Traffic Balancing: A Method for Exploiting System Capacity in Wireless Ad Hoc Networks

by

Xiaojing Tao, B.Eng, M.A.Sc

A thesis submitted to
the Faculty of Graduate Studies and Research
in partial fulfillment of
the requirements for the degree of

Doctor of Philosphy

Ottawa-Carleton Institute for Electrical and Computer Engineering
Faculty of Engineering
Department of Systems and Computer Engineering
Carleton University
Ottawa, Ontario, Canada, K1S 5B6

November, 2005

The undersigned recommend to

the Faculty of Graduate Studies and Research

acceptance of the thesis

Traffic Balancing: A Method for Exploiting System Capacity in Wireless Ad Hoc Networks

submitted by Xiaojing Tao, B.Eng., M.A.Sc.

in partial fulfillment of the requirements for

the degree of Doctor of Philosophy

____

Chair, Department of System and Computer Engineering
Dr. Rafik Goubran

____

Thesis Co-Supervisor
Dr. David Falconer

____

Thesis Co-Supervisor
Dr. Thomas Kunz

____

External Examiner
Dr. Hossam Hassanein

Carleton University

November 2005

Abstract

Wireless Ad Hoc technology, which has received a rapidly increasing amount of attention over the last few years, provides a viable means of ubiquitous, untethered communication that could radically alter the way we work, learn, consume, and entertain. The routing protocols aim to set up connections and reestablish connections under a frequently changing topology. Current routing protocols generally search for the shortest path between sender and receiver, which usually results in fast response for route setup and a small number of hops. However, the shortest path algorithm has a high probability that traffic concentrates in the middle area of the network so that the system utilization is poor. Realizing the reality that the available system resources such as bandwidth in wireless Ad Hoc networks are limited and precious, research efforts have been undertaken to improve efficiency of resource utilization by means of traffic balancing. Because not all interferences are considered, these solutions fall short in collecting comprehensive traffic load information. Furthermore, some solutions generate a large amount of extra control packets that consume the available bandwidth.

The research in this thesis focuses on exploring the unused and wasted system capacity of wireless Ad Hoc networks, by providing comprehensive and accurate traffic load information to the routing protocols with minimum complexity. Proposed Traffic Balancing, a routing algorithm revised on the basis of reactive routing protocols, routes traffic load from congested areas to idle or lightly-loaded areas, such that network resources are allocated more efficiently. Knowing the number of observed collisions and past medium usage, every relay node provides the traffic load information into the route request in a simple way. After receiving route replies, a sender can choose the path with the least number of busy relay nodes. The simulation results illustrate that the Traffic Balancing approach is capable of decreasing the packet loss rate and average delay dramatically when some areas of the network start experiencing congestion with traditional on-demand routing protocols. Under certain scenarios, the improvement in system performance could exceed 50%. Furthermore, Traffic Balancing provides a solution to the problem of uneven traffic load that occurs at the access points of wireless mesh networks. The improvement is noticeable from the results of the simulation carried out on wireless mesh networks, even when the overall traffic load is light.

In this thesis, system capacity is first defined and investigated based on the interference range of a transmitting node and the size of network. The results show that the system utilization is far below the estimated capacity in the case where nodes are in movement. The major reason for the deficiency in system utilization is that movement and heavy traffic load cause collisions in congested areas, and the resulting long backoff periods after collisions waste a large amount of bandwidth.

It is also worth noting that traffic load tends to cluster, due to the routing and MAC protocols. Traffic Balancing is designed to provide accurate information about the traffic state along the paths, and to choose the path that has the least number of busy intermediate nodes. Numerical results indicate that the system performance is improved significantly when traffic load is deviated away from busy areas. With slight modifications, Adaptive Traffic Balancing is developed to change the verification rule of the traffic state at each node dynamically, according to the number of collisions detected by the node.

Furthermore, we investigate the problems that exist in wireless mesh networks, which have many similarities to wireless Ad Hoc networks. One severe problem is that performance is degraded by an uneven traffic load at the access points, whenever there is more than one access point in the network. After deploying Traffic Balancing, performance is improved and the system resources are used more efficiently.

Acknowledgements

During my stay at Carleton University I had the opportunity to associate with a number of truly remarkable people. I would like to deeply thank my supervisors, Prof. Thomas Kunz and Prof. David Falconer, for giving me the opportunity to work with them and guiding me during the research. Over the years that I had privilege of working with them, I grew to respect them not only for their great intuition and intellect, but also, equally important, for their unbounded enthusiasm for their work, and for their dedication to their students. When I first took Prof. David Falconer's class, I was instantly impressed by his intelligent and kind speech, but also by his personality. This encourages me applying for Ph. D. at Carleton University.

I also wish to thank Prof. Halim Yanikomeroglu, Prof. Hossam Hassanein, Prof. Ivan Stojmenovic, Prof. Evangelos Kranakis, Prof. Jerome Talim, Prof. S. Mahmoud for giving me unselfish help and serving as a reader of this dissertation.

My interaction with members and visitors of my lab has been extremely beneficial. Many thanks to Gama, Vrang, Fayez, Xiaobin, Huining, Naren, Jeff, John Knox, Michel, Yongyi, and John Boyer. In particular, I wish to express my appreciation to Naren for his computer administration work.

I wish to thank my wife Ou for her patience and support, particularly during the less exciting periods of my studies. Last, I would like to thank my parents for the unconditional lave and support that they have showed over the years.

Table of Contents

List of Figures

List of Tables

List of Acronyms

ACK	**Acknowledgment**
AP	**Access Point**
APR	**Alternative Path Routing**
ATB	**Adaptive Traffic Balancing**
ATM	**Asynchronous Transfer Mode**
bps	**bits per second**
CAC	**Channel Access Control**
CBR	**Constant Bit Rate**
CPU	**Central Processing Unit**
CSMA/CA	**Carrier Sense Multiple Access/Collision Avoidance**
CTS	**Clear To Send**
CW	**Contention Window**
DARPA	**Defense Advanced Research Projects Agent**
dBm	**dB gain with respect to a milliwatt**
DCF	**Distributed Coordination Function**
DIFS	**DCF Interframe Space**
DLAR	**Dynamic Load-Aware Routing**
DSL	**Digital Subscriber Line**
DSR	**Dynamic Source Routing**
EIFS	**Extended Interframe Space**
ETX	**Expected Transmission Count Metric**
GPS	**Global Positioning System**
IEEE	**Institute of Electrical and Electronics Engineers**

IETF	**Internet Engineering Task Force**
IFS	**Interframe Spaces**
LAN	**Local Area Network**
LBAR	**Load-Balanced Wireless Ad Hoc Routing**
LSR	**Load-Sensitive Routing**
LWR	**Load aWare Routing**
MAC	**Medium Access Control**
MANET	**Mobile Ad Hoc Network**
NS2	**Network Simulator Version 2**
OLSR	**Optimized Link State Routing**
PCF	**Point Coordination Function**
PDA	**Personal Digital Assistant**
PHY	**Physical Layer**
PIFS	**PCF Interframe Space**
PLR	**Packet Loss Rate**
QoS	**Quality of Service**
RFC	**Request For Comments**
RTS	**Request To Send**
SIFS	**Short Interframe Space**
SINR	**Signal to Interference and Noise Ratio**
SNR	**Signal to Noise Ratio**
TB	**Traffic Balancing**
TCP/IP	**Transport Control Protocol/Internet Protocol**
UDP	**User Datagram Protocol**

WLAN	**Wireless Local Area Network**
WMN	**Wireless MESH Network**
ZRP	**Zone Routing Protocol**

Chapter 1. Introduction

The idea of forming an on-the-fly Ad Hoc network of mobile nodes dates back to the days of the DARPA packet radio network. More recently, interest in this subject area has grown due to the availability of license-free, wireless communication devices with which users of laptop computers or PDAs can communicate with each other. Interest within the Internet Engineering Task Force (IETF) is also growing, as evidenced by the formation of a working group (MANET: Mobile Ad Hoc Network), whose charter is to develop a solution framework for routing in Ad Hoc networks and standardize routing protocols for MANETs.

The motivation of mobile Ad Hoc networks is to support a robust and efficient operation in wireless networks by incorporating routing functionality into mobile nodes. Each node in the network not only acts as a sender/receiver, but also as a router, forwarding data for other nodes. Thus, a wireless Ad Hoc network is one of the best candidates in the case where a fixed communication infrastructure, wired or wireless, does not exist or has been destroyed. For example, when Hurricane Katrina devastated many communication facilities, only 35 of over 200 cellular base stations set up by Sprint Nextel were functional in the New Orleans area. It took weeks to restore the whole system, when wireless Ad Hoc networks could have been used to support necessary communications. The implementation of mobile Ad Hoc networks has raised a lot of research interest.

However, the implementation of wireless Ad Hoc networks faces some challenges with respect to the mobility of nodes and the adverse wireless environment. First of all, the routing protocols need to react fast to the topology changes to ensure seamless communication. The MAC (Medium Access Control) protocol has to solve wireless-related problems such as hidden nodes. Since the capacity of a wireless system is precious and limited, efficiency is the crucial issue in system design. The tradeoff between the overhead caused by control information and the reaction to topology changes is worth analyzing. In addition, the system capacity may not be explored fully due to the routing and medium access protocols. The objective of this thesis is to provide a solution, called Traffic Balancing, that is based on an on-demand routing protocol, which collects traffic load information from the physical layer and selects paths according to this information, so as to explore system capacity or utilize it more efficiently. Therefore, the system performance can be improved considerably in certain aspects.

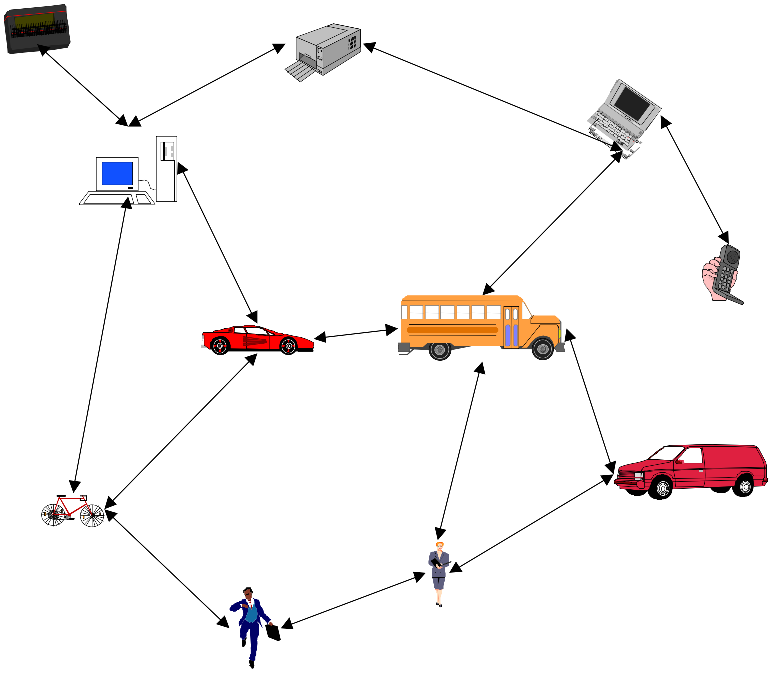


Figure 1.1 An Example of a Wireless Ad Hoc Network

This chapter includes background information about wireless Ad Hoc networks, with respect to their applications, structures and characteristics. Some conventional routing algorithms are also introduced briefly. Finally, some problems caused by the routing and MAC protocols are addressed, along with the contributions of this thesis.

1.1 Wireless Ad Hoc Network

A wireless Ad Hoc network is defined as a collection of mobile platforms or nodes, in which each node is free to move around arbitrarily without losing its connection to the rest of the world (as shown in Figure 1.1). Supporting this form of host mobility (or nomadicity) requires address management, enhancements of protocol interoperability and so on. The goal of mobile Ad Hoc networking is to extend mobility into the realm of autonomous, mobile, wireless domains, where a set of nodes—which may be combined routers and hosts—form the network routing infrastructure in an Ad Hoc fashion [CM99].

System Characteristics	Requirements for Routing Protocol
Dynamic Topologies	• Distributed operation • Fast response to link failure • Loop freedom
Bandwidth Constrained, Variable Capacity Links	• QoS requirements awareness • Load fairness • Unidirectional link support
Energy Constrained Operation	• Load Fairness • Unidirectional link support
Limited Physical Security	• Security provision

Table 1.1 Wireless Ad Hoc Network System Characteristics vs. Requirements for Routing Protocols

MANET RFC2501 [CM99] describes the system characteristics and presents some requirements for the routing protocols. The relationship between the system characteristics and the requirements for the routing protocols is listed in Table 1.1. A more detailed description is given later, to emphasize the importance of designing proper routing protocols.

Unlike wired systems, MANETs experience severe impairments due to the wireless environment and mobility. The collected information may not lead to a correct conclusion for the upper layers, if no modifications are made. Therefore, to operate properly, the different layers need to provide each other with more precise information and coordinate with each other.

1.2 Applications

Dynamic Ad Hoc networking technology has current and future needs. The emerging field of mobile and nomadic computing, with its current emphasis on mobile IP operation, should gradually broaden and require highly-adaptive mobile networking technology to effectively manage multi-hop Ad Hoc network clusters that can operate autonomously or, more likely, be attached at some points to the fixed Internet.

Some usages of the wireless Ad Hoc network include military, emergency, and commerce applications that involve data exchange in local groups. Using an Ad Hoc network to relay information provides robustness to the loss or movement of nodes. In emergencies, such as earthquakes or avalanches, rescue or emergency personnel have to send information about the environment and victims out through relays when the fixed infrastructure is destroyed or not available.

Business people can also use wireless Ad Hoc networks for meetings and other events, anywhere and at any time, without worrying about whether they will be able to find an infrastructure to exchange information.

With the development of the 4G wireless network, the idea of the wireless Ad Hoc network can be adopted to extend the coverage or improve the service quality of cellular networks. Wireless Ad Hoc networks can also be seen as an extension of wireless LANs (WLANs) to support more users and to cover a larger area in home networks.

1.3 Protocol Stack

As shown in Figure 1.2, the basic structure of a wireless Ad Hoc network is similar to the ordinary wireless system, but the functionality of each layer is different, in order to support the unique characteristics of wireless Ad Hoc networks.

The functionality of the physical layer includes modulation, demodulation, equalization, coding, and decoding. Each node continuously senses the medium usage in order to extract the packets for itself, and to provide information for the upper layer protocols. Due to the mobility of the nodes, a transmission may be interrupted when the received power is below a required level. The physical layer needs to forward this information on to the upper layer functions for necessary actions, such as retransmission, or sending new routing requests. Depending on the local environment and the node mobility, the physical layer may have to have the ability to combat fast-fading and shadowing to make the system more stable.

The data link layer consists of the MAC sublayer and the Logical Link Control sublayer. The IEEE 802.11 MAC protocol is the most popular choice for wireless LANs, and may also be considered for wireless Ad Hoc networks in order to avoid collisions caused by simultaneous transmission in a certain range. It uses the carrier sense multiple access protocol with the collision avoidance (CSMA/CA) medium-sharing mechanism to avoid hidden station problems. The Logical Link Control sublayer provides the function to transfer the data packet from and to the upper layer protocols. The data link layer also provides flow control and error control, if required. There is a possible sublayer in the Data Link Layer, below the MAC sublayer; this is called the Channel Access Control (CAC) sublayer. The function of this sublayer is to deal with the channel access signaling and protocol operations required to support packet priorities.

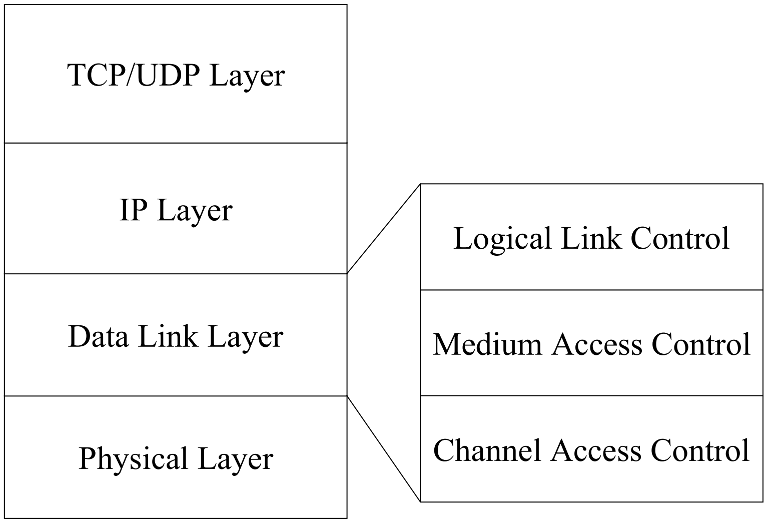


Figure 1.2 The Architecture of a Wireless Ad Hoc Network

The IP layer (also called the network layer in other models) is responsible for routing and some other functions related to resource allocation, traffic shaping, and so on. Depending on the application, the routing decision can be based on IP addresses or other machine identifiers. The frequent topology changes in wireless Ad Hoc networks require the routing protocols to react fast to link failure, with low control information overhead. In general, the routing protocol at the IP layer has a major impact on wireless Ad Hoc networks. Implementing an efficient routing algorithm at the IP layer is critical.

The TCP/UDP layer sets up the connections for particular applications with necessary parameters, such as transmission rate and so on. As mobility and wireless links can cause frequent transmission errors and long delay, which do not happen on wired networks, more care is needed to control the transmission based on the information collected from the lower layers.

1.4 Routing Protocols

The routing protocol is the major component of the IP layer that attracts a lot of research attention. The protocol takes care of finding a suitable path for each data packet. Unlike in wired systems, a wireless Ad Hoc network tends to experience many more topology changes and link failures, which creates challenges for the routing protocols.

First, the status of each route should be updated frequently so that the protocol can react to the frequent topology changes. However, frequent information updating will generate a large amount of control traffic, which may reduce system efficiency. Second, some applications have delay and jitter requirements that cannot tolerate long connection interruptions. Each node should receive every kind of transmitted packet (even if the packet is not destined to it) to extract new information at all times and send information about any connection changes in the control packets. The key characteristic of routing protocols in wireless Ad Hoc networks is to make transmission smooth and determine the tradeoff between control information overhead and system efficiency.

Since this proposal also looks at the optimal usage of the limited capacity of wireless Ad Hoc networks by enhancing the routing protocols, it is important to discuss the proposed routing protocols and their characteristics. There are different ways of classifying the routing protocols, depending on their properties. Here we simply divide the proposed Ad Hoc routing protocols into two major groups: reactive and proactive.

Reactive (on-demand): In this category a node finds a route only if it has a packet to be sent out. It also learns about the other paths when it forwards the packets of other nodes and maintains these paths for a certain period. One of the most popular on-demand routing protocols is the DSR (Dynamic Source Routing) protocol [BJM99].

Proactive: Proactive routing protocols construct the routing table in each node by vector or full path before any data transmission takes place. They update the routing table automatically and periodically. The OLSR (Optimized Link State Routing) protocol is a typical proactive routing protocol [CJ03].

A more detailed classification of routing protocols in wireless Ad Hoc networks can be found in [I01] and [M00]. The following table (Table 1.2) lists the advantages and disadvantages of the above categories. A comparison of the pros and cons of these protocols shows that it is easier to implement on-demand routing protocols than proactive protocols. On-demand routing protocols always try to find the shortest and least weighted path if each link is associated with a weight. In addition, they consume less system capacity, increasing the chances of improving system utilization. Taking these advantages of the on-demand routing protocols into consideration, finding a better solution based on these protocols may be easy and straightforward.

	Advantages	**Disadvantages**
Reactive (On-Demand)	• Scale to large networks • Less control traffic when mobility is low • Short path can be found • Small cache for routing information	• Long delay to find the route and recover from a bad link • Flooding the network with control packets
Proactive	• Fast route finding and recovering • Least weighted path can be found	• Huge information cache • Large overhead in control information • Difficult to scale to large network

Table 2.2 The Pros and Cons of Different Types of Routing Protocols.

1.5 Problems and Contributions

Although routing protocols offer the convenience of finding an effective path in this dynamic network promptly, they also have some disadvantages. First, in order to react to the frequent topology changes, routing protocols have to generate a large amount of control information to notify the nodes of these changes (by exchanging link information periodically or rediscovering the path through flooding). No matter what kind of routing protocol is adopted, the amount of bandwidth consumed by the control information is considerable and has a significant impact on the system performance. With the formulation of system utilization and throughput described in Chapter 3, it is obvious that decreasing the control information bandwidth consumption, especially when the network is saturated, will improve system performance.

Second, existing routing protocols typically search for the shortest path, and this frequently leads to congestion in the central area of a network, while other areas are far from saturated. This results in a degradation of system performance (uneven traffic load).

Third, when the MAC layer protocol is employed to solve problems in wireless networks, such as discovering hidden nodes, some other side effects arise. Traffic through a particular area tends to use the same node as the relay node. Once a link failure occurs at this relay node, all connections passing through this node suffer. Under a heavy traffic load, collisions may cause several nodes to be backed up for an extremely long time, which wastes bandwidth. In the following chapters, these problems are addressed in more detail.

As traffic tends to be concentrated in the middle of a network, due to the algorithm of traditional routing protocols, which use the shortest path as the default path, system utilization is far below the network capacity, which results in an unacceptable performance overall. Thus, research is underway to balance the traffic load evenly. The system resources at the edge of the network, which are only lightly used, could be used more efficiently. Meanwhile, less congestion in the middle leads to a smaller number of collisions and a lower number of backoff periods or shorter backoff periods at the MAC layer. The efficiency in the middle area of the network then also increases, thereby improving system performance. In order to distribute traffic effectively, more system information is required to help routing protocols select the appropriate paths. Except for the number of hops, the traffic load state along the path is also very important for assisting routing protocols to weigh each path. In Chapter 2, several existing traffic distribution solutions are discussed, and in Chapter 6, several are compared to our approach.

In Chapter 3, a simple method is proposed to estimate the system capacity, and simulation results indicate that poor system utilization results in poor overall system throughput. One of the major contributions in this thesis is the proposal of a solution based on the concept of traffic distribution, so that better system performance is achieved by distributing traffic load to idle or lightly-loaded areas of the network, based on the traffic load state of the network. With respect to a node, the local area traffic load state consists of three parts: 1) the packet transmissions of the node itself; 2) the packet transmissions of the neighboring nodes, which the node can reach through direct radio links; and 3) the interference from the packet transmissions of all the other nodes in the network. In this case, the traffic load information may need to be collected from the queue at the IP layer, the neighbor nodes' activity at the data link layer, and the medium state at the physical layer. Unlike other proposed load-balancing approaches (discussed in Chapter 2), which collect traffic load information at the IP layer and the data link layer, our solution determines traffic information by measuring the power level of the medium at the physical layer. It therefore avoids extra hello messages and collects information in the area instead of at particular nodes. With a better understanding of the traffic load state of the network, our proposal explores the unused system resources ignored by reactive routing protocols to increase utilization. The throughput may then be increased proportionally.

Furthermore, a MESH network is implemented, and it is shown that Traffic Balancing can significantly improve system performance if there is more than one access point. Through Traffic Balancing, all traffic to the outside network will be distributed more evenly to each access point.

Our contributions per chapter are as follows:

- In Chapter 3, a simple method is defined to calculate system capacity, based on network size and interference range of a transmitting node. Thus, this computed system capacity can be used as a rough upper boundary to determine how network resources are used under different routing protocols.

- Based on the problem analysis from the aspect of MAC and routing protocols in Chapter 4, in Chapter 5 we propose a routing method called Traffic Balancing to deviate newly generated traffic or rerout traffic away from congested areas. This enables the unused resources at the edge of the network to be used. In addition, the number of collisions in the busy area is decreased due to the lower traffic load. Fewer collisions means fewer backoff periods, which can waste a lot of network resources. The bandwidth utilization at both the middle and the edge of the network is then increased, and the overall performance is improved. Furthermore, we find that Traffic Balancing can change one of its parameters dynamically, so that Traffic Balancing further improves system performance with respect to node mobility.

- Chapter 6 analyzes the characteristics of wireless Ad Hoc networks and further explains where and under what conditions Traffic Balancing can improve system performance. This proposal is also compared to other proposals, which focus on how to balance the traffic load. The analysis emphasizes that in wireless data communication, more cross-layer cooperation is required to give a more precise illustration of network states in order to improve the efficiency of network utilization.

- In Chapter 7, some problems in MESH networks are presented and discussed. This chapter also illustrates how Traffic Balancing can benefit performance.

- During the research, several conference papers were published:

    - "Traffic Balancing in Wireless Ad Hoc Networks: Extending System Capacity and Improving System Performance" [TFK02] presents the first version of the proposed Traffic Balancing;

- "Implementing Traffic Balancing for Exploring System Capacity in Wireless Ad Hoc Networks" [TKF04] presents the second version of Traffic Balancing;

- "Adaptive Traffic Balancing in Wireless Ad Hoc Networks" [TKF05a] presents Adaptive Traffic Balancing.

- "Throughput Maximizing Routing in a MANET: Protocols and Analysis" [TKF05b] presents a linear programming solution for estimating capacity and the results of Traffic Balancing in this regard.

- "Traffic Balancing in Wireless MESH Networks" [TKF05c] discusses the ways in which Traffic Balancing solves the uneven traffic loads that occur in wireless mesh networks.

1.6 Summary

In wireless Ad Hoc networks, each node can be a sender, receiver or router. As nodes may be mobile, the network topology can change frequently, requiring routing protocols to react fast and to handle traffic fairly. Wireless Ad Hoc networks are well suited to emergency scenarios, temporary business cases, or extensions of fixed infrastructure wireless networks. Its protocol stack follows the traditional TCP/IP four-layer stack, with three sublayers in the Data Link layer to deal with logic control, medium access, and channel access. There are two types of routing protocols in wireless Ad Hoc networks: proactive and reactive. Due to the nature of routing protocols and MAC layer protocols, system efficiency is very low. In this thesis, a solution based on the reactive routing protocol (Traffic Balancing) is proposed to explore unused system capacity. This solution selects the path based on traffic load information collected from the physical layer.

Chapter 2. Related Work

The maximum throughput of a wireless Ad Hoc network depends on the network topology and the traffic pattern. How to reach maximum throughput depends partly on the path selection of the routing protocols. Traditional routing protocols select paths based on the metrics of number of hops, end-to-end delay, jitter at each hop, and so on. In a wireless Ad Hoc network, the selection of paths according to the number of hops or end-to-end delay may result in the shortest path between two connection peers being chosen. The shortest path usually has the lowest number of hops between sender and receiver. The end-to-end delay for routing packets, which have high priority and are always inserted into the head of the outgoing queue, tends to be shortest due to shorter transmission time and processing time (because of the smaller number of hops) by traveling through the shortest path. Geometrically, the shortest path tends to go through the middle of the network and so causes more congestion in this central region. Congestion results in more collisions in the middle, so that more bandwidth is wasted during the backoff periods.

Traffic distribution seeks to move some traffic from the middle to the edge of the network to increase utilization at the edge and relieve congestion in the middle. Connections moved to the edge may have a greater number of hops and use the bandwidth that is provided by the improved utilization at the edge. However, the decreased traffic load in the middle reduces the chance of congestion, and bandwidth is used more efficiently so that the overall performance is still improved.

This chapter introduces the basic methods of traffic distribution, after which the aspects that affect the traffic load state in wireless networks are discussed. Several proposed traffic distribution solutions for wireless Ad Hoc networks are listed, and some of the disadvantages are included and discussed.

2.1 Traffic Distribution

In wired networks, path selection may depend on the weight of each path. The weight can be assigned by the performance metric per path, or be added by the performance metric per hop. In the latter case, if the performance metric is based on the traffic load at each relay node, the weight of each path represents the traffic load along the path. The routing protocol can then select the path with the lowest traffic load, so that newly incoming traffic is added to the path with the lightest load and traffic distribution takes place.

There is a tradeoff between information accuracy and implementation complexity with regard to how to use weights to reflect the traffic load of each intermediate node. The more information is included in the route request, the better the route selection will be. However, more information requires more overhead and more computation in route discovery.

The simple solution [S94] for realizing traffic distribution in wired networks is based on calculating the weight of each link, according to the queue length, processing time, transmission rate, and so on. By adding the weight of each link together, the weight of each path is ascertained. Choosing the path with the lowest weight can then distribute traffic more evenly in the network. For example, if the traffic load of each link is calculated as shown in Figure 2.1, the best path from node 1 to node 6 is 1-4-5-6, which has the lowest cumulative weight of 6. Although the shortest path is 1-3-6 (with a weight of 13), the traffic load in this path is extremely high, and adding more traffic to it may overload node 3 and cause packets to be dropped (due to a full queue) or experience long delays. By using path 1-4-5-6, the traffic load in the network is more even.

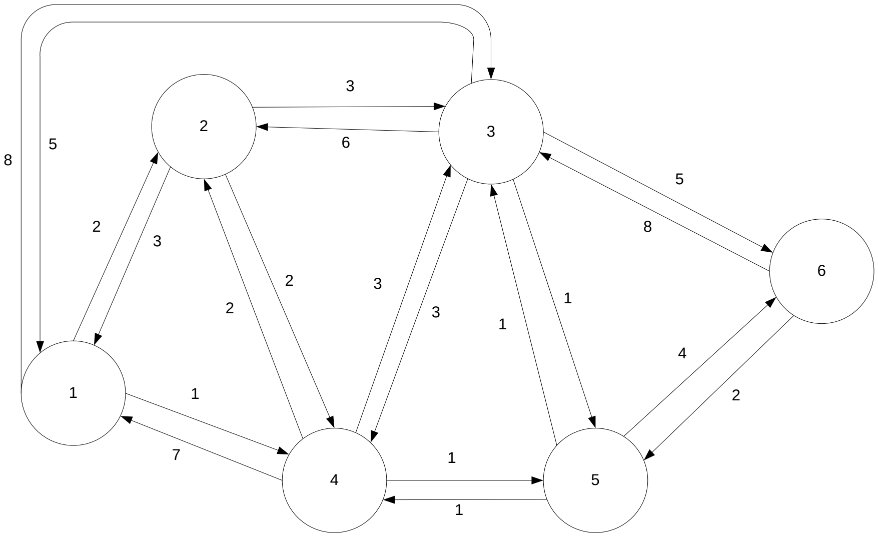


Figure 2.1 Example of Traffic Distribution in a Wired Network.

However, simply adding up the weights of each link may sometimes not provide enough information to select an optimal path. If one link in the path has extremely high traffic load and the weight of the path is smallest, selecting this path may still have poor performance. In this case, if route discovery could record the weight of each link separately instead of adding it to one record, this would allow for better choices to be made. On the other hand, if the number of links is large, the overhead brought by the records could also be large. Thus, implementing traffic distribution requires more consideration of traffic pattern and network topology to maximize its improvement.

In wireless networks, all users share the medium, and MAC layer protocols are required to solve the access competition among nodes in an area. It is well known that the performance (throughput) of some MAC layer protocols, especially random access protocols, is optimal for a certain traffic load [T96]. Beyond this load, performance decreases dramatically. Traffic distribution aims to move traffic from the areas that are above the optimal load to less loaded areas, so that all areas achieve better performance, as shown in Figure 2.2. Without traffic distribution, some areas in a network are under heavy load; the system throughput in these areas is indicated by the solid line in throughput vs. the traffic load picture (the "busy area" in Figure 2.2), which is above the optimal point. Meanwhile, some other areas have a light load, and the system throughput is below optimal (the "idle area" in Figure 2.2). If some traffic can be moved from the busy area to the idle area, both throughput points of the busy and idle areas tend to be close to the optimal point (shown by the direction of the arrow). Thus, with the same traffic load, higher throughput can be achieved.

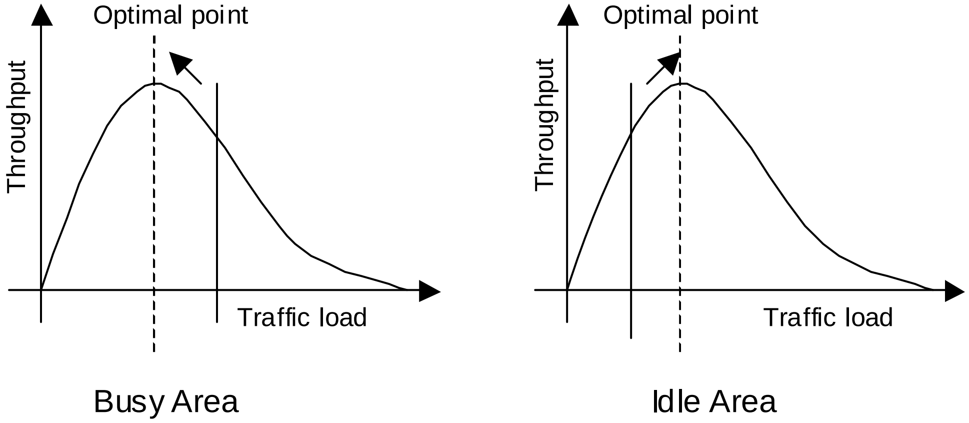


Figure 2.2 A Conceptual Illustration of Traffic Distribution.

Node mobility in wireless Ad Hoc networks may cause a large number of collisions in busy areas and dramatically degrade system performance. This phenomenon has a throughput curve similar to that shown in Figure 2.2, so that it is necessary to distribute the traffic load evenly in wireless Ad Hoc networks in order to maximize the throughput. Each link in a path can be assigned a weight depending on the traffic load in the area, and the path with the lowest weight will be selected in order to realize traffic distribution. The throughput improvement by traffic distribution relates to how the traffic load at each node should be calculated, and how a weight should be assigned according to the traffic load. In the next section, the traffic load of wireless Ad Hoc networks is analyzed in detail.

Although distributing traffic evenly throughout the network may increase the number of hops for certain connections, if the gain from unused resources is higher than the waste in the longer paths, it is still a good choice. In addition, as Traffic Balancing reduces congestion, the chance of collision becomes lower and the bandwidth wasted by backoff time is saved for data transmission.

2.2 Traffic Load of Wireless Ad Hoc Networks

The traffic load of each intermediate node or link in wireless Ad Hoc networks is more complicated than in wired networks and relates not only to the traffic going through the node itself, but also the transmission all over the network. For any node in a wireless network, a packet can be received correctly if the signal to noise and interference ratio—SINR—is above a certain level *k'* , as presented in the following inequality:

$$\frac{P_{intended}}{P_{noise} + P_{unintended}} \geq k'$$

where $P_{intended}$ is the received power of the intended packet, P_{noise} is the noise power level and $P_{unintended}$ is the received power from all other transmissions in the network at the same time. Thus, when the sensed medium power level on a node is over a certain threshold (*k* dBm), the node believes that the medium is occupied by other nodes. Since $P_{unintended}$ is presumably not known, *k* would be $10log_{10}P_{noise}$ (where P_{noise} is in milliwatts). If the node starts a transmission, other nodes, or the node itself, will detect a collision.

The sensed signal power consists of the white noise and the sum of the received signals from all the transmissions in the network (interference and intended signal). The node cannot sense the medium when it is transmitting. Thus, when the node is in transmission, the sensed signal power is treated as being definitely larger than the threshold and the medium is in a busy state or occupied.

No matter how far apart the transmitting nodes are, their transmissions have an impact on the sensed power level of a node in the network. Thus, the sensed power level (P_{sense}) for a node can be presented as:

$$\begin{aligned} P_{sense} &= P_{noise} + P_{intended} + P_{unintended} \\ &= P_{noise} + \sum_{i=1}^{n} P_i \end{aligned}$$

where n is the number of concurrent transmissions at one time and P_i is the received power from the i_{th} transmission. Assume medium can only be sensed by a node when it is not transmitting, the P_i has to also include the transmission from the node itself. When P_{sense} is over the threshold (k dBm), the node believes that the medium is occupied and it will not allow the medium to be accessed. The formulas indicate that in wireless Ad Hoc networks, traffic load is related to the activities of all the nodes. Transmissions that are multiple hops away from a node may also block the transmission of the node.

Because whether the medium is idle or busy relates to all the transmissions in the network at any time, a node needs to know the queue size and the position of all nodes in the network in order to predict the traffic load around itself. Depending on the size of the network, the queue size information of some nodes may have to go through several hops to reach the node. It may take some time to disperse the information, which causes the information to become obsolete. In addition, carrying information about the queue size or other parameters (such as position of a node) brings overhead to the network, which consumes the bandwidth for data transmission. Acquiring the accurate position of a node may require help from other devices installed in the node, such as a GPS. Even knowing the positions of all nodes in the network, the power level prediction is still only approximate due to the dynamics of the wireless link. Thus, the traffic load prediction may not be accurate enough to provide efficient information for the routing protocols.

Furthermore, P_{sense} could be used as a good indicator of the traffic load information of the area around the node because it covers all activities in the network in the past and can reflect the activities in near future; this is due to the strong correlation among traffic loads. Over a period of time, the percentage that P_{sense} is over the threshold (k dBm) can indicate how busy the medium is. A weight can be figured out based on this percentage. Alternatively, this percentage could be combined with other information, such as the queue length of the node and neighbor nodes, to provide the weight of the link. The percentage will change with movement because signal level varies due to distance change, shadowing and fading.

2.3 Proposed Solutions for Traffic Distribution in Wireless Ad Hoc Networks

Recently, several load-balancing routing protocols have been proposed for wireless Ad Hoc networks, in order to combat unevenly distributed traffic loads and maximize system performance. Most of them are based on reactive routing protocols and evaluate the routes by weights, which depend on the traffic load information collected by each intermediate node. The load state information can be collected from the queue size of each node, the transmission quality of each radio link, or the number of connections held by each node. Based on the collected information, each hop in a route is assigned a corresponding weight and the route with the lowest weight is selected as the default route for the connection. A brief description of these load-balancing methods follows:

- DLAR (Dynamic Load-Aware Routing) in [LG01] evaluates the weight by looking at the queue length of each intermediate node. Three schemes are proposed: counting the queue length of all intermediate nodes, the average queue length over the path, or the number of intermediate nodes that have a heavy-loaded queue. In the first scheme, the overall weight of each path is the sum of the queue length of each intermediate node. The overall weight of each path in scheme 2 equals the overall weight in scheme 1, divided by the number of intermediate nodes. In scheme 3, a threshold value τ is defined. When the queue length of an intermediate node exceeds τ, the node is defined as being in a state of congestion. The overall weight of the path is the number of the intermediate nodes in this congestion state.

- LBAR (Load-Balanced Wireless Ad Hoc Routing) in [ZH01, HZ01] counts the number of connections in each intermediate node and its neighboring nodes as the weight. The weight of each intermediate node is the sum of the number of connections going through the node, and the number of connections going through the neighboring nodes. The overall weight of the path is the sum of the weights of each intermediate node. The connection information from the neighboring nodes can be piggybacked onto the data packets or collected from periodical hello messages broadcasted by the neighboring nodes. The hello messages also help the node to identify connectivity to neighboring nodes. When a link failure occurs, an error message is generated and propagated to the destination node. The destination node tries to find an alternative path to the upstream node of the broken link. If the alternative path is not available, the destination node sends an ACK message to the source node, and the source node uses the path of the ACK message. Otherwise, the destination node propagates an error message to the source node and the source node restarts route discovery.

- LSR (Load-Sensitive Routing) in [WH01] calculates two overall weights for each path: the total path load, which is the sum of the traffic load of each intermediate node, and the standard deviation of the traffic load along the path. The traffic load of each intermediate node includes the queue length of this node and the queue length from the neighboring nodes. Without using hello messages, LSR only allows the node to collect the traffic load information from those neighboring nodes that are currently transmitting packets to the node or receiving packets from the node. Two parameters α and β are used for route comparison. If the total path load of a path is α smaller than the other paths, this path is selected. Otherwise, among the paths with an absolute difference of the total path loads being less than α, the path with β-smaller standard deviation is selected. In other cases, when the absolute difference of the total path loads is less than α and the absolute difference of the standard deviation is less than β, LSR chooses the path randomly. LSR recalculates the path load information during the route reply, then chooses the path with updated information. During data transmission, path load information is calculated continuously and piggybacked into the data packets. When the path becomes much worse than its initial state, a new route discovery starts.

- APR (Alternative Path Routing) in [PHST00] is based on strategies like deflection routing and shortest path first with emergency exit. The hybrid proactive and reactive Zone Routing Protocol (ZRP) is used to evaluate local network connectivity [PH99]. Each node keeps listening to the Hello messages from its neighbors to track link state information, and computes the weights in the form of route coupling: once it knows the local network topology, a node can figure out how many nodes will be blocked when it is transmitting. Then, for any path, the routing coupling equals the sum of the number of blocked nodes by each transmitting node in the path (including the source node), divided by the number of hops of the path. A path that minimizes the level of route coupling and the number of hops is selected.

- ETX (Expected Transmission Count metric) in [CABM03] predicts the number of data transmissions required to send a packet over a link, including retransmissions, and uses this number as the weight of the link. Then it selects the path with the lowest overall weight, which will go through the link with less interference. The prediction of the link state is based on the measurement of dedicated link probe packets in both directions. The primary goal of the ETX design is to find paths with high end-to-end throughput, despite losses. By checking dedicated link probe packets during the last *w* seconds, the delivery ratios d_f and d_r are measured for forward and reverse directions. The weight of each link in ETX is then calculated as $1/(d_f \times d_r)$. The overall weight of each path is the sum of the weight of each link in the path. ETX measures both directions of a link, because it assumes that the nodes use the IEEE 802.11 MAC layer protocol, and the hand-shaking procedure in IEEE 802.11 MAC requires frame exchange in both directions. The designers of ETX believe that the delivery ratios affect throughput directly, so that ETX makes fine-grained selection among routes. Furthermore, they also think that ETX penalizes routes with more hops and tends to minimize spectrum use, which should maximize overall system capacity.

- LWR (Load aWare Routing) in [YKG01] uses the channel utilization, the queue size, the number of neighbors, and the backoff timer to ascertain the load level of each intermediate node. An intermediate node drops the route request if it is under a heavy load. The heavy load state is reached when the channel utilization is “full-heavy,” or all four values of the channel utilization, the queue length, the number of neighboring nodes and the duration of backoff period are large enough. The threshold for LWR to decide if the channel utilization reaches “full-heavy” is selected as the maximum performance experienced in the scenario at hand, and generally depends on the collision probability and the number of neighboring nodes. When some nodes occupy the channel longer than certain periods under a high load situation, the channel utilization is also assumed to be full-heavy.

Recently, some researchers have proposed that using multiple paths for connections can also realize load balancing in wireless Ad Hoc networks [JKW05]. The multiple paths solution was originally proposed for improving the reliability of connections in wireless Ad Hoc networks [YKT03]. By finding link or node disjoint paths, the reliability is increased, because the chance that all these paths are broken is small (no common link or node). In addition, having multiple paths may eliminate the extra routing packets for route rediscovery when a link breakage occurs. In [JKW05], the routing protocol tries to find multiple paths in which the middle nodes of the multiple paths are apart, at least over the interference range, which may be twice the transmission range (in Chapter 3, Section 3.1.3, the interference range is described again). After multiple paths have been added, the throughput can be doubled or more, because the maximum throughput of one path is limited by a quarter of the direct radio link capacity when the number of hops exceeds three (more details are discussed in Chapter 3, Section 3.1.3).

2.4 Deficiency of the Proposed Solutions

In a wireless Ad Hoc network, a node competes for access to the medium with other nodes in the area and the usage of the medium relates to all activities in the network. As a result, the node’s own queue size information only reflects part of the load in the area. In addition, the interference may come from nodes beyond the transmission range of a node, so that information collected from the neighboring nodes also does not show the full traffic load of the area when it is combined with local queue size information. Some nodes in the busy areas may have an empty queue, and may have a small weight when just the queue length or number of connections is counted. The selected paths passing through these nodes may not drop packets because the queue is empty, but the medium in these areas might already be under a heavy load. The number of collisions is still high and retransmission cannot be avoided. Consequently, system performance will be poor if these paths are selected. The delay will still be very long if these paths are used, and may cause problems for the upper layer protocols, such as TCP. Thus, weights based on this information will not be accurate enough for selecting optimal paths that distribute traffic more evenly in the network. With limited information about the traffic load, the selected routes still go through the congested area, because some nodes in the congested areas have little traffic going through them. Meanwhile, extra hello messages for collecting information from the neighbor nodes or probe messages for measuring link quality occupy bandwidth that could be used for transmitting data packets. Processing extra hello messages and probe messages also requires more computational resources.

Besides interference, node mobility and the quality of the wireless radio link also affect data transmission in wireless Ad Hoc networks. High node mobility tends to be associated with more collisions in the network, which decreases system capacity. An area with low node mobility has more capacity than an area with high node mobility and could accommodate more traffic. Poor wireless radio links cause more packets or frames to be corrupted and to be retransmitted, which also decreases network capacity. Even with a lower traffic load, poor wireless radio links may still not be a good choice to be part of the path. Poor wireless radio links may be caused by a severe environment, long distance between two transmitting peers, or high node mobility. Lack of consideration of node mobility and the quality of wireless links along the selected paths will still lead to poor performance.

Table 2.1 lists the deficiencies of the proposed traffic distribution routing protocols.

Proposed Routing Protocols	**Limited Traffic Load Information**	**Extra Information Packets**	**Considering Node Mobility**	**Considering Quality of Wireless Radio Link**
DLAR	Yes, only local queue length	No	No	No
LBAR	Yes, only local activities and activities in neighboring nodes	Yes, periodic hello messages	No	No
LSR	Yes, only local queue length and queue length of neighboring nodes connected currently	No	No	No
APR	Yes, number of nodes could be blocked at most two hops away	Yes, periodic hello messages	No	No
ETX	No	Yes, periodic probe packets	No	Yes
LWR	No	Yes, periodic hello messages	Yes	No

Table 2.1 Deficiencies in Proposed Traffic Distribution Routing Protocols in Wireless Ad Hoc Networks

The multiple-paths solutions may rely on location-based routing protocols to acquire the overall network topology or positions of all nodes, and the distance between the sender and receiver must be longer than the interference range in order to take advantage of load balancing. The chance of finding such multiple paths is low when the network size is small (related to the transmission range), or the node density is low.

As local queue length and queue length of neighboring nodes do not reflect the overall network traffic load, the route decision made by DLAR, LBAR, LSR and APR can hardly balance traffic load from busy areas to idle areas, i.e. the chance of making an efficient route selection is very small. LWR does measure the channel utilization that covers all activities in the network but it drops the route requests when "full-heavy" condition exists. A connection having to go through "full-heavy" nodes or areas will be blocked even if some traffic in "full-heavy" nodes or areas can be rerouted to yield some bandwidth for the new connection. The efficiency of LWR will be very low when such cases happen. Among all proposed traffic distribution solutions in wireless Ad Hoc networks, ETX acquires the overall network traffic load information through an indirect measurement, which also considers the quality of wireless link. The route decision will be more optimal comparing to the rest. However, as measurement in ETX relies on the probe message, the number of probe messages per time unit decides how accurate the information is. Large number of probe messages per time unit provides an accurate information about quality of wireless link and the network traffic load but a large amount of bandwidth is consumed that decrease the available bandwidth for transmitting data packets. Thus, a method that could provide accurate traffic load information with less consumed system resources is researched and proposed in following chapters.

2.4 Summary

Traditional traffic distribution solutions in wired networks assign a weight to each link according to the transmission rate, processing ability and the queue length. The weight of a path is the sum of the weights of each link in the path, and the path with the smallest weight is selected in order to carry out traffic distribution. It also works in wireless Ad Hoc networks to distribute traffic more evenly, but traffic load information relates to all activities in wireless Ad Hoc networks and cannot be acquired without help from the physical layer. Several routing protocols have been proposed to distribute traffic load efficiently in wireless Ad Hoc networks, based on different ways of evaluating the traffic state of the network. Some of these protocols generate extra hello messages or probe packets to acquire the load information or link state. Others evaluate the traffic load based on limited local information. Node mobility and the quality of the wireless radio links are rarely taken into consideration.

Chapter 3. System Capacity and Utilization

Obviously, if there is a way to ascertain system capacity that identifies the theoretical upper and lower bounds of the achievable performance of networks and the utilization that is realized by the protocols employed at the routing and MAC layers, we can determine whether there is space to improve the system performance and how much bandwidth is required for extra information to achieve this improvement. In this chapter, several different methods of analyzing network capacity are discussed. Due to the complexity and computation limitations, these methods are not suitable for calculating the system capacity of a real wireless Ad Hoc network. Thus, a simple solution has been adopted and developed to estimate the network capacity, based on the network size and the interference range of a transmitting node. In addition, the relationship between system throughput and traffic patterns is identified and can be used to calculate system utilization.

Before the methods are discussed, a definition of system capacity and utilization will be given. **System capacity** (or **network capacity**) refers to the aggregated transmission rate of a wireless Ad Hoc network when the network topology and traffic pattern are known, the optimum paths between sender and receiver have been found, and an ideal schedule has been given to each node regarding when to transmit packets. **System utilization** (or **network utilization**) measures the achieved aggregated transmission rate of a wireless Ad Hoc network by deploying a routing protocol and a MAC protocol without the network topology being known. This counts the transmission rate per hop. **System throughput** (or **network throughput**) measures the achieved transmission rate of a wireless Ad Hoc network by deploying a routing protocol and a MAC protocol without knowledge of the network topology. This counts the transmission rate per connection. All are measured in bits per second (bps).

One ultimate objective of increasing the efficiency of wireless Ad Hoc networks is to maximize their throughput under limited resources. As the system utilization is bounded by the system capacity and the relationship between the system utilization and throughput can be formulated, we can find possible solutions for increasing system throughput.

3.1 System Capacity

Unlike wired networks and one-hop wireless networks (such as WLAN and traditional cellular systems), a capacity analysis of a wireless Ad Hoc network is more complicated because it combines the attributes of both wired networks and cellular systems. In this case, both routing decisions and the signal to noise and interference ratio need to be considered, since they affect each other when calculating system capacity. In general, there are two major methods for calculating network capacity: linear programming and geometrical analysis.

3.1.1 Linear Programming

Efficiency is an important issue to consider with regard to the performance of a network or evaluating a network. Efficiency here refers to the maximum throughput of a network under the same capacity and traffic patterns. It relates to access protocols, routing protocols, and transmission probability. When the routing algorithm only is considered, the system design tries to find the best way to assign suitable resources to each link, or make the best routing decision to support a maximum number of connections when the capacity of each link is fixed.

In wired networks, because topology changes rarely occur, designers can assign a suitable capacity to each link based on the bandwidth requirement constraints. When there are requirements in delay and blocking probability, some research in [GW90, GN89] provides solutions based on linear programming to design the network. When the capacity of each link cannot be changed, a refined routing protocol is needed to find paths for each connection so that the network can reach its maximum throughput. Researchers in [MS87] present an adaptive solution for both routing and flow control to fulfill the requirements in delay, throughput and so on. [JL98] provides a method based on linear programming to optimize ATM network performance. In [CF96, YS91], researchers propose similar ideas implemented in global networks.

In general, the problem of efficiency for wired networks can be formulated as:

Maximize:

$$F = \mathrm{sum}(\ x\ )$$

Subject to:

$$Ax \leq C$$

$$x \geq 0$$

where:

F is the objective function value that is to be maximized. Here it refers to the total throughput of the network.

x is the column vector of independent variables. Here each element of the vector is the throughput of an individual connection.

A is the matrix of coefficients representing network connectivity and the traffic parameters.

C is the column vector of right-hand-side terms, associated with the constraints on link capacity.

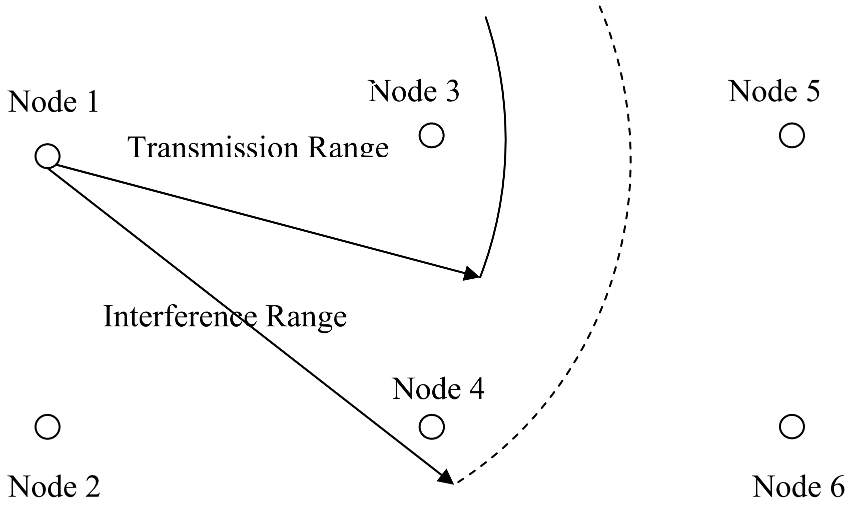


Figure 3.1 Example of a Static Wireless Ad Hoc Network

A similar approach can be used in wireless ad hoc networks to estimate the maximum system capacity. For the sake of simplicity, a static scenario is considered first, in a situation where the network topology is unchanged. Assuming there is no time-varying fading and MAC layer functions are simple (only listening and sending, no handshaking), the interference pattern will not be changed at any time and each node is associated with a matrix that reflects its transmission occupation. For example, as shown in Figure 3.1, there are six nodes in a wireless ad hoc network and the interference range is assumed to be less than twice the maximum transmission range. When node 1 is transmitting, nodes 2, 3, and 4 cannot transmit their packets. Hence the transmission matrix of node 1 is as follows:

$$\begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 0 \\ 0 \end{bmatrix}$$

The 1s in the matrix mean that at this time nodes 1, 2, 3, and 4 are blocked by the transmission of node 1. The 0s mean that at this moment, nodes 5 and 6 are free to transmit. Similarly, the matrices for the other nodes are:

$$\begin{bmatrix}1\\1\\1\\1\\0\\0\end{bmatrix},\begin{bmatrix}1\\1\\1\\1\\1\\1\end{bmatrix},\begin{bmatrix}1\\1\\1\\1\\1\\1\end{bmatrix},\begin{bmatrix}0\\0\\1\\1\\1\\1\end{bmatrix},\begin{bmatrix}0\\0\\1\\1\\1\\1\end{bmatrix}$$

when the nodes are transmitting. Thus, the overall matrix of the entire network is as follows:

$$A=\begin{bmatrix}1&1&1&1&0&0\\1&1&1&1&0&0\\1&1&1&1&1&1\\1&1&1&1&1&1\\0&0&1&1&1&1\\0&0&1&1&1&1\end{bmatrix},$$

The whole matrix illustrates the interference relationship among all the nodes. Assuming that maximum capacity is reached, node 1 can transmit x_1 bits per second and x_2, x_3, x_4, x_5, x_6 for nodes 2, 3, 4, 5 and 6 (x_1 includes all the traffic generated by node 1 and the traffic routed through node 1). In addition, the maximum bandwidth for a link is b bits per second. The inequality of the network is presented as:

$$AX \le B, where \qquad X=\begin{bmatrix}x_1\\x_2\\x_3\\x_4\\x_5\\x_6\end{bmatrix}, B=\begin{bmatrix}b\\b\\b\\b\\b\\b\end{bmatrix}. \tag{1}$$

The maximum capacity $C_{network}$ of the network can be calculated by:

$$C_{network} = MAX(\sum_{i=1}^{6} x_i)$$

with a feasible *X* that conforms to inequality (1). More than one feasible *X* may be used to calculate $C_{network}$, and some of them may not be realistic from a communication point of view. Thus, more constraints can be added to reflect the reality of wireless communications. In [JPPQ03], wireless interference is incorporated into the formulation through the definition of a conflict graph, whose vertices correspond to the links in the connectivity graph. However, it is NP-hard to find the optimal throughput under the protocol interference model, and it is NP-hard to approximate the optimal throughput. After defining the independence vectors, which include several links that can transmit concurrently, the lower and upper bounds can be found. Some researchers use flow balance, capacity, and scheduling constraints together to estimate the overall end-to-end capacity [TKF05b]. In [RM02], a linear program is formulated based on game theory, and concentrates on the core capacity region. It considers both many-to-one and one-to-one traffic models. The solution in [TG03] defines a capacity (rate) region based on power, channel gain, noise, signal-to-interference and noise ratio (SINR), and time-division schedule first. It is then converted into linear programming in order to calculate the system capacity.

When considering mobility, the topology of the network becomes a time-dependent variable. Every time the topology is changed, the system capacity may change. Thus, we need to treat the network in every time period during which the topology is static. We then sum up the capacity in all these time periods to define the overall capacity as follows:

$$C_{network} = \left. \sum_{t_1}^{t_n} C_t \middle/ \sum_{t_1}^{t_n} t \right. .$$

However, the linear program method of estimating system capacity imposes a scalability problem. When the network size increases and the number of nodes rises, computation for a feasible X may take a very long time. If mobility is high, the frequency with which the topology changes also goes up. The computation for $C_{network}$ will then become even more complex.

3.1.2 Geometrical Analysis

In [GK00] and [NTV02], the per node capacity to be expected in an Ad Hoc network is estimated in a geographical manner. If node density is constant, this means that the total one-hop capacity is O(n), where n is the total number of nodes. As network size increases, the number of hops between each source and destination may also grow, depending on the communication patterns. One might expect the average path length to grow with the spatial diameter of the network, or equivalently the square root of the area, or O($\sqrt{n}$). With this assumption, the total end-to-end capacity is roughly O(n/$\sqrt{n}$), and the end-to-end throughput available to each node is

$$O\left(\frac{1}{\sqrt{n}}\right) \qquad 3.1$$

The analysis in [GK00] also demonstrates the existence of a global scheduling scheme that achieves a throughput of Θ(W/$\sqrt{n\log n}$) for a uniform random network with a random traffic pattern (W indicates radio channel capacity in bps).

To verify the asymptotic bounds, several simulations were run in NS2 (network simulator version 2) with varying network sizes and traffic load. DSR was selected as the routing protocol. In the contour plot in Figure 3.2, each line represents a specific packet loss rate. The packet loss rate increases from the bottom line to the top line. Thus, if the performance (or packet loss rate) remains constant, a larger network accommodates fewer communication pairs. Because a connection with more hops consumes more resources, the total bandwidth used remains the same within a certain range. The end-to-end capacity of the network roughly follows O(n/$\sqrt{n}$).

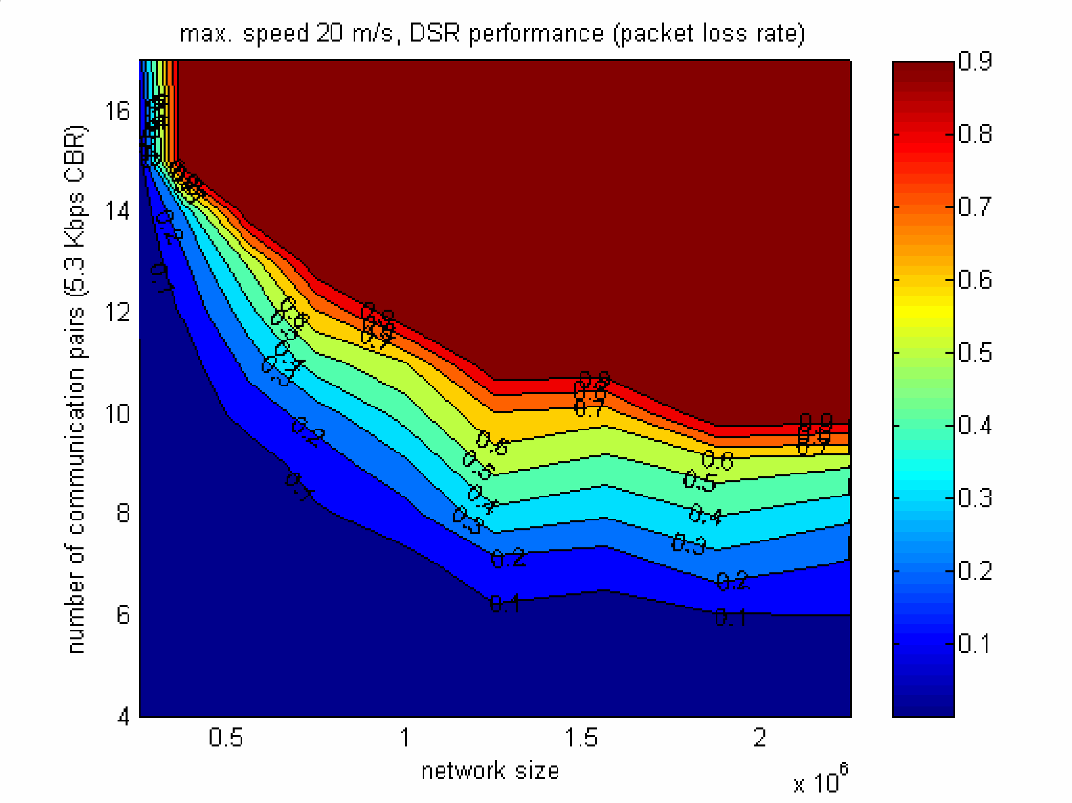


Figure 3.2 Contour Plot of Packet Loss Rate vs. Network Size and Number of Communication Pairs (Dynamic Source Routing)

Research in [GZ03] incorporates the walk/talk ratio, which is the ratio of the node connection rate to the overall acquired packet rate of the node. The asymptotic capacity of an additive white Gaussian wireless network under a relay traffic pattern is derived in [GV02]. The upper and lower bounds of system capacity are defined as: $\frac{1}{4}\log_2\frac{P}{D_1} \le C \le \frac{1}{4}\log_2(1+\frac{\|\alpha\|^2 P}{N})$, where P is the transmission power, N is the noise power, and α is a vector of weights of each receiver antenna. The multi-channel scenario is analyzed in [LHS02], and the system capacity with TCP regulated traffic is discussed in [BSMCGK02].

Geometrical analysis also has limitations when it is used to compute the system capacity of a real wireless Ad Hoc network. When considering the node mobility and random position of each node, it is very difficult to formulate the system capacity in a simple way. In most cases, only an upper bound is given under certain assumptions.

3.1.3 A Simple Solution for Calculating Capacity

Due to the computational limitations of linear programming and the complexity of geometrical analysis when dealing with random node positions and mobility, it is not easy to calculate the system capacity of an arbitrary wireless Ad Hoc network using the above methods. Hence, a simple method is provided to roughly estimate the network capacity via a geometrical solution, which is independent of the network topology and connection patterns. As mentioned in [LBCLM01] and seen in Figure 3.3, the interference range in IEEE 802.11 is far beyond the transmission range (for example, if a node can transfer data up to 250 meters away, its transmission will interfere with other nodes up to 550 meters away due to the SNR—signal-to-noise ratio [LBCLM01]). Thus, when node A is sending to node B, node C will not be able to send any data, because node C realizes the medium is busy. Only once node A has finished can node C start its transmission. For a two-hop connection (A to B and B to C), only one transmission takes place at any one time (A to B or B to C). For a three-hop connection (A to B, B to C, and C to D), there is still only one transmission at any one time (when A is sending packets to B, C will listen to the medium and ascertain whether it is busy). Therefore, in a three-hop connection, the throughput of the connection is just one-third of the total transmission capacity and the throughput of a two-hop connection is half.

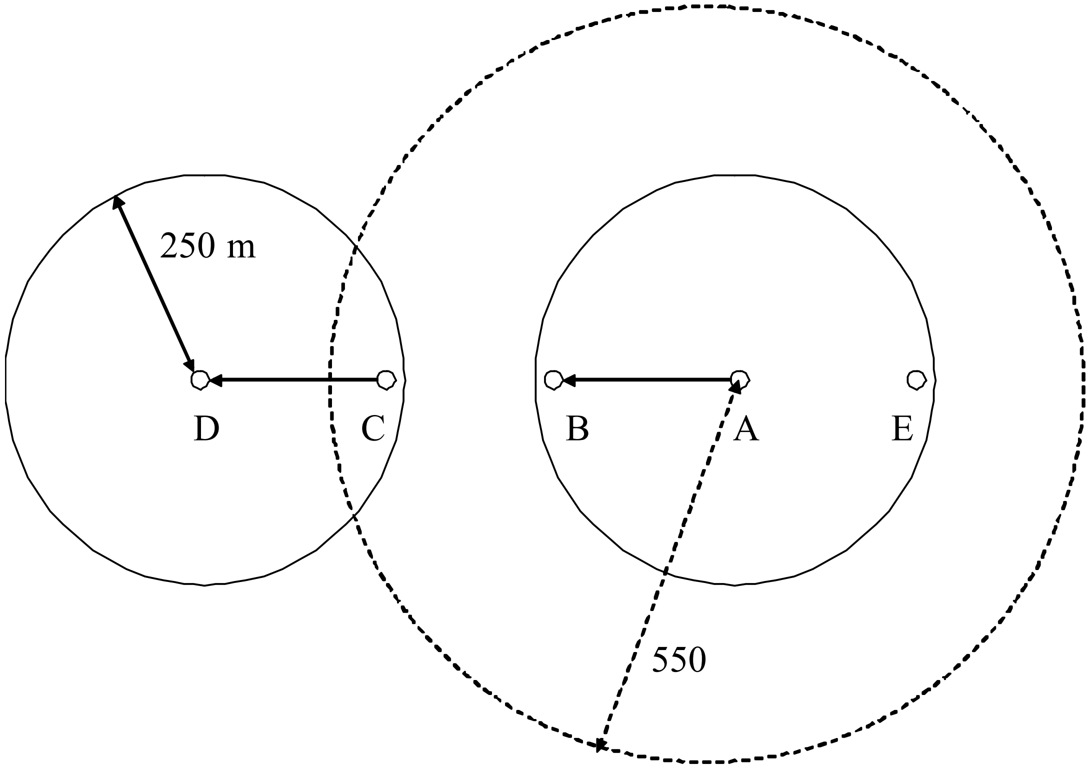
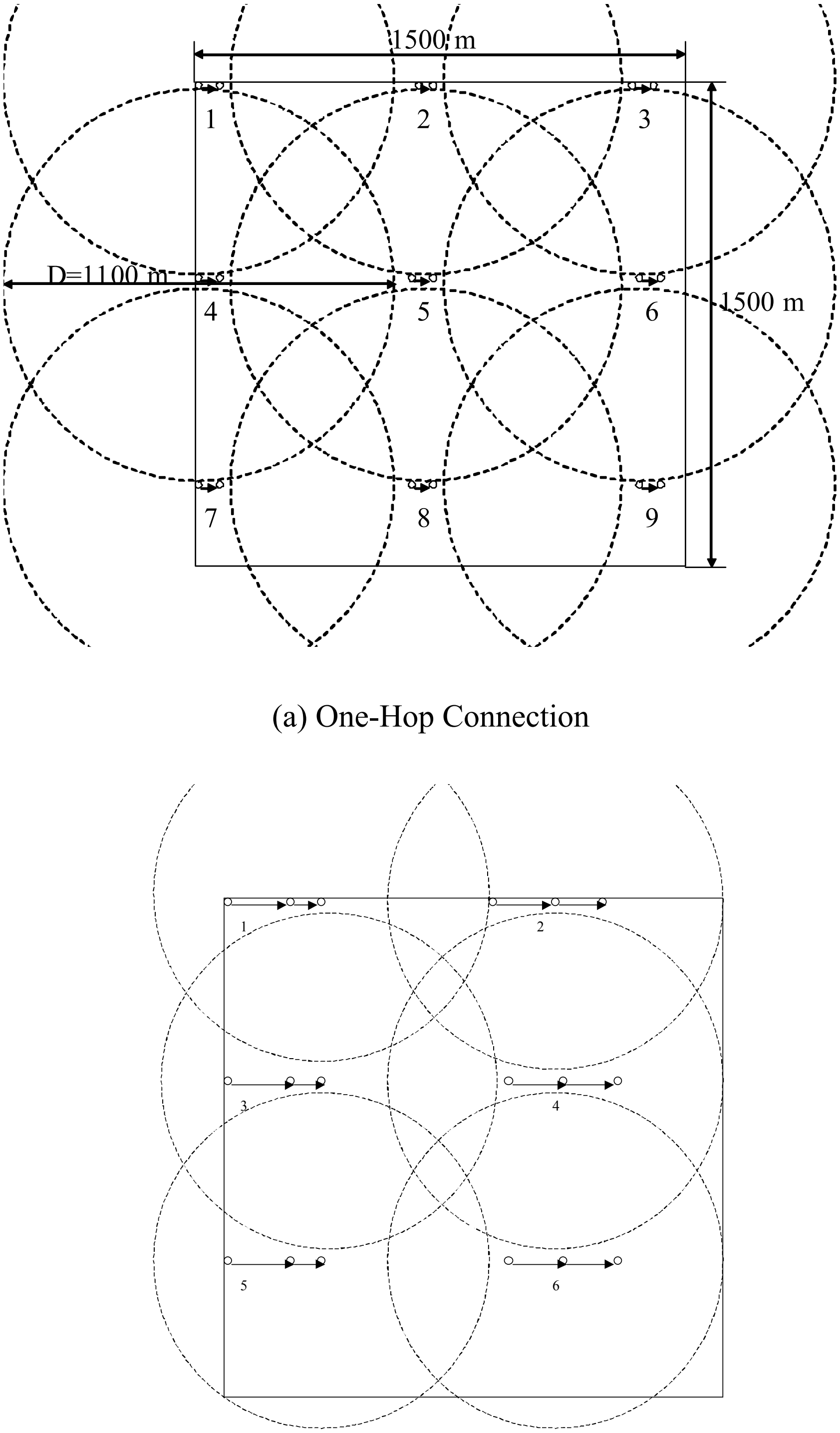


Figure 3.3 MAC Layer Transmission and Interference Range (Bold Circle Represents Transmission Range; Dotted-line Circle Indicates Interference Range).



(a) One-Hop Connection

(b) Two-Hop Connection

Figure 3.4 Optimal Number of Connections in a 1500x1500 Meter Square Area, Assuming Transmission Range is 250 meters and Interference Range is 550 Meters.

Looking at a network in a 1500x1500 meter square (m^2) area with the above deduction, assuming the transmission range is 250 meters (m) and the interference range is 550 meters and no power control is deployed, the largest number of concurrent transmissions can only be nine, if each pair of sender and receiver is just one hop away and the distance between them is very short (as shown in Figure 3.4 a; the interference range of each connection is a circle with a radius of 550). Thus, the total reachable system capacity is 9*W Mbps (W Mbps is the maximum transmission rate for each connection and depends on the allocated bandwidth for the system). If each connection is at least two hops away, the transmission range of a connection is bigger than 250 meters. The interference range of each connection is therefore at least a circle with a radius of 550+250=800 meters, and the maximum number of concurrent transmissions is six (Figure 3.4 b). Usually, the chance of nodes staying at the edge of the network is low. [BRS03, JBAS03, SM04] point out that all nodes in a wireless Ad Hoc network tend to be in the middle of the network for typical mobility models. Along with more hops per connection, random transmission direction, random positioning of nodes and the mobility of the nodes, the number of concurrent transmissions will decrease more than the previous number. The simulations showed that the number of concurrent transmissions was around two to three in a network of this size, because the chance of connections being along the edge of the network is small, and most connections go through the centre of the network. Assuming the maximum transmission rate of one direct link is 2 Mbps, the available system capacity will be around 3*2Mbps=6 Mbps.

Roughly, we could calculate the upper bound of network capacity by the following formula:

$$C_{network} = \frac{A_{network}}{\pi R^2} \times W \qquad \textbf{3.1.3}$$

where $A_{network}$ is the area of the whole network and R is the interference range of each node.

Thus, a 1500x1500m^2 network with an interference range of 550m has a capacity of around 2.36*W Mbps. If W = 2 Mbps, the upper bound of network capacity is around 4.72 Mbps.

The maximum node moving speed (m/s)	0	5	20
System utilisation (Mbps)	4.4	2.0	1.8

Table 3.1 System Utilization with Nodes Moving at Different Speeds for DSR

Table 3.1 provides one example of system utilization, measured by simulation in an overloaded, wireless Ad Hoc network of the abovementioned size and 100 nodes. The maximum capacity per link is 2 Mbps. Each connection generates 5.3 Kbps of traffic, with a packet payload of 64 bytes; the number of connections is increased until the packet drop rate reaches 10%. If all nodes in the network are static, the system utilization is very close to the estimated upper bound of the network capacity mentioned earlier. While the nodes are in movement, system utilization drops dramatically, to less than one-half of the estimated system capacity. This degradation in system utilization is caused by the topology changes due to the movement of the nodes. An ongoing communication is interrupted when a link is broken. A long backoff time before communication is resumed introduces a waste of bandwidth. In addition, the movement causes more collisions during transmission. That also leads to a long backoff time. A detailed analysis is given in the next chapter, where the routing protocol and MAC layer are discussed.

If traffic concentrates in certain areas of the network (sometimes several communications go through the same link), a broken link or collision will cause all traffic to be backed up, wasting even more bandwidth. When we look at the routing protocols and MAC layer protocols, we find that all traffic tends to concentrate or go through the same link.

The upper bound of network capacity estimated by this simple solution is based on a very simple physical layer model. The interference range and transmission range of a node only obey the exponential law for propagation, and both the interference area and the transmission area of a node are perfect circles. With shadowing, neither the interference range nor the transmission range of a node is a regular geometrical shape, and the size of the area covered changes. Then the formula used to calculate the upper bound of network capacity must be modified as:

$$C_{network} = \frac{A_{network}}{A_{interference}} \times W$$

Here $A_{interference}$ is the interference area of a transmitting node.

Furthermore, we only consider the interference of one transmitting node at a time. In fact, the combined interference from all concurrently transmitting nodes may lead to a larger interference area. Thus, the upper bound of the network capacity may be even smaller. In addition, after implementing some techniques, such as power control and directional antenna, at the physical layer, the interference area changes again and as a result, network capacity also changes.

3.2 Throughput and Utilization

After ascertaining the estimated network capacity, we need to calculate system utilization and throughput to see how the different routing and MAC protocols explore system capacity. The throughput of a network is the sum of the traffic load between any communication pairs, and improving it is our ultimate objective. This traffic load only captures the traffic generated by the sender. Therefore, it does not care about how many hops this connection needs. Assuming there are *n* connections in a wireless Ad Hoc network with *m* nodes in it, the total throughput of the network can be written as:

$$T_{network} = \sum_{k=1}^{n} \left(\prod_{j=1}^{h_k} \left(1 - PL_{j,k}\right) \right) R_k \qquad \textbf{3.2.1}$$

where h_k is the number of hops of the *k*th connection and R_k is the traffic load of the *k*th connection. $PL_{j,k}$ is the packet loss rate of the *k*th connection at the *j*th hop.

Unlike the throughput, the utilization of a network not only considers the total bandwidth occupied by each connection, but also includes the bandwidth that is used for control information and retransmissions. Thus, its expression is written as follows:

$$U_{network} = \sum_{k=1}^{n} \left(\left(\prod_{j=1}^{h_k} \left(1 - PL_{j,k}\right) \right) h_k R_k + \sum_{j=1}^{h_k} j PL_{j,k} R_k \right) + \Delta(m,n)$$

3.2.2

where *Δ(m,n)* refers to the traffic load that is generated by control information. The amount of *Δ(m,n)* is roughly exponential in the number of nodes *m* and roughly linear in the number of connections *n* for typical reactive routing protocols. The exact formulation of *Δ(m,n)* is based on the specific routing protocol, the traffic pattern, and the mobility pattern. Because it is not a primary focus of our analysis, we leave it as *Δ(m,n)* to indicate the amount of network resources consumed by the control packets.

It is obvious that the throughput of the network only depends on the packet loss rate of each connection from Formula 3.2.1. However, the utilization of the network takes into consideration the number of hops per connection and where the packet is lost.

The system capacity is limited, which means the utilization of the network is limited. So $U_{network}$ is bounded after the number of connections *n* reaches a certain value. Adding more communications to the system will only lead to a higher $PL_{j,k}$ to keep $U_{network}$ constant. A well-designed network should support more traffic load (more throughput) when utilization is at its the limit, or explore more system capacity that is wasted or unused. In order to do this, the relationship between the throughput and the utilization needs to be found. The relationship between the throughput and the utilization can be represented as:

$$
\begin{aligned}
U_{network} &= \sum_{k=1}^{n}\left(\prod_{j=1}^{h_k}\left(1-PL_{j,k}\right)\right)h_k R_k + \sum_{k=1}^{n}\sum_{j=1}^{h_k} jPL_{j,k}R_k + \Delta(m,n), \\
&= h_{\min}T_{network} + \sum_{k=1}^{n}\left(\prod_{j=1}^{h_k}\left(1-PL_{j,k}\right)\right)\left(h_k - h_{\min}\right)R_k + \sum_{k=1}^{n}\sum_{j=1}^{h_k} jPL_{j,k}R_k + \Delta(m,n), \\
&\quad (h_{\min} \geq 1)
\end{aligned}
$$

3.2.3

where h_{min} is the number of hops of the shortest path out of all the communications.

Obviously, from the above formula, the best way to improve network throughput is to increase network utilization, if the network utilization is far below the network capacity.

Formula 3.2.3 shows that there are several ways to improve throughput when the utilization of the network has reached its limit. First of all, *$\Delta(m,n)$* occupies a considerable amount of bandwidth for the control information. This control information includes the routing packets, update information packets, and so on. The amount of control information depends on the topology change rate and the way the protocol updates the information. Thus, the design of the protocol directly affects system performance. Secondly, where a packet gets lost plays an important role in how much bandwidth is wasted by the lost packets. It is obvious that packets dropped early will waste less bandwidth. Thus, if a method could let each node drop the packet based on the packet lifetime or number of hops, more system capacity could be reserved for successful communication. In addition, the number of hops has an impact on bandwidth occupancy. The lower the number of hops, the more connections can be supported. However, wireless Ad Hoc networks are designed to relay information over several hops. It is inevitable for them to support connections over several hops.

In most cases, because they have poor knowledge of the network topology, routing protocols might not explore system capacity fully. A subsequent bad routing decision may lead to large packet loss at some place in the network, which wastes a large amount of network resources. System performance can be improved if a routing protocol is able to make good routing decisions that explore unused network resources and decrease the amount of wasted resources, at the cost of more control packets, and if the resources occupied for these extra contol packets are less than the sum of the wasted and unused but explorable resources. The following chapters will discuss why system capacity is not fully explored and how it can be used more efficiently by an enhanced routing protocol.

3.3 Summary

One method of evaluating system efficiency is to look at how network capacity is used. Theoretical deductions for network capacity are mostly based on two methods: linear programming and geometrical analysis. Linear programming can give a good estimation of network capacity, but has the limitation of scaling for a large network. By making proper assumptions, geometrical analysis can formulate the relationship between network capacity and number of nodes. However, due to mobility and different traffic patterns, it is not easy to formulate an estimation of network capacity. To avoid scalability and complexity, a simple geometrical solution is given to calculate the network capacity by size of network and interference range of a node. Furthermore, formulas for computing system throughput and utilization are derived when all routes are known. Based on these formulas, several ideas for performance improvement are deduced, such as dropping data packets based on lifetime when interface queues are full.

Chapter 4. Problems and Challenges

Because wireless resources are precious, the efficiency of wireless networks becomes a crucial issue. One way to measure efficiency is to ascertain whether system resources are used evenly or fairly. If evenness and fairness are lacking in a wireless Ad Hoc network, some nodes in the network may take on some extra work in relaying packets for other nodes. The transmission of their own communication will be affected and the battery life of these nodes will be shortened. The latter case will lead to more topology changes, because the node goes down once the battery is exhausted. If all users share system resources fairly and provide service to other users fairly, wireless Ad Hoc networks can provide an efficient solution for communications.

Meanwhile, when traffic load goes up in a wireless Ad Hoc network, some areas become congested while other areas are in an idle state. Deviating traffic or forcing part of the traffic to areas with a light load may increase system utilization and throughput by decreasing wasted bandwidth and exploring unused bandwidth. It also provides a solution for wireless Ad Hoc networks to support communications with different requirements.

In this chapter, a detailed description of reactive routing protocols and the IEEE 802.11 MAC sublayer protocol are given, to illustrate the problems that may exist in the network. The reasons for these problems are discussed thereafter.

4.1 DSR (Routing Protocols)

Reactive routing protocols will be illustrated first, as these form the basis of our proposal. Our first implementation is based on DSR, which is an example of a reactive routing protocol, so DSR will be described here to explain the functionalities of reactive routing protocols. DSR includes the whole route in the packet header of each data packet, which allows for more options for optimization.

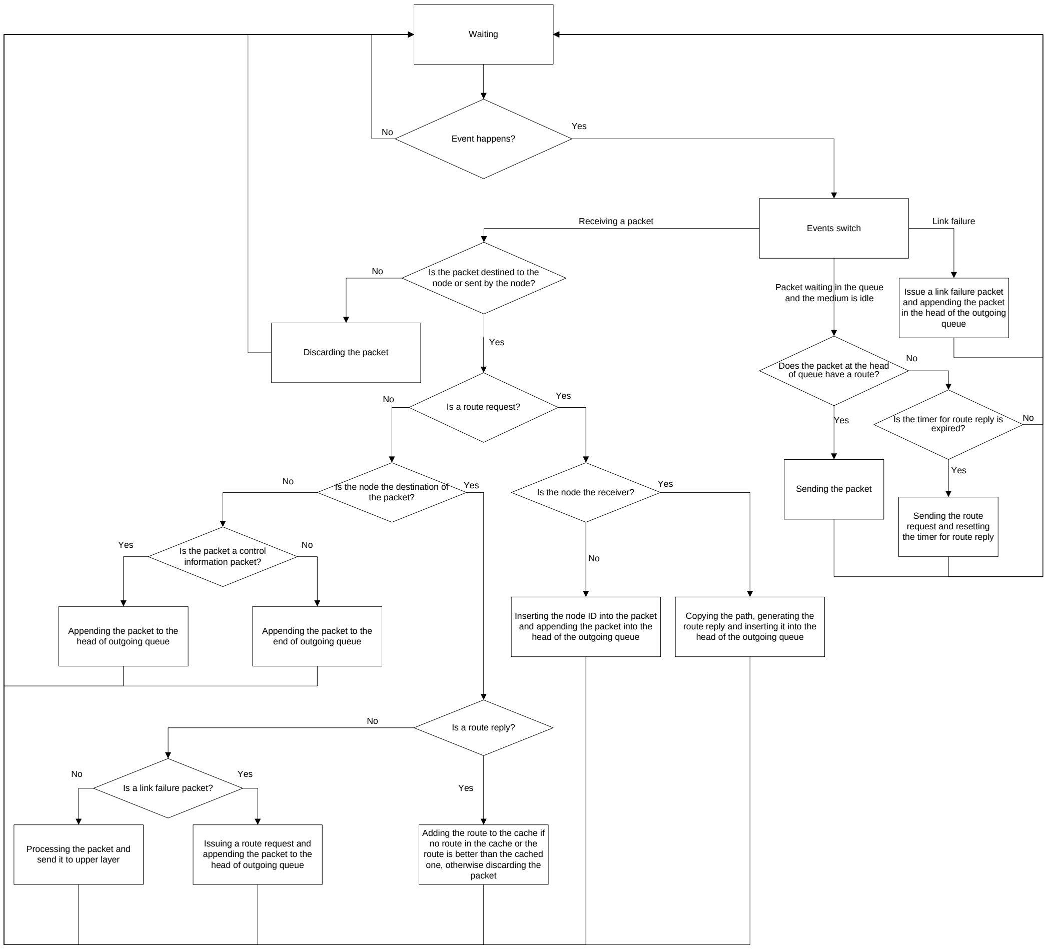

Waiting
No
Event happens?
Yes
Receiving a packet
Events switch
Link failure
No
Is the packet destined to the node or sent by the node?
Packet waiting in the queue and the medium is idle
Issue a link failure packet and appending the packet in the head of the outgoing queue
Discarding the packet
Yes
No
Does the packet at the head of queue have a route?
No
Is a route request?
Yes
Yes
Is the timer for route reply is expired?
No
No
Is the node the destination of the packet?
Yes
Is the node the receiver?
Yes
Yes
Sending the packet
Sending the route request and resetting the timer for route reply
Yes
Is the packet a control information packet?
No
No
Appending the packet to the head of outgoing queue
Appending the packet to the end of outgoing queue
Inserting the node ID into the packet and appending the packet into the head of the outgoing queue
Copying the path, generating the route reply and inserting it into the head of the outgoing queue
No
Is a route reply?
No
Yes
Is a link failure packet?
Yes
Processing the packet and send it to upper layer
Issuing a route request and appending the packet to the head of outgoing queue
Adding the route to the cache if no route in the cache or the route is better than the cached one, otherwise discarding the packet

Figure 4.1 Basic Flow Chart of DSR

DSR has two major functions: Route Discovery and Route Maintenance [BJM99]. Route Discovery works by flooding a request through the network in a controlled manner. When sending or forwarding a packet to some destination, Route Maintenance is used to detect if the network topology has changed, and whether the route used by this packet has broken. Each node along the route is responsible for detecting if its link to the next hop is broken. When the sender knows the route is broken, it can attempt to use another route that is already in its route cache, or it can invoke Route Discovery again to find a new route.

Figure 4.1 illustrates the basic flow chart of DSR. In route discovery, when a node needs to send a data packet without knowing its route, route requests are flooded to the network first. The protocol caches several routes (from different replies), but only the shortest will be chosen as the default route. The protocol caches these routes for future use. When any link is broken, a route error packet is generated and sent to the source node. A new route discovery process may be initiated by the source if the route is still in use. As the route request and reply packets have higher priority than the data packets, the first reply usually goes through the shortest or near shortest paths because the number of hops is lower.

DSR uses salvaging, gratuitous route repair, and promiscuous listening to optimize and maintain the routes:

- Usually, when a node forwards a packet and detects that the next hop is broken, it will drop the packet and send a Route Error message back to the source. However, if the node has an alternative route to the destination, it can attempt to salvage the packet by forwarding it to the destination via that alternative route, while still informing the source of the broken link via a Route Error message. When a source receives a Route Error message for a packet that it originated, the source propagates this Route Error to its neighbors by piggybacking it on its next Route Request.

- In Gratuitous Route Replies, the node overhears a packet not addressed to it and checks whether the packet's header contains its address in the unprocessed portion of the source route. If so, the node knows that the packet could bypass the unprocessed hops that precede it on the source route. The node then sends a gratuitous Route Reply message to the packet's source, giving it a shorter route without these hops.

- When a node forwards a data packet, it "snoops" on the unprocessed portion of the source route and adds to its cache the route from the node to the final destination listed in the source route, if this route is not already in the cache. In addition, the node overhears all the packets without filtering the address. Before being discarded, these packets are scanned for a useful source route or Route Error messages.

The implementation of DSR in NS2 (network simulator version 2) also includes other features, such as bidirectional source routes and one-hop route requests. Bidirectional source routes allow the sender and receiver to use the same route to transmit packets if the communication is bidirectional. If the generated traffic is bidirectional, switching this property off worsens performance, because more control information is inserted into the system. The benefits of a one-hop route request depend on the topology, mobility, and traffic pattern of the network. In a one-hop route request, the sender first sends a route request with a hop count equal to 1; thus, only the neighbors of the sender will receive this request. If they can find a route from their cache or they are the receiver, no more requests will be sent out. In this case, bandwidth can be reserved for data transmission instead of control information.

Like all reactive routing protocols, DSR discovers the new route or rediscovers the broken route by flooding the request to the network. As routing requests and replies have a higher priority than data packets, routing packets are always placed at the head of the queue and sent out as soon as possible. In any case, routes with a lower number of hops will typically be found earlier than those with long paths. In addition, only the addresses of intermediate nodes are included in the routing requests and replies. Nothing in the route requests and replies reflects the state of this particular path. As a consequence of flooding, the route found tends to be the route that has the least number of hops between sender and receiver. For example, as shown in Figure 4.2 after request flooding, the default routes between communicating pairs (S1-R1, S2-R2, and S3-R3) are indicated by the solid line, with arrows in both directions. The path length or the number of hops for these routes is close to optimal. In this situation (lots of senders and receivers are not in the middle of the network), most routes cross the middle of the network. In the example, all three connections go through node I3, and two connections go through nodes I6 and I4. All three intermediate nodes are around the middle of the network. Statistically, the chance that a connection will go through a node in the middle is much higher than through the nodes at the edge of the network. The result of this kind of routing algorithm is that traffic tends to concentrate in the middle of the network.

When traffic load goes up, congestion will occur. Because of the limited buffer size in each node, some packets (mostly data packets) have to be dropped when the buffer is full. Some packets have delay requirements and may be dropped due to the long waiting period in the queue. Congestion also leads to more collisions, which cause all transmitting nodes to back up. More bandwidth is then wasted during the backoff and the probability that this area stay congested is higher because there are fewer useful resources.

As the routing algorithm tends to use the shortest paths, which tend to go through the middle of the network, congestion in the middle will occur sooner than in other areas.

Because of this central congestion, system performance may not be maintained when more connections come in. In addition, as other areas are not saturated, network utilization may be far below the network capacity.

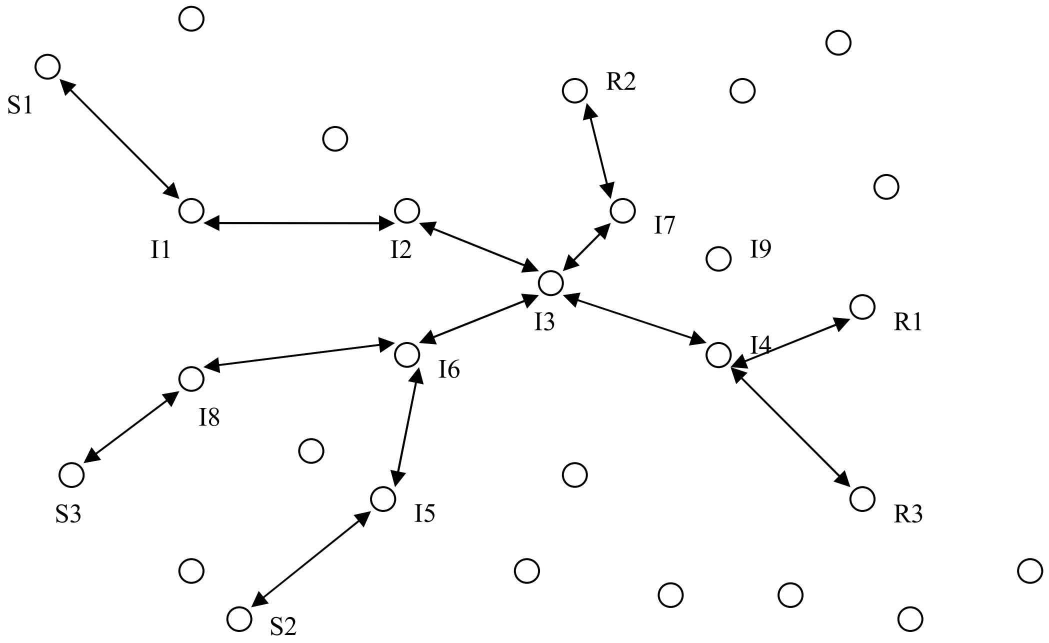


Figure 4.2 Example of a DSR Routing Result

Reactive routing protocols can use metrics other than the number of hops to decide the default path. For example, DSR can choose the default path based on the delay experienced by the route request and route reply. The variants of DSR can still use the same procedures in DSR. At the IP layer, where the routing protocol resides, nodes could put their IDs and timestamps, which are available to the routing protocols, onto the routing packets. The timestamp records the time when the node receives the packet. From the information in the routing packets, number of hops, end-to-end delay, and jitters can be ascertained. DSR is able to record the routing information in the past, and some other information, such as the expected hop count metric, can be deducted and used to select the proper path. However, because routing packets have a higher priority than data packets, routing packets are always inserted into the head of the queue and processed first, once they are received. Thus, for instance, routing packets experience a different delay, which is usually shorter than the delay experienced by data packets. The routing request and reply packets that have gone through the shortest path usually experience the shortest delay, due to a small number of transmissions and processing at intermediate nodes. As the default path selected on delay tends to be the same as the path selected by looking at the number of hops, the phenomenon mentioned previously still exists. DSR treats routing packets and data packets in different ways, so that the information included in the routing packets is not enough to decide on the traffic states in the network. All variants may lead to the same conclusion, because the routing packets only provide the topology information, or part of the load information, which is not enough for the routing protocol to avoid congestion.

DSR could be modified to retrieve extra information from the data link layer or the physical layer, in order to find out more about the states of the network. This extra information can include the interface queue length, the number of connections going through a node, the number of neighboring nodes, or the delivery ratios of the wireless link, which are acquired by some of the traffic distribution solutions described in Chapter 2 from the data link layer or the physical layer. As these solutions need to interact with the data link layer or the physical layer and to add extra information onto the routing packets, it is more suitable to call them an extension of DSR other than the variants of DSR, even through some routing functions are the same. The objective and deficiency of the DSR extensions are discussed in Chapter 2.

4.2 IEEE 802.11 (MAC layer protocols)

The MAC sublayer deals with resource access and plays a very important role in the wireless environment. Unlike wired networks, the MAC sublayer has to solve problems like hidden nodes, which may lead to numerous collisions, and high bit error rates, which may lead to the loss of many packets.

So far, IEEE 802.11 is the most popular MAC sublayer protocol, adopted by most proposals to support medium access in a wireless environment. It solves the hidden station problem, based on the IEEE 802.3 protocol. Typically, in a wireless Ad Hoc network, the distributed coordination function (DCF) of the IEEE 802.11 MAC layer protocol is used. In the following paragraphs, a detailed description of DCF is given. The benefits and problems in wireless Ad Hoc networks are then explained.

There are four types of interframe spaces (IFS) between transmitted frames:

- SIFS: short interframe space
- PIFS: PCF interframe space
- DIFS: DCF interframe space
- EIFS: extended interframe space

The relationship of the first three IFS is indicated in Figure 4.3. PIFS is used only by the Point Coordination Function (PCF), and DCF uses EIFS whenever the PHY (physical layer) has indicated to the MAC that a frame transmission has started that does not result in the correct reception of a complete MAC frame with a correct FCS value. As PCF is not suitable for wireless Ad Hoc networks, PIFS is not relevant in view of routing protocols and related problems. In addition, as we have assumed a simple physical model, there is no bit error during the transmission due to fading, and EIFS is excluded.

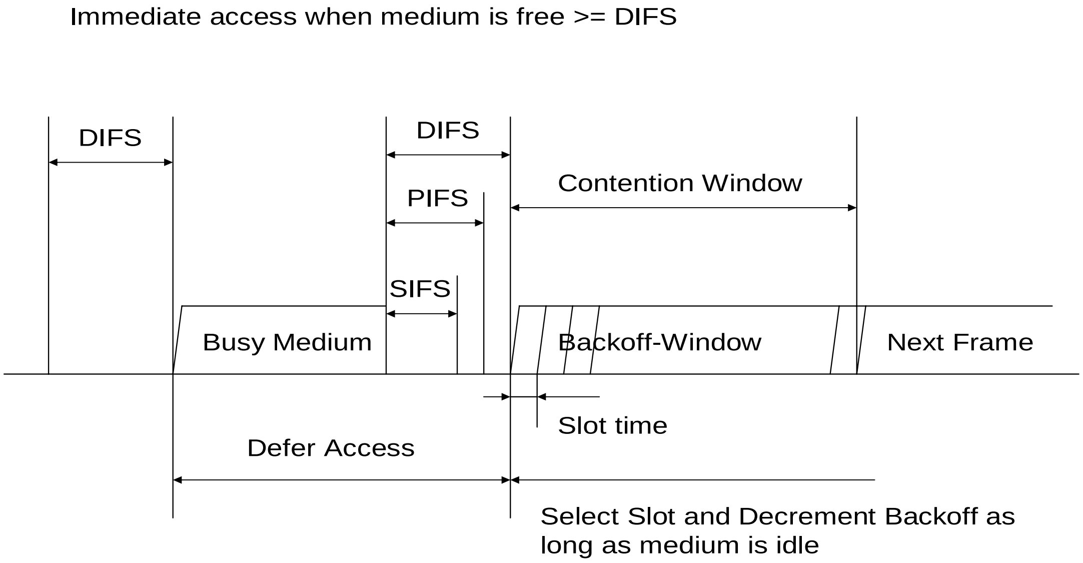

Immediate access when medium is free >= DIFS
DIFS
DIFS
PIFS
SIFS
Contention Window
Busy Medium
Backoff-Window
Next Frame
Slot time
Defer Access
Select Slot and Decrement Backoff as long as medium is idle

Figure 4.3 Some IFS Relationships

SIFS is the shortest interframe space. It is used when a node has seized the medium and needs to keep it while the frame exchange sequence is performed. Using the smallest gap between transmissions within the frame exchange sequence prevents other nodes, which wait for the medium to be idle for a longer gap (usually DIFS plus some random time period), from attempting to use the medium, thus giving priority to the completion of the frame exchange sequence in progress. A node uses DIFS to determine that the medium is idle.

The general procedure of transmitting a frame is as follows: A source first listens to the channel. If the medium is idle for at least a DIFS period of time, the source sends a transmission request (Request to Send, RTS). After receiving this request, the destination waits for a SIFS time to send back a reply (Clear to Send, CTS). The source then waits for another SIFS time to start sending data. Upon receiving the data frame, the destination sends an acknowledgement (ACK) back before waiting for another SIFS time.

After having transmitted a sequence of frames, the source returns to the contention period. All nodes that have packets to send will wait for a random period before attempting transmission. This random time period is chosen from 1 to CW (contention window size, which is a multiple of a slot time). The node that chooses the smallest value will start a new sequence of frames (Figure 4.4).

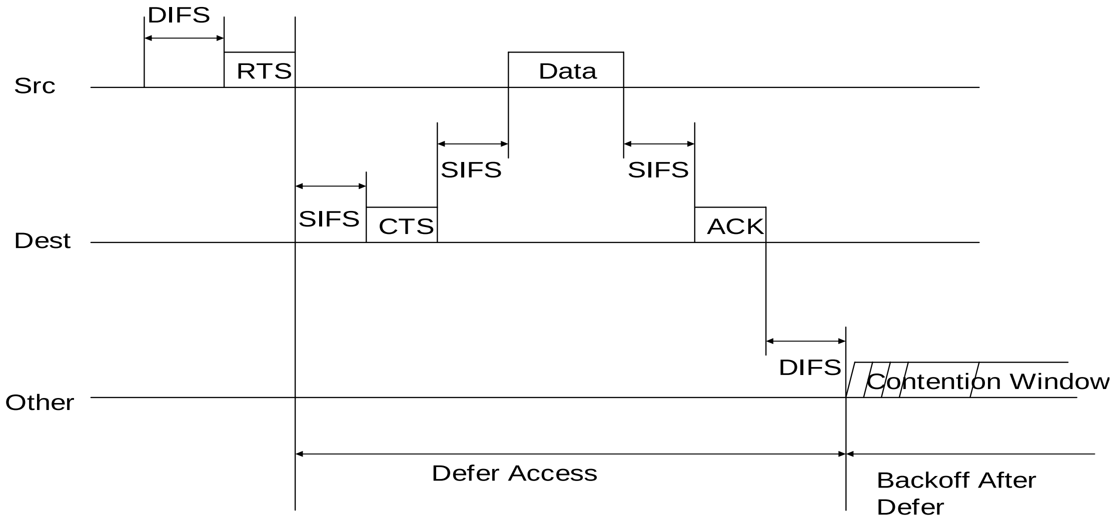


Figure 4.4 RTS/CTS/Data/ACK

If a node that has packets to send senses that the medium is busy, the node defers its access until a DIFS amount of time has passed after receiving an ACK. If the node has already backed off once, the value of CW has to be incremented according to the following formula:

$$\mathbf{CW_{new} = 2 * CW_{old} + 1;} \qquad \mathbf{4.1}$$

A similar situation occurs after a collision. In this case, two or more nodes try to send a frame in the same time slot. As lost packets are not acknowledged, the senders have to defer access and wait for a random amount of time. If a collision happens again, the value of CW increases exponentially.

When the traffic load goes up, more nodes are trying to transmit at the same time. More collisions take place and more nodes wait for the chance to transmit. Some of them have to increase CW more than once. Some time-sensitive packets may be dropped during the backoff time. However, there is a side effect, which may improve performance. Assuming there are several nodes in one area, trying to transmit more than one data packet, one node gets to transmit first and recaptures the medium after having completed the first sequence of frames. Thus, the CW of this node is kept at a low value, while all other nodes have increased the value of CW once. This node then has a high probability of continuing to occupy the channel if it has packets to send. Since the route request is a high priority packet, it always goes first. In this scenario, the node that is transmitting tends to forward the route request first and becomes the relay node for all traffic that goes through this area. (This is shown in Figure 4.2; although a route that passes through I9 for S1-R1 has the same number of hops, the default route still passes through I4 due to its fast-forwarding of the route request when I4 is already in a transmission state). As this node takes over all the traffic, the competition in this area decreases to a very low level. Thus, the chances of a backoff decrease, and the bandwidth utilization goes up.

However, all traffic going through one node will lead to buffer overload, which causes more data packets to be dropped. Although the resource utilization of this area will increase, the overall system performance may degrade. Moreover, if this node defers access, all traffic will experience this backoff time. If traffic is distributed to different nodes in the area, only some of them will have to wait for a backoff time, and the delay in the overall traffic may be less.

In any case, if the traffic load exceeds a certain level and this area of the network becomes congested, degradation in system performance is unavoidable. It is best to prevent this situation from happening, if possible. In the next chapter, a solution based on this idea is presented.

4.3 Challenges

Although [GT02] theoretically shows that mobility increases system capacity, nodes need to have a queue with infinite size, and delay is not considered to be a performance issue. In Section 3.1.3, system capacity is estimated by a simple solution, and it is illustrated through the example in that section that system utilization is far below system capacity when nodes are in movement. Node movement causes two major problems: 1) Link breakage, when the distance between two nodes with an ongoing connection is above the maximum allowed range (the received signal falls below the required power level), which generates extra error messages to inform the sender about the broken route. The sender then has to rediscover the routes, generating more extra control packets. System utilization may not change much in this case, but a large amount of bandwidth is consumed for routing information, and the throughput of the network is degraded. 2) Collisions, when two nodes in transmission step into each other’s interference range. The collided packets need to be retransmitted, occupying extra bandwidth. Due to the existence of the MAC layer protocol (IEEE 802.11), both nodes wait for a long backoff time after a collision, and if the next transmission is not successful, the backoff time is even longer. These backoffs consume system resources. Moreover, if the traffic load is very high in the area, collisions take place more often.

When links break, to avoid or reduce rediscovering the route, nodes should listen to the medium continually and cache all routes for future usage. However, as mentioned in the previous section, the IEEE 802.11 MAC layer protocol causes traffic to go through the same node or the same direct radio link, if possible. The nodes may then not have any alternative paths in the cache and a route rediscovery will have to be performed. In addition, continually checking the route in all packets adds burden to mobile equipment when battery power is precious. Some routing protocols may not allow the intermediate node to change the path of the packet, so that extra packets may be needed to inform the senders about alternative paths. The benefit from having a smaller number of route rediscovery messages may then disappear.

The number of collisions depends on the traffic load around the area. The heavier the traffic is, the more the collisions are. Limiting the traffic load to decrease the number of collisions may sacrifice the utilization of the system. What is more, limiting traffic load in this distributed system is not a simple task. Coordination among the network nodes requires bandwidth for information exchange.

Furthermore, more network information is required in order to support QoS in wireless Ad Hoc networks. Because of the limited resources and congestion in certain areas, a node cannot just admit a new connection into the network. Before data transmission, the node needs to get permission from the relay nodes. Some negotiations need to be implemented, such that some ongoing connections can yield some bandwidth by degrading their QoS to establish the new connections. The extra information can be included in the route requests/replies, or it can be carried in particular control packets. Deciding what kind of information is essential for supporting QoS in wireless Ad Hoc networks is also a challenging issue.

4.4 Summary

Low system utilization is mainly caused by traffic concentrating in certain areas of the network. Unevenly distributed traffic leads to a large amount of collisions in congested areas, and low usage of the medium in other areas of the network. Both situations waste network resources. With respect to routing algorithms such as DSR, the shortest path between sender and receiver uses the lowest amount of system resources and may need the least amount of time for packets traveling end-to-end. However, the shortest path algorithm results in traffic going through the central area of the network. The MAC layer protocols, especially IEEE 802.11, try to solve the problems inherent in wireless medium access, such as hidden terminals, while at the same time resulting in the transmitting nodes having an advantage over the nodes that are waiting for transmission. Thus, traffic concentrates even more on several nodes. Once these nodes experience problems such as link breakages and collisions, the traffic as a whole suffers.

Chapter 5. Traffic Balancing

The network topology keeps changing because of mobility and signal fading, and this causes routing protocols to generate a large amount of overhead in the form of control messages (routing update information). Thus, system performance decreases dramatically under these protocols when the rate of movement and temporary link failures in a system goes up. Furthermore, system efficiency also goes down, due to more collisions, congestion and control information packets.

The analysis of wireless Ad Hoc networks shows that the traffic load in a system is not evenly distributed. The central areas tend to carry more traffic, and therefore transmission is congested. In congested traffic areas, data packets may be dropped due to the fixed length of interface queues and excessive delays caused by the waiting time in the queue.

Thus, a new solution must be sought to explore those underutilized resources, without losing the virtues of routing protocols that support the dynamic environment of the system. As routing protocols such as DSR are reactive routing protocols, the routing overhead is lower compared to other types of protocols, especially when the topology changes frequently. In addition, system performance under the reactive routing protocols is similar to that under the proactive routing protocols, while the computing resources, such as CPU times and memory, are lower due to the simple implementation of reactive routing protocols and the fact that less information is required. The convergence time of reacting to broken links is also shorter, compared to that of the proactive protocols. The chance of routing loops under reactive routing protocols is much lower than the other protocols. Based on these issues, a method called Traffic Balancing, which is implemented based upon reactive routing protocols, is proposed to expand system capacity and improve system performance by migrating the traffic load from the congested areas to the idle areas.

5.1 Motivation

As mentioned earlier, wireless communication experiences severe impairment from the environment. Even if all portable equipment is stable at all times, the direct radio connection may be interrupted due to propagation impairments or objects moving inside the network area. Meanwhile, the mobility of the nodes causes the radio connection to be lost when the distance between two transceivers is far. All these scenarios lead to network topology changes and the system has to restart route discovery or use other methods to keep ongoing communication from being interrupted.

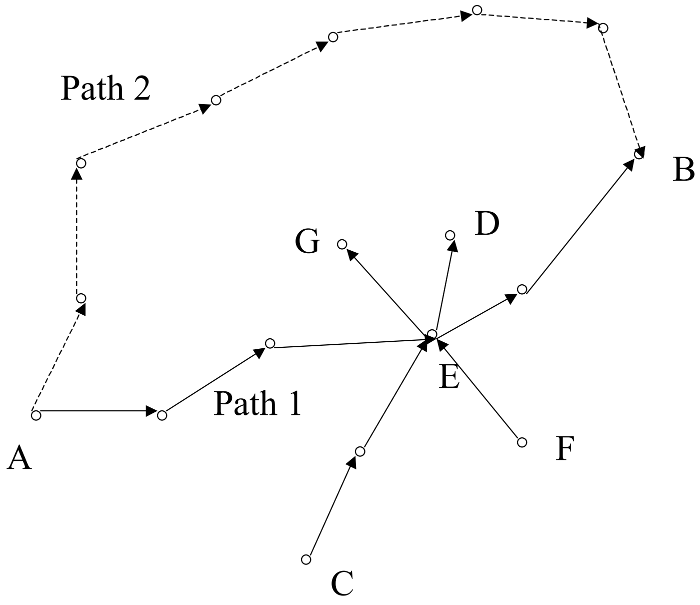


Figure 5.1 Selected Paths According to the Routing Protocols.
(Bold line indicates those chosen by DSR)

By default, most proposed routing protocols try to discover the shortest path between the sender and the receiver. The path through which a route reply goes back to the route requester first may not be the shortest path, but usually it is close. In addition, as routing protocols usually give higher priority to control information packets than data packets, a node with a full queue may still be chosen as part of the path. Due to the MAC layer protocol, traffic loads tend to go through the same node or direct radio link in an area.

For example, in Figure 5.1, nodes A, C, and F communicate with nodes B, D, and G respectively. According to the reactive routing algorithm, the broadcast route requests and replies have higher priority than data packets. Usually the route with fewer hops will return the reply to the sender faster than other possible routes, and be accepted by the sender as the default route, and the selected paths are likely to be the ones indicated by the bold lines. All of them will pass through node E in order to achieve the shortest distance. When the traffic load is high, the middle of the network (the area around node E) will be congested (more collisions may take place and data packets have to be retransmitted), while other areas (such as the area around the dotted line in Figure 5.1) remain in a less loaded state. Because of the poor path selection, the overall system utilization is far below the theoretical limit, even if the traffic load becomes very high. Because the queue length of each node is fixed, when the queue is full, new incoming data packets have to be dropped. Moreover, frequent transmission collisions lead to a large number of retransmissions and longer backoff time, which cause packets to wait longer in queues. Some packets are dropped because they exceed their allowed lifetime.

Figure 5.2 plots one example of a measured traffic load in a simulated 1000x1000m^2 network with 100 nodes. The routing protocol is DSR and the maximum node mobility is 20m/s. As discussed in the previous paragraph, due to the routing protocol and MAC protocol, the traffic density in the center is much higher than at the edge, as shown in

Figure 5.2. Thus, the chance of collision in the center is quite large, and bandwidth is wasted during the backoff period and for retransmission.

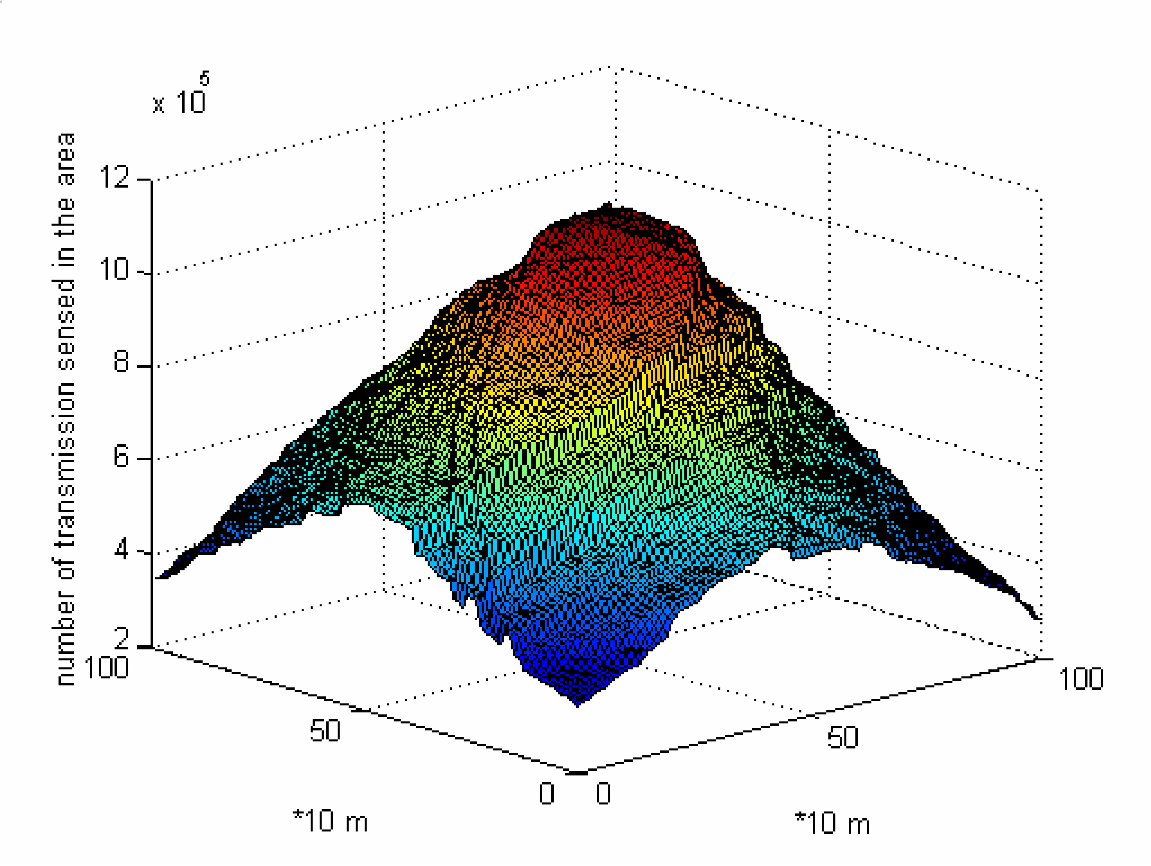


Figure 5.2 Traffic Load Measured in a 1000x1000m^2 Network, Simulated in NS2 (Routing Protocol: DSR; Node Mobility: 20m/s)

5.2 Traffic Balancing

Traffic Balancing is a routing approach that explores unused resources around areas that have a light load. For example, instead of selecting path 1, indicated by the solid line between nodes A and B in Figure 5.1, path 2—indicated by the dotted line—is used for packet delivery for node A and B, and gives node E a good chance to fully support communication between nodes C and D, and nodes F and G. This can either expand the system's ability to support more communications under the same performance level, or improve the performance of the system. The solution is based on a reactive routing protocol and it can be implemented with any reactive or proactive routing protocol.

A few issues need to be addressed before discussing the solution. First of all, each node should have the ability to record the usage of the medium around itself. In fact, all information required can be collected from the MAC layer, because it is monitoring the medium at all times in order to send or receive packets. The measured results will help the node decide if the medium in its area is overloaded or not. Our protocol function is implemented in such a way that each node records the state of the medium in the past *n* milliseconds. If a node detects that the received power is greater than the noise threshold for a certain period (equal to or longer than the time needed to transmit the smallest frame), the medium is recognized as being used by other nodes. The duration of the state that the medium is occupied is recorded and accumulated in order to calculate the percentage that the medium is busy. In the MAC layer of each node, a linked list is required to record every busy period of the medium. The procedure for **medium measurement** is as follows: Each node continues to sense the received power level. Once the power level has gone over a certain threshold (noise threshold), the medium is assumed to be used for a transmission, which could be a transmission by the node itself. The MAC layer functions then record the start time (T_{start_time}) and the duration of the transmission ($T_{duration}$), and add them to the linked list as a new element. Meanwhile, all the elements in the linked list are checked, and those elements that do not satisfy the following condition are removed:

$T_{current\_time} – (T_{start\_time} + T_{duration}) < n,$

It also means that the medium occupations happened *n* milliseconds ago are removed from the linked list because they are not counted for calculating the medium usage. When the node forwards a route request, the MAC layer functions check the linked list again and accumulate the elements that obey the previous condition, as follows:

$T_{busy} = \Sigma\ T_{duration},$

The node can then verify the percentage of medium usage in the past n milliseconds time periods by:

Measured Medium Usage $= T_{busy} / n,$

Now that it has the value of the measured medium usage, the node can decide if it is in a busy area or not. Overall, the MAC layer protocol updates the linked list once the medium is used, and calculates the medium usage when a route request needs to be forwarded. The choice of the value of *n* depends on the types of traffic in the network. If the traffic is bursty, the *n* (measurement period) should be small. Otherwise, the *n* should be large.

The second issue is the introduction of an additional bit or byte to indicate the medium usage in the header of the route request, which is set by the sender. The reason why this bit or byte is used will be explained later, when the details of the solution are illustrated.

Thus, two parameters have to be defined before the implementation of Traffic Balancing: the medium usage threshold *p* and the measurement time period *n*. The measurement time period has already been mentioned. The medium usage threshold is used to verify whether the medium is busy or not. The choice of this parameter is based on the MAC layer protocol and the routing protocol. The traffic model also has an impact on this parameter.

There are two ways to implement Traffic Balancing, depending on who makes the routing decision: the sender or the intermediate nodes.

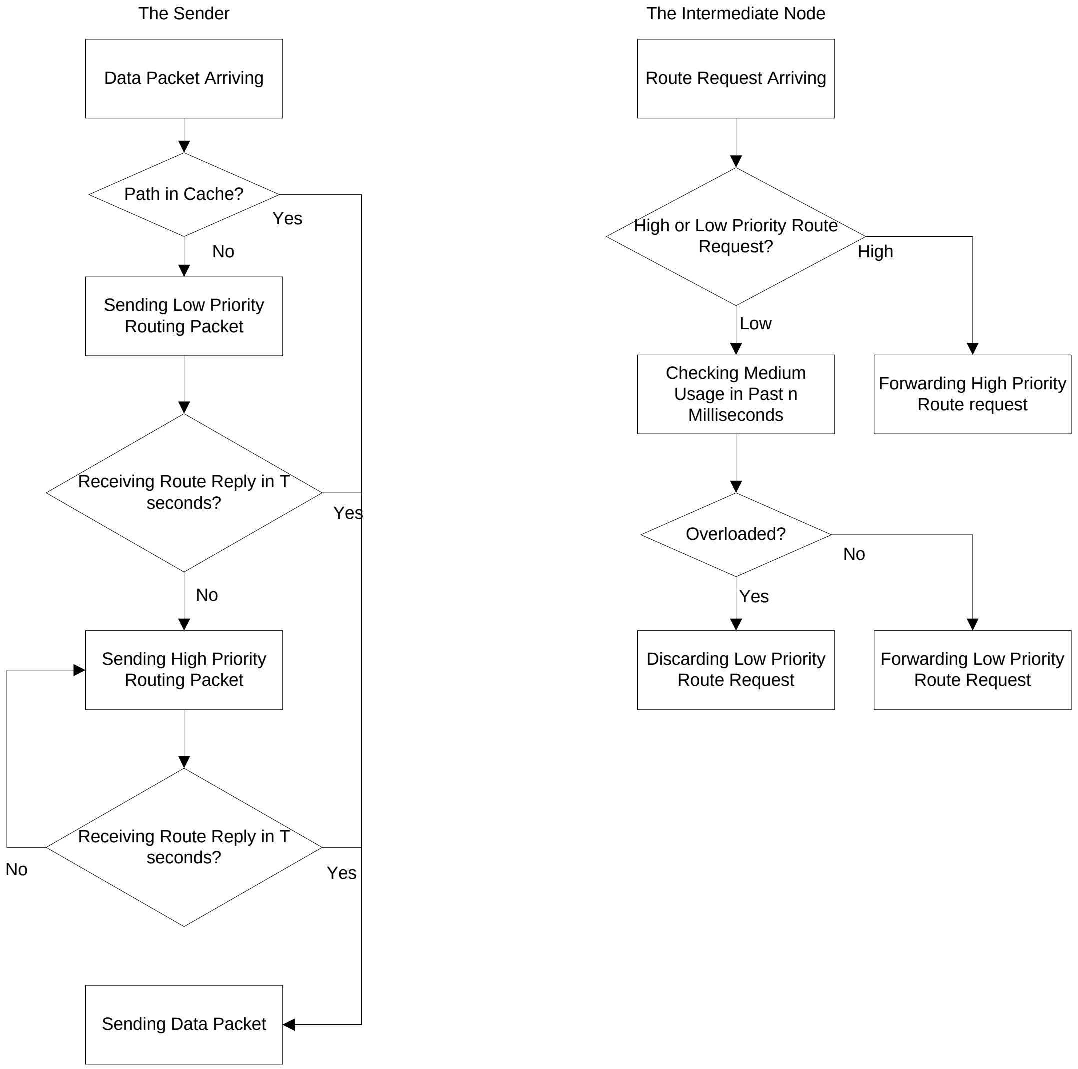

The Sender
Data Packet Arriving
Path in Cache?
Yes
No
Sending Low Priority Routing Packet
Receiving Route Reply in T seconds?
Yes
No
Sending High Priority Routing Packet
Receiving Route Reply in T seconds?
No
Yes
Sending Data Packet
The Intermediate Node
Route Request Arriving
High or Low Priority Route Request?
High
Low
Checking Medium Usage in Past n Milliseconds
Forwarding High Priority Route request
Overloaded?
No
Yes
Discarding Low Priority Route Request
Forwarding Low Priority Route Request

Figure 5.3 Flow Chart, Traffic Balancing Solution 1.

Solution 1 (decision made by intermediate node): This solution tries to find a path without any congested intermediate nodes and needs two types of routing requests: low or high priority, requiring one bit in the packet header to distinguish them (0 for a low priority route request and 1 for a high priority route request). Figure 5.3 illustrates the flow chart for Traffic Balancing Solution 1, which is different from the basic DSR routing protocol in functionality. For the intermediate nodes: When a node receives a general route request (low priority route request), it first checks the medium usage around it in the past *n* milliseconds. If the medium usage is higher than the medium usage threshold, the node will ignore this route request. Otherwise, it processes this route request the way it usually does for any reactive routing protocol. When a node receives a high priority route request, it will process it regardless of the medium usage around it. For the sender: To initiate a path, the node first broadcasts a low priority route request. If there is no response after a back-off time (*T*), a high priority route request is generated and broadcasted [TFK02].

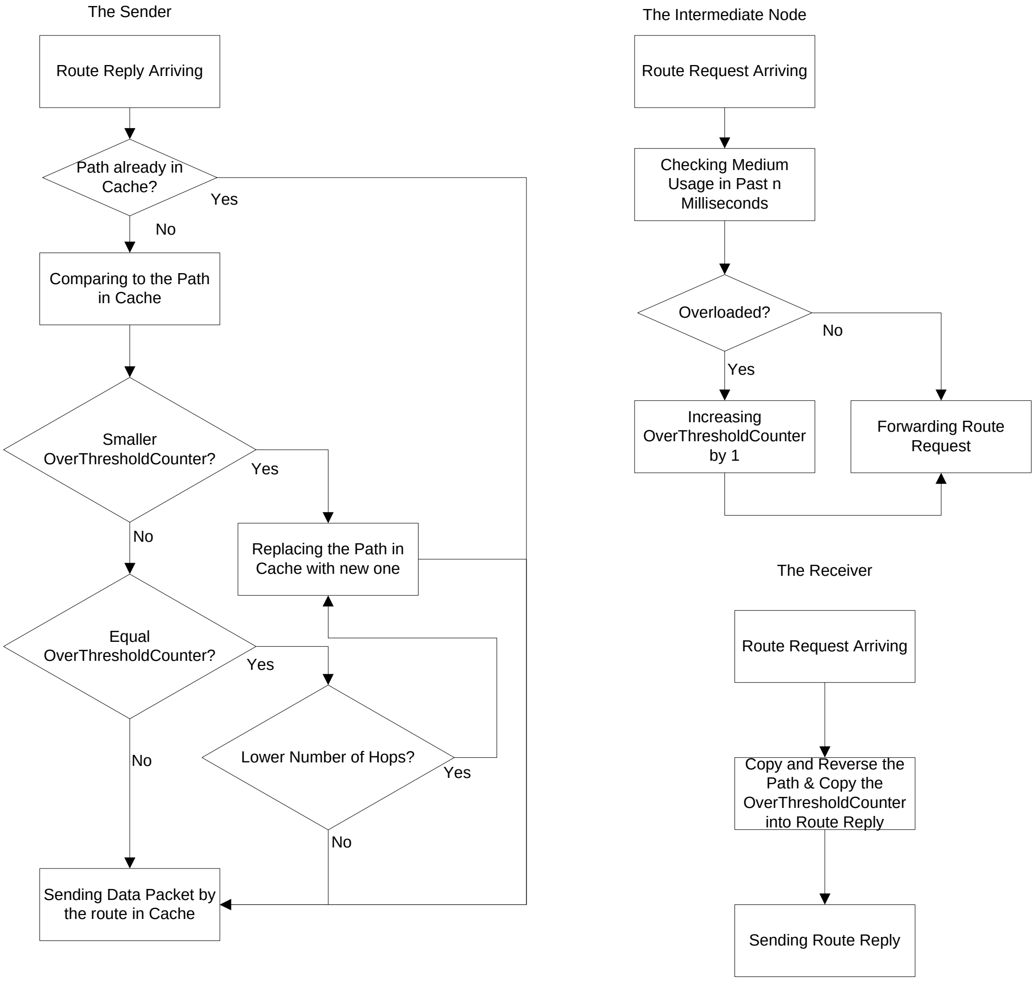

The Sender
Route Reply Arriving
Path already in Cache?
Yes
No
Comparing to the Path in Cache
Smaller OverThresholdCounter?
Yes
No
Replacing the Path in Cache with new one
Equal OverThresholdCounter?
Yes
No
Lower Number of Hops?
Yes
No
Sending Data Packet by the route in Cache
The Intermediate Node
Route Request Arriving
Checking Medium Usage in Past n Milliseconds
Overloaded?
No
Yes
Increasing OverThresholdCounter by 1
Forwarding Route Request
The Receiver
Route Request Arriving
Copy and Reverse the Path & Copy the OverThresholdCounter into Route Reply
Sending Route Reply

Figure 5.4 Flow Chart, Traffic Balancing Solution 2.

Solution 2 (decision made by the sender): This solution tries to find a path with the smallest number of congested intermediate nodes, and needs one byte in the route request packet header (*overThresholdCounter*) to indicate the number of nodes in the path that are overloaded (assuming a maximum number of nodes on any given path $< 2^8$). The *overThresholdCounter* is set to 0 by the sender before the route request is sent out. Figure 5.4 shows the functionality of Traffic Balancing Solution 2, which is different from the functionality of DSR. For the intermediate nodes: When a node receives a route request, it calculates the medium usage around it. If the medium usage is over the medium usage threshold, the node increases the *overThresholdCounter* by one. Otherwise, it simply follows the regular procedure and forwards it. For the receiver: After having received the route request, the receiver feeds back the threshold counter to the sender via the route reply. For the sender: The sender initiates the route request with an *overThresholdCounter* equal to 0. Upon receiving the route replies, the sender chooses the route with the smallest *overThresholdCounter*. If more than one path has the same smallest *overThresholdCounter*, the one with fewer hops is chosen. If more than one path has the same *overThresholdCounter* and the same number of hops, the sender chooses the one that arrives first [TKF04].

5.3 Performance Study

Simulators like Opnet Modeller, NS2 (Network Simulator 2), or GloMoSim support the definition of mobile Ad Hoc scenarios and are used for evaluating the behavior of wireless Ad Hoc routing protocols. The different simulators have different ways of implementing the functionality at each layer, and it is hard to ascertain which is best. In this research, the solutions are simulated in NS2 version 2.1b8 because many researchers around the world use it. In [KCC05], a research is conducted that shows over 40% researches using NS2 to simulate wireless Ad Hoc networks. Furthermore, it also shows that the simulation is an effective way to explain the benefits of the proposed solutions. Thus, it is convincing to compare the simulation results with those of other proposed solutions, which are also simulated in NS2. A very simple physical model is adopted in NS2 that used only exponential radio propagation law, as follows:

$$P_r(d) = \frac{P_t G_t G_r \lambda^2}{(4\pi)^2 d^4 L} \qquad \mathbf{5.1}$$

where G_t and G_r are antenna gains at transmitter and receiver, λ is the wavelength in meters, d is the distance between transmitter and receiver, L is the system loss factor not related to propagation ($L \geq 1$), P_t is the transmitted power, and P_r is the received power. No fading and shadowing were considered, which limited the reality of the simulations. NS2 also had scalability problems. When the number of nodes in the network was very large, the simulation became very slow and the size of the generated trace file was too big for the operating system to handle. IEEE 802.11 MAC layer protocol was chosen for the simulation. Constant bit rate (CBR) was selected as the traffic model, modeling voice communication. Assuming that two users were having a voice communication, the communication included two nodes sending data at a rate of 5.3 Kbps symmetrically to each other [CM99]. Each data packet had a payload of 64 bytes. The simulation time period was 300 seconds and the measurement was started after the first 100 seconds, because the system state became stable after 100 seconds (as is explained in later paragraphs). The maximum moving speed of each node was 5 or 20m/s, and all nodes continued to move around (no pause at any place). The original node position was set uniform-randomly in the network.

The medium usage threshold was set to 0.7 (in 70% of cases the medium was sensed to be busy), because during the early measurements it was found that when medium usage reached 70%, the collision rate increased dramatically. In fact, the best medium usage threshold depends on certain network characteristics, such as node mobility, and how medium usage threshold can be adjusted dynamically is discussed later in this chapter. The time period for the measurement was two seconds (i.e., when a node needed to forward a route request, it calculated the medium usage during the past two seconds).

Number of nodes	100
Maximum moving speed (m/s)	0, 5, 20
Pause time (seconds)	0
Network size (m^2)	500x500 to 1500x1500
Generated traffic	CBR at 5.3 Kbps
Connection type	Bi-directional
Data packet size (bytes)	64
Simulation period (seconds)	300
Medium usage threshold	0.7
Medium measurement period *n* (seconds)	2
MAC	IEEE 802.11

Table 5.1 Major Parameters Used in the Simulation

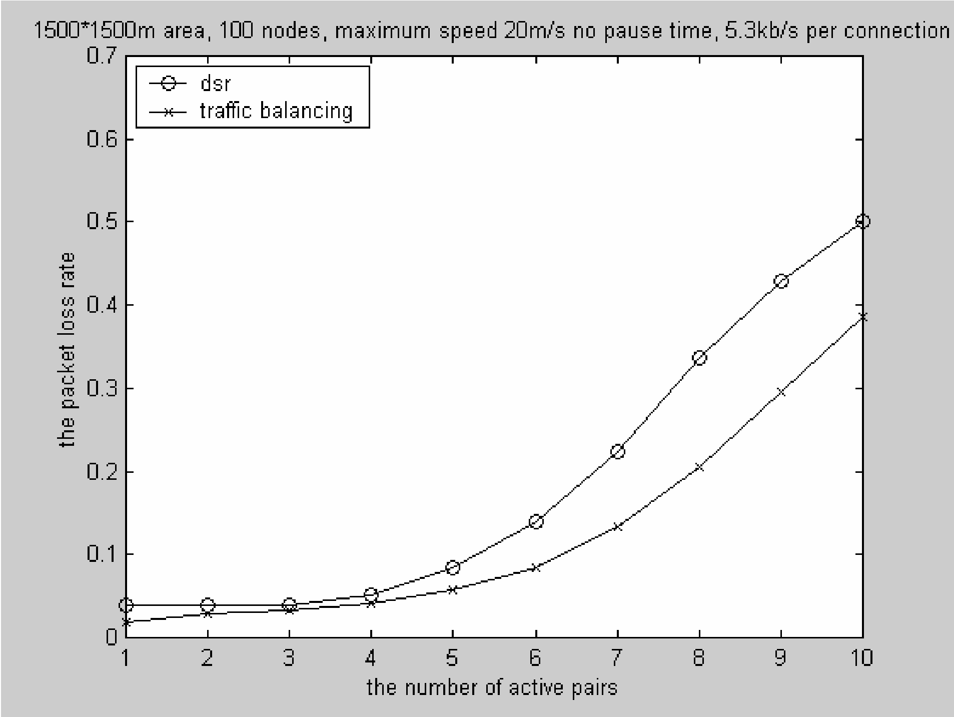
Table 5.1 lists all the major parameters that might have affected the results in the simulations. In the following sections and in the next chapter, we will discuss how these parameters impact on the performance of wireless Ad Hoc networks. In all the simulations, the nodes started generating a traffic load from 0 to 50 seconds and the performance was measured after the simulation had been running for 100 seconds. After running for 100 seconds, the network was in a steady state. For example, when the total traffic load was medium, the average queue length for transmitting nodes stayed at about 25 (the maximum queue size was 50 packets). In addition, if all the nodes in the network were stationary, the difference in average queue length and delay for every 20 seconds was less than 5% after the first 100 seconds. Table 5.2 lists the measured average queue size and delay in the fixed wireless Ad Hoc network simulated in NS2, after 100 seconds. It indicates that the system performance after 100 seconds is relatively stable.

Time Period	Average Queue Size	Average Delay (seconds)
From 100 to 120 seconds	24.6	2.1585
From 120 to 140 seconds	25.1	2.2073
From 140 to 160 seconds	24.3	2.1331
From 160 to 180 seconds	25.2	2.2115
From 180 to 200 seconds	24.9	2.1878

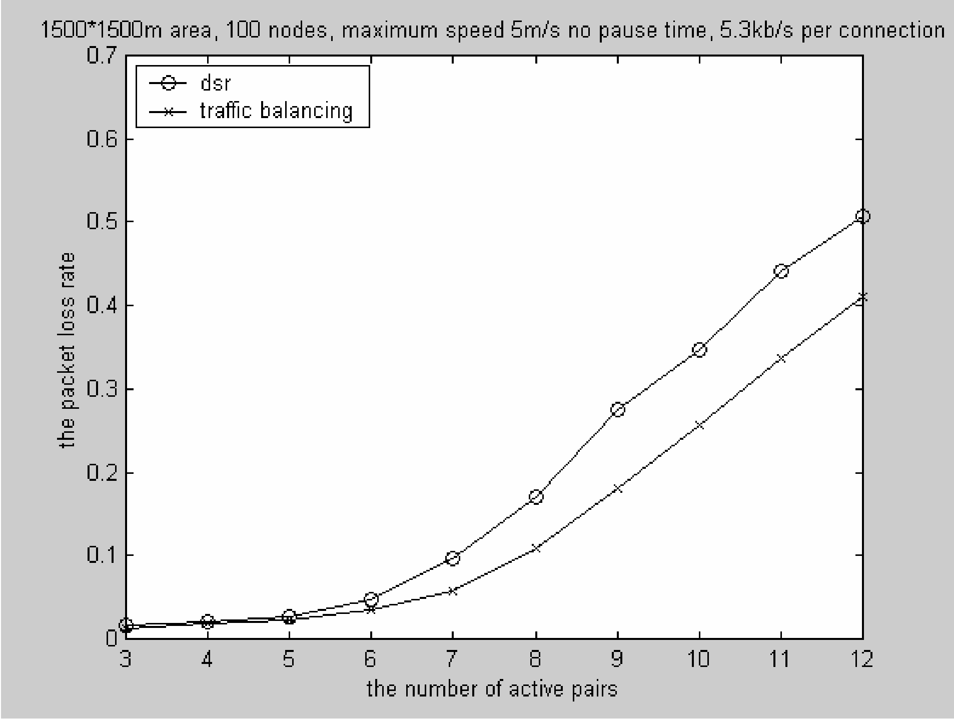
Table 5.2 Measured Average Queue Size of Transmitting Nodes and Average Delay Experienced by Data Packets after Network Running 100 Seconds.

5.3.1 Solution 1 (Intermediate Node Decision)

The network area was set to 1500 meters in length and 1500 meters in width, and 100 nodes were located randomly in this area. Each data point was collected from an average over 10 runs (i.e. 10 scenarios). The average packet loss rate and packet delays were measured for different numbers of active communication pairs to compare the performance of DSR and our solution. As mentioned earlier (see Table 3.1), there is not a lot of room for improvement in static scenarios, and in 10 scenarios, only two of them showed that performance improved using Traffic Balancing; the rest performed similarly to DSR. In addition, Traffic Balancing needed more control packets to deliver the required information when all nodes were static. Thus, even though Traffic Balancing finds better paths in a static scenario, the explored resources are consumed by the extra control information and no improvement can be made to system performance.



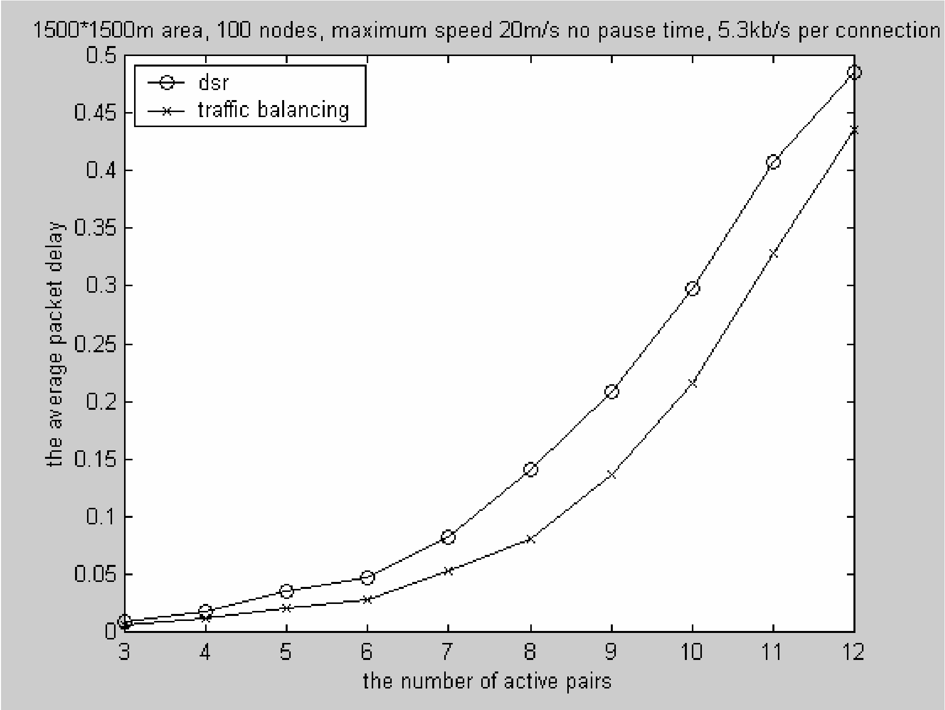
(a) 20 m/s



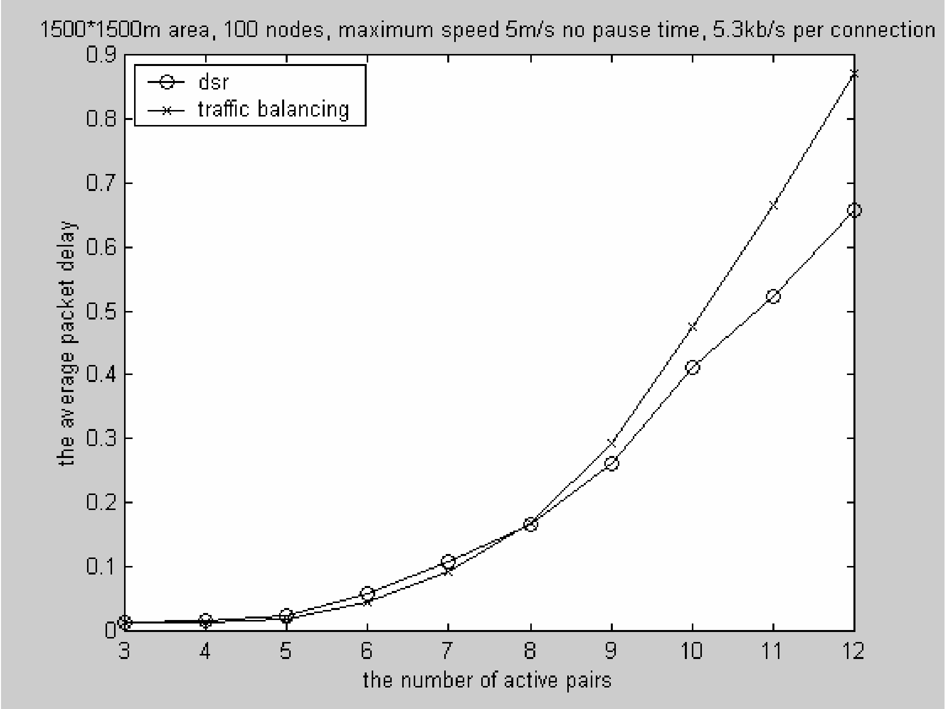
(b) 5 m/s

Figure 5.5 Solution 1: Packet Loss Rate vs. Number of Communication Pairs.

Figure 5.5 shows the packet loss rate for both DSR and Traffic Balancing. When the traffic is heavy (more users are active), traffic balancing can support more users for a given threshold (such as packet loss rate less than 20%). The average packet delay (in seconds) is shown in Figure 5.6. For a maximum speed of 20m/s, the overall delay with Traffic Balancing is lower, because there are fewer collisions in the congested area. Even though some connections require more hops to reach the destination, the time consumed by retransmission and the backoff period is longer. For a maximum speed of 5m/s, a heavy traffic load may cause longer packet delays, but the delay caused by additional hops is greater than the delay caused by congestion. One reason is that if there is no alternative path, DSR functions are implemented to find a shortest path, which adds an extra delay because of the second routing request. Overall, the largest improvement occurs when the DSR packet loss rate is from 5% to 10% (the improvement reaches close to 50%).



(a) 20 m/s



(b) 5 m/s

Figure 5.6 Solution 1: Average Packet Delay vs. Number of Communication Pairs.

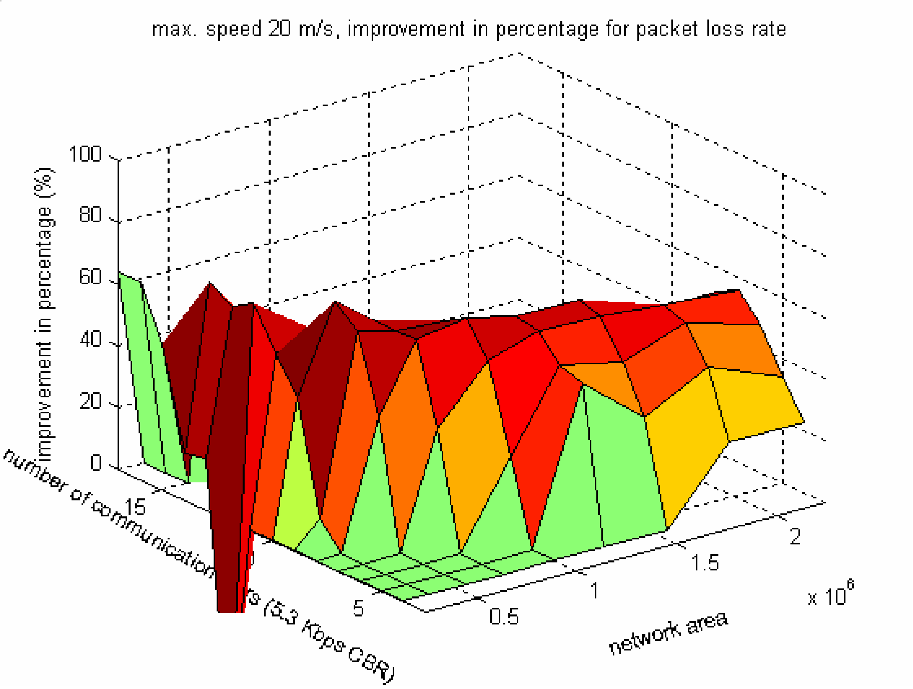
5.3.2 Solution 2 (Sender Decision)

Again, 100 nodes were randomly distributed in the network. The maximum speed of each node was set to 20m/s. Each result set was collected from an average over five runs (i.e. five scenarios). The number of communication pairs as well as the network size is considered when comparing Traffic Balancing with DSR. The network size was chosen from 1500x1500m^2 to 500x500 m^2.

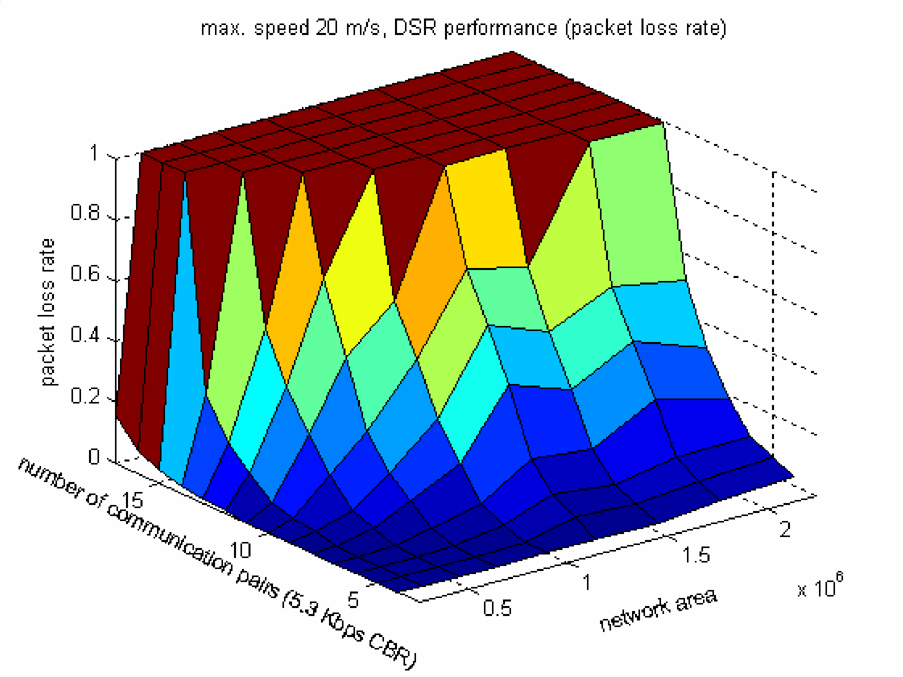
Figure 5.7 illustrates the performance of DSR with respect to packet loss rate and the improvement due to Traffic Balancing. The improvement is calculated by the following formula:

$$improvement = \frac{PLR_{DSR} - PLR_{TB}}{PLR_{DSR}} \qquad \textbf{5.2}$$

When the packet loss rate of DSR is about 20%–40%, Traffic Balancing gains around 50%, i.e. the packet loss rate of Traffic Balancing is around 10%–20%. A similar trend shows in the average delay (in Figure 5.8). The results show that Traffic Balancing moved the traffic load from the congested area to the idle area so that the utilization of all areas increased.

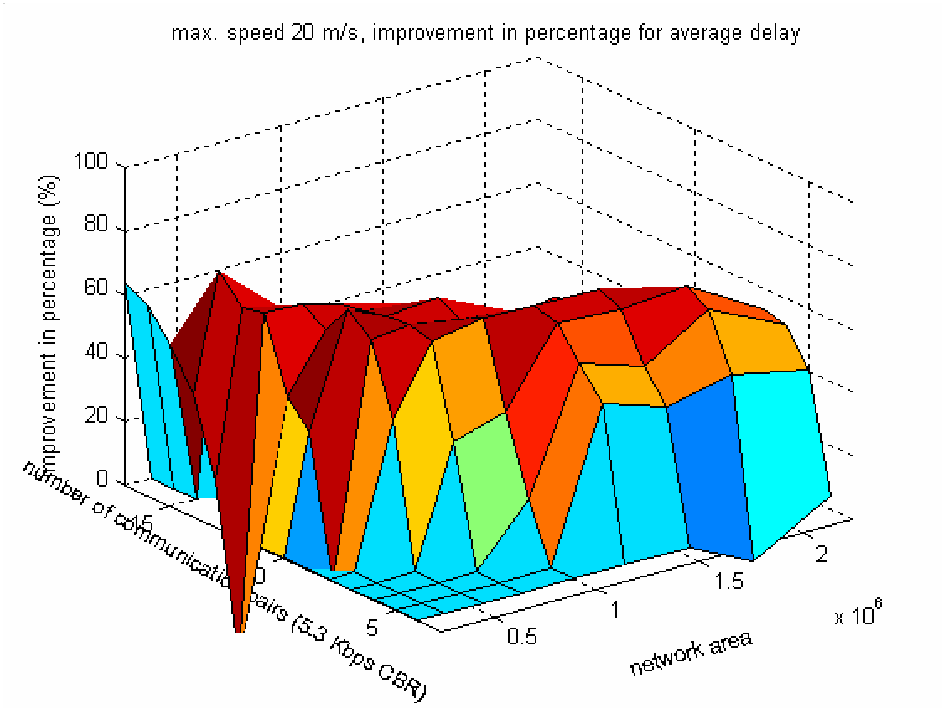


(a) Improvement of Traffic Balancing for Packet Loss Rate

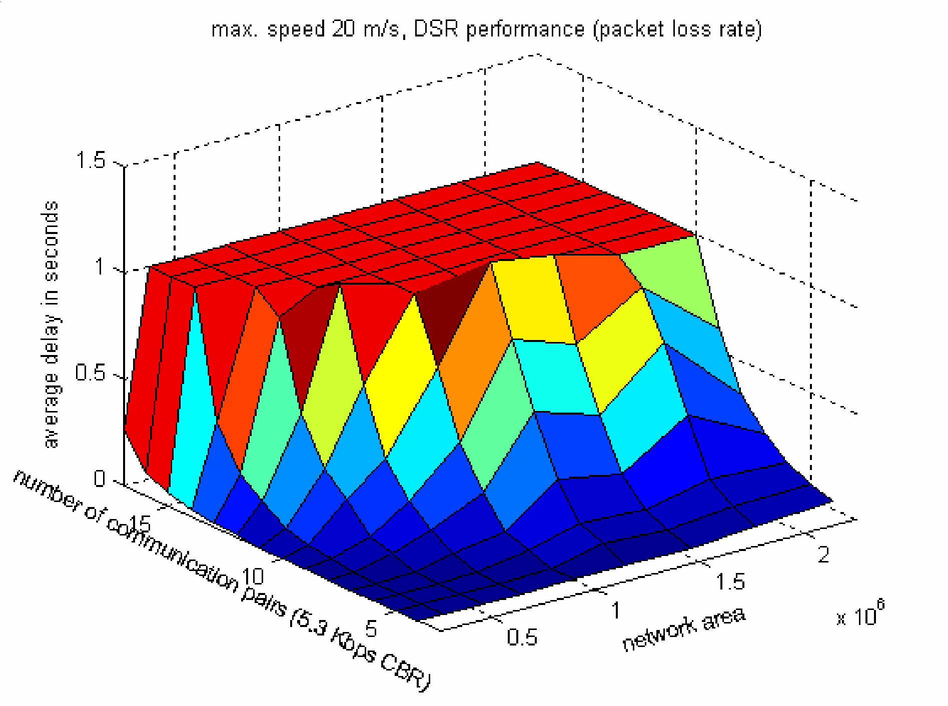


(b) DSR Performance

Figure 5.7 Solution 2: Packet Loss Rate vs. Number of Communication Pairs and Network Size.



(a) Improvement of Traffic Balancing for Average Delay

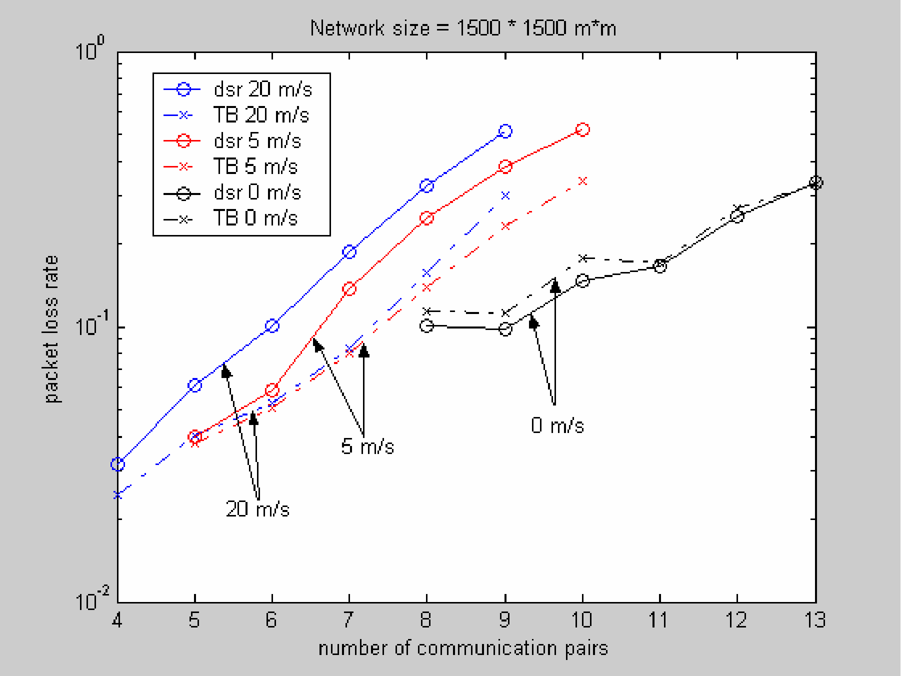


(b) DSR Performance

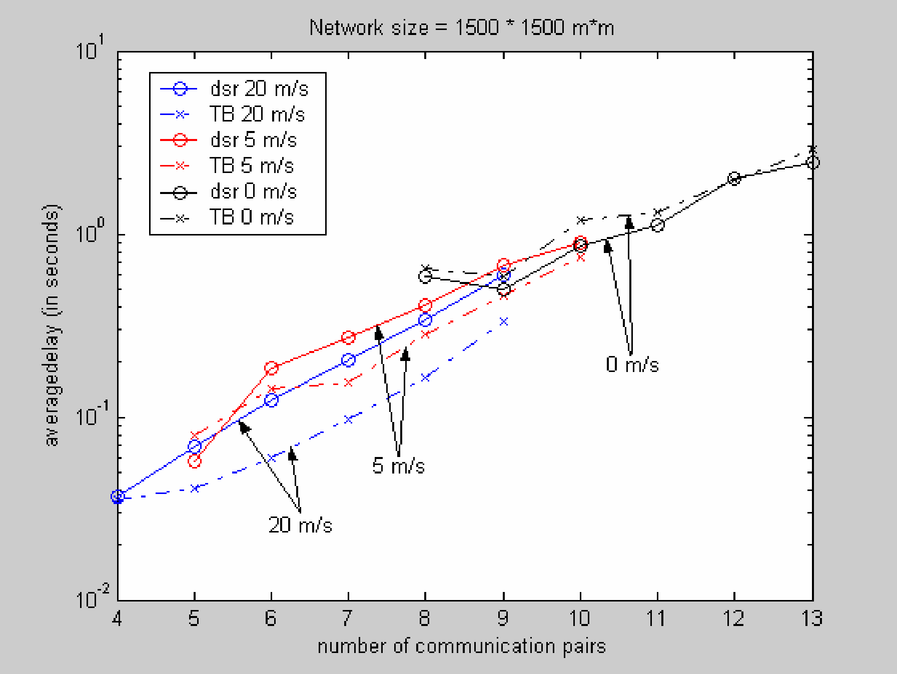
Figure 5.8 Solution 2: Average Delay vs. Number of Communication Pairs and Network Size

When the traffic load is light, Traffic Balancing can do little. When the traffic load is too high, Traffic Balancing cannot find any area that has a light load, and no traffic can be rerouted. So, as shown in Figures 5.7 and 5.8, in both the heavy load and light load networks, the performance of Traffic Balancing is only slightly better than DSR. When the traffic load is at an intermediate range, certain areas experience congestion, and Traffic Balancing has been shown to extend system utilization and improve performance.

In Figures 5.9, 5.10 and 5.11, the performances of Traffic Balancing (Solution 2) under different network sizes are given, compared to DSR. Figure 5.9 (a), 5.10 (a) and 5.11(a) plot the percentage of packets lost in transmission, and Figure 5.9 (b), 5.10 (b) and 5.11 (b) illustrate the average delay experienced by data packets. As the transmission range of each node is a circle with a radius of 250m, the average numbers of hops required by the connections is about 4.9, 3.3, and 1.9 for network sizes of 1500x1500, 1000x1000, and 500x500m^2 respectively, based on the results of the simulations. In a small network and with a small number of hops, the chance that an alternative path with fewer overloaded intermediate nodes exists is low. Thus, the improvement under this kind of scenario is lower (as shown in Figure 5.11), especially with respect to average delay. On the contrary, in a large network and with a large number of hops, the shortest path, through the middle of the network, is just one or two hops less than the other paths. With a large number of hops, the selected route by the shortest path algorithm can usually be any path from all possible paths. Relatively speaking, the traffic is evenly distributed into the network by the shortest path algorithm. In this way, the performance improvement is a bit lower than in a medium-sized network. From Figure 5.10, the improvement in both packet loss rate and average delay is over 50% when the nodes are moving.

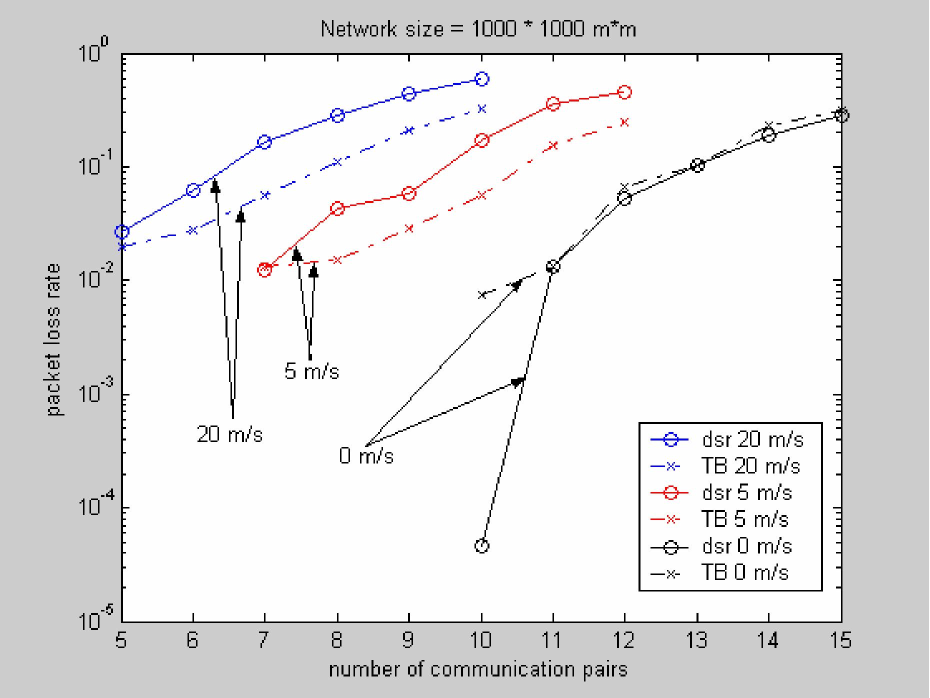


(a) Packet Loss Rate

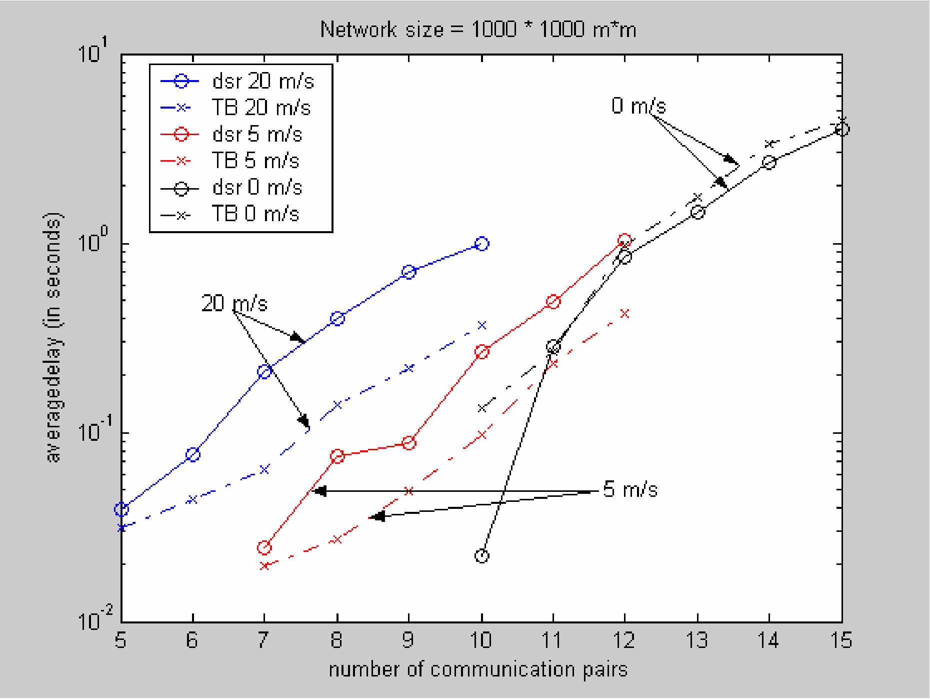


(b) Average Delay

Figure 5.9 Solution 2: System Performance with Network Size of 1500x1500m^2

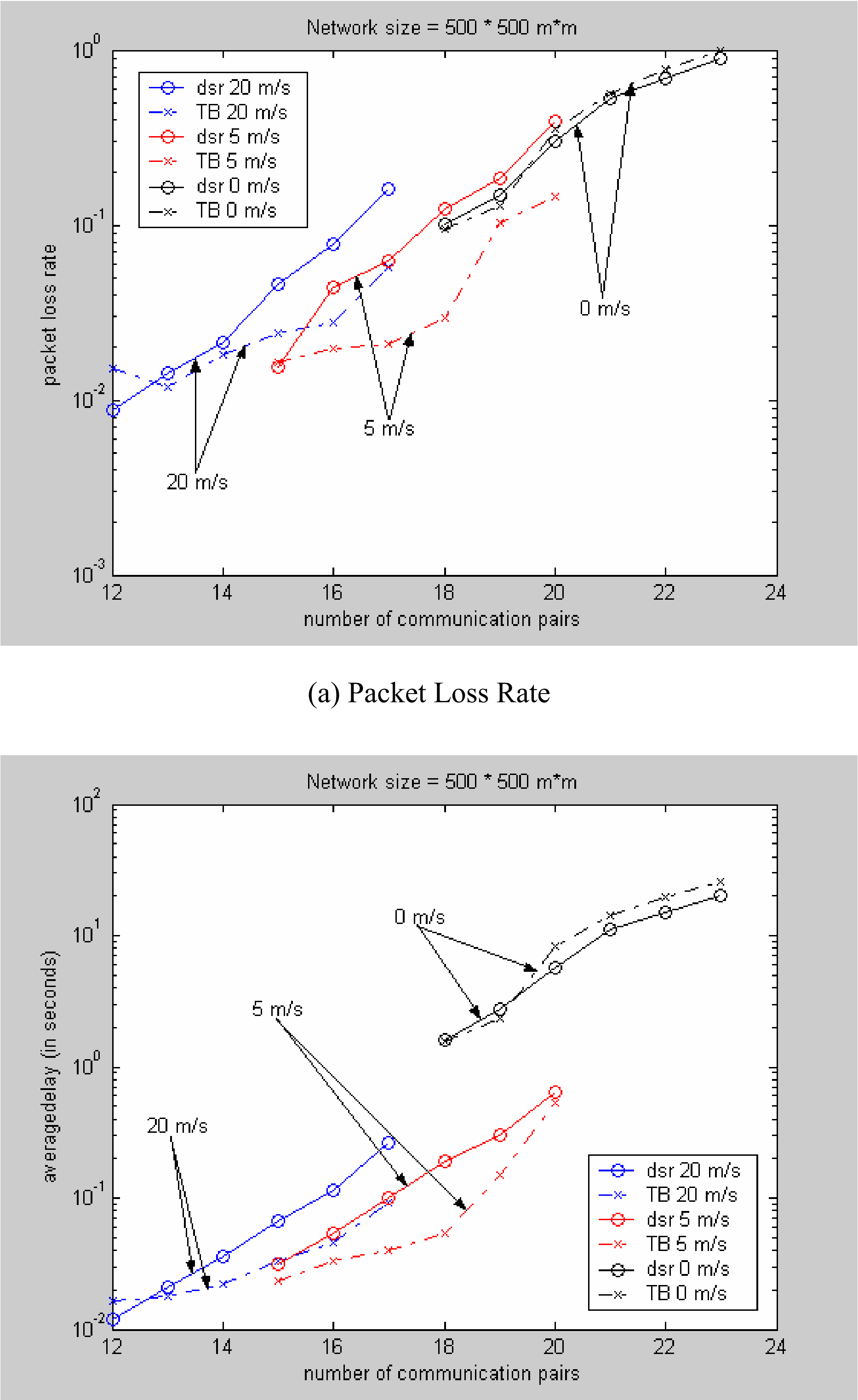


(a) Packet Loss Rate



(b) Average Delay

Figure 5.10 Solution 2: Performance with Network Size of 1000x1000m^2



(a) Packet Loss Rate

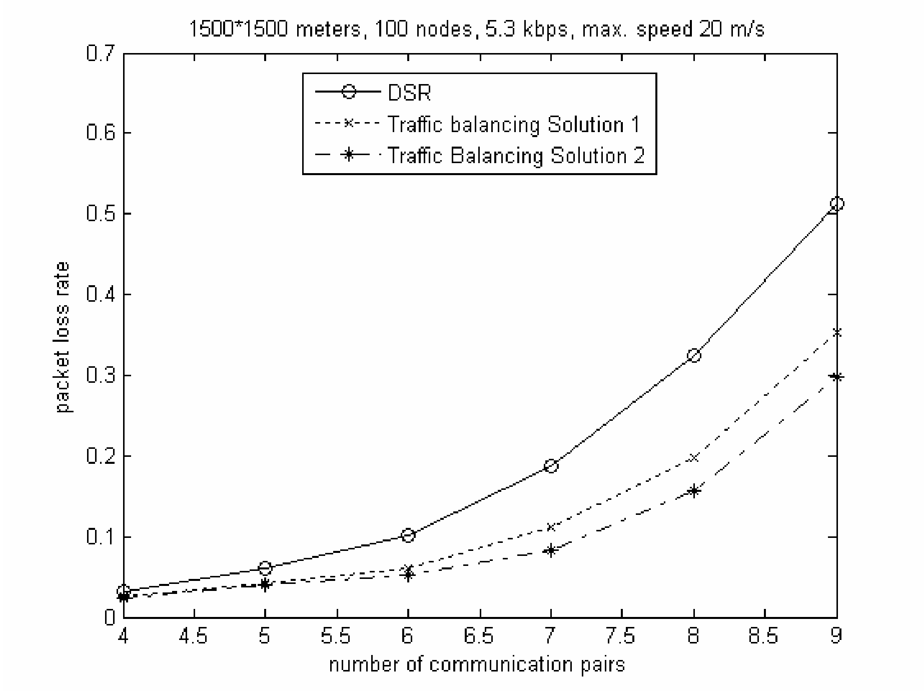
(b) Average Delay

Figure 5.11 Solution 2: Performance with Network Size of 500x500m^2

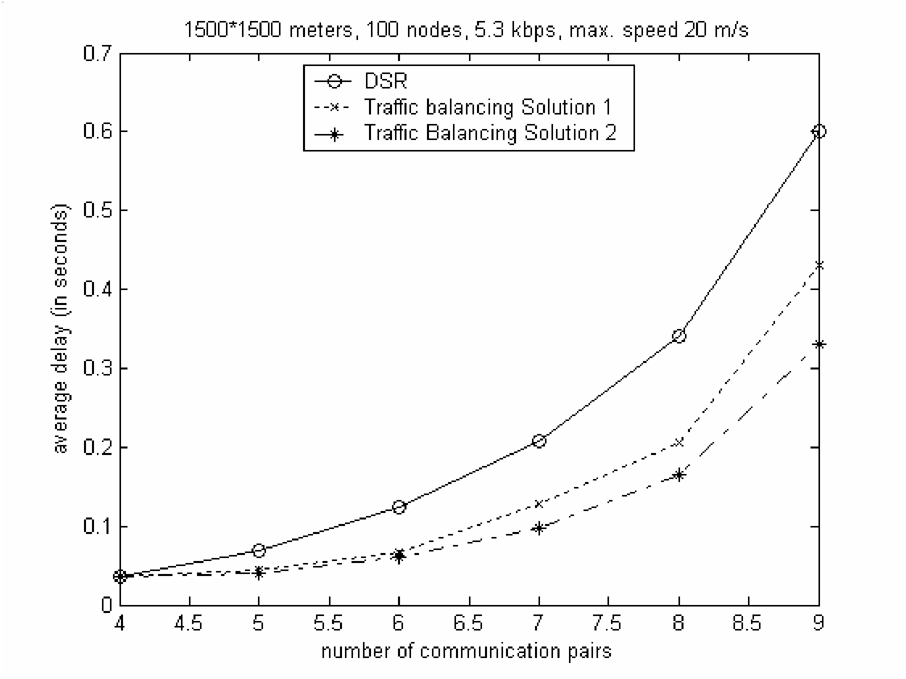
5.3.3 Solution 1 vs. Solution 2

It was expected that Solution 2 would perform better than Solution 1, because Solution 1 only works if there is a path with all intermediate nodes not overloaded. Indeed, the simulations showed that Solution 1 is slightly worse than Solution 2 with respect to packet loss rate (as shown in Figure 5.12). In Solution 1, a sender has to wait a long time to send a high priority request if there is no reply to the low priority request. The delay in Solution 1 was longer than in Solution 2. However, this conclusion is based on simulations in which communication was bi-directional and the transmission rate in each direction was the same. As shown in Figure 5.13, when the transmission was asymmetrical (unidirectional transmission), Solution 1 performed better than Solution 2 with a maximum node speed of 10m/s.

From Figure 5.13, it can be seen that traffic model and the mobility pattern have a big impact on the performance of the two solutions. Although both Figures 5.12 and 5.13 show that Solution 1 and 2 outperform each other in different scenarios, Solution 2 is shown to perform better in most cases, especially when communication is symmetrical. In addition, Solution 2 has a better chance of finding an alternative route under a heavy traffic load. Solution 2 outperforms Solution 1, especially when all nodes are static and traffic load is not light, because Solution 2 generates fewer control packets. As a result, the remainder of this thesis will focus on Solution 2.

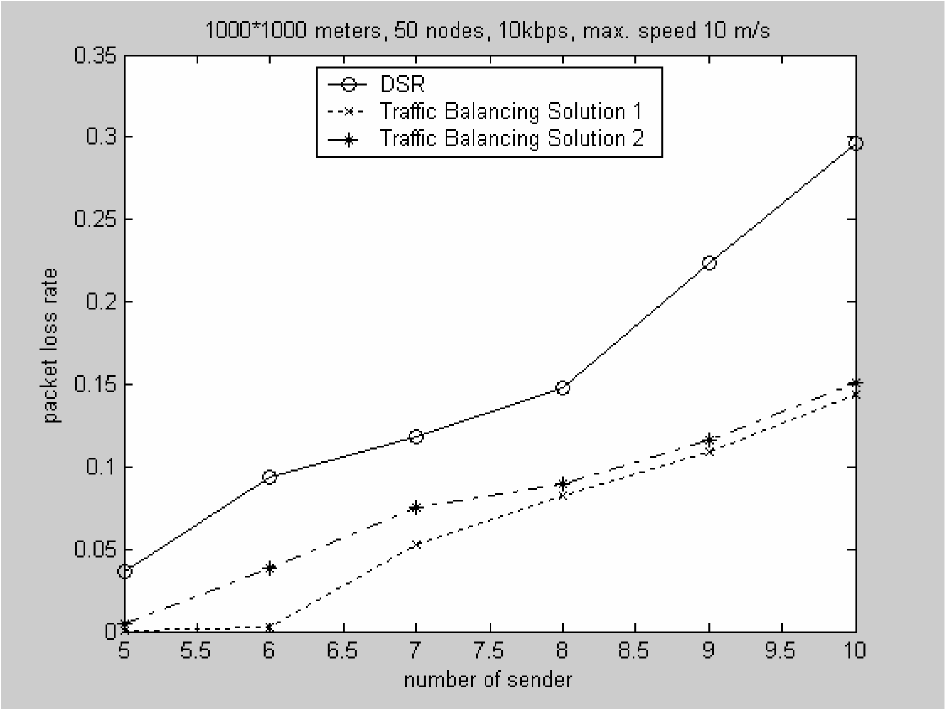


(a) Packet Loss Rate

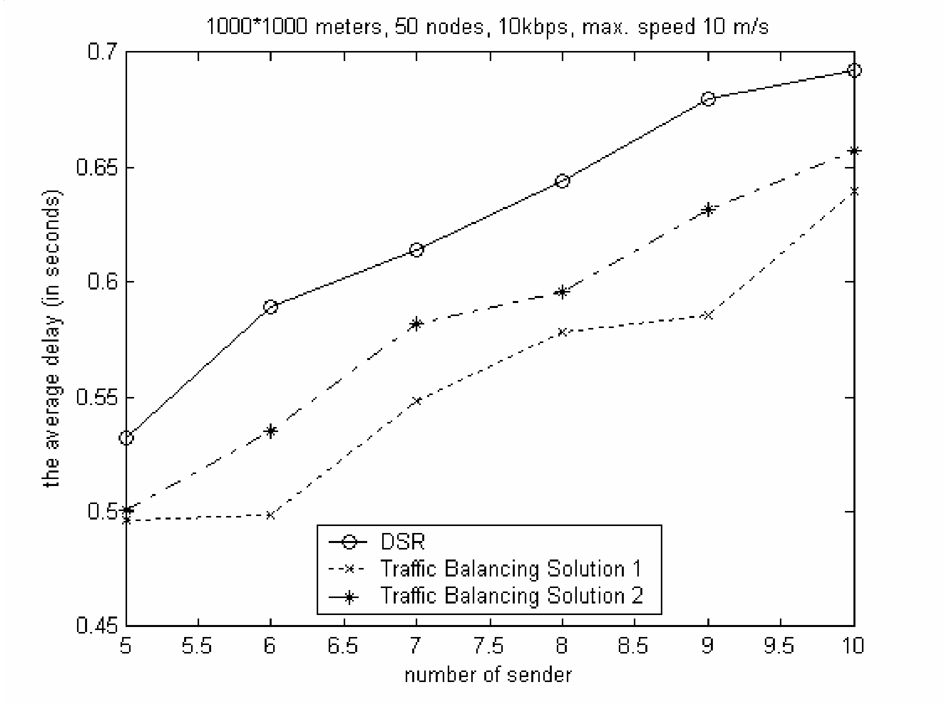


(b) Average Delay

Figure 5.12 Comparison Between Traffic Balancing Solution 1 and Solution 2 (Case 1).



(a) Packet Loss Rate



(b) Average Delay

Figure 5.13 Comparison Between Traffic Balancing Solution 1 and Solution 2 (Case 2).

5.3.4 Threshold of Medium State

The threshold of the medium state is used to decide at what value the medium is thought to be busy. Its value depends on the MAC layer protocol and the mobility of the nodes. In the previous simulations, the threshold was set to 0.7 (70% of the time during the last interval the medium was occupied for transmission). If not stated otherwise, the default value of the threshold is 0.7 in this research. This value was chosen by checking the measured medium usage and the packet collision rate for a certain period. (When the measured medium usage is above 0.7, the observed collision rate increases dramatically. Here the maximum speed of the node is 10m/s and the traffic model is CBR at a rate of 5.3 Kbps).

In fact, because mobility causes collision and transmission loss, the measured medium usage under a busy state may vary. For example, in a static scenario, after all transmissions have run for a while, few collisions will take place because the nodes know about their neighbors and their transmissions. If a node constantly sends packets with a packet size of 64 bytes, the measured medium usage is (IEEE 802.11, the MAC layer protocol):

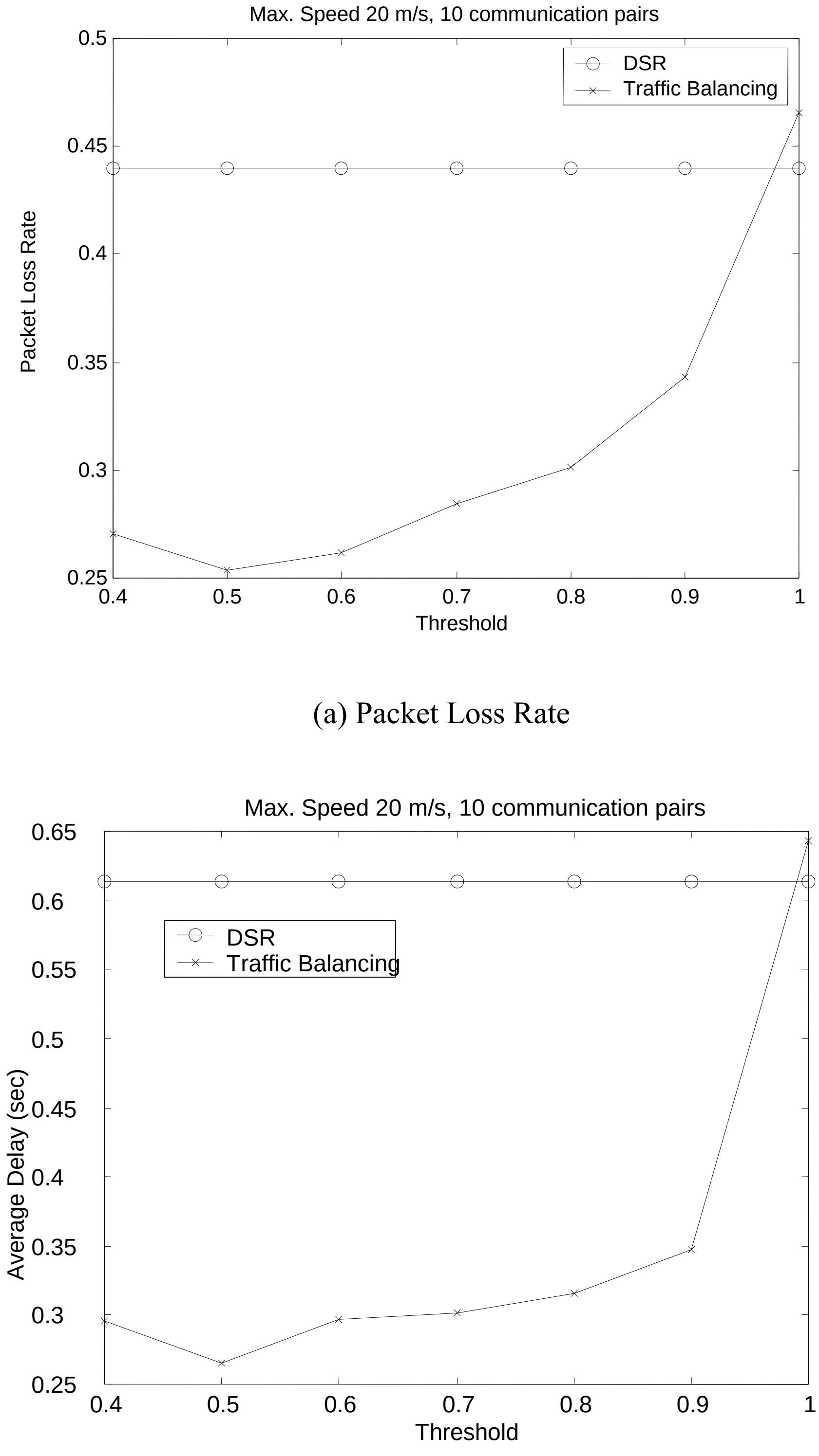
$$
\begin{aligned}
&medium\_usage \\
&= \frac{RTS + CTS + Data + ACK}{RTS + SIFS + CTS + SIFS + Data + SIFS + ACK + SIFS} \\
&= (0.06 + 0.06 + 0.32 + 0.06)/(0.0\ 6 + 0.01 + 0.06 + 0.01 + 0.32 + 0.01 + 0.06 + 0.01) \\
&= 92.59\%
\end{aligned}
$$

where the transmission rate is 2Mbps and it takes around 0.06ms to transmit a RTS or CTS or ACK. The data packet needs 0.32ms to be sent, and the short interframe period is 0.01ms. On the other hand, if mobility is very high, the collision rate and transmission loss rate go up. When the maximum moving speed is 5m/s and the measured collision rate reaches 10% of transmitted packets (including RTS, CTS and ACK), the medium becomes very busy. In this case, the measured medium usage is around

$$
\begin{aligned}
&medium\_usage \\
&= \frac{RTS + CTS + Data + ACK}{RTS + SIFS + CTS + SIFS + Data + SIFS + ACK + SIFS + 4 \times backOff \times P_{collision}} \\
&= \frac{0.06 + 0.06 + 0.32 + 0.06}{0.06 + 0.01 + 0.06 + 0.01 + 0.32 + 0.01 + 0.06 + 0.01 + 4 \times 0.02 \times 15 \times 0.1} \\
&= 76.22\%
\end{aligned}
$$

where the backoff time is estimated at 0.02x15 (CW_{MIN}=31 and the slot time is 0.02ms). If collisions occur continuously, the backoff time is longer and the number of the backoff is higher. Thus, the measured medium usage could be even lower and using a fixed threshold value may therefore not be adequate for different scenarios.

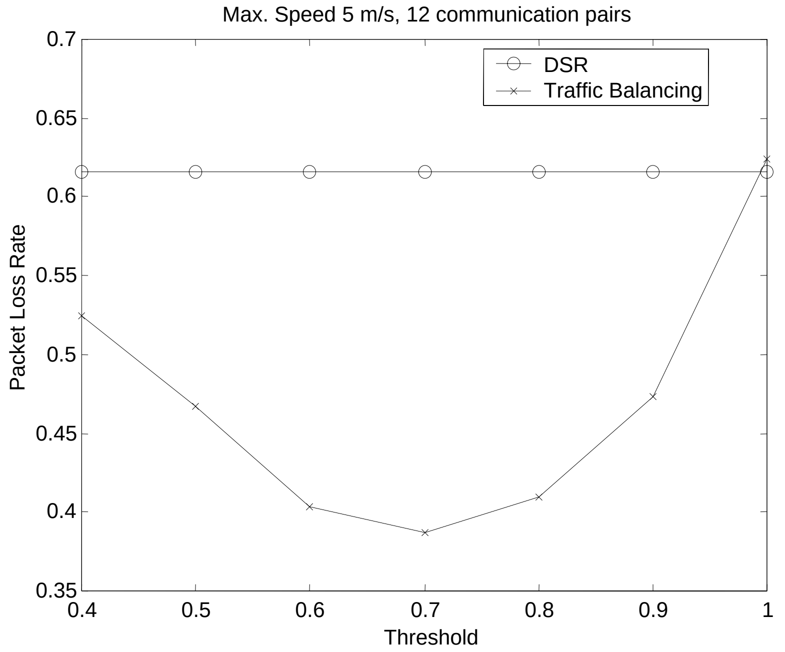
In Figures 5.14, 5.15 and 5.16, the results show that at different node mobility rates, the best performance is achieved for different threshold values. When the mobility is very high (20m/s), the chance of collision and transmission loss occurring is also very high. That causes longer backoff time and more backoffs (in 50% of the time, nodes sense an idle medium). Thus, the performance reaches its maximum when the threshold is set to around 0.5. Meanwhile, in a static scenario, the chance of collision and transmission loss is very low, and nodes continue transmitting the packets. The best performance is obtained when the threshold is set around 0.9.



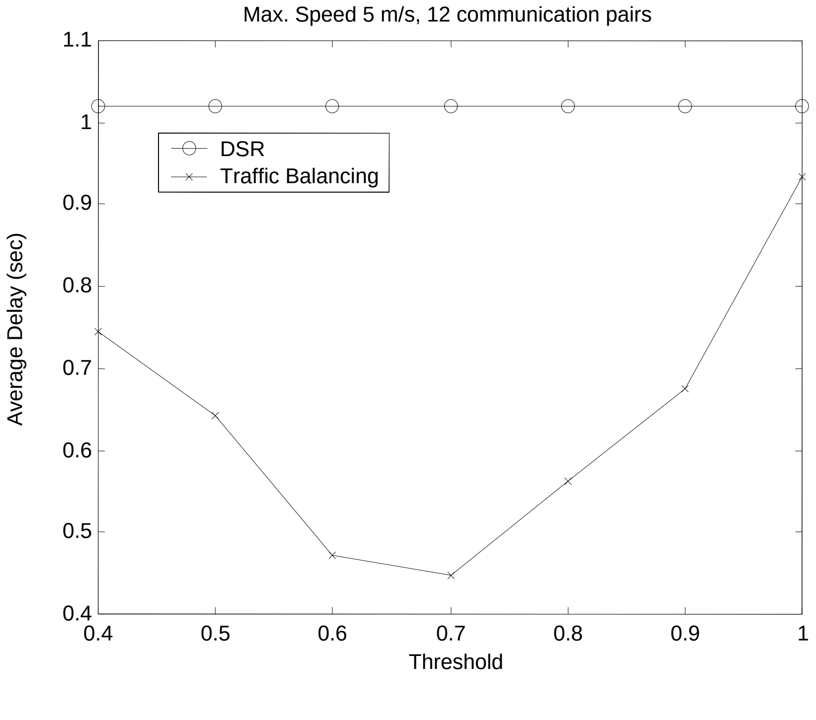
(a) Packet Loss Rate

(b) Average Delay

Figure 5.14 Threshold vs. System Performance at a Max. Moving Speed of 20m/s

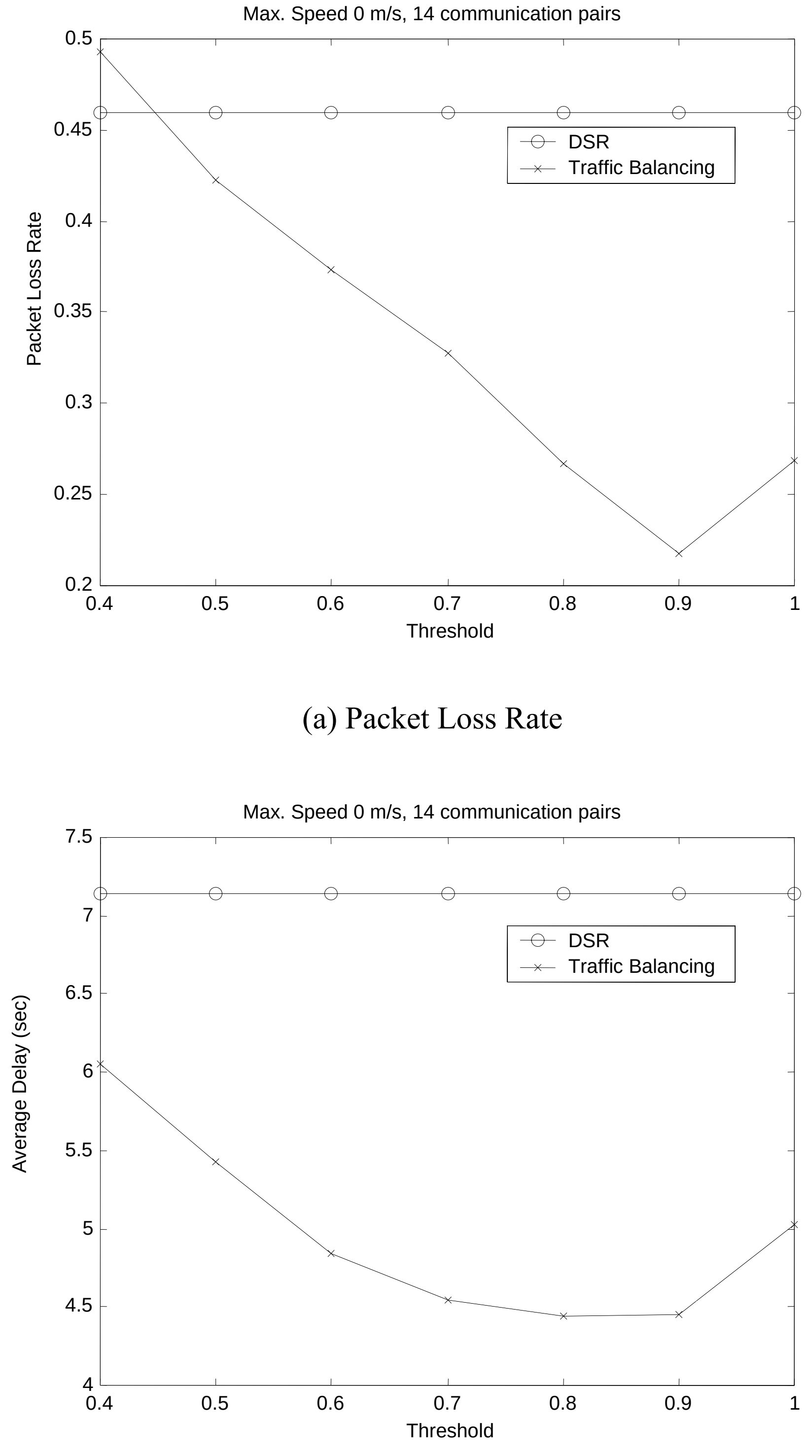


(a) Packet Loss Rate



(b) Average Delay

Figure 5.15 Threshold vs. System Performance at a Max. Moving Speed of 5m/s



(a) Packet Loss Rate

(b) Average Delay

Figure 5.16 Threshold vs. System Performance at a Max. Moving Speed of 0m/s

It is obvious that for higher mobility, a lower threshold value needs to be set to achieve better performance. However, it is not easy for nodes to decide the mobility rate of the network. A node may check its queue length or the collision rate in its area to get the estimation of its mobility. From Figures 5.14, 5.15 and 5.16, a threshold of 0.7 gives moderate network performance under different mobility rates. Section 5.4 discusses and implements a simple solution to change the medium usage threshold dynamically in Traffic Balancing.

5.3.5 Information Accuracy

Traffic Balancing Solution 2 uses one byte in the routing request to record the number of overloaded intermediate nodes. However, this byte cannot indicate how heavily these intermediate nodes are loaded. Out of all the routes, the one with only one overloaded but very heavily loaded intermediate node may not be better than a route with more than two overloaded but not very heavily loaded intermediate nodes. Including more detailed information with the route requests can allow the sender to choose a better path, with more bandwidth available. However, when doing this, several issues need to be resolved:

- As mentioned in the previous chapter, transmission over the last link has an impact on the transmission over the current link, and even the next link. How much bandwidth is available at each intermediate node cannot be calculated in any simple way. It is related to the number of links that are affected by the current link.

- Recording the available bandwidth in each intermediate node requires more bytes to be sent in the route request. Simply accumulating the available bandwidth in each intermediate node may not make any difference for Traffic Balancing. As a routing request is flooded to the network, the overhead caused by adding more bytes may introduce a high cost to Traffic Balancing, especially when the average number of hops is large.

- Evaluating the paths is a challenge. For example, if one path has one intermediate node with 20% of available bandwidth, and the other one has two intermediate nodes with 30% of available bandwidth, it is hard to decide which one is better, unless the new connection needs more than 20% of bandwidth.
- Other ways of calculating the weight of each direct radio link face problems similar to those listed above. If the weight is related only to the available bandwidth (i.e., can be deducted by the available bandwidth), how to record the weight of each hop can still be a problem, as mentioned previously.

In conclusion, the complexity and cost of including more detailed information may not be worth the effort.

5.4 Adaptive Traffic Balancing

The aforementioned Traffic Balancing determines the medium state only by the percentage of time period that the measured signal strength exceeds the threshold. Some important issues, such as node mobility and congestion state, are not considered although they also influence the state of the medium. Section 5.3.4 shows that medium usage is varied when the mobility of the nodes changes. Moreover, if Traffic Balancing is able to adjust the medium usage threshold according to node mobility or congestion state, the performance can be largely improved. This section proposes Adaptive Traffic Balancing to adjust the medium usage threshold dynamically, according to the number of collisions seen by the node. As a result, traffic load can be distributed more evenly and system resources can be utilized more efficiently.

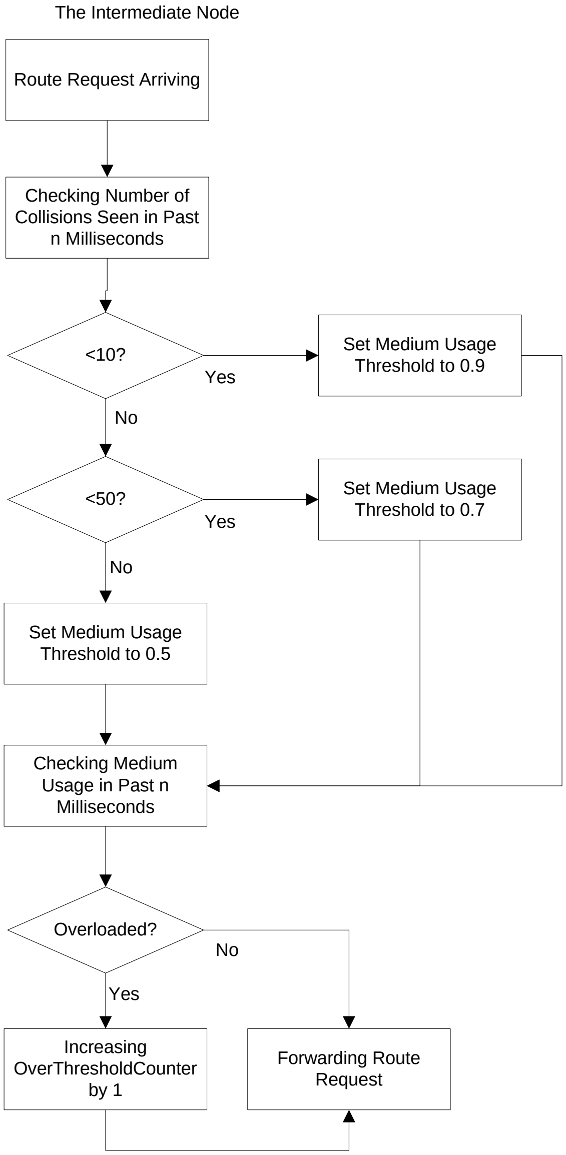

The Intermediate Node
Route Request Arriving
Checking Number of Collisions Seen in Past n Milliseconds
<10?
Yes
Set Medium Usage Threshold to 0.9
No
<50?
Yes
Set Medium Usage Threshold to 0.7
No
Set Medium Usage Threshold to 0.5
Checking Medium Usage in Past n Milliseconds
Overloaded?
No
Yes
Increasing OverThresholdCounter by 1
Forwarding Route Request

Figure 5.17 Flow Chart of Adaptive Traffic Balancing

For the purpose of changing the medium usage threshold dynamically, according to node mobility, certain parameters are required to designate node mobility or usage condition of the medium. During the simulations, it was noticed that the number of collisions that can be seen by a given node over a certain period can be applied to specify the node mobility or medium usage conditions more precisely than other parameters. Although a node cannot perceive all collisions that happen in its interference range, the frequency of observed collisions is still a good parameter for indicating the mobility around it.

Therefore, Adaptive Traffic Balancing measures the medium usage and number of collisions over a certain period. Due to a variety of traffic models, the reality is that the measurement period for the medium usage and the measurement period for the number of detected collisions may not be the same. Further research needs to be done to determine the compromise measurement periods that are suitable for both. Based on the number of collisions, a suitable threshold is selected and compared with the measured medium usage. If the medium usage exceeds the threshold, the overThresholdCounter is increased. Otherwise, the route request is simply forwarded to the next node.

Adaptive Traffic Balancing records the number of collisions in the past *n* milliseconds in a similar way that the medium usage at the MAC layer is recorded. In this case, a linked list is required. When a collision or a capture is detected, an element is added to the linked list to note the time (A capture happens when over two transmissions starts simultaneously and one transmission can be received correctly while others are not). Meanwhile, the linked list is updated to remove those elements that satisfy the following condition:

$T_{current\_time} - T_{start\_time} > n,$

It also means the elements, i.e. the collisions or captures, which is over *n* milliseconds old, are removed from the linked list because they are not counted for deciding the medium usage threshold. When a route request needs to be forwarded, the linked list is checked again, and the number of elements that obey the previous condition is calculated as well. If the number of collisions in the past *n* milliseconds is known, a suitable medium usage threshold is then selected.

Based on the simulation results in Section 5.3.4, Adaptive Traffic Balancing uses a simple method to adjust the medium threshold, as described below (shown in Figure 5.17). **For intermediate nodes**: Each time a route request is received by a relay node, the node checks the number of collisions in the past *n* milliseconds. If the number is less than 10, the medium usage threshold is then set to 0.9, assuming that the node mobility in the network is low (static). If this number is between 10 and 50, the medium usage threshold is set to 0.7. That assumes that the node mobility is moderate (node mobility is less than 5m/s). Otherwise, the medium usage threshold is set to 0.5 (assuming node mobility is high, with a maximum moving speed of 20m/s). After the medium usage threshold has been set, the node compares the measured medium usage with the threshold and decides whether or not to increase the *overThresholdCounter*.

In Figures 5.18, 5.19 and 5.20, the results show that Adaptive Traffic Balancing improves system performance further in forms of both packet loss rate and average delay. The improvement with a maximum moving speed of 20m/s is up to 70%–80% for Adaptive Traffic Balancing in the packet loss rate, compared to 50% for Traffic Balancing. Due to the number of collisions seen by the nodes, the medium usage thresholds vary at different nodes. With this additional information at each node, Adaptive Traffic Balancing is able to obtain more information during route discovery and recovery. The chosen path is then more feasible in terms of bandwidth availability. When the maximum moving speed of the nodes is 5m/s, 0.7 is almost the best value for the medium usage threshold, on average; thus the improvement is lower compared to the case when the maximum moving speed is 20m/s. When the node mobility decreases, the improvement due to Adaptive Traffic Balancing is also decreased, because Adaptive Traffic Balancing and Traffic Balancing consume more bandwidth for control information to ascertain a better path. In addition, in the 0m/s scenario, there is almost no collision at all. Adaptive Traffic Balancing stays at one level (here it is 0.9). The improvement in the 0m/s scenario is caused by Traffic Balancing using 0.7 as the medium usage threshold.

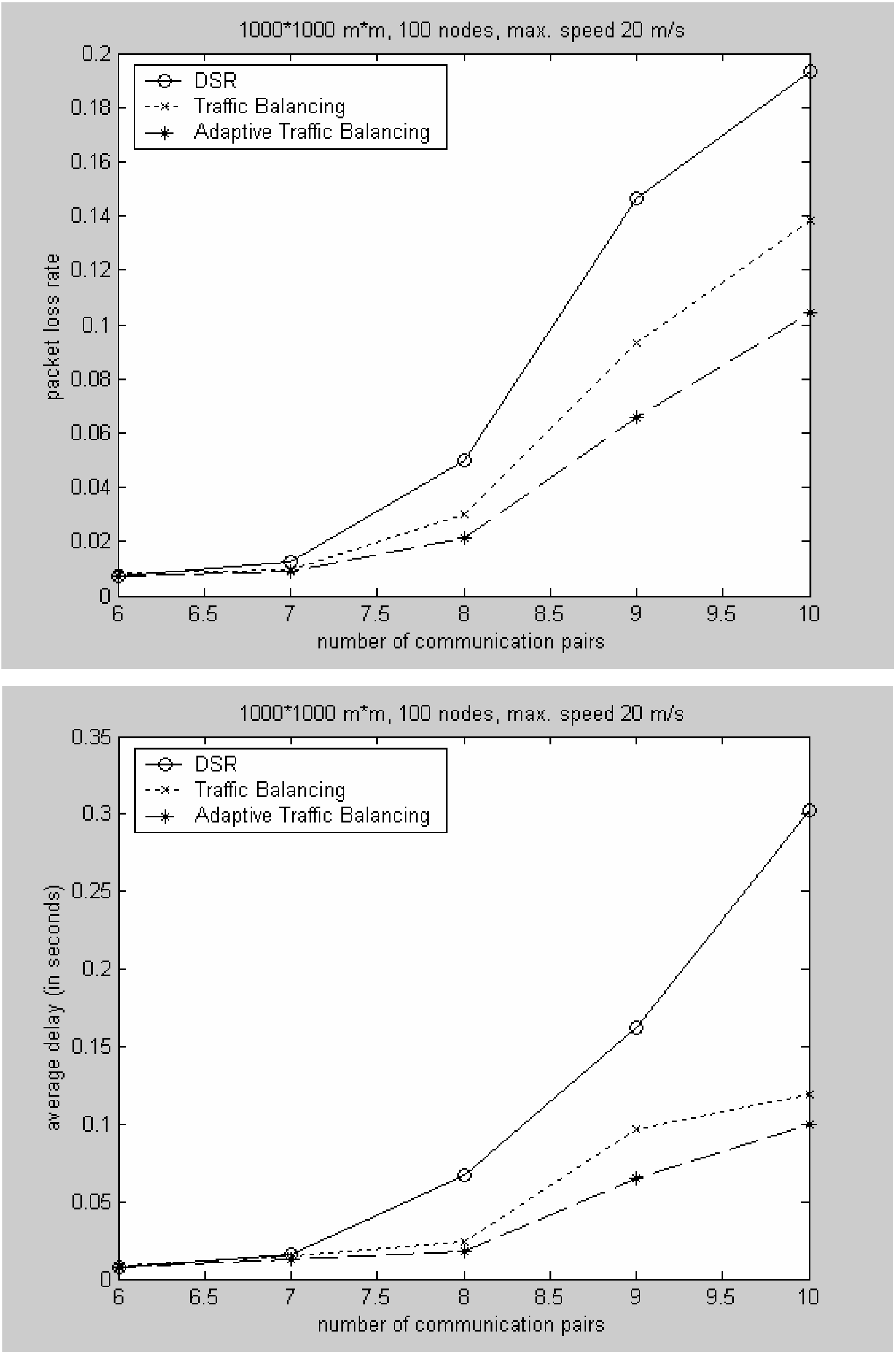


Figure 5.18 System Performance of Adaptive Traffic Balancing at a Max. Moving Speed of 20m/s. (Traffic Balancing uses a fixed threshold of 0.7 for all cases.)

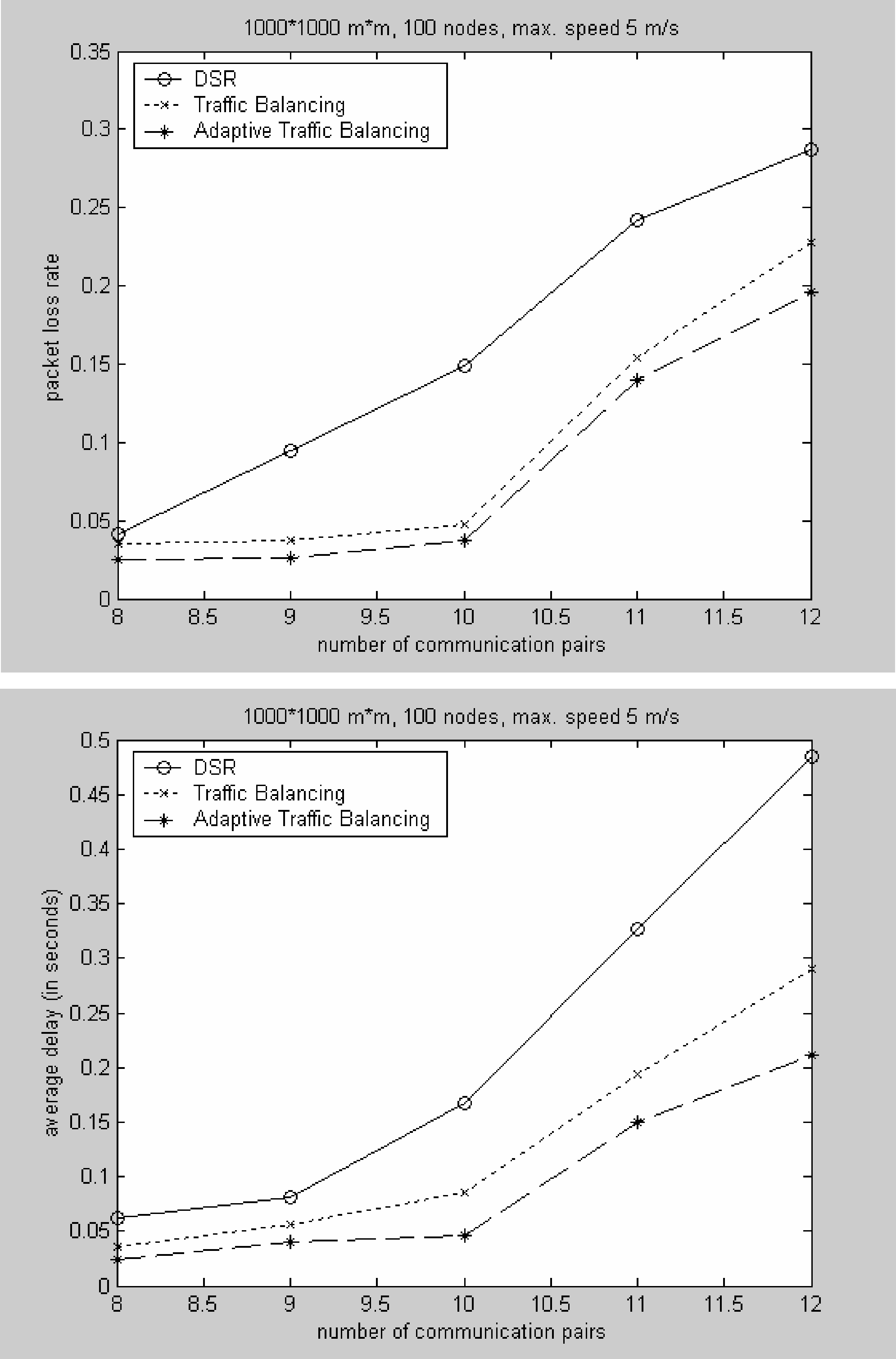
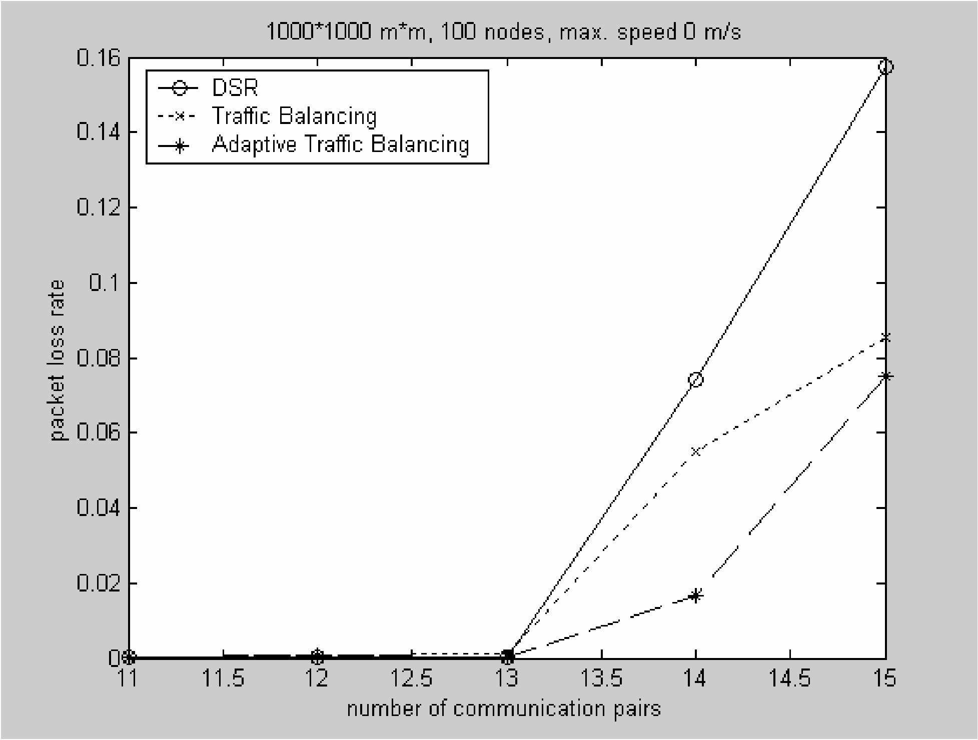


Figure 5.19 System Performance of Adaptive Traffic Balancing at a Max. Moving Speed of 5m/s. (Traffic Balancing uses a fixed threshold of 0.7 for all cases.)



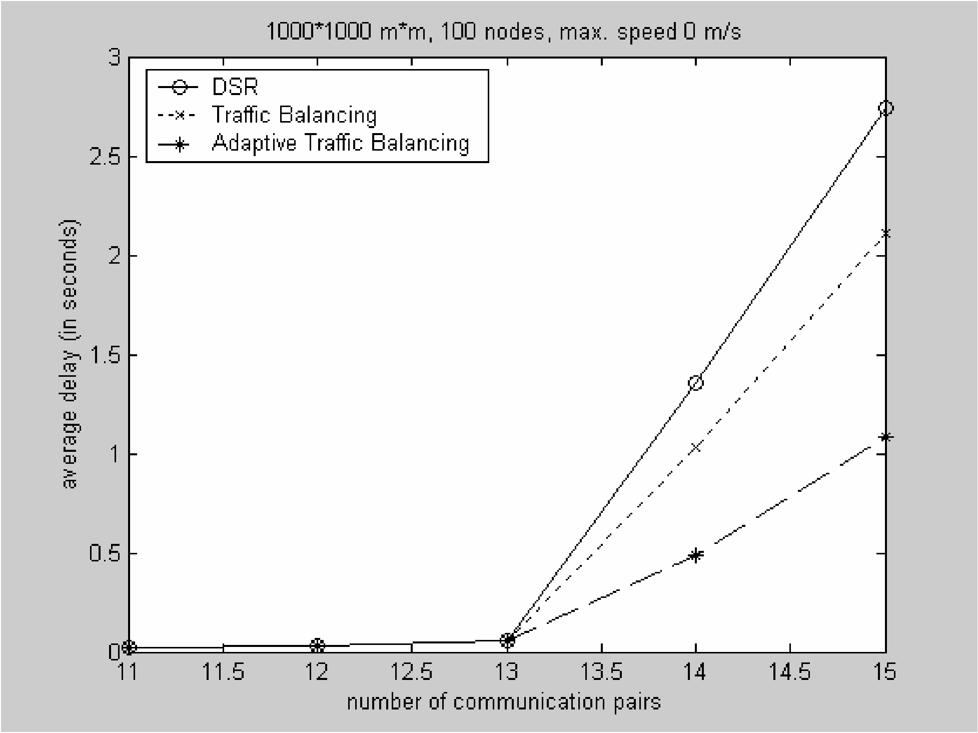
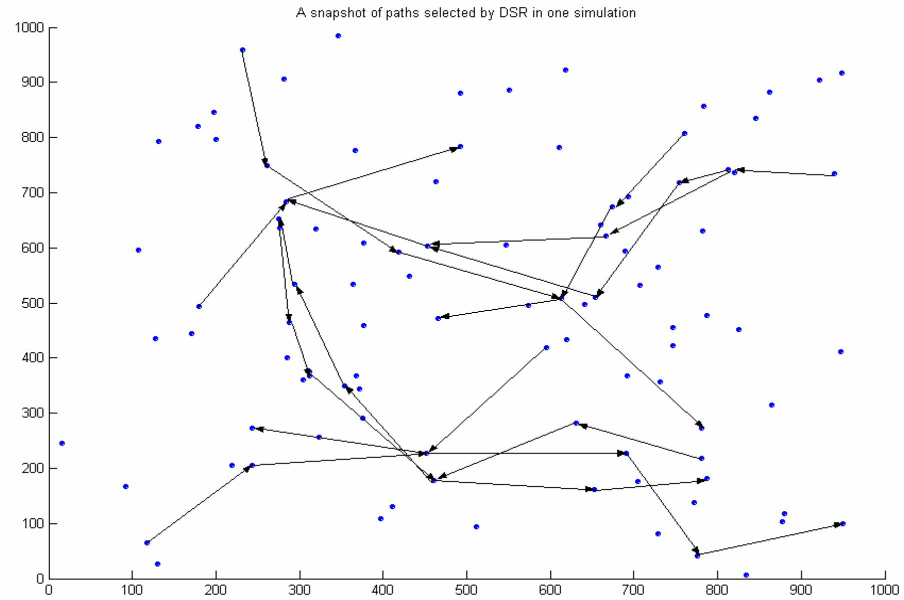


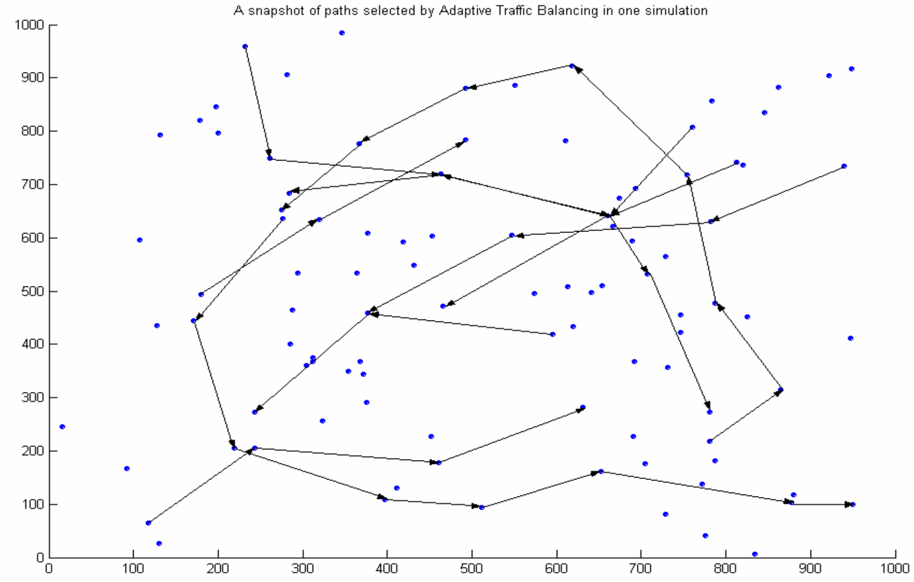
Figure 5.20 System Performance of Adaptive Traffic Balancing at a Max. Moving Speed of 0m/s. (Traffic Balancing uses a fixed threshold of 0.7 for all cases.)

As only three levels of medium threshold are used in Adaptive Traffic Balancing, the improvement is still limited. A more accurate relationship between the mobility and collision rate may be realized by carrying out more simulations and observations. Therefore, Adaptive Traffic Balancing may be implemented in a more elaborate manner, with an increased number of levels.

Figure 5.21 shows a simulated example of the route selection by DSR and Adaptive Traffic Balancing. The paths selected by DSR have a tendency to go through the center of the network so as to have fewer hops. Thus, the traffic load is concentrated in the center and the number of collisions is high. In contrast, Traffic Balancing aims to distribute the traffic load more evenly into the network, so some connections are routed towards the edges of the network. In this way, the possibility of congestion is smaller, and the number of collisions is smaller as well.



(a) Routes Selected by DSR



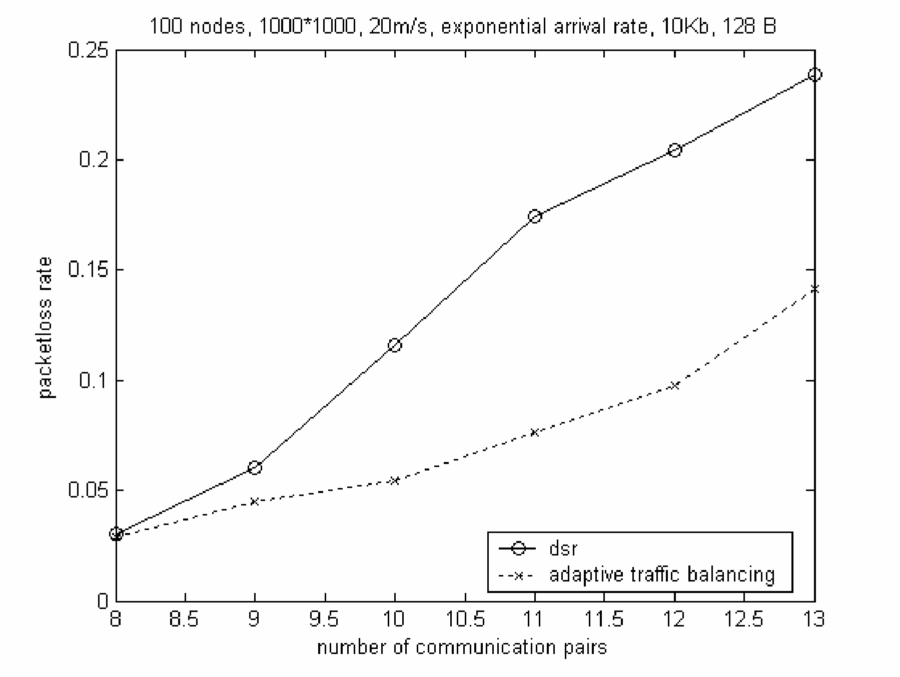
(b) Routes Selected by Adaptive Traffic Balancing

Figure 5.21 Snapshot of Paths Selected by DSR and Adaptive Traffic Balancing at the Same Time in the Simulation

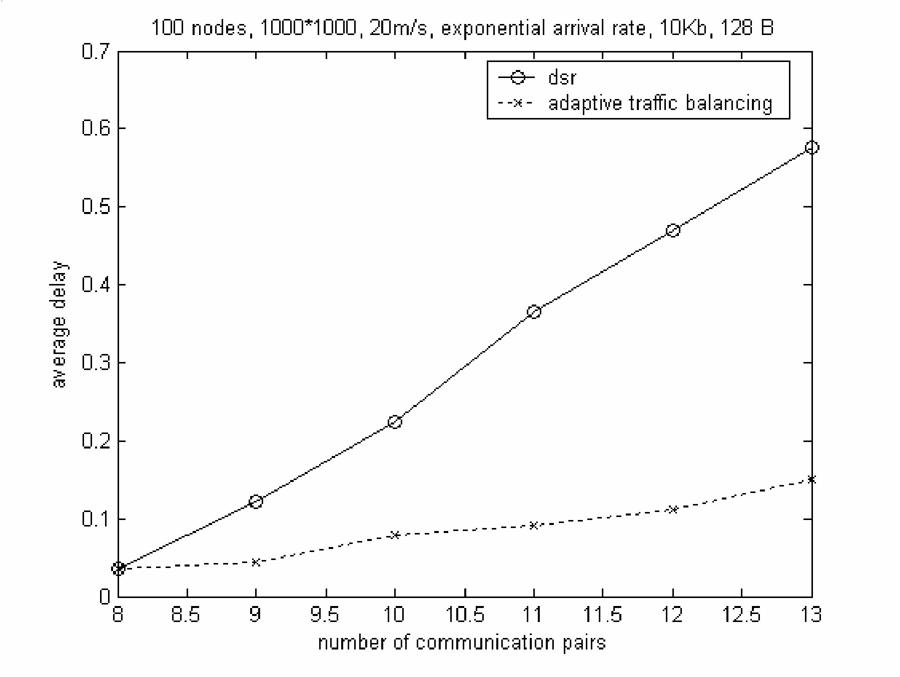
5.5 Simulation Results for other Traffic Models

This section presents the simulation results for different traffic types. The Poisson distribution is popular for describing network traffic. Under this assumption, the interval between the arrival times of consequent packets follows an exponential distribution. Figures 5.22, 5.23 and 5.24 show the system performance in a wireless Ad Hoc network with a size of 1000x1000m^2. The packet size was 128 bytes, the measurement period was fixed at two seconds and each communication pair generated a traffic load of 2x10Kbps. Adaptive Traffic Balancing used three levels (0.5, 0.7 and 0.9) to indicate the medium state, based on the collision rate. The results were averaged over five runs (five scenarios).

The results show that the trend is similar to the CBR results. Improvement increases with the mobility of the nodes. When the packet loss rate was around 10% for DSR, Adaptive Traffic Balancing decreased the packet loss rate by more than 50%. Meanwhile, in both the 20m/s and 5m/s cases, the average delay to transmit a packet went up slower using Adaptive Traffic Balancing when the traffic load increased. In static scenarios, when the traffic load went up extremely high (the number of connection increases), DSR started outperforming Traffic Balancing. First, extreme congestion caused the routing protocol to generate more routing traffic, and Traffic Balancing generated more routing traffic than DSR. Second, with an extremely high traffic load, the network had already reached its capacity and there was no room for Traffic Balancing to find alternative paths to increase performance. These two reasons together explain why Traffic Balancing performed poorly at an extremely high traffic load.

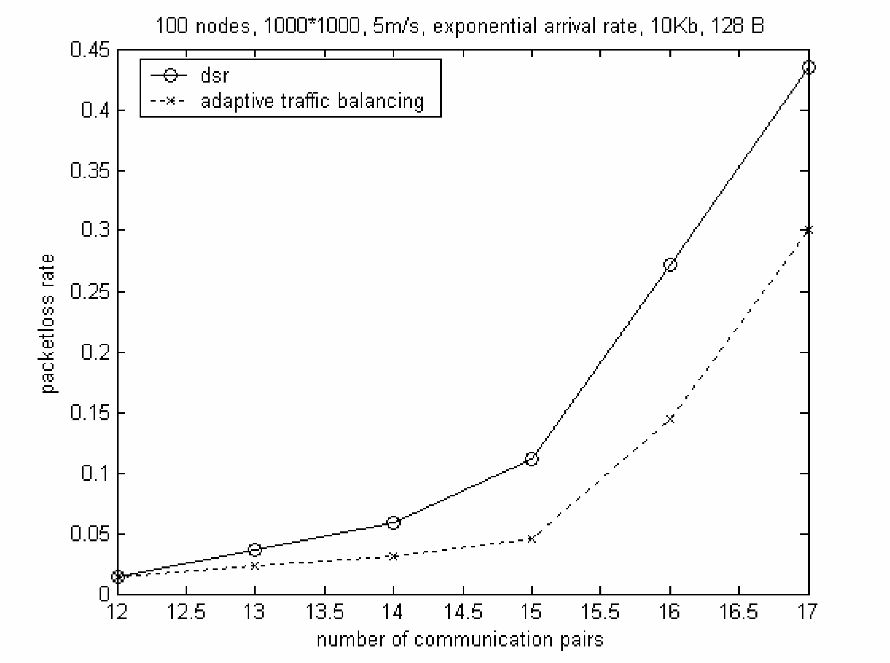


(a) Packet Loss Rate

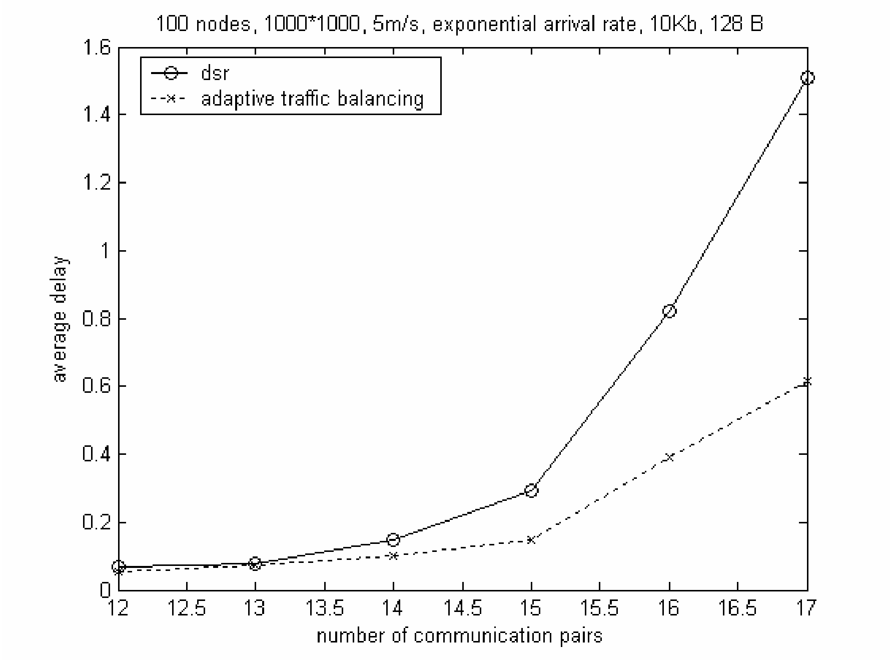


(b) Average Delay

Figure 5.22 System Performance with Poisson Traffic Model at a Max. Moving Speed of 20m/s

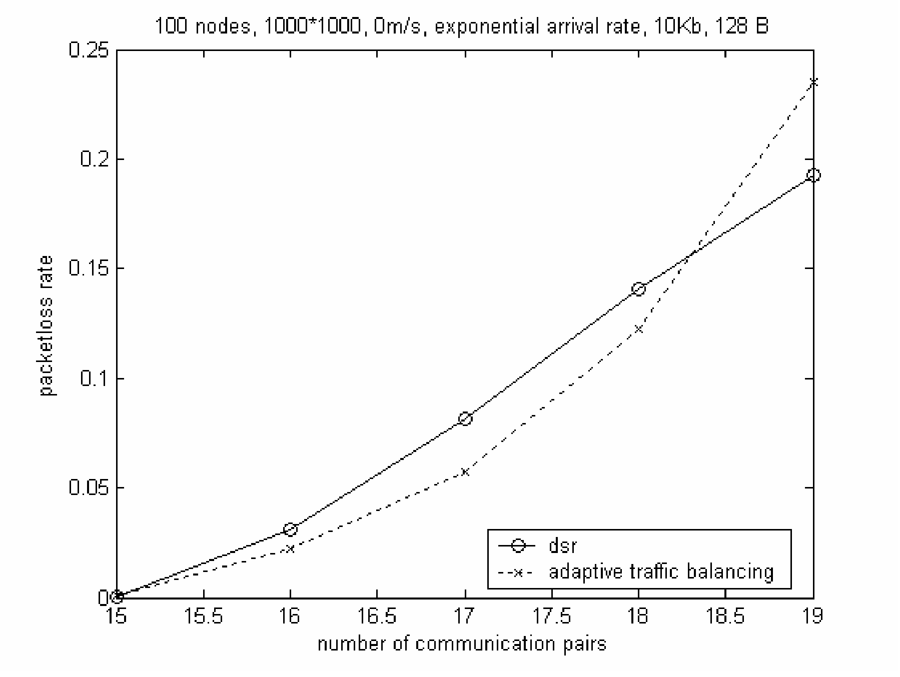


(a) Packet Loss Rate

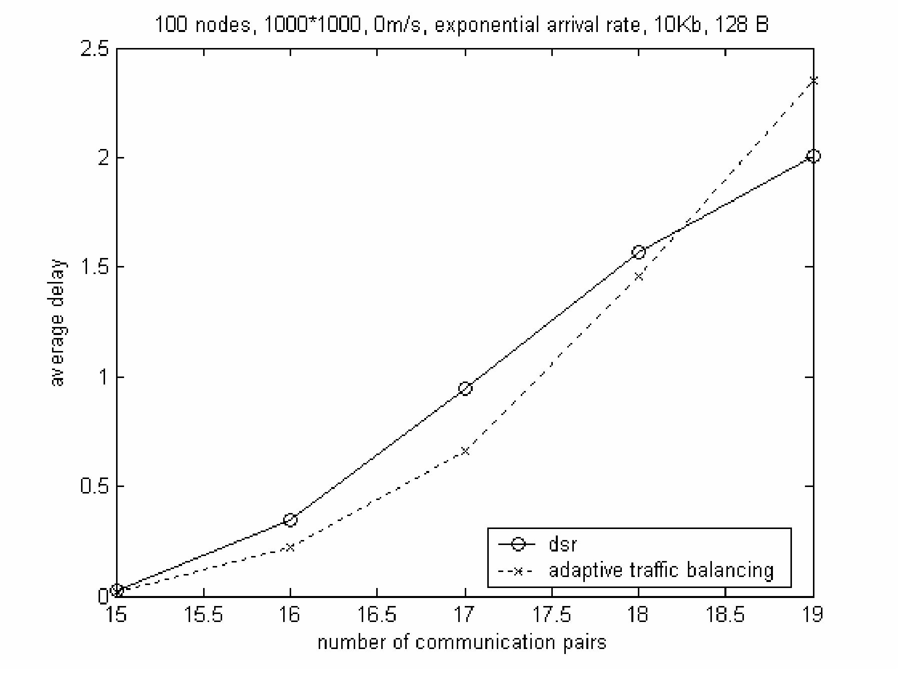


(b) Average Delay

Figure 5.23 System Performance with Poisson Traffic Model at a Max. Moving Speed of 5m/s



(a) Packet Loss Rate

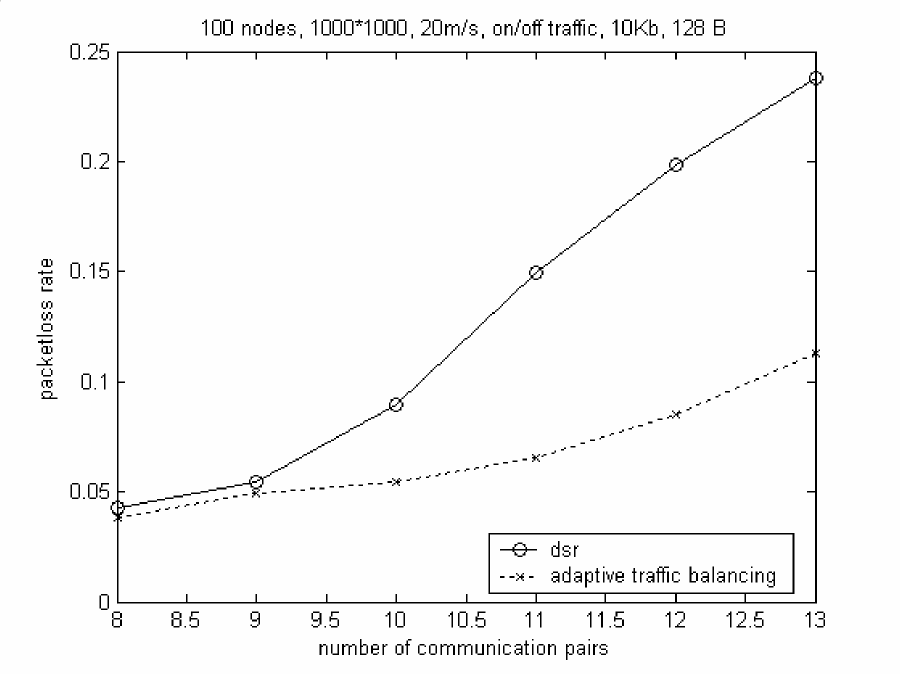


(b) Average Delay

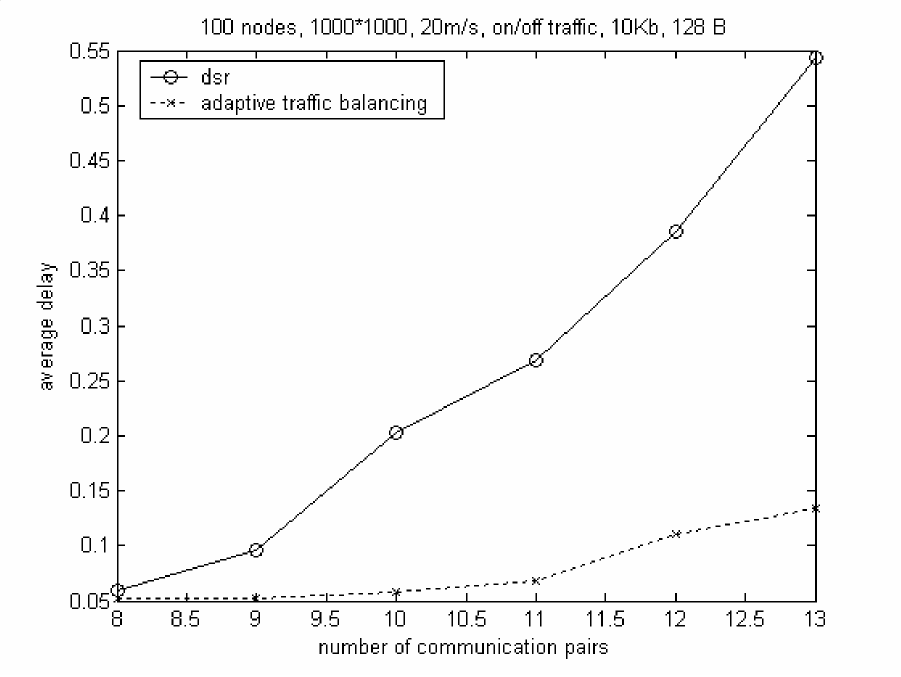
Figure 5.24 System Performance with Poisson Traffic Model at a Max. Moving Speed of 0m/s

More than a decade ago, some researchers observed that real data traffic on the Internet has long-range dependence [C84, CB97, BSTW95]. This type of traffic is also termed "self-similar traffic." However, generating long-range dependence traffic artificially is challenging work, and researchers are still trying to find a solution for this. One simple method of generating traffic that is similar to long-range dependence traffic is to add an on/off property to Poisson traffic. When the communication is in the ON state, the nodes generate Poisson traffic. Otherwise, the nodes keep quiet. The period of the on/off state follows a uniform distribution over 100 seconds. Here the same measurement period and thresholds as in the pure Poisson case are adopted.

Figures 5.25, 5.26 and 5.27 show the simulation results from an on/off traffic model. At high mobility, Adaptive Traffic Balancing shows dramatic improvements in performance, especially with respect to average delay. The performance improvement of Adaptive Traffic Balancing is higher than in the case of CBR and Poisson traffic models. One of the reasons is that after the OFF state, the former path may become obsolete and senders have to request a new path. The new path may be better than the older ones due to changes in the topology and traffic pattern.

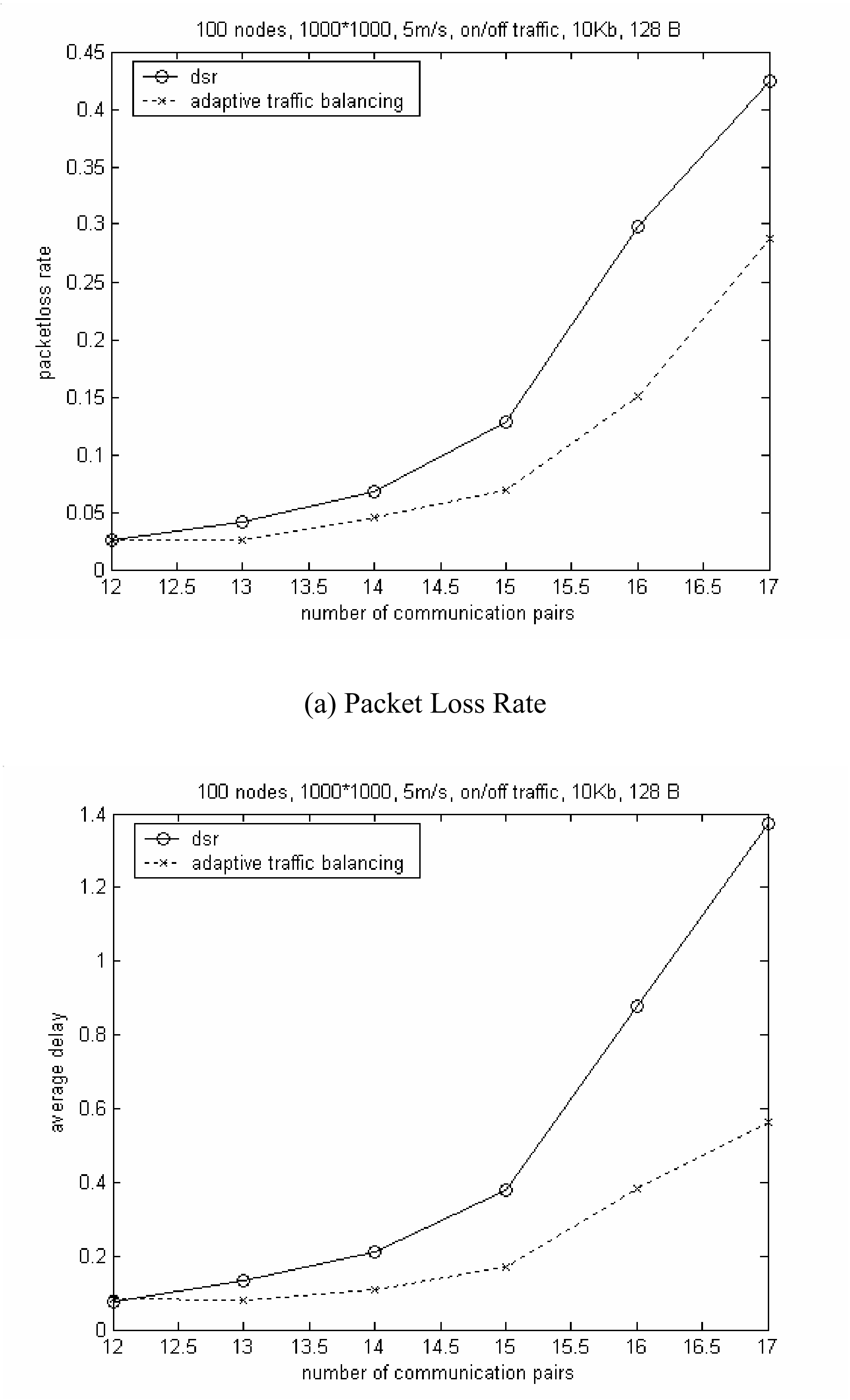


(a) Packet Loss Rate



(c) Average Delay

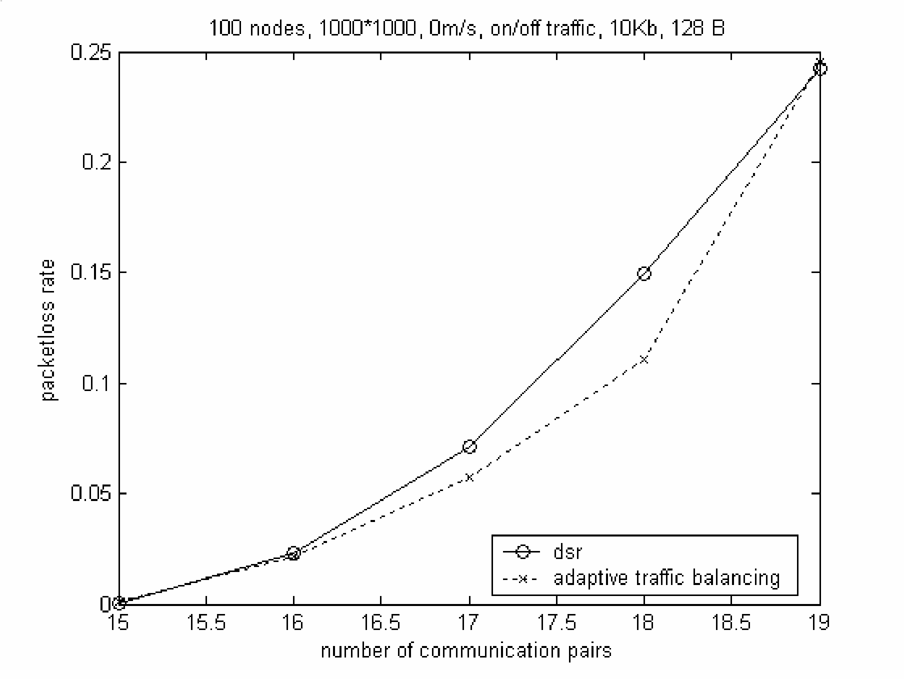
Figure 5.25 System Performance with On/Off Traffic Model at a Max. Moving Speed of 20m/s



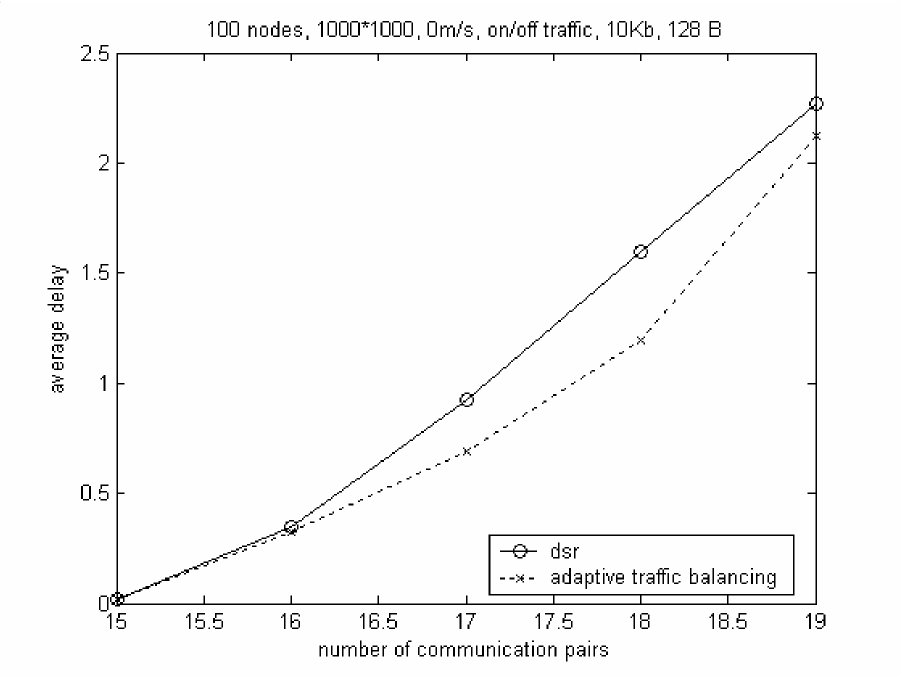
(a) Packet Loss Rate

(b) Average Delay

Figure 5.26 System Performance with On/Off Traffic Model at a Max. Moving Speed of 5m/s



(a) Packet Loss Rate



(b) Average Delay

Figure 5.27 System performance with On/Off Traffic Model at a Max. Moving Speed of 0m/s

Under both the Poisson and the on/off traffic models, Adaptive Traffic Balancing performs better than DSR when the nodes are static and the traffic load is not extremely high. Unlike the CBR traffic model, the interval between consecutively transmitted packets may be larger than the effective time period of the paths in the cache. For this reason, the next packet cannot find a path in the cache, and a new path has to be found using new route requests. The new path is located based on the updated network state, and may be better than the older path that was found during the initial time period. The long interval allows the network a chance to adjust the overall route selection and to reach a better traffic load distribution. This property can also provide a feasible improvement to the applications that continue sending packets in very short intervals using periodic route discovery. Although periodic route discovery may consume more bandwidth for routing information, it could balance the traffic more efficiently using updated information on current traffic and network topology.

In the future, multimedia traffic will become the major input for wireless Ad Hoc networks, such as streaming video (Video on Demand). Some traffic may be bursty and some might require low end-to-end delay. Based on the different requirements, traffic can be divided into different classes. For those classes that care about end-to-end delay, paths with a small number of hops are attractive. For those classes with a bursty traffic load, paths with empty queues are preferable, to decrease the chance of packet-dropping. Traffic Balancing may provide a solution to deal with these traffic classes and their different requirements. Traffic Balancing could set up a medium usage threshold or set of thresholds (in Adaptive Traffic Balancing) for each class, according to its QoS. For example, setting the threshold close to 100% will allow the sender to choose the shortest path that suits the classes that require a low end-to-end delay.

5.6 Summary

The results of the simulation indicate that the traffic load in the center of the network is much heavier than in other areas. When the node density reaches a certain level, some connections can find alternative paths that bypass the center. As a result, the congestion in the central area is decreased, and the utilization of light-loaded areas is improved as well. Of the different parameters, the measured medium usage around the nodes provides more accurate information about traffic load in the area. In Solution 1, Traffic Balancing allows the intermediate nodes to decide whether or not to forward route requests, based on the measured medium usage, so that an upcoming connection may find a path without having to go through a congested area. In Solution 2, Traffic Balancing embeds the measured medium usage of every intermediate node into the routing requests, so that a sender is able to choose the path with the lowest number of congested intermediate nodes and hops. Although Solution 1 outperformed Solution 2 in some scenarios, Solution 2 generally offers a larger possibility for a sender to allocate an alternative path. The performance of Traffic Balancing relies on two parameters: the medium usage threshold and the measurement period. Under different network conditions, such as node speed, the medium usage characterizing the heavy-loaded state is different due to the number of backoff periods and the length of the backoff period. Thus, the medium usage threshold also needs to be adjusted. Adaptive Traffic Balancing alters the medium usage threshold based on the number of collisions seen by the nodes, because collision is a major factor in backoff and is caused by node movement. Besides CBR, Poisson-distributed and on/off traffic was also simulated, and the results showed that long off periods will force new route discovery based on updated network state and that the traffic is distributed more evenly, resulting in increased protocol performance.

Chapter 6. Analysis and Discussion

So far, the performance of wireless Ad Hoc networks has been analyzed with regard to both DSR and Traffic Balancing. Even when the simple physical model and only the IEEE 802.11 MAC layer are applied, some valuable results have been obtained through simulations. The relationships between system capacity, number of users, network size and transmission power (also interference range, i.e. the noise-to-signal ratio threshold) can be ascertained as well.

The resources in wireless communication system are precious and limited. Using them efficiently is crucial as more and more multimedia applications are equipped with mobile devices. As an outcome of our simulation, a more efficient approach to utilizing system resources and co-operating the related protocols will be designed for real wireless Ad Hoc networks. Other research efforts in wireless Ad Hoc networking are also listed in this chapter. In order to evaluate network performance more precisely, the impact of a more sophisticated physical layer model, advanced wireless communication techniques—such as directional antenna—and more traffic models—such as real video streaming—will be briefly discussed.

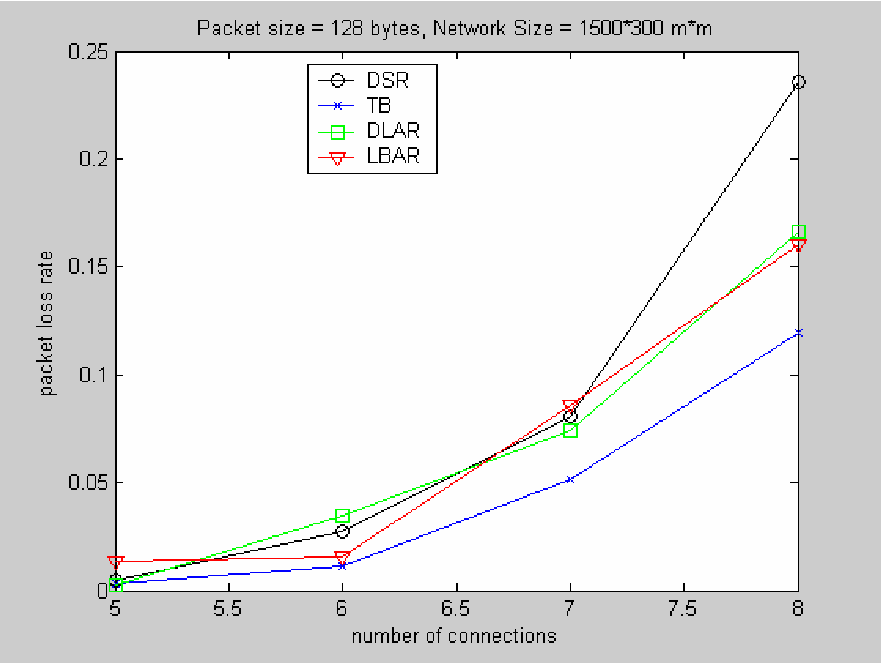
Moreover, an evaluation is undertaken to verify the effectiveness of Traffic Balancing. The comparison with existing techniques that aim to improve system performance illustrates that the Traffic Balancing approach performs better because more traffic load information is acquired from the physical layer. Some results from the simulations are presented in order to support our conclusion.

6.1 Performance Comparison

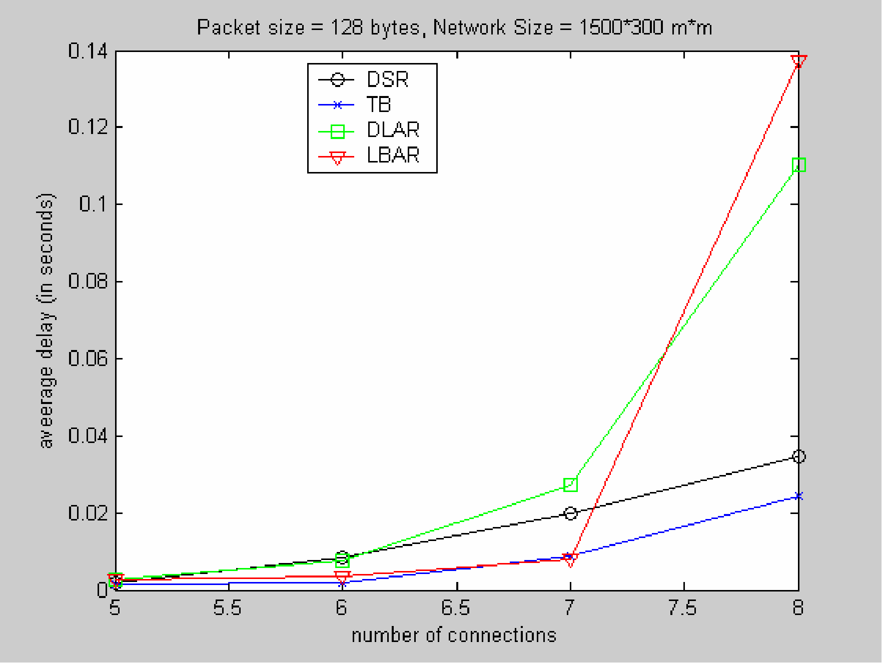
For the purpose of distributing traffic rationally and efficiently in wireless Ad Hoc networks, Traffic Balancing offers an extension to reactive routing protocols in improving system performance with only minor additional complexity. In the previous chapter (Chapter 5), the improvement of Traffic Balancing over DSR was presented. The packet loss rate and the average delay were decreased by exploring unused system capacity. Meanwhile, Traffic Balancing was also shown to be superior over other proposals that try to distribute traffic based on incomprehensive local load information.

As described in Chapter 2, some researchers have proposed methods that take into account the traffic load information of each node locally or partially globally, so as to make better routing decisions. These solutions assign a weight to each hop based on the collected traffic information, and the sender chooses the path by evaluating the sum of the weights. Alternative Path Routing (APR) in [PHST00] is based on the Zone Routing Protocol (ZRP), which is implemented differently to other reactive routing protocols. For the same reason that APR is not implemented to simulate in NS2, it is not feasible to compare the system performance between Traffic Balancing and APR through simulation. Load aWare Routing (LWR) in [YKG01] allows intermediate nodes to drop route requests whenever the nodes have a heavy-loaded status. In some cases, a connection has to go through the busy areas and will not be able to locate a path. The packet loss rate in these cases is extremely high. As a result, the performance under LWR is not acceptable (the packet loss rate is often extremely high). For this reason, we have not run a simulation to compare it with Traffic Balancing. The retransmission rate in Expected Transmission Count (ETX) [CABM03] also collects information about the physical quality of the wireless radio link, which is not a case in NS2 due to the simple physical layer model. Using simulation in NS2 to compare Traffic Balancing and ETX would ignore this aspect of ETX, and so we did not run a comparison in the NS2 environment. As discussed early in Chapter 2, the overhead caused by probe messages could be very high depending on how accurate the measurement needs to be. When network starts to be congested in middle area, the overhead of probe messages will accelerate the congestion and the performance will be very poor. In the future work mentioned in Chapter 8, a more sophisticated physical layer model will be developed and the comparison between ETX and Traffic Balancing will be conducted.

Two approaches (Dynamic Load-Aware Routing (DLAR) [LG01] and Load-Balanced Wireless Ad Hoc Routing (LBAR) [ZH01]) were implemented in the NS2 simulation environment to compare performance in terms of packet loss rate and average delay, because both protocols are based on reactive routing protocols. As Load-Sensitive Routing (LSR) [WH01] collets the traffic information more than DLAR and less than LBAR, LSR's performance may not be much different from them, and LSR was not simulated.



(a) Packet Loss Rate



(b) Average Delay

Figure 6.1 Performance Comparison in Packet Delivery Rate among DSR, Traffic Balancing, LBAR and DLAR.

The simulation results illustrate that more accurate traffic load information leads to the routing protocol making better route decisions and better performance being achieved. Through progressive simulations with different scenarios and packet sizes (as shown in Chapter 5), it was noticed that these two protocols have the following requirements: the packet size should be larger than 128 bytes, communication should be unidirectional, and the number of connections cannot be larger than 10 pairs in order for acceptable performance to be achieved. As shown in Figure 6.1, even when all these requirements are satisfied, their performance is still worse than Traffic Balancing. LBAR has an advantage in average delay compared to DSR when traffic load is light. As the traffic load goes up, both protocols perform better in terms of packet loss rate, but their average delays are much longer than in DSR. Except for the reason mentioned above—i.e. that the information collected from each intermediate node is not comprehensive—other reasons impede the improvement in performance. In DLAR, the queue size of each node is recorded. For those paths with a large number of hops, even if each node has a light-load status, the cumulative value may still be very large, and larger than the shortest path with only one intermediate node that has a full queue. Under these conditions, a better path may not be located. Nodes in the congested area that do not carry any traffic have a very small weight. The overall weight of the paths going through these nodes may be very small. However, choosing these paths may still add traffic load to the congested areas. A similar situation occurs in LBAR. Furthermore, LBAR records the number of connections in neighboring nodes. If a node is close to two intermediate nodes, the number of connections is counted twice. There is then a possibility that the long paths might double count the number of connections in the neighboring nodes, resulting in a large sum and an incorrect decision might be made in this situation.

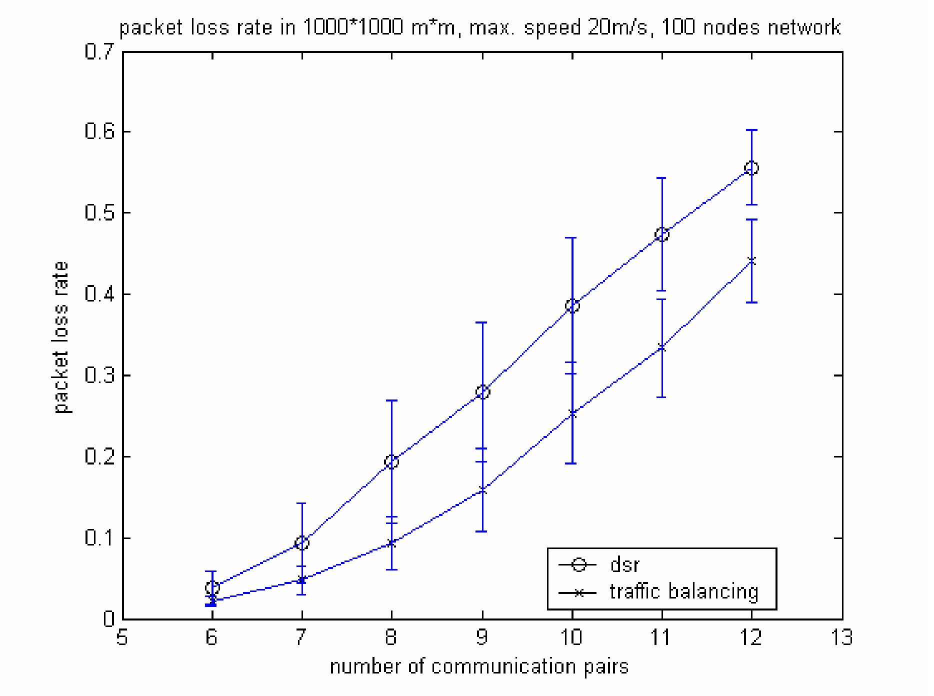
As DLAR only checks the local queue length, the collected traffic information is very limited. The improvement in performance is negligible. LBAR includes the traffic information of neighboring nodes and therefore has a better understanding of the traffic state than DLAR. Thus, LBAR can locate better paths and its performance outperforms

DLAR, as shown in Figure 6.1. Traffic Balancing effectively uses the traffic information over the entire network and the routing decision is improved. Therefore, Traffic Balancing performs better than both DLAR and LBAR.

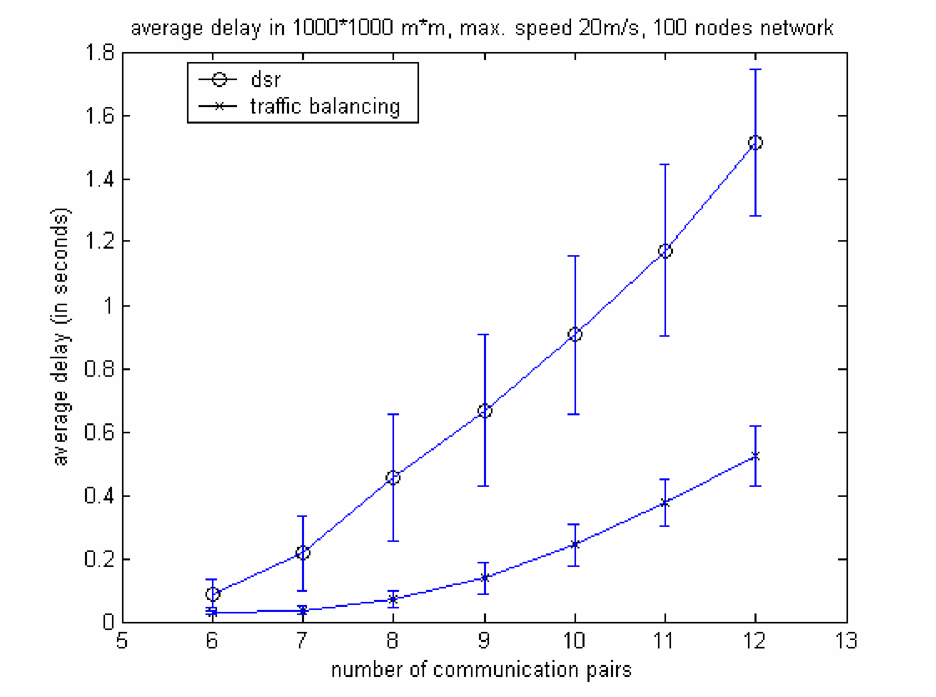
6.2 Statistical Analysis

In the previous chapter, we showed and compared the simulation results for both DSR and Traffic Balancing under different scenarios. Although Traffic Balancing outperformed DSR in terms of both packet loss rate and average delay, the comparison was made based on an average over five or ten workloads, and further statistical analysis is required to show whether Traffic Balancing really is better than DSR from a statistical point of view.

In fact, it is not possible to acquire a perfect estimate of the mean of the packet loss rate and the average delay from a finite number of samples. What we can do is to get the confidence interval for the mean [J91]. Based on the simulation results from 15 runs in 1000x1000m^2 networks with 100 nodes at a maximum moving speed of 20m/s, confidence intervals of 90% for both DSR and Traffic Balancing have been calculated and plotted in Figure 6.2. The connection is bi-directional and generates CBR traffic at a rate of 5.3Kb/s. With a 90% confidence interval, Traffic Balancing demonstrates its superior performance to DSR in terms of average delay. Even though the traffic load is extremely light, the average delay with Traffic Balancing is still smaller than the delay with DSR. The 90% confidence intervals for the packet loss rate in DSR and Traffic Balancing overlap when the traffic load is not overwhelming (the packet loss rate is less than 40%). Thus, by looking only at the 90% confidence interval of the packet loss rate from Figure 6.2 (a), it is not convincing to argue that Traffic Balancing performs better than DSR statistically, and more complex analysis is required to demonstrate whether or not Traffic Balancing outperforms DSR.



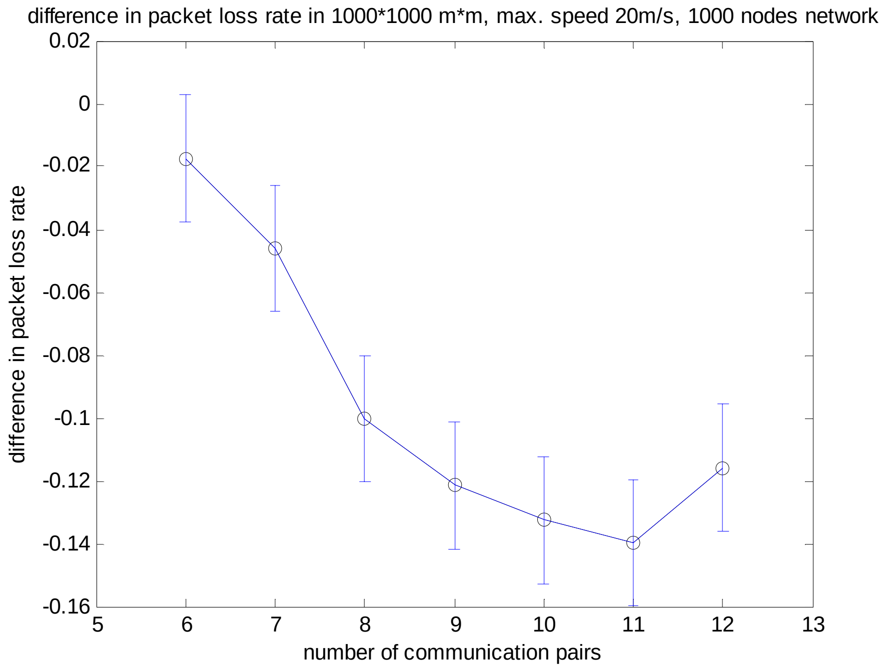
(a) Confidence Interval of 90% for Packet Loss Rate



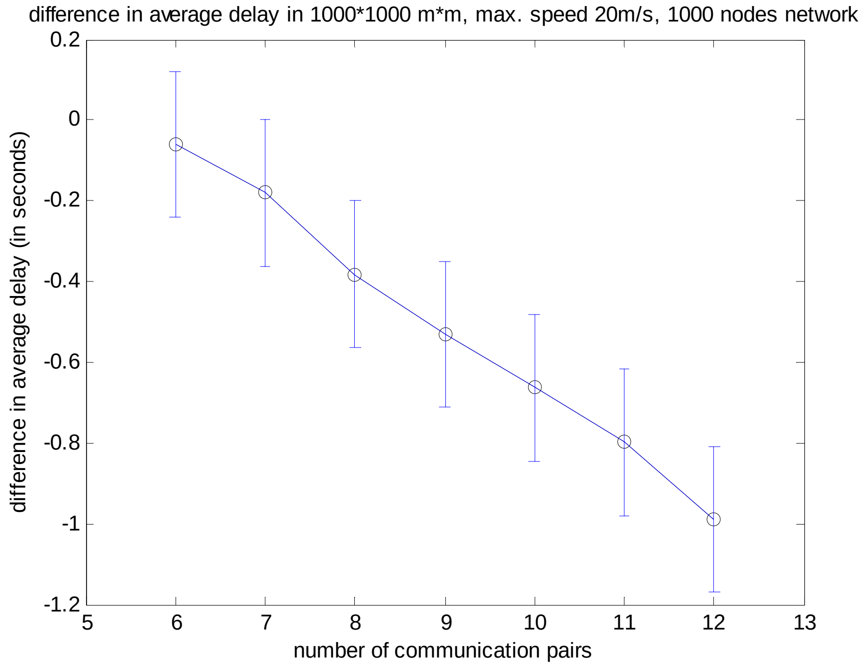
(b) Confidence Interval of 90% for Average Delay

Figure 6.2 Confidence Interval of 90% for Performance of DSR and Traffic Balancing

We conducted 15 experiments on both DSR and Traffic Balancing, such that there was a one-to-one correspondence between the two protocols. Thus the observations were paired and the analysis is relatively straightforward. The difference in performance was computed as the performance of Traffic Balancing, minus the performance of DSR. The 90% confidence intervals for the difference were constructed as shown in Figure 6.3. Except for the case of six communication pairs, the 90% confidence intervals for other cases do not include zeros for differences in both average delay and packet loss rate. This means that the differences are statistically meaningful and that Traffic Balancing has a smaller packet loss rate and average delay with 90% confidence, when traffic load in the network is not too light.



(a) Difference in Packet Loss Rate



(b) Difference in Average Delay

Figure 6.3 90% Confidence Intervals for Difference in Performance

6.3 Network Size vs. System Capacity

Because the transmission power of mobile equipment in wireless Ad Hoc networks has an upper limit, the direct radio connection distance is limited. When the distance between sender and receiver is too great, some intermediate nodes are needed to relay the packets. The number of hops that a connection uses to transmit packets is equal to or greater than the value of the distance divided by the maximum transmission range.

If two radio links are far apart (over interference range), both transmissions can take place simultaneously without interference to each other. Therefore, if the network size increases, more radio transmissions can occur at the same time. However, when a network size increases, the distance between sender and receiver also increases, and so does the number of hops that a connection experiences. The resources required by a connection are equal to the value of the number of hops, multiplied by the data rate.

As shown in the contour plot (Figure 6.4) of packet loss rate in DSR, if performance remains constant, a larger network will accommodate fewer communication pairs. Because a connection with more hops consumes more resources, the total consumed bandwidth is restricted to a certain range. The relationship between the end-to-end capacity and the network size roughly follows O(n/$\sqrt{n}$), as described in Chapter 3.

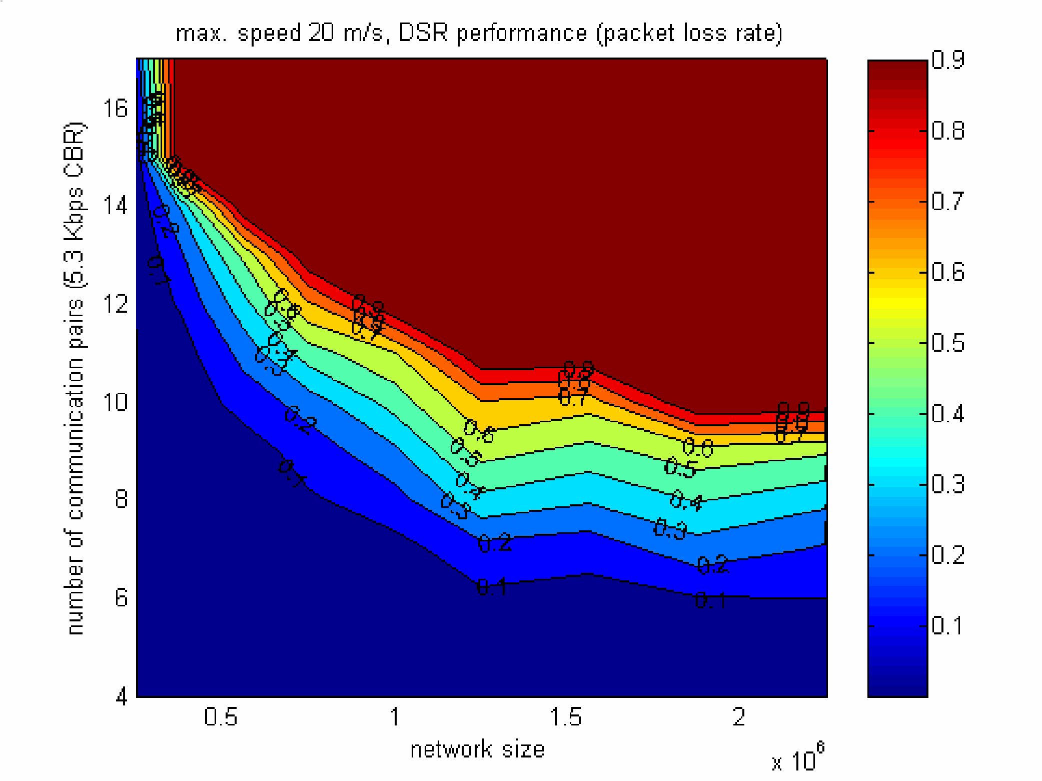


Figure 6.4 DSR: Contour Plot of Packet Loss Rate vs. Network Size and Number of Communication Pairs

After applying Traffic Balancing (see Figure 6.5), the upper limit bound of the contour picture is almost the same. This means when the traffic load is high, all of the system capacity is utilized and it is almost impossible for a routing protocol to find more resources. Traffic Balancing pushes the lower lines up so more communication pairs can be supported under the same performance requirements, or better performance can be reached with the same number of communication pairs when the total traffic load is medium.

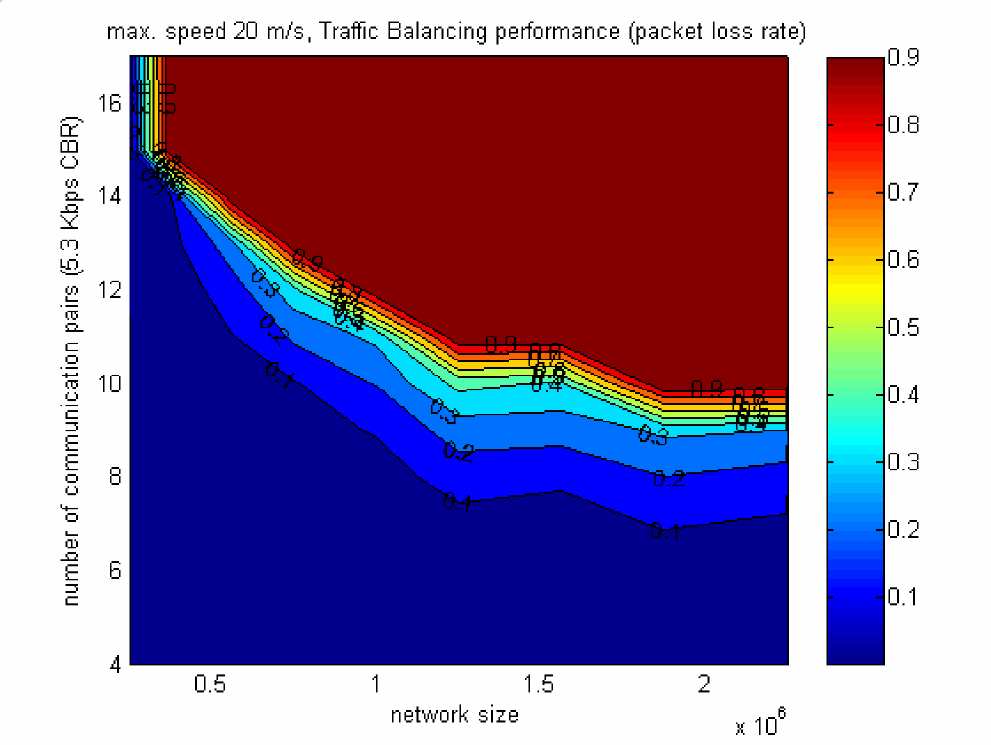


Figure 6.5 Traffic Balancing: Contour Plot of Packet Loss Rate vs. Network Size and Number of Communication Pairs

6.4 Node Mobility vs. System Capacity

The mobility of the nodes leads to topology changes and causes link failures. If nodes move at a high speed, more system bandwidth is used for the control information that is required for route rediscovery and packets retransmission after collision (backoff period). Looking at the 1000x1000m^2 network with 100 nodes, to provide at most a 10% packet loss rate for CBR traffic (5.3Kbps) with a maximum link capacity of 2Mbps, nodes with a maximum speed of 0, 5, and 20m/s can support 13, 10, and 9 simultaneous communication pairs for Traffic Balancing, and 13, 9, and 7 pairs for DSR, respectively. The relationship between maximum speed and number of communication pairs is not linear, because when the nodes are fixed in one place, the control overhead occurs only at the beginning of the simulation and route discovery will not happen after a route has been set up. Therefore, plenty of bandwidth is available for data transmission.

However, the physical model implemented in NS2 is simple. No fading or shadowing is considered. Thus, when the nodes are fixed, the nodes in the communication range will never experience route failure after the route has been set up. In the real world, a radio link may suffer temporary power degradation, and that may cause link failure. Depending on the environment, even with nodes in fixed position, the topology of the network may change frequently and route rediscovery may happen often.

Table 6.1 shows the system utilization under different speeds (both the data packet and the control information packet are included) with network size at 1500x1500 m^2. The estimated system capacity of the network is around 4.72Mbps (see Chapter 3). The utilization of the system is already very high when all nodes are motionless. While there is movement, the system utilization is lowered due to the collisions and following the backoff period. Decreasing the possibility of collision may greatly improve system performance.

Maximum moving speed (m/s)	0	5	20
System utilisation (Mbps)	4.5	2.4	2.2
Percentage of increased utilisation compared to DSR	2%	20%	22%

Table 6.1 System Utilization with Different Node Moving Speeds for Traffic Balancing.

When Traffic Balancing (fixed medium usage threshold at 0.7) is implemented, system utilization increases, especially for mobility scenarios (see Table 6.1). The utilization increases by 20% and 22% for a maximum speed of 5m/s and 20m/s, respectively. Obviously, the performance improves because Traffic Balancing uses more system resources and increases the system utilization.

6.5 Diversity vs. System Capacity

In Chapter 5, our simulations showed that Traffic Balancing does improve system performance. Even when all nodes are static, Traffic Balancing may improve system performance under certain scenarios (on/off and Poisson traffic model). This improvement in static scenarios depends on the possibility that an alternative path with low traffic load is available, and the senders have to rediscover a path after a certain long interval. Unlike in the static scenario, when there is movement, the nodes may migrate to any place in the network, and the possibility of locating a better path is fairly high. For this reason, Traffic Balancing can always be advantageous when there is movement.

To illustrate that Traffic Balancing can bring more improvements if more alternative paths are available, a three-dimensional area network was simulated. The number of nodes in the network remained at 100. All senders and receivers were on the second floor of the theoretical building, while some other non-active nodes were randomly distributed on the first and third floors so that some alternative paths could be found through the nodes on these floors. Ceilings and floors added 20dB attenuation to the receiver power, and the distance between the two floors was set to 3 meters.

To show the benefits of diversity, the simulations were carried out for both 2D and 3D networks with the same number of nodes. In the 3D network, our protocol explored available paths on different floors, so the possibility that a connection could find multiple lightly loaded paths increased. The average number of paths found in the 3D scenario simulation was 3.7, while only 3.2 paths were found in the 2D scenario. Similar to our deduction in the 2D scenario, only one case of improvement in five scenarios occurred when the network size was 750x500 m^2.

It is obvious that Traffic Balancing can benefit the system if more alternatives available. Achieving this diversity is the challenging issue. When designing an Ad Hoc wireless network, some nodes that do not generate any traffic and are used only to forward packets for other nodes can purposely be put into the network to make sure that all areas in the network are covered and more alternative paths found.

If there exists more than one radio channel in a wireless Ad Hoc network, more diversity is provided to the system and Traffic Balancing can benefit from this. In route requests, each radio channel has its own *overThresholdCounter*. The sender broadcasts the route request in only one of the channels and the relay nodes increase the *overThresholdCounter* of each radio channel—or not, depending on the medium usage in each radio channel. Then either the sender or the receiver could compare all the paths in all the radio channels to choose the path with the least number of congested intermediate nodes and hops. As more choices (or paths) exist in multi-radio channel networks, Traffic Balancing could balance the traffic load more evenly and increase efficiency.

Hypothetically, diversity comes not only from the space and resource allocation, but also from the traffic. Non-uniform traffic also leads to unevenly distributed traffic in the system. Traffic balancing can control the system load intelligently in order to reach a better throughput.

6.6 Packet Loss and Routing Overhead

The reasons for a packet being dropped can provide a more detailed description of system performance when DSR is deployed. The bottleneck of the system will be determined and some refinements may be proposed based on this information.

Total number of packets generated	94734
Total number of packets dropped	18297
1. Number of packets dropped by IFQ (Full Queue)	18033
2. Number of packets dropped by RTR (Route Agent)	264
a. Number of packets dropped because of NRTE (No Route)	262
b. Number of packets dropped because of CBK (Call Back)	2

Table 6.2 Number of Packets Dropped in Stationary Scenario

Table 6.2 shows the number of packets that are dropped in one simulation run. In this scenario, all nodes are stationary. The reason why most packets are dropped (approximately 98.56% in total) is queue overflow (IFQ). Due to heavy traffic load, packets being transmitted are always in collision and have to be retransmitted. An increasing number of incoming packets will find a full queue and are dropped. As the traffic load in this area exceeds the network capacity and congestion continues for a while, the number of packets dropped continues to grow. A small number of packets were dropped by the router (around 1.44% of the total dropped packets). Most were dropped because no route was found. Two of them were dropped by the MAC layer. Overall, the main impairment was due to congestion.

Table 6.3 compares the packet loss statistics between DSR and Traffic Balancing for the same scenario (maximum speed: 5m/s; network size: 1000x1000m^2). The total number of packets dropped decreased from 21.71% to 5.85%. Most of the improvement results from reducing the packets dropped by IFQ (queue overflow). Traffic balancing does well in distributing the traffic load evenly into the network. The number of packets dropped by a router was reduced only minimally. This reduction may also be caused by traffic load distribution. Less congestion leads to less retransmission, and the control information can therefore be accessed by the sender on time.

Protocol	DSR	Traffic Balancing
Total number of packets generated	63155	63155
Total number of packets dropped	13708	3693
1. Number of packets dropped by IFQ	11477	1881
2. Number of packets dropped by RTR	2231	1812
a. Number of packets dropped because of NRTE	2171	1812
b. Number of packets dropped because of CBK	60	0

Table 6.3 Packet Loss Statistics of DSR and Traffic Balancing

As Traffic Balancing compares all possible paths between the sender and receiver, it cannot use salvaging, gratuitous route repair, and promiscuous listening to decrease the amount of routing packets. The results of the simulations show that if all nodes in the network have no routing information at the beginning, Traffic Balancing generates 50%

more routing packets compared to DSR. Here, the size of network is 1000x1000m^2, and the average number of hops is about 2.6 for DSR. However, as Traffic Balancing decreases the chance of congestion in certain areas, the chances of a link being broken due to the collision and backoff period are reduced. Hence, the chance of route rediscovery is smaller, and the total number of routing packets is decreased when Traffic Balancing is used. As shown in Figure 6.6, after the initiation of the simulation, Traffic Balancing only required 50% of routing packets that were generated by DSR.

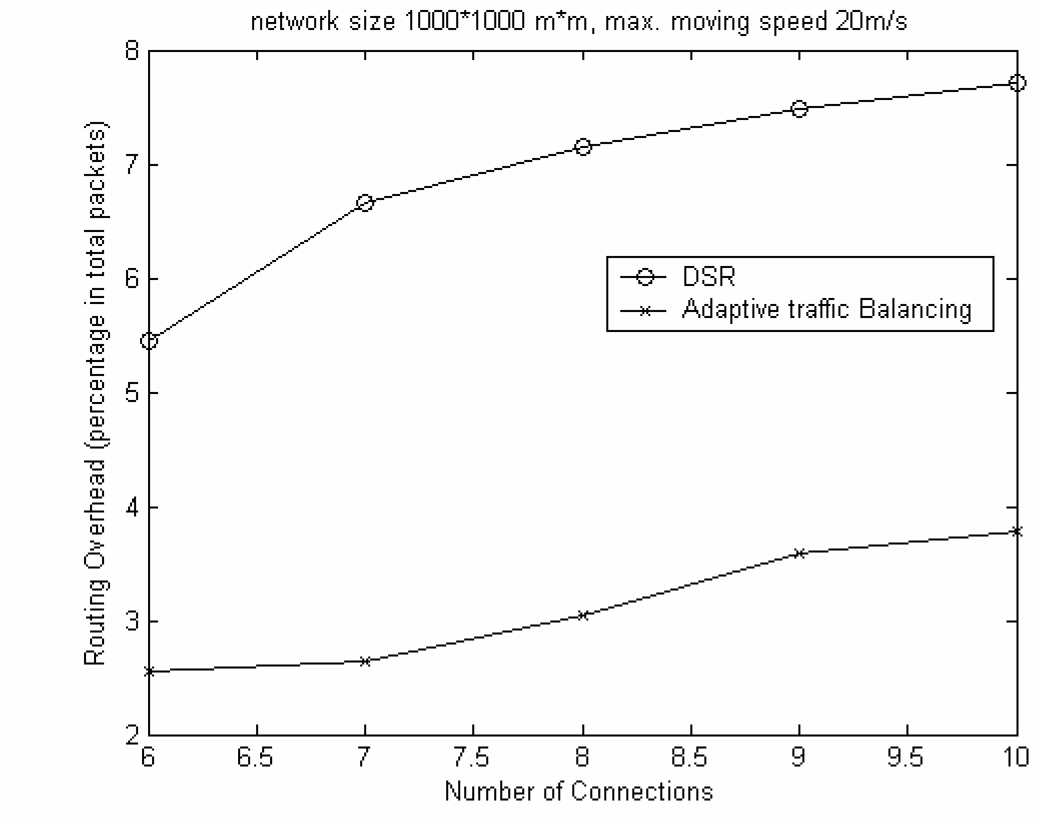


Figure 6.6 Percentage of Routing Packets to Total Packets for DSR and Traffic Balancing.

6.7 Correlation with PHY/MAC/APPLICATION

As previously described (see Formula 5.1), a plain physical layer model was applied in the simulations in order to simplify the system and accelerate the simulation. However, in a real network, a radio connection experiences severe fading and shadowing. The transmission range and interference range are no longer perfect circles. They are irregular shapes and vary with time.

The irregular shape of the transmission range and the interference range forms extraordinary topologies of the networks. For example, a network may be divided into two sub-networks, which are connected through two separated links. One of the links might be used by most users due to the characteristics of the shortest path routing algorithm. Therefore, even if the overall traffic load in the network is not heavy, the system performance may be unacceptable, for the reason that one link cannot handle all the traffic. With the benefits of Traffic Balancing, a partial load can be balanced from this heavily loaded link to the other link. As a result, the performance will be improved and the network resources will be utilized more efficiently.

Due to variation in the wireless environment, such as objects moving, the shape of the transmission range and interference range also changes. This change may lead to a change in the topology of the network and may have an effect similar to node movement. Traffic Balancing could then improve system performance by redirecting the traffic load after the link has broken and route rediscovery has begun.

Fading is usually the consequence of multipath propagation, node movement, and the movement of the surrounding objects. The effect of fading is a rapid change in signal strength over a small travel distance or time interval. Thus, a direct radio link is not stable and a high bit error rate may occur from time to time. A high bit error in a link results in packet corruption and retransmission, and the capacity of the link is decreased as well. Therefore, in a realistic wireless environment, Traffic Balancing may adjust the medium usage threshold based on each radio link, to reflect the real capacity of each radio link.

The faded radio link will then have a low threshold of medium usage and traffic will be routed to avoid traveling through the faded radio link.

To combat the unstable wireless environment and decrease interference, some advanced techniques, such as directional antenna and power control, are proposed. These techniques may be applied along with Traffic Balancing to further extend system utilization. It is essential to provide QoS to applications in a wireless Ad Hoc network to make it comparable to other types of networks. Traffic Balancing has a significant potential to offer QoS for wireless Ad Hoc networks. A brief description of these techniques is given below, together with some existing methods and their challenges when they are deployed in wireless Ad Hoc networks.

6.7.1 Directional Antennas

The electromagnetic energy of the signal in omnimode transmission is spread over an enormous region of space, while only a small portion of it is received by the intended receiver. A directional antenna is a technology proposed for dealing with this problem. By using M elements, this kind of antenna transmits in directional mode, which means that the electromagnetic waves are enhanced in certain directions while they are canceled in others, resulting in an amplified signal that is directed to certain positions. Because they incorporate these main characteristics, directional antennas constitute an attractive component for all wireless devices.

Directional antennas have many potential benefits for wireless Ad Hoc networks. The nature of the directional transmission results in spatial reuse, as multiple transmissions can take place in the same neighborhood without destroying the transmitted packets. At the same time, the directional transmission increases the signal energy in the direction of the receiver, resulting in an expanded coverage area. These two benefits result in an increase of the channel capacity.

Research in [PS03, SR03, YPK03, BPSUHP03] analyzes the capacity improvement of wireless Ad Hoc networks through the use of directional antennas. Based on the approach in [GK00], [YPK03] finds that the improvement can reach a factor of $\frac{2\pi}{\sqrt{\alpha\beta}}$. Here α and β are the beamwidths in radians of transmission and receiving directional antennas, respectively. [PS03] formulates the capacity as a multi-commodity flow problem and gives the gain of maximum stable throughput to be, at most, $\Theta(\log^2(n))$ (here n is the number of nodes in the network), due to the usage of multiple simultaneous arbitrarily narrow beams. However, the result from [BPSUHP03] shows that the overall system performance in every context cannot be improved if only a directional antenna is applied.

Although a directional antenna can improve system capacity, it can also cause problems in wireless Ad Hoc networks, e.g. increasing the instances of hidden terminals, deafness in routing and control information, and difficulty in determining neighbors' location. Because traditional MAC protocols have been designed for wireless Ad Hoc networks with omnidirectional antennas that do not properly support the directional antennas, several new MAC protocols are proposed in [KJT03, RSBUT03, HS02]. Problems still remain regarding how to evaluate the different paths and how to discover a path quickly. As an example, with a directional antenna, each node has to send or forward route requests in every direction one by one, and it might take a very long time to set up a connection. If the nodes return to the omni-directional antenna when sending or forwarding route requests, some advantages of the directional antenna are lost.

6.7.2 Power Control

Power control has been studied extensively in the context of cellular radio systems, both channelized and CDMA-based. Distributed iterative power control algorithms have been introduced for cellular systems, and convergence results have been established. Power control has received a lot of attention recently, for two primary reasons. First, power control has been shown to increase spatial channel reuse, hence increasing the overall channel utilization. This issue is particularly critical given the ever-increasing demand for channel bandwidth in wireless environments. Second, power control improves the overall energy consumption in a wireless Ad Hoc network, consequently prolonging the lifetime of the network [MK04].

The basic suggested power control scheme is based on the IEEE 802.11 MAC protocol and works in each RTS/CTS/DATA/ACK loop [AKKD01, FTK02, MBH00]. The RTS/CTS uses the maximum power level, while DATA/ACK uses the minimum required power level. However, decreasing transmission power shrinks the sensing zone of the node, and it may lead to packet collisions during the DATA/ACK period. In order to eliminate this asymmetrical link phenomenon, [LKL03] proposes using a short power control packet in a separate power control channel to inform the neighbors of the transmission power level. [MK04] chooses to control the power level of routing requests so that a power-efficient network topology can be built. [EE02, YSL04, PS02] coordinate a group of nodes into a cluster and set the power level of each cluster, thereby eliminating the asymmetrical link problem.

As the IEEE 802.11 MAC protocol may not be the best multiple access control protocol for wireless Ad Hoc networks from other aspects, it is difficult to adapt some power control schemes to new MAC protocols that may not have an RTS(request to send)/CTS(clear to send)/DATA/ACK loop. Some schemes require extra traffic or extra bandwidth to handle information concerning power level. In addition, no scheme takes into account traffic models and node movement patterns that have a significant impact on system performance under power control.

The implementation of both the directional antenna and power control requires more sophisticated routing algorithms. Furthermore, if traffic concentrates on certain links due to the routing algorithms, a directional antenna and power control still won't be able to improve system performance, since the bottleneck is the capacity of certain links. Thus, Traffic Balancing is needed to balance the traffic load over the entire network before a directional antenna or power control can further improve system performance.

6.7.3 Quality of Service (QoS)

In most applications, and especially if wireless Ad Hoc networks are to be commercially viable, quality of service is an essential component. QoS aspects include bandwidth, delay and delivery guarantees. It has been proposed that to achieve reliable QoS, wireless Ad Hoc networks will require traffic engineering capabilities, and providing these capabilities will require the cooperation of three components: a QoS-capable medium access control protocol; a resource reservation scheme; and a QoS routing protocol [PH02].

Most early research on QoS focused on designing QoS routing protocols [CM01, CN99, LL99]. Later, more researchers found that it is difficult to satisfy QoS guarantees solely in higher layers without support from the MAC layer [XL04, SK99]. [SV04, CP03] move further into the physical layer to ensure certain QoS requirements. Research in [PP03] provides a theoretical analysis for QoS implementation in wireless Ad Hoc networks, if global information can be acquired.

IEEE 802 has defined a new MAC protocol (IEEE 802.11e) to support QoS, and more research groups are still working on the MAC layer to address multichannel problems and be more efficient [RP03, CSC03, CCG00]. A lot of research still needs to be done to address QoS implementation, which may combine techniques from all layers.

Traffic Balancing selects paths based on the number of relay nodes over the medium usage threshold. If different medium usage thresholds are set for different applications, some application requirements will be guaranteed. For example, the medium usage threshold for a delay-sensitive application should be larger than the medium usage thresholds for other applications. Traffic Balancing will then allow delay-sensitive applications to select paths that are shorter than the other applications. Traffic Balancing can match different service classes against different thresholds to meet their QoS requirements.

6.8 Evaluation of Traffic Balancing

Reactive routing protocols aim to find a path as quickly as possible when a packet with an unknown route needs to be delivered. However, a fast response may not reflect the overall system state. The routing packets have a higher priority and are processed in advance of the data packets, and it then takes a shorter period for the routing packets from sender to destination. To overcome this shortcoming, more traffic load information is required to make better routing decisions.

By measuring the medium usage around the node, Traffic Balancing adds accurate traffic load information to the routing packets. This measured medium usage includes all activities in the network so that the route decision will be more optimal. Traffic Balancing shows fairly constant improvements in terms of both packet delivery rate and average delay, even compared to other traffic distribution solutions, which only collect some traffic load information. In some cases, keeping away from the congested area leads to longer paths, which may increase the average delay for these rerouted connections. However, decreasing the traffic load in congested areas reduces the delay of connections going through the congested areas, and the overall average delay is smaller.

When comparing system performance with two other traffic distribution proposals (DLAR and LBAR), it is worth noting that more traffic load information can help routing protocols make better route decisions. As Traffic Balancing acquires most of the traffic load information in the network, it outperforms the other proposals. Statistical analysis shows that Traffic Balancing outperforms DSR in term of average delay, with 90% confidence. By analyzing the difference in performance for both Traffic Balancing and DSR, Traffic Balancing does better than DSR with respect to packet loss rate, with 90% confidence. Although the network capacity is limited, Traffic Balancing allocates system resources more efficiently. The packet loss rate contour plots (Figure 6.4 and 6.5) show that Traffic Balancing extends the good-performance region under the limitation of network capacity. Traffic Balancing improves system performance dramatically under high node mobility, because high node mobility causes a huge number of collisions in the congested areas and Traffic Balancing reduces the number of congested areas. With fewer congested areas, the number of packets dropped due to full queues is diminished, as shown by the statistics of dropped packets in Tables 6.2 and 6.3. In addition, fewer congested areas lead to a lower number of links broken, with less control traffic then generated for route rediscovery in Traffic Balancing (as shown in Figure 6.6). More diversity from space, resource allocation, and traffic pattern enable Traffic Balancing to further increase in efficiency. Several advanced techniques, such as directional antennas and power control, are proposed to increase system capacity, but unacceptable performance may still occur without assistance from Traffic Balancing. Overall, Traffic

Balancing provides more advantages than other proposals, in terms of diversity, complexity or topology information. Meanwhile, it also brings out ideas for offering QoS in the IP layer and extending system capacity further, by combining it with other techniques at the physical layer or data link layer.

6.9 Summary

Compared to other load balance proposals for wireless Ad Hoc networks, Traffic Balancing is more stable and fits with most scenarios and for various types of traffic model, because the information collected by Traffic Balancing includes all activities in the network. Statistically, Traffic Balancing shows its superior to DSR, with 90% confidence. In general, the size of a wireless Ad Hoc network determines the average number of hops between sender and receiver. The bigger the network is, the larger the average number of hops is. More resources are required for each connection and the number of control packets will increase. Although the capacity also increases with the size, the overall throughput still degrades. Node mobility causes collisions, especially in congested areas. Traffic Balancing moves the traffic load from the congested areas, so that the number of collisions is decreased and system utilization is improved. Traffic Balancing improves system performance by locating alternative paths through fewer congested intermediate nodes. More paths provide more choices for Traffic Balancing. The number of data packets dropped due to full queues in congested areas is large for typically on-demand routing protocols, such as DSR. With Traffic Balancing, the prospect of overflowing queues decreases dramatically. A realistic physical layer model is complex to simulate, but may provide more diversity, from which Traffic Balancing could benefit. Recently, there has been more research on directional antennas and power control with the aim of improving the system capacity. In addition, researchers are working on how to provide QoS in wireless Ad Hoc networks. Some advanced physical layer techniques could improve system performance, but without Traffic Balancing, some routing problems still limit the potential for improvement.

Chapter 7. Wireless MESH Network

In a wireless Ad Hoc network, information is exchanged mostly between users, such as users playing interactive video games. In the future, more and more people will demand Internet access through wireless Ad Hoc networks. Several access points may be deployed in wireless Ad Hoc networks for people connecting to the WWW or other Internet services. This kind of wireless network is called a wireless mesh network (WMN), and still has nodes that need to relay packets for other users.

The meshed topology provides an efficient alternative to other broadband technologies, including cable, digital subscriber line (xDSL), broadband wireless local loop, and satellite Internet access. In downtown areas with a high density of mobile users, WMNs have significant advantages in Internet service offering. As more access points are installed in the same area, network capacity could be increased.

Despite the recent startup surge in WMNs [W00], a great deal of research remains to be done before WMNs realize their full potential. In this chapter, a brief description of wireless mesh networks is presented. The issues that may affect the performance of WMNs are discussed, and the ways in which Traffic Balancing could solve some of them are also addressed.

7.1 Overview of WMN

Figure 7.1 depicts a possible wireless mesh network scenario. In a WMN, several access points (G1 to G4), which have other communication methods of connecting to the Internet, are deployed, and they do not have to cover all the areas of the network. If the node density is at certain level, most nodes can reach the access points in a few hops to access the Internet. Therefore, WMNs have several significant advantages [W00]:

- Very high coverage levels with very low initial investment.
- Excellent spectral efficiency.
- Complete flexibility in service delivery.

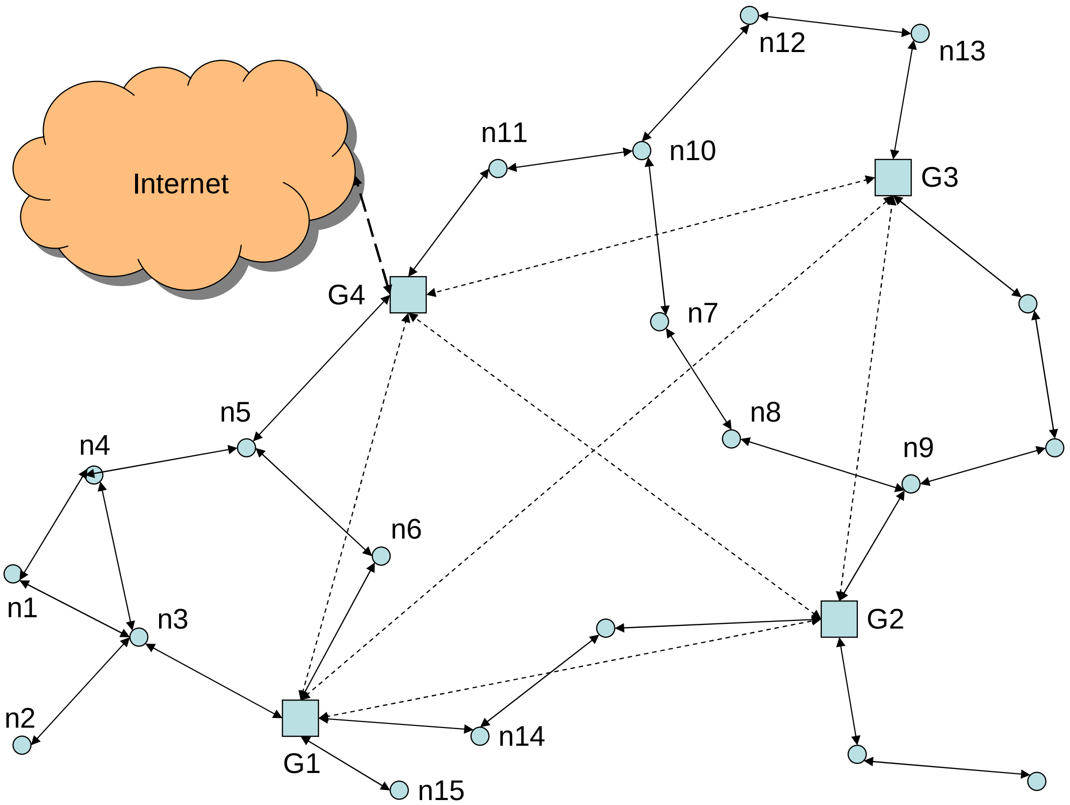


Figure 7.1 Example of Wireless Mesh Networks

Unlike wireless Ad Hoc networks, traffic in WMNs concentrates in the area around the access points and the throughput of each node decreases as $O(1/n)$, where n is the total number of nodes in the network [JS03]. However, if there is more than one access point, a mobile node is able to choose a feasible path leading to one of the access points. Here we assume that a protocol takes responsibility for coordinating among the access points. Therefore, a node can change the path that may lead to different access points, without worrying about the interruption of ongoing communications (similar to handover in cellular systems). If desired, repeater nodes (or pure wireless routers) can be deployed to extend the coverage, or to improve the performance of the network. The access points in WMNs can be added one at a time, as needed. Adding more access points will increase not only the capacity of the network but also its reliability.

7.2 Problems with Implementation

Several problems must be addressed in order to offer proper services to users. These problems are related to the fairness in throughput and efficiency in resource allocation.

7.2.1 MAC Layer Protocol

The IEEE 802.11 MAC layer protocol is the most popular choice for wireless Ad Hoc networks. Under the definition of the IEEE 802.11 MAC layer protocol, all users have the same level of priority in accessing the channel, so that the bandwidth can be allocated fairly to the nodes with packets in the queue. However, this property is not suitable for WMNs when the IEEE 802.11 MAC layer protocol is deployed at the access points. In WMNs, practically all traffic is either to or from the access points. When all mobile nodes are one hop away from the access points, the traffic load from the access points to the mobile nodes is heavier than to the neighboring mobile nodes (if the traffic from the mobile nodes to outside the WMNs is equal to the traffic from outside to mobile users, access points need half the bandwidth, similar to base stations in cellular networks). For example, access point G1 in Figure 7.1 has to compete for bandwidth with mobile nodes n3, n6, n14, and n15. If all of them have packets to be sent out, G1 only has about a 20%

chance of accessing the medium. Due to the asymmetric throughput around the access points, the delay experienced at the access points for those packets addressed to the mobile nodes is larger. If the overall traffic load becomes bursty, the data packets from the access points to the mobile nodes may be dropped because of full queues. When the overall traffic load is over the system limit, the number of packets from access points to the mobile nodes is far below the number of packets from the mobile nodes to the access points.

Number of Communication Pairs		Packet Loss Rate (in %)	Average Delay (in seconds)
10	uplink	0	0.0225
	downlink	0	0.0226
15	uplink	0	0.0409
	downlink	1.3145	0.7568
20	uplink	0.3257	0.0595
	downlink	15.6389	1.4513

Table 7.1 Performance Comparison Between Inside and Outside Traffic for WMNs with Two Access Points.

Table 7.1 shows the results taken from simulations where all nodes in the network were static. All the results were averaged over 1,000 seconds. The size of the network was $1000\times1000m^2$, and the number of mobile nodes was 100. Each connection was bidirectional CBR traffic with a rate of 5.3Kbps. At very low level of traffic load, both directions performed approximately the same. When traffic load reached an intermediate level (15 pairs), the downlink packets accumulated in the queues at the access points and the average delay became great (from 0.0226 to 0.7568). Some packets were even dropped due to full queues. Meanwhile, the performance of the uplink packets (from mobile users to outside) was retained. At a heavy traffic load level, the performance became even more asymmetrical. The performance of the downlink was totally unacceptable, while the quality of service on the uplink was still fine (the packet loss rate was less than 1% and the average delay was around 50ms).

If access points are given more chance to access the medium, downlink data packets might not be dropped. Moreover, the delay at the access points can be decreased to satisfy the QoS requirements for those time-sensitive applications. One simple solution is to keep the CW in the MAC of the access points to the smallest value (CW_{min}). Then the access points can occupy nearly half the bandwidth if the nodes are mobile. However, a more rational method is needed to achieve a stable performance in any situation. Local coordination may be required to assign access rights to each node or access point. Meanwhile, the new protocol should not waste too many resources doing so, and the performance of nodes far from the access points should not be impacted too much.

7.2.2 Location of Access Points

Unlike wireless Ad Hoc networks, the throughput of WMNs is also related to the location of the access points, especially when more than one access point is deployed. Obviously, fewer hops for each connection lead to a better throughput in the WMN. The best location for the access points will be the place where the overall average number of hops for all connections is the smallest.

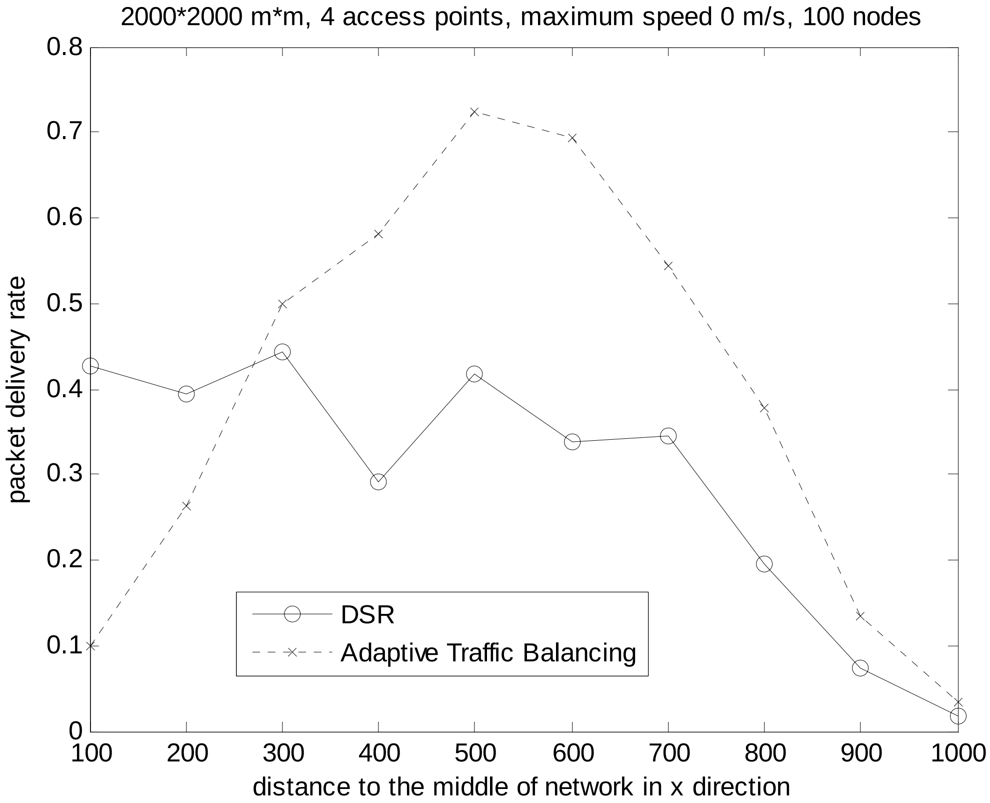


Figure 7.2. Performance Depends on Location of Access Points in WMNs.

Figure 7.2 plots the relationship between the position of the access points and the packet delivery rate (percentage of generated data packets that are received by the users) when the number of connections is kept constant. The network size is 2000×2000 m^2 and the maximum moving speed for each node is 0m/s (static). There are four access points in the network, as shown in Figure 7.3, and the x-axis in Figure 7.2 indicates the distance from each access point to the center of the network in a horizontal direction. Similar behaviors were noted for networks with two, three, or more access points.

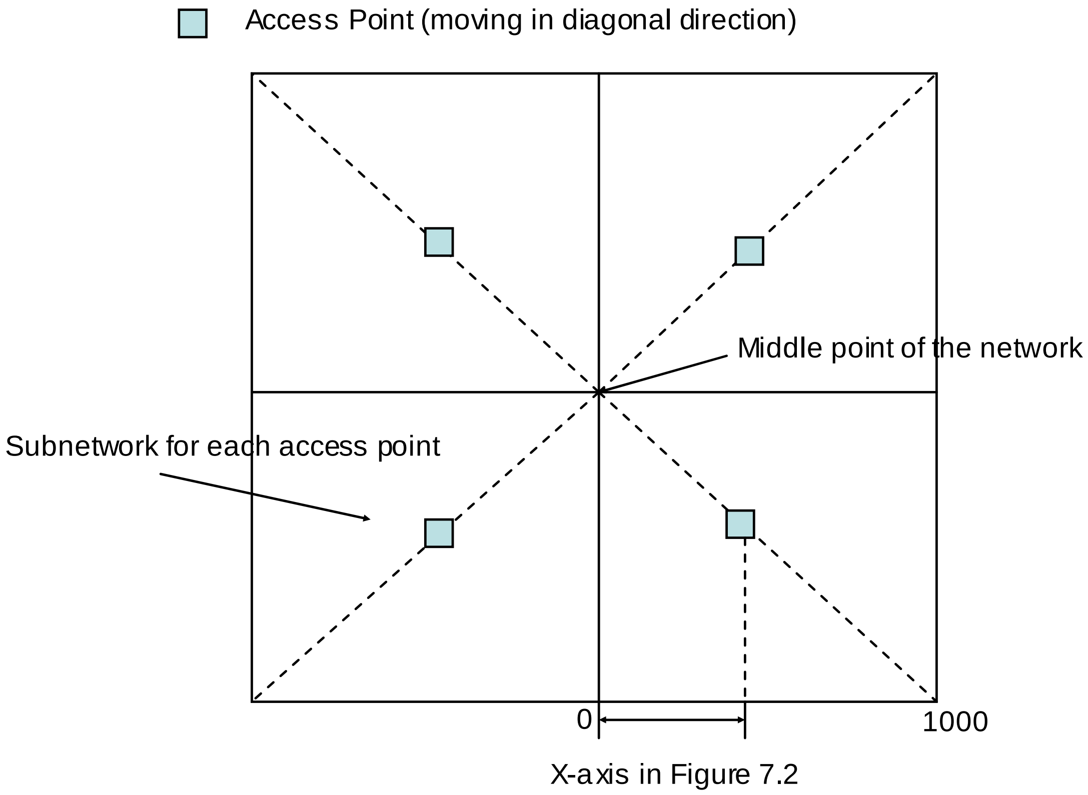

Access Point (moving in diagonal direction)
Middle point of the network
Subnetwork for each access point
0
1000
X-axis in Figure 7.2

Figure 7.3 Position of Access points in Figure 7.2.

DSR's performance showed minor differences when access points were located in the 100 to 600m range. When the access points moved further away, the performance decreased accordingly. The reason for this phenomenon is that at the corner of the network, traffic comes from a quarter of all possible directions (90 degrees), while in the middle of the network traffic comes from all directions (360 degrees). Therefore, interference at the corners is higher, and all traffic in each sub-network may go through the same few intermediate nodes before reaching the access points. On the other hand, performance also degrades when all access points move close together (<100 m). In this case, as all access points are very close, as if there is only one access point, no additional capacity is obtained for the network.

However, Traffic Balancing shows a different relationship between the location of the access points and system performance. Figure 7.2 shows that the best performance occurred when the access points were located around the middle of the sub-network. This location may result in the overall average number of hops being smallest. When the access points move farther away, performance decreases, for the same reason mentioned above. When the access points move closer, Traffic Balancing has a reduced chance of finding a better path, because the traffic load will concentrate in the middle anyway. As Traffic Balancing requires more control packets, its performance worsens compared to DSR when the access points are allocated in the network center.

The above discussion shows that finding the best place for access points is nontrivial. A good location can optimize system performance. However, where to locate access points may depend on node density, traffic pattern and other issues. In addition, performance can be improved by deploying Traffic Balancing, as illustrated in Figure 7.2. In Section 7.3, we will show how Traffic Balancing benefits WMNs.

7.2.3 Uneven Traffic Load

There is another problem related to the routing protocol when more than one access point is deployed in the network. As all mobile nodes choose the shortest path to one of the access points, traffic may concentrate at one access point, due to the position of mobile nodes and the traffic pattern. For instance, in Figure 7.4, most traffic goes to access point G1, due to the traffic pattern and the routing protocol, even though some of them could be routed to access point G2, which is nearby.

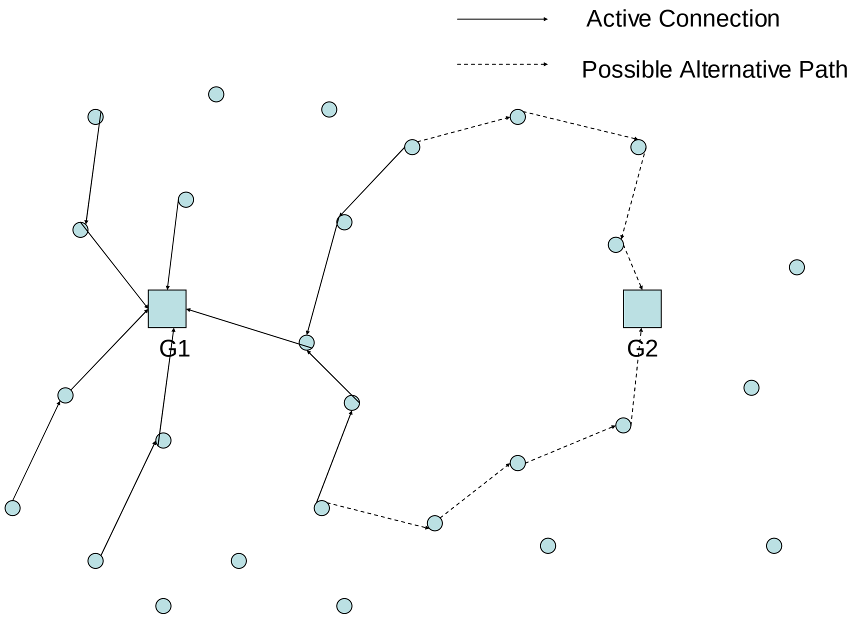


Figure 7.4 Uneven Traffic Load to Access Point

Table 7.2 shows one simulation result that indicates that access points have different incoming traffic when the performance reaches a critical point (packet loss rate reaches 5%). Meanwhile, Traffic Balancing allows some traffic to be routed away from the busy access point to the idle access point, resulting in better overall performance.

	Number of Access Points	Packet Loss Rate	Number of Packets (Access Point 1)	Number of Packets (Access Point 2)
DSR	2	7.26%	3993	1104
Traffic Balancing	2	1.32%	4347	2552

Table 7.2 Uneven Traffic Load at Access Point

Besides the above problems, there are other issues with the implementation of WMNs.

- As a node can choose any access point, route discovery has to be broadcast to all access points in the network. Afterwards, the connection becomes a peer-to-peer one.

- If the communication is started outside the WMN, all access points need to flood the route request in order to find the best path to the receiver, even though the receiver may only register with one of the access points.

7.3 Traffic Balancing in WMN

As mentioned in the previous sections, the efficiency of WMNs is far below optimal when there is often an uneven traffic load. Traffic Balancing solves this problem efficiently. For example, the node n1 in Figure 7.1 has two available paths to reach the access points (path (n1, n3, G1) leads to the access point G1, and path (n1, n4, n5, G4) leads to the access point G4). If the nodes n1, n2, n3, n15 and n14 all intend to use G1 as an access point, traffic around G1 will be high and the throughput for each user will be very low. If Traffic Balancing is used, n1 will choose path (n1, n4, n5, G4) because the traffic around G4 is low. As a result, the throughput of all the nodes can be improved.

The results from Figure 7.2 and Table 7.2 already show that Traffic Balancing performs better than traditional DSR in the cases where access points are placed in an appropriate position. After balancing traffic from busy access points to idle ones, all access points will achieve a higher throughput, as indicated in Table 7.2. For busy access points, reducing the traffic load decreases the possibility of collisions, and fewer resources are wasted in backoffs. Figures 7.5, 7.6 and 7.7 compare performance in packet delivery rate between Traffic Balancing and DSR. The network size was 2000×2000 m^2 and two access points were placed at (500, 1000), (1500, 1000). Again, the results illustrate that the system performance improved dramatically when the nodes were in movement. When the traffic load was at an intermediate level (eight–ten connections), Traffic Balancing supported over 20% more traffic than DSR did. The improvement in the static scenario was negligible because Traffic Balancing required more bandwidth for its control packets than DSR did. In addition, a CBR connection was maintained until the end of the simulation, so that connections set up early had the advantage of taking the shortest path and gave Traffic Balancing less of a chance to find an alternative path for upcoming connections.

As WMNs have various traffic patterns, the traffic load around the access points tends to be heavy. Thus, the hops closer to an access point should acquire more bandwidth when a route decision is made. This information should then be included in the route reply. Revisions to Traffic Balancing are necessary in order to fit with the new features of WMNs.

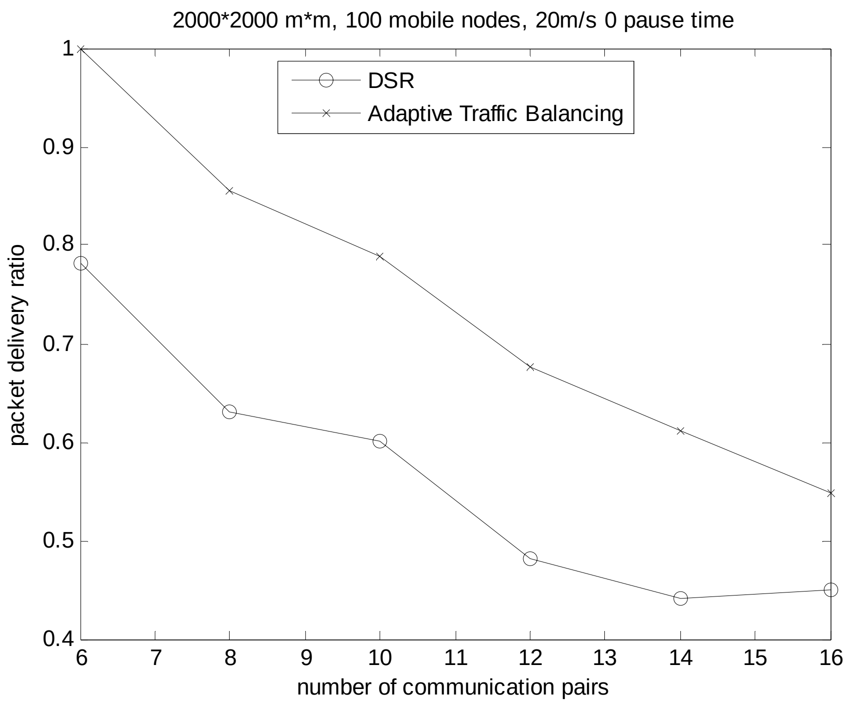


Figure 7.5 Traffic Balancing for WMN at a Max. Moving speed of 20m/s

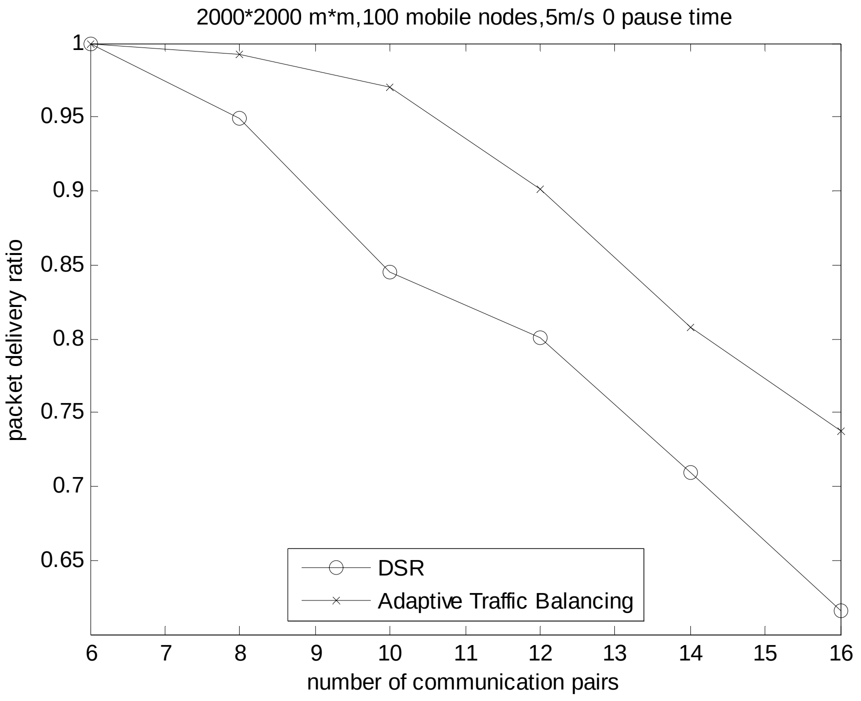


Figure 7.6 Traffic Balancing for WMN at a Max. Moving speed of 5m/s

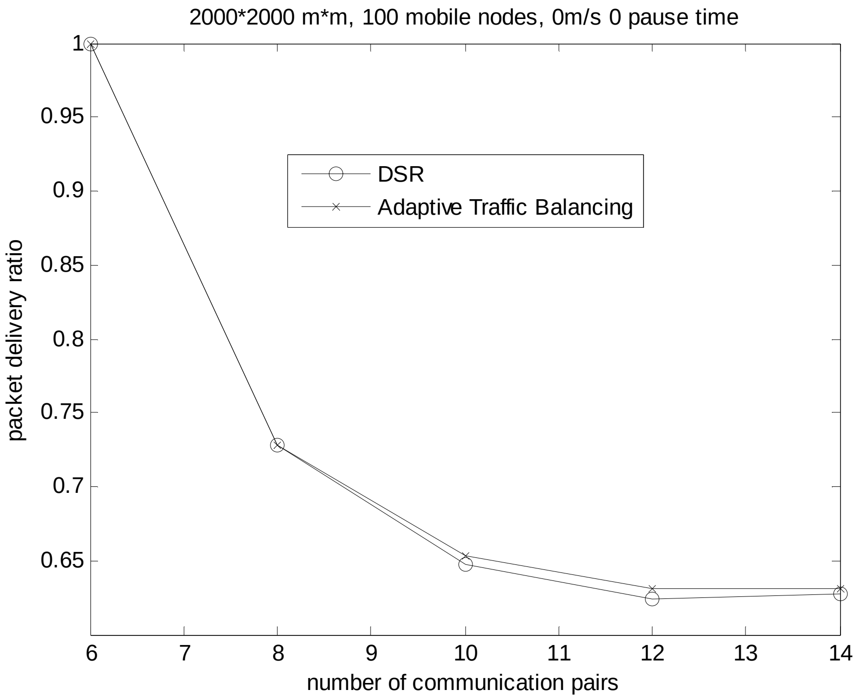


Figure 7.7 Traffic Balancing for WMN at a Max. Moving speed of 0m/s

7.4 Summary

WMNs provide Internet services to wireless users and have some similarities to wireless Ad Hoc networks. Some nodes in WMNs relay packets for other users, in the case where a direct radio link to the access points cannot be found. Because in WMNs most traffic comes from and goes to access points, simply adapting protocols from wireless Ad Hoc networks to WMNs is not enough. The typical MAC layer protocols give all nodes the same level of priority to access the medium, while access points and nearby nodes in WMNs may need more access to the medium, especially when the traffic load is heavy. Depending on traffic patterns and mobility patterns, traffic may concentrate temporarily at one access point. The access point is then overloaded, packets are dropped, and long delays are experienced. Simulation results indicate that Traffic Balancing solves this temporary overloaded access point problem effectively. However, in order to implement

Traffic Balancing in WMNs, the position of the access points is crucial. Usually, the center of each subnetwork is the best location to place the access points.

Chapter 8. Conclusion and Future Work

In this work, an enhanced routing protocol called Traffic Balancing was proposed to improve system performance in wireless Ad Hoc networks. With knowledge of the load on the medium along all paths, Traffic Balancing exploits unused network resources and routes packets through the appropriate paths. In the proposed Adaptive Traffic Balancing, nodes are able to change the medium usage threshold intelligently by checking the collision rate in the area, so that the sender collects more accurate network information and chooses a better path. The simulation results show that both the data packet loss rate and the average end-to-end delay can be decreased by over 50% during congestion. With the benefits brought by Traffic Balancing, more connections could be supported with no deterioration in quality of service.

We started this research by defining the capacity of the network. With respect to the interference range of each transmission node and the size of network, the capacity of the network is estimated by a simple geometrical solution. After comparing the estimated system capacity and the system utilization, measured through simulations, we found that the system utilization is far below the system capacity when nodes are in movement. Some of the system resources are wasted by the backoff time after the nodes detect that the medium is busy, or a collision is detected. Some of them are never used due to the decisions of the routing protocols. Most reactive routing protocols choose the shortest path as the default path, and this causes traffic to concentrate in the middle area of a network. The resources at the edge of network are then never fully utilized.

As a result of realizing this major shortcoming with existing routing protocols, Traffic Balancing was developed to deviate part of the traffic load away from the heavy-loaded or central area of the network, if possible. For this purpose, each node continuously measures the medium usage around itself. The measured medium usage includes not only the traffic load going through the node and its neighboring nodes, but also all other activities in the network, because all activities in the network impact on the accessibility of the medium. When a route request arrives, the node appends information about whether or not the traffic load in its area is heavy to the request. The sender then chooses the best path from all the possible paths. The best path has the least number of heavily-loaded intermediate nodes, so that new traffic will use the under-utilized system resources. Moving traffic from the congested areas to the non-congested areas also decreases the possibilities of backoff in the congested areas, and bandwidth wasted by backoff periods can thus be recovered.

As the busy state of an area changes under some conditions, such as node mobility, the state of medium usage also needs to change, in accordance with these conditions. Traffic Balancing is extended to enable nodes to verify the medium state, depending both on usage measured and the number of collisions detected in the area. The concept of adaptation to congestion states in Traffic Balancing improves the system performance to a great extent. If a more precise method of determining the medium state is available, senders will be able to obtain more information about the network, so that Traffic Balancing can make better decisions.

As an extension of wireless Ad Hoc networks, wireless mesh networks will become more realistic and useful for end-users in the future. The current reality is that the implementation of WMNs needs to solve the problems related to MAC and routing protocols. A severe uneven traffic load problem to access points exists, causing poor utilization. Traffic Balancing provides an efficient way of solving this specific problem by forcing traffic from busy access points to idle ones.

The simulation results show that Traffic Balancing improves system performance with respect to packet delivery rate and average delay. However, its maximum potential is still under investigation. As shown in Section 5.4, Adaptive Traffic Balancing only allows three levels to be adjusted, and the selected threshold may not precisely represent the current state of medium. In addition, the duration of the measurement period affects the estimation of system utilization; its value can change dynamically according to packet size, inter-arrival time between successful data packets, and other information. Setting a reasonable value for this period offers a good estimation of the system state, so that Traffic Balancing can make better routing decisions. Unfortunately, there is no simple way of realizing this optimization. As discussed in Section 5.5, in the cases of Poisson and on/off traffic models, simulation results indicate that a cached route may be deleted and a new route discovery may start due to the long interval between consecutive data packets. Therefore, traffic is balanced upon the updated network state (topology and traffic distribution) and the problem of early start-up connections holding onto short paths is eliminated. For applications that generate data packets frequently, if the network is stationary, forcing route rediscovery after a certain period can balance traffic in a more efficient manner. However, there exists a tradeoff between the overhead generated by route rediscovery and the interval between route rediscoveries. The shorter the interval is, the better the balance of the traffic is. In this case, more bandwidth is consumed by routing information.

If more medium usage information or other related information is placed into the routing requests, thereby providing a more accurate depiction of the network state, Traffic Balancing will make better routing decisions. More research is required to ascertain what information can be used or derived from *a priori* information, and how to use this information. Although Traffic Balancing has been developed on the basis of reactive routing protocols, it may also be implemented for proactive routing protocols. For example, in the Optimized Link State Routing (OLSR) protocol [CJ03], nodes can put their current medium usage into the broadcast HELLO message, and eventually all other nodes can reach this information. When a node needs to find a path to another node, it could use the information to make a choice that has a lower number of overloaded intermediate nodes. However, as it takes some time for medium usage information to propagate to all the nodes in the network, the information received by a faraway node may become obsolete. Further modifications are required in this case to make Traffic Balancing more efficient.

Some other techniques can be adopted to increase the benefits of Traffic Balancing. For example, splitting bandwidth into separate channels for different purposes may bring more diversity to the network, thus giving Traffic Balancing more of a chance to find better communication paths. For QoS requirements, Traffic Balancing can match different service classes against different thresholds to meet their requirements.

A very simple physical layer model was adopted during the simulations. In order to reflect the real wireless world, channel models need to be revised to cope with fading and shadowing. Moreover, various applications imply a more complex traffic model. Formal evaluations and usability studies are yet to be conducted for studying the effectiveness of Traffic Balancing in more complex traffic load conditions. The impact of advanced wireless communication techniques, such as directional antenna and power control, need to be considered in future work, to evaluate the possibility of Traffic Balancing combining with them and the resulting level of system performance.

References

[AKKD01] S. Agarwal, S. V. Krishnamurthy, R. H. Katz, S. K. Dao, “Distributed Power Control in Ad-hoc Wireless Networks,” Proceedings of the 12th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications, Vol. 2, 30 Sept.-3 Oct. 2001, pp. F-59–F-66.

[BJM99] J. Broch, D. Johnson, and D. Maltz, “The Dynamic Source Routing Protocol for Mobile Ad Hoc Networks,” http://www.ietf.org/internet-drafts/draft-ietf-manet-dsr-10.txt, IETF Internet draft, Oct. 1999.

[BPSUHP03] S. Bandyopadhyay, M. N. Pal, D. Saha, T. Ueda, K. Hasuike, R. Pal, “Improving System Performance of Ad Hoc Wireless Network with Directional Antenna,” Proceedings of IEEE International Conference on Communications ICC 2003, Vol. 2, May 2003, pp. 1146–1150.

[BRS03] C. Bettstetter, G. Resta, P. Santi, “The Node Distribution of the Random Waypoint Mobility Model for Wireless Ad Hoc Network,” IEEE Transactions on Mobile Computing, Vol. 2, No. 3, July—Sept. 2003, pp. 257–269.

[BSMCGK02] S. Bansal, R. Shorey, A. Misra, S. Chugh, A. Goel, K. Kumar, “The Capacity of Multi-Hop Wireless Networks with TCP Regulated Traffic,” Proceedings of IEEE Global Telecommunications Conference GLOBECOM 2002, Vol. 21, No. 1, Nov. 2002, pp. 133–137.

[BSTW95] J. Beran, R. Sherman, M. S. Taqqu, W. Willinger, “Long-range dependence in variable-bit-rate video traffic,” IEEE Trans. Commun., Vol. 43, 1995, pp. 1565–1579.

[C84] D. R. Cox, “Long-Range Dependence: A Review,” In H. A. David and H. T. David, editors, *Statistics: An Appraisal*, Ames, Iowa, 1984, pp. 55–74.

[CABM03] D. S. J. De Couto, D. Aguayo, J. Bicket, R. Morris, “A High-Throughput Path Metric for Multi-Hop Wireless Routing,” Proceedings of the Ninth Association for Computing Machinery (ACM) Annual International Conference on Mobile Computing and Networking MobiCom’03, Sept. 14–19, 2003, pp. 227-240.

[CB97] M. Crovella, A. Bestavros, “Self similarity in world wide web traffic: Evidence and possible causes,” IEEE/ACM Trans. Networking, Vol. 5, 1997, pp. 835–846.

[CCG00] F. Cali, M. Conti, E. Gregori, "Dynamic Tuning of the IEEE 802.11 Protocol to Achieve a Theoretical Throughput Limit," IEEE/ACM Transaction on Networking, Vol. 8, No. 6, Dec. 2000, pp. 785–799.

[CF96] S. Civanlar, L. D. Fossett, "Survivable International Networks: Architecture and Dimensioning", Proceedings of the IEEE Conference on Communication ICC 93, Vol. 1, May 23-26, 1993, pp. 261–266.

[CJ03] T. Clausen, P. Jacquet, "Optimized Link State Routing Protocol (OLSR)," http://www.ietf.org/rfc/rfc3626.txt, IETF Internet Drafts, October 2003.

[CM99] M. S. Corson and J. Macker, "Mobile Ad Hoc Networking (MANET): Routing Protocol Performance Issues and Evaluation Considerations," ftp://ftp.isi.edu/in-notes/rfc2501.txt, Jan. 1999.

[CM01] S. Chakrabarti, A. Mishra, "QoS Issues in Ad Hoc Wireless Networks," *IEEE Communications Magazine*, Feb. 2001, pp. 142–148.

[CN99] S. Chen, K. Nahrstedt, "Distributed Quality-of-Service Routing in Ad Hoc Networks," IEEE Journal on Selected Areas in Communications, Vol. 17, No. 8, Aug. 1999, pp. 1488–1505.

[CP03] C. Comaniciu, H. V. Poor, "QoS Provisioning for Wireless Ad Hoc Data Networks," Proceedings of the 42nd IEEE Conference on Decision and Control, Dec. 2003, pp. 92–97.

[CSC03] N. Choi, Y. Seok, Y. Choi, "Multi-Channel MAC Protocol for Mobile Ad Hoc Networks," Proceedings of Vehicular Technology Conference 2003, VTC 2003 Fall, Oct. 6–9 2003, pp. 1379–1382.

[EE02] T. Elbatt, A. Ephremides, "Joint Scheduling and Power Control for Wireless Ad-hoc Networks," Proceedings of Twenty-First Annual Joint Conference of the IEEE Computer and Communications Societies, Vol. 2, 23–27 June 2002, pp. 976–984.

[FTK02] N. S. Fahmy, T. D. Todd, and V. Kezys, "Distributed Power Control for Ad Hoc Networks with Smart Antennas," Proceedings of Vehicular Technology Conference, Fall 2002, Vol. 4, pp. 2141–2144.

[GK00] P. Gupta, P. R. Kumar, "The Capacity of Wireless Networks," IEEE Transactions on Information Theory, Vol. 46, No. 2, March 2000, pp. 388–404.

[GN89] B. Gavish, I. Neuman, "A System for Routing and Capacity Assignment in Computer Communication Networks," IEEE Trans. Commun., Vol. 37, Apr. 1989, pp. 360–366.

[GT02] M. Grossglauser, D. N. C. Tse, "Mobility Increases the Capacity of Ad Hoc Wireless Networks," IEEE/ACM Transactions on Networking, Vol. 10, No. 4, August 2002, pp. 477–486.

[GV02] M. Gastpar, M. Vetterli, "On the Capacity of Wireless Networks: The Relay Case," Proceedings of IEEE, the Conference on Computer Communications INFOCOM 2002, pp. 1577–1586.

[GW90] A. Gersht, R. Weihmayer, "Joint Optimization of Data Network Design and Facility Selection," IEEE Journal on Selected Areas in Communications, Vol. 8, No. 9, Dec. 1990, pp. 1667–1681.

[GZ03] E. W. Grundke, A. Nur Zincir-Heywood, "A Uniform Continum Model for Scaling of Ad Hoc Networks," Proceedings of ADHOC-NOW 2003, pp. 96–103.

[HS02] Z. Huang, C. Shen, "A Comparison Study of Omnidirectional and Directional MAC Protocols for Ad hoc Networks," Proceedings of Global Telecommunications Conference, 2002, Globecom'02, IEEE Vol. 1, 17-21 Nov. 2002, pp. 57–61.

[HZ01] H. S. Hassanein, A. Zhou, "Routing with Load Balancing in Wireless Ad hoc networks," Proceedings of ACM (Association for Computing Machinery) Workshop on Modeling, Analysis and Simulation of Wireless and Mobile Systems, July 2001, pp. 89–96.

[I01] Mohammed Ismail, "Routing Protocols for Ad Hoc Wireless Networks," M.Sc. Project, Carleton University, Ottawa, Ontario, Canada, August 2001, http://kunz-pc.sce.carleton.ca/Thesis/IsmailEssay.pdf.

[J91] R. Jain, *The Art of Computer Systems Performance Analysis: Techniques for Experimental Design, Measurement, Simulation, and Modeling*, Wiley, ISBN: 0-471-50336-3, 1991, pp. 203–219.

[JBAS03] A. Jardosh, E. M. Belding-Royer, K. C. Almeroth, S. Suri, "Towards Realistic Mobility Models for Mobile Ad Hoc Networks," Proceedings of the Ninth Association for Computing Machinery (ACM) Annual International Conference on Mobile Computing and Networking MobiCom'03, Sept. 14–19 2003, pp. 217–229.

[JKW05] E. P. C. Jones, M. Karsten, P. A. S. Ward, "Multipath Load Balancing in Multi-hop Wireless Networks," Proceedings of WIMOB 2005—IEEE International Conference on Wireless and Mobile Computing, Networking and Communicating, Aug. 2005, pp. 158–167.

[JL98] P. Jiyapanichkul, J. Lambert, "Optimal Resource Management with Dynamic Virtual Paths in ATM Networks," Proceedings of ICC1998—IEEE International Conference on Communication, No. 1, June 1998, pp. 1728–1732.

[JPPQ03] K. Jain, J. Padhye, V. Padmanabhan, L. Qiu, "Impact of Interference on Multi-hop Wireless Network Performance," Proceedings of the 9th Annual International Conference on Mobile Computing and Networking, 2003, pp. 66–80.

[JS03] J. Jun, M. L. Sichitiu, "The Nominal Capacity of Wireless Mesh Networks," IEEE Wireless Communications, October 2003, pp. 8–14.

[KCC05] S. Kurkowski, T. Camp, M. Colagrosso, "MANET Simulation Studies: The Incredibles", Mobile Computing and Communications Reviews, Volume 9, Number 4, 2005, pp. 50-61.

[KJT03] T. Korakis, G. Jakllari, L. Tassiulas, "A MAC protocol for full exploitation of Directional Antennas in Ad-hoc Wireless Networks," Proceedings of the Fourth Association for Computing Machinery (ACM) International Symposium on Mobile and Ad Hoc Networking and Computing MobiHoc'03, June 1–3, 2003, pp. 98–107.

[LBCLM01] J. Li, C. Blake, D. S. J. De Couto, H. I. Lee, R. Morris, "Capacity of Ad Hoc Wireless Networks," Proceedings of Association for Computing Machinery (ACM) Special Interest Group on Mobility of System User, Data, and Computing (SIGMOBILE) 7/01, Rome, Italy, pp. 61–69.

[LG01] S. J. Lee, M. Gerla, "Dynamic Load-Aware Routing in Ad Hoc Networks," Proceedings of IEEE International Conference on Communications, 2001 (ICC 2001), Vol. 10, pp. 3206–3210.

[LHS02] J. Li, Z. J. Haas, M. Sheng, “Capacity Evaluation of Multi-Channel Multi-Hop Ad Hoc Networks,” Proceedings of IEEE International Conference on Personal Wireless Communications ICPWC 2002, pp. 211–214.

[LKL03] X. Lin, Y. Kwok, V. K. N. Lau, “A new Power Control Approach for IEEE 802.11 Ad Hoc Networks,” Proceedings of the 14th IEEE 2003 International Symposium on Personal, Indoor and Mobile Radio Communication, pp. 1761–1765.

[LL99] C. R. Lin, J. S. Liu, “QoS Routing in ad hoc wireless networks,” IEEE Journal on Selected Areas in Communications, Vol. 17, Aug. 1999, pp. 1454–1465.

[M00] R. Maatta, Wireless Ad Hoc Routing Protocols, a Taxonomy, Defence Forces Research Institute of Technology, Electronics and Information Technology Section, May 11, 2000.

[MBH00] J. P. Monks, V. Bharghavan, W. W. Hwu, “Transmission Power Control for Multiple Access Wireless Packet Networks,” Proceedings of the 25th Annual IEEE Conference on Local Computer Networks, November 2000, pp. 12–21.

[MK04] A. Muqattash, M. Krunz, “A Distributed Transmission Power Control Protocol for Mobile Ad Hoc Networks,” IEEE Transactions on Mobile Computing, Vol. 3, No. 2, April–June 2004, pp. 113–128

[MS87] K. H. Muralidhar, M. K. Sundareshan, “Combined Routing and Flow Control in Computer Communication Networks: A Two-Level Adaptive Scheme,” IEEE Trans. Automatic Control, Vol. AC–32, No. 1, Jan. 1987, pp. 15–25.

[NTV02] G. Nemeth, Z. R. Turanyi, A. Valko, “Throughput of Ideally Wireless Ad Hoc Networks,” Mobile Computing and Communications Review, Volume 5, Number 4, 2002, pp. 40–46.

[PH99] M. R. Pearlman, Z. J. Haas, “Determining the Optimal Configuration of the Zone Routing Protocol,” IEEE Journal of Selected Area on Communication, Vol. 17, Num. 6, Aug. 1999, pp. 101-117.

[PH02] D. Perkins, H. Hughes, “A survey of quality-of-service support for mobile ad hoc networks,” Wireless Commun. Mobile Comput., vol. 2, no. 5, Sept. 11, 2002, pp. 503–513.

[PHST00] M. R. Pearlman, Z. J. Haas, P. Sholander, S. S. Tabrizi, “On the Impact of Alternate Path Routing for Load Balancing in Mobile Ad Hoc Networks,” Proceedings of 2000 First Annual Workshop on Mobile and Ad Hoc Networking and Computing, pp. 3–10.

[PP03] A. Pandya, G. Pottie, “QoS in Ad Hoc Networks,” Proceedings of IEEE Vehicular Technology Conference, VTC Fall 03, Vol. 5, Oct. 2003, pp. 3089–3093.

[PS02] S. Park, R. Sivakumar, “Load-Sensitive Transmission Power Control in Wireless Ad-Hoc Networks,” Proceedings of IEEE Global Telecommunications Conference GLOBECOM 2002, Vol. 21, No. 1, Nov. 2002, pp. 42–46.

[PS03] C. Peraki, S. D. Servetto, “On the Maximum Stable Throughput Problem in Random Networks with Directional Antennas,” Proceedings of the Fourth Association for Computing Machinery (ACM) International Symposium on Mobile and Ad Hoc Networking and Computing MobiHoc ’03, June 1–3, 2003, pp. 76–87.

[RM02] V. Rodoplu, T. H. Meng, “Core Capacity of Wireless Ad Hoc Networks,” Proceedings of the 4th International Symposium on Wireless Personal Multimedia Communications, Vol. 1, Oct. 2002, pp. 247–251.

[RP03] M. Realp, A. I. Perez-Neira, “Decentralized Multiaccess MAC Protocol for Ad-Hoc Networks,” the 14th IEEE 2003 International Symposium on Personal, Indoor and Mobile Radio Communication Proceedings, pp. 1634–1638.

[RSBUT03] S. Roy, D. Saha, S. Bandyopadhyay, T. Ueda, S. Tanaka, “A Network-Aware MAC and Routing Protocol for Effective Load Balancing in Ad Hoc Wireless Networks with Directional Antenna,” Proceedings of the Fourth Association for Computing Machinery (ACM) International Symposium on Mobile and Ad Hoc Networking and Computing MobiHoc ’03, June 1–3, 2003, pp. 88–97.

[S94] W. Stallings, *Data and Computer Communications*, Fourth edition, Macmillan, ISBN: 0-02-415441-5, 1994, pp. 288–326.

[SK99] J. L. Sobrinho, A. S. Krishnakumar, “Quality-of-Service in Ad Hoc Carrier Sense Multiple Access Wireless Networks,” IEEE Journal on Selected Areas in Communications, Vol. 17, No. 8, Aug. 1999, pp. 1353–1368.

[SM04] G. Sharma, R. R. Mazumdar, “Scaling laws for capacity and delay in ad hoc wireless network with random mobility,” Proceedings of IEEE International Conference on Communications ICC 2004, June 2004, pp. 152–161.

[SR03] A. Spyropoulos, C. S. Raghavendra, “Capacity Bounds For Ad-Hoc Networks Using Directional Antennas,” Proceedings of IEEE International Conference on Communications ICC 2003, Vol. 1, May 2003, pp. 348–352.

[SV04] J. A. Stine, G. de Veciana, “A Paradigm for Quality-of-Service in Wireless Ad Hoc Networks Using Synchronous Signaling and Node States,” IEEE Journal on Selected Areas in Communications, Vol. 22, No. 7, Sept. 2004, pp. 1301–1321.

[T96] Andrew S. Tanenbaum, *Computer Networks,* third edition, Prentice Hall, ISBN: 0-13-066102-3**,** 1996, pp. 243–333.

[TFK02] X. Tao, D. Falconer, T. Kunz, “Traffic Balancing in Wireless Ad Hoc Networks: Extending System Capacity and Improving System Performance,” Proceedings of Wireless 2002, July 2002, pp. 418–423.

[TG03] S. Toumpis, A. J, Goldsmith, “Capacity Regions for Wireless Ad Hoc Networks,” *IEEE Transactions on Wireless Communication*s, Vol. 2, No. 4, July 2003, pp. 736–748.

[TKF04] X. Tao, T. Kunz, D. Falconer, “Implementing Traffic Balancing for Exploring System Capacity in Wireless Ad Hoc Networks,” Proceedings of World Wireless Research Forum (WWRF) 12, Nov. 2004.

[TKF05a] X. Tao, T. Kunz, D. Falconer, “Adaptive Traffic Balancing in Wireless Ad Hoc Networks,” Proceedings of World Wireless Research Forum (WWRF) 13, March 2005.

[TKF05b] X. Tao, T. Kunz, D. Falconer, “Throughput Maximizing Routing in a MANET: Protocols and Analysis,” Proceedings of the International Conference on Wireless Networks, Communications, and Mobile Computing (WirelessCom 2005), Maui, USA, June 2005.

[TKF05c] X. Tao, T. Kunz, D. Falconer, “Traffic Balancing in wireless MESH networks,” Proceedings of the International Conference on Wireless Networks, Communications, and Mobile Computing (WirelessCom 2005), Maui, USA, June 2005.

[W00] P. Whitehead, “Mesh Networks: a new Architecture for Broadband Wireless Access Systems,” Proceeding of IEEE Radio & Wireless Conference (RAWCON) 2000, pp. 43–46.

[WH01] K. Wu, J. Harms, “Load-Sensitive Routing for Mobile Ad Hoc Networks,” Proceedings of 26th Annual IEEE Conference on Local Computer Networks, 2001, pp. 568–575.

[XL04] Y. Xiao, H. Li, “Local Data Control and Admission Control for QoS Support in Wireless Ad Hoc Networks,” IEEE Transactions on Vehicular Technology, Vol. 53, No. 5, Sept. 2004, pp. 1558–1572.

[YKG01] Y. Yi, T. J. Kwon, M. Gerla, *“*A Load aWare Routing (LWR) based on Local Information*,*” Proceedings of 12th IEEE International Symposium on Personal, Indoor and Mobile Radio Communication, Vol. 2, 2001, pp. G-65–G-68.

[YKT03] Z. Ye, S. Krishnamurthy, S. Tripathi, “A framework for reliable routing in mobile ad hoc networks,” Proceedings of the IEEE Conference on Computer Communications INFOCOM 2003, pp. 78-89.

[YPK03] S. Yi, Y. Pei, S. Kalyanaraman, “On the Capacity Improvement of Ad Hoc Wireless Networks Using Directional Antennas,” Proceedings of the Fourth Association for Computing Machinery (ACM) International Symposium on Mobile and Ad Hoc Networking and Computing MobiHoc’03, June 1–3, 2003, pp. 108–116.

[YS91] J. R. Yee, F. Shiao, “An Algorithm to Find Global Optimal Routing Assignments for a Class of PRNS,” Proceedings of IEEE International Conference on Communication ICC’91, pp. 1604–1608.

[YSL04] C. Yu, K. G. Shin, B. Lee, “Power-Stepped Protocol: Enhancing Spatial Utilization in a Clustered Mobile Ad Hoc Network,” IEEE Journal on Selected Areas in Communications, Vol. 22, No. 7, Sept. 2004, pp. 1322–1334.

[ZH01] A. Zhou, H. Hassanein, “Load-Balanced Wireless Ad Hoc Routing,” Proceedings of Canadian Conference on Electrical and Computer Engineering 2001, Vol. 2, pp. 1157–1161.