp. 1222-1231.

[Ross95] Ross, K.W, "Multiservice Loss Models for Broadband Telecommunication Networks", *Springer-Verlag London Limited,* 1995, pp. vii-viii

[Royer99] Royer, E.M, Perkins, C, "Multicast operation of ad-hoc, on-demand distance vector routing protocol," *ACM/IEEE MobiCom '99,* Aug. 1999.

[Royer01] Royer, E.M, Melliar-Smith, P.M, Moser, L, "An Analysis Of The Optimum Node Density for Ad Hoc Mobile Networks", *Proceedings of the IEEE International Conference on Communications (ICC),* 2001.

[Sanchez98] Sanchez, M, "Mobility Models", [Online] http://www.geocities.com/SiliconValley/Campus/4251/mobmodel.htm (Last Accessed : November 2002)

[Shukla01] Shukla, D, Iyer, S, "Mobility Models in Ad Hoc Networks", [Online] www.it.iitb.ac.in/~deepanshu/downloads/MobilityModels.ps (Last Accessed: November 2002)

[Sinha99] Sinha, P, Sivakumar, R, Bharghavan, V, "MCEDAR: Multicast Core-Extraction Distributed Ad hoc Routing", *Wireless Communications and Networking Conference*, 1999, pp.1313-1318

[Siva97] Sivakumar, R, Das, B, Bharghavan, V, "Routing in Ad Hoc Networks Using Minimum Connected Dominating Sets", *IEEE International Conference on Communications (ICC 97),* Jun. 1997.

[Siva98] Sivakumar, R, Das, B, Bharghavan, V, "Spine Routing in Ad Hoc Networks", *ACM/Baltzer Cluster Computing Journal,* 1998.

[Siva99] Sivakumar, R, Sinha, P, Bharghavan, V, "CEDAR: A Core-Extraction Distributed Ad Hoc Routing Algorithm," *IEEE Journal on Selected Areas in Communications, Special Issue on Wireless Ad Hoc Networks*, vol.17, no.8, Aug. 1999, pp.1454-1465.

[Talpade98] Talpade, R, Bornmaiah, S, Liu, M, McAuley, A, "AMRoute: Ad Hoc

Multicast Routing Protocol," *Internet Draft,* [Online] http://www.ietf.org/internet-drafts/draft-talpade-manet-amroute-00.txt, 1998, Work in Progress (Last Accessed: August 2001)

[Toh96] Toh, C-K, "A Novel Distributed Routing Protocol To Support Ad Hoc Mobile Computing", *Proceedings IEEE 15thAnnual International Conference on Computers and Communication*, Mar. 1996, pp. 480-486.

[Toh99] Toh, C-K, "Long-Lived Ad-Hoc Routing Based on the Concept of Associativity", *IETF Draft* March 1999, [Online] http://www.ietf.org/internet-drafts/draft-ietf-manet-longlived-adhoc-routing-00.txt (Last Accessed: June 2001)

[Tolety99] Tolety, V, "Load Reduction In Ad Hoc Networks Using Mobile Servers", M.Sc. Thesis, Department of Mathematical and Computer Sciences, Colorado School of Mines, Colorado, 1999.

[Viterbi93] Viterbi, A. M, Viterbi, A. J, "Erlang Capacity of a Power Controlled CDMA System", *IEEE Journal of Selected Areas in Communications,* vol. 11, no. 6. Aug. 1993. pp. 892-900.

[Viterbi95] Viterbi, A. M, "CDMA Principles of Spread Spectrum Communication", *Addison Wesley Publishing Company,* 1995.

[Want92] Want, R, Hopper, A, "The Active Badge Location System", *ACM Transactions on Information Systems,* Jan. 1992, pp. 91-102.

[Wu99] Wu, C, Tay, Y, Toh, C-K, "Ad Hoc Multicast Routing protocol utilizing Increasing id-numberS (AMRIS) Functional Specification," *Internet Draft,* [Online] http://www.ietf.org/internet-drafts/draft-ietf-manet-amris-spec-00.txt, Nov. 1998, Work in Progress. (Last Accessed: August 2001).

[Zachary99] Zachary, S, Ziediens, I, "Loss Networks and Markov Random Fields", *Journal of Applied Probability*, vol. 36, 1999.

[Zonoozi97] Zonoozi, M, Dassanayake, P, "User Mobility Modeling and Characterization of Mobility Patterns", *IEEE Journal on Selected Areas in Communications,* Vol. 15, No. 7, Sept. 1997, pp. 1239-1252.

Appendix

Lemma: The number of routes between each node pair in an N node network, where the maximum length of a route is H intermediate hops, is given by

$$M = 1 + \sum_{h=2}^{H} \frac{(N-2)!}{(N-h-1)!} \quad ; \; H \leq N-1$$

Proof: The number of permutations of a distinct objects is

$$a! = a(a\text{-}1)(a\text{-}2)\ldots(3)(2)(1)$$

The number of permutations of a distinct objects taken b at a time is

$${}_aP_b = \frac{a!}{(a-b)!}$$

For an ad hoc network with N nodes the possible routes between a node pair are composed of intermediate nodes, which can be any combination of the nodes remaining in the network. Therefore a = N-2. If there are h hops between the node pair on a route under consideration, there are b = h-1 ways that the intermediate nodes can be arranged. The number of h hop routes between a node pair is then

$${}_{N-2}P_{h-1} = \frac{(N-2)!}{(N-h-1)!}$$

The maximum number of hops that can be taken is $H \leq N\text{-}1$ which will be routes that include all nodes in the network. To consider routes of all lengths, the sum of the number of routes of each length needs to be considered, remembering that the single hop route has been neglected in the above calculations.

Hence the number of routes between each node pair in an N node network, where the maximum length of a route is H intermediate hops, is given by

$$M = 1 + \sum_{h=2}^{H} \frac{(N-2)!}{(N-h-1)!} \quad ; \; H \leq N-1$$

Q.E.D.