HOW TO EFFICIENTLY UTILIZE MULTI-HOP WIRELESS AD HOC NETWORKS - THROUGHPUT IMPROVEMENT AND MOBILITY SUPPORT

By

RONGSHENG HUANG

A DISSERTATION PRESENTED TO THE GRADUATE SCHOOL OF THE UNIVERSITY OF FLORIDA IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

UNIVERSITY OF FLORIDA

2011
To my mother in heaven,
and to my wife and kids
ACKNOWLEDGMENTS

I owe thanks to a lot of people upon presenting this dissertation. I cannot accomplish this task without them. Among them, I need to first thank my advisor, Prof. Yuguang Michael Fang. All my research achievements are the outcome of his lab, our lab. Without the chances and advices he gave to me, I would not accomplish any of them. He showed me to the world of research. He guides me whenever I am lost. I am so lucky of being one member of his lab. I also want to thank my committee members, Prof. Liuqing Yang, Prof. Dapeng Wu, Prof. Janise McNair and Prof. Shigang Chen, for their advices and supports.

Great thanks to my fellow members in WINET lab. My research deeply relates to everyone’s effort. I would thank Dr. Hongqiang Zhai, Dr. Jianfeng Wang, Dr. Yanchao Zhang, Dr. Shushan Wen, Dr. Yun Zhou, Dr. Xiaoxia Huang for showing me how to do research. Thanks to Dr. Pan Li, Dr. Feng Chen, Dr. Yang Song, Dr. Jinyuan Sun for the great research interaction we had. The same thanks to all the current fellow WINET members. I want to express my special thanks to Chi Zhang for giving me advices when I struggled with my research. He helped me to build my own research methodology when I needed it most. I also want to thank Dr. Sunmyeng Kim and all the visiting members for their help.

Personally, I want to take this chance to thank my father and my sisters. Family is always the motivation of my diligence. Their supports have carried me through my tough PhD years and their love soothes my pain of losing mom.

At the 8th year of marriage, we are expecting our second baby. I guess I owe more than thanks to my wife, for keeping me away from loneliness, for bringing me this family and for accompanying me in this long long journey.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACKNOWLEDGMENTS</td>
<td>4</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td>9</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>10</td>
</tr>
<tr>
<td>ABSTRACT</td>
<td>12</td>
</tr>
<tr>
<td><strong>CHAPTER</strong></td>
<td></td>
</tr>
<tr>
<td>1 OVERVIEW OF RESEARCH TOPICS</td>
<td>15</td>
</tr>
<tr>
<td>1.1 Wireless Ad hoc Networks</td>
<td>15</td>
</tr>
<tr>
<td>1.2 Observations and Motivations</td>
<td>17</td>
</tr>
<tr>
<td>1.3 My Research</td>
<td>18</td>
</tr>
<tr>
<td>1.3.1 Throughput Improvement</td>
<td>18</td>
</tr>
<tr>
<td>1.3.2 Mobility Support</td>
<td>20</td>
</tr>
<tr>
<td>1.4 Outline</td>
<td>21</td>
</tr>
<tr>
<td>2 SAM-MAC: AN EFFICIENT CHANNEL ASSIGNMENT SCHEME FOR MULTIPLE CHANNEL AD HOC NETWORKS</td>
<td>23</td>
</tr>
<tr>
<td>2.1 Background</td>
<td>23</td>
</tr>
<tr>
<td>2.2 Related Works</td>
<td>24</td>
</tr>
<tr>
<td>2.3 Motivations and Basic Idea</td>
<td>27</td>
</tr>
<tr>
<td>2.3.1 Motivations</td>
<td>27</td>
</tr>
<tr>
<td>2.3.2 Basic Idea</td>
<td>29</td>
</tr>
<tr>
<td>2.4 Protocol Description of SAM-MAC</td>
<td>30</td>
</tr>
<tr>
<td>2.4.1 Basic Messages</td>
<td>30</td>
</tr>
<tr>
<td>2.4.2 Basic Operational Procedures</td>
<td>31</td>
</tr>
<tr>
<td>2.4.2.1 Choosing traffic channel (Initial channel selection)</td>
<td>31</td>
</tr>
<tr>
<td>2.4.2.2 Transmitting to a node listening on an unknown traffic channel</td>
<td>31</td>
</tr>
<tr>
<td>2.4.2.3 Transmitting to a node listening on a known traffic channel</td>
<td>32</td>
</tr>
<tr>
<td>2.4.2.4 Adjusting to a different traffic channel when busy/collision (Self-adjustment procedure)</td>
<td>33</td>
</tr>
<tr>
<td>2.4.2.5 Transmitting data on the common channel</td>
<td>34</td>
</tr>
<tr>
<td>2.4.3 Other Issues</td>
<td>35</td>
</tr>
<tr>
<td>2.5 Performance Analysis</td>
<td>36</td>
</tr>
<tr>
<td>2.5.1 Multi-channel Hidden Terminal Problem</td>
<td>36</td>
</tr>
<tr>
<td>2.5.2 Control Channel Saturation Problem</td>
<td>37</td>
</tr>
<tr>
<td>2.5.3 Missing Receiver Problem</td>
<td>37</td>
</tr>
<tr>
<td>2.5.4 Overhead Comparison and Throughput Improvement</td>
<td>38</td>
</tr>
</tbody>
</table>
3. EXPLOITING THE CAPACITY OF MULTI-CHANNEL MULTI-RADIO WIRELESS MESH NETWORKS

3.1 Background .................................................. 47
3.2 Related Works ................................................. 50
3.3 Problem Formulation ........................................... 52
   3.3.1 Portal Capacity .......................................... 52
   3.3.2 Problem Formulation .................................... 53
   3.3.3 Problem Simplification .................................. 55
3.4 Proposed Solution ............................................. 57
   3.4.1 Part I: Centralized Part ................................. 58
      3.4.1.1 Overview .......................................... 58
      3.4.1.2 Centralized resource allocation and channel assignment ........................................... 60
      3.4.1.3 Actual total resource for allocation ......................... 63
      3.4.1.4 Revision of weighted resource allocation ................. 66
   3.4.2 Part II: Distributed Part ................................. 67
      3.4.2.1 Overview .......................................... 67
      3.4.2.2 Distributed channel assignment ......................... 68
      3.4.2.3 Dynamic adjustment mechanism ........................ 71
      3.4.2.4 Infrastructure formation ............................... 72
   3.4.3 Part III: Support of Multi-gateway ........................ 73
3.5 Performance Evaluation ...................................... 73
3.6 Summary ..................................................... 76

4. UTILIZING MULTI-HOP NEIGHBOR INFORMATION IN SPECTRUM ALLOCATION FOR WIRELESS NETWORKS ........................................... 77

4.1 Background .................................................. 77
4.2 Related Works ................................................. 79
4.3 Spectrum Usage Issues in Multi-hop Wireless Networks .... 81
   4.3.1 Ideal Spectrum Usage .................................... 81
   4.3.2 Traffic Dependency ...................................... 83
   4.3.3 Allocation Inefficiency Ratio ............................. 84
   4.3.4 Asymmetric Neighborhood ................................ 85
4.4 Proposed Scheme ............................................. 85
   4.4.1 Overview of The Proposed Scheme ......................... 85
   4.4.2 Scheme Description ....................................... 88
      4.4.2.1 Data structure and message format ....................... 89
      4.4.2.2 Procedure ........................................ 91
4.5 Performance Evaluation ...................................... 94
4.6 Summary ..................................................... 98
5 A MOBILITY MANAGEMENT SCHEME FOR WIRELESS MESH NETWORKS

5.1 Background .......................................................... 99
5.2 Related Works ......................................................... 100
5.3 \( M^3 \) Description ...................................................... 102
   5.3.1 Model Description ............................................ 102
   5.3.2 Proposed Solution ............................................. 104
      5.3.2.1 Power-up ................................................ 104
      5.3.2.2 Handling downstream packets ......................... 105
      5.3.2.3 Handling upstream packets ............................ 105
      5.3.2.4 Handling handoff ........................................ 105
      5.3.2.5 Periodic location update ............................... 106
   5.3.3 Extended Discussion of The Protocol ...................... 107
5.4 Performance Analysis ............................................... 108
5.5 Summary ................................................................ 112

6 IMPROVING HANDOFF PERFORMANCE BY UTILIZING AD HOC LINKS IN MULTI-HOP CELLULAR SYSTEMS .................................................. 113

6.1 Background .......................................................... 113
6.2 Related Works ......................................................... 116
6.3 System Model and Basic Idea ..................................... 117
   6.3.1 System Model .................................................. 117
   6.3.2 Basic Idea ..................................................... 121
6.4 ANHOA: Ad-hoc-Network-Embedded Handoff Assisting Scheme .................................................. 122
   6.4.1 Overview ....................................................... 122
   6.4.2 Architecture and Roles ...................................... 123
   6.4.3 Information Maintenance of EAN ......................... 124
   6.4.4 Handoffs via ANHOA ......................................... 127
   6.4.5 Handoff Decision Algorithm ............................... 130
6.5 Resource Management for Multi-hop Handoff .................. 132
   6.5.1 Minimum Reservation ....................................... 132
   6.5.2 Load Sharing among Adjacent Cells ..................... 135
6.6 Performance Evaluation ............................................. 137
6.7 Summary ................................................................ 140
6.8 Appendix ................................................................ 140

7 RELIEVING THE LOCATION MANAGEMENT CONGESTION VIA AGGREGATION IN MULTI-HOP CELLULAR NETWORKS .................................................. 142

7.1 Background .......................................................... 142
7.2 Related Works ......................................................... 144
7.3 System Model ........................................................ 146
7.4 Grouping To Save The Signaling Cost .......................... 147
   7.4.1 Overview ....................................................... 147
   7.4.2 Procedures of Group Location Management ............. 148
## LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2-1 Simulation Parameters For Single-hop Topology</td>
<td>42</td>
</tr>
<tr>
<td>3-1 Algorithm of Centralized Resource Allocation and Channel Assignment</td>
<td>62</td>
</tr>
<tr>
<td>4-1 Local Parameter Set</td>
<td>89</td>
</tr>
<tr>
<td>4-2 Neighbor Parameter Set</td>
<td>90</td>
</tr>
<tr>
<td>5-1 Typical Value Set of Average Path Length</td>
<td>111</td>
</tr>
<tr>
<td>6-1 Portal Nodes’ Information Table</td>
<td>125</td>
</tr>
<tr>
<td>6-2 Attaching Nodes’ Information Table</td>
<td>126</td>
</tr>
<tr>
<td>6-3 Subordinate Nodes’ Information Table</td>
<td>126</td>
</tr>
<tr>
<td>6-4 Simulation Setting for Each Cell</td>
<td>138</td>
</tr>
<tr>
<td>7-1 Parameters’ Settings For Evaluating The Grouping Scheme</td>
<td>162</td>
</tr>
<tr>
<td>7-2 Parameters’ Settings For Evaluating The Generic Aggregation Scheme</td>
<td>163</td>
</tr>
</tbody>
</table>
**LIST OF FIGURES**

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2-1</td>
<td>Comparison of DCC Type and SP Type</td>
<td>27</td>
</tr>
<tr>
<td>2-2</td>
<td>Illustration of SAM-MAC Procedures: Unaware Case</td>
<td>32</td>
</tr>
<tr>
<td>2-3</td>
<td>Illustration of SAM-MAC Procedures: Aware Case</td>
<td>33</td>
</tr>
<tr>
<td>2-4</td>
<td>Illustration of SAM-MAC Procedures: Self-Adjustment Case</td>
<td>34</td>
</tr>
<tr>
<td>2-5</td>
<td>Multi-channel Hidden Terminal Problem</td>
<td>36</td>
</tr>
<tr>
<td>2-6</td>
<td>Throughput Gain in Single-hop Topology with 3+1 Channel, with and without</td>
<td>43</td>
</tr>
<tr>
<td></td>
<td>common channel transmitting data</td>
<td></td>
</tr>
<tr>
<td>2-7</td>
<td>Throughput Gain in Single-hop Topology with Multiple Channels</td>
<td>44</td>
</tr>
<tr>
<td>2-8</td>
<td>Throughput Gain in Multi-hop Topology with Multiple Channels</td>
<td>45</td>
</tr>
<tr>
<td>3-1</td>
<td>The Common Structure of a WMN</td>
<td>47</td>
</tr>
<tr>
<td>3-2</td>
<td>An Example of Link Distribution in WMNs and the Proposed Frequency Reuse</td>
<td>54</td>
</tr>
<tr>
<td></td>
<td>Pattern</td>
<td></td>
</tr>
<tr>
<td>3-3</td>
<td>Ideal Case When Frequency Reuse is Perfect</td>
<td>59</td>
</tr>
<tr>
<td>3-4</td>
<td>Illustration of Channel Busyness Indicator Phase (CBIP)</td>
<td>70</td>
</tr>
<tr>
<td>3-5</td>
<td>Comparison of Portal Capacity among 3 Schemes: CW (our solution), SC(single</td>
<td>74</td>
</tr>
<tr>
<td></td>
<td>channel), RM (randomly-choosing multi-channel)</td>
<td></td>
</tr>
<tr>
<td>3-6</td>
<td>Portal Capacity with Frequency Reuse</td>
<td>76</td>
</tr>
<tr>
<td>4-1</td>
<td>Inefficiency of Fair Random Access</td>
<td>81</td>
</tr>
<tr>
<td>4-2</td>
<td>Flow Chart for Message Processing</td>
<td>92</td>
</tr>
<tr>
<td>4-3</td>
<td>Flow Chart for Spectrum Allocation</td>
<td>93</td>
</tr>
<tr>
<td>4-4</td>
<td>The topologies for evaluation</td>
<td>95</td>
</tr>
<tr>
<td>4-5</td>
<td>Comparison of end-to-end throughput (a), fairness index (b) and AIR (c) for</td>
<td>96</td>
</tr>
<tr>
<td></td>
<td>RFA, DSS and 2hopMAC schemes</td>
<td></td>
</tr>
<tr>
<td>5-1</td>
<td>Mesh Model and Illustration of $M^3$ Scheme</td>
<td>104</td>
</tr>
<tr>
<td>5-2</td>
<td>Gain of $M^3$ over previous schemes</td>
<td>111</td>
</tr>
<tr>
<td>6-1</td>
<td>Model of Multi-hop Handoffs</td>
<td>118</td>
</tr>
<tr>
<td>6-2</td>
<td>Embedded Ad hoc Network in Cellular System</td>
<td>123</td>
</tr>
</tbody>
</table>
6-3 EAN-involved Handoffs .................................................. 128
6-4 HOD rate ................................................................. 138
6-5 Channel Reservation ...................................................... 139
6-6 Markov Chain Model for Handoff Reservation System ................. 140
7-1 Location Management Preliminary ..................................... 146
7-2 Group Association Procedure .......................................... 151
7-3 Generic Aggregative Location Management .......................... 157
7-4 Grouping Gains of Signaling Cost With Different Group Sizes (K) .... 162
7-5 Grouping Gains of Signaling Cost With Different Trip Length (\bar{L}) .... 163
7-6 The Aggregation Interval's Ranges with Different LU Traffic .......... 164
7-7 The Aggregation Signaling Gains with Different LU Traffic .......... 165
Due to its low-cost, flexible in construction and self-organizable features, wireless ad hoc networks (WANETs) have provided many applications in a lot of areas. However, the multi-hop transmissions and lack of infrastructure bring the low throughput performance and poor mobility support. Aiming at efficiently utilizing WANETs, I strive to tackle these two problems in the recent years.

Throughput performance in WANETs mainly relates to the spectrum wastage caused by improper MAC design. For multi-channel MAC schemes, due to the limited number radios, the spectrum wastage is mostly created by the channel assignment signalings. Observing that channel assignment is not necessary before each transmission, we can let nodes cache the previous channel assignment and lift the signalling burden of the control channels. We proposed this self-adjustable multi-channel MAC scheme (SAM_MAC) to increase the MAC throughput in multi-channel systems. Only high MAC throughput cannot guarantee high end-to-end throughput in multi-hop wireless networks because if the area with high traffic load cannot acquire more spectrum congestions will appear and lead to poor overall throughput performance. We proposed a semi-centralized semi-distributed resource allocation scheme for wireless mesh networks, with which central areas can easily acquire more spectrum and the outer areas can easily reuse the spectrum. Within one neighborhood, nodes with high traffic load should acquire more spectrum. Previous
distributed scheduling schemes which aim at solving this problem are too complicated in algorithm and cannot suit multi-hop, contention-based networks. Moreover, they ignore the fact that the traffic demands of neighboring nodes are correlated with each other. We proposed “2-hop MAC” scheme for WANETs to allocate spectrum resource among neighboring nodes in a more intelligent way. Spectrum allocated previously will not be wasted by packet droppings in the following hops. Using the collected traffic dependency information, each node can allocate resource accordingly so that a locally optimized end-to-end throughput can be achieved. In this scheme, although the detailed flow information is not necessary for MAC layer since only the traffic in the neighborhood is considered, a cross-layer information exchange is required.

Mobility issues are not well addressed in WANETs due to its lack of infrastructures, compared to cellular systems. By adding some functional entities or modules to the system, we can support mobility in WANETs similarly. Mobile IP can be used as the solution for macro-mobility (inter-domain roaming). For micro-mobility, we proposed the mesh mobility management ($M^3$) scheme. This scheme stores the locations (serving mesh routers) of mesh nodes in the gateway as the anchors and use temporary routing entries between the anchors and the exact locations to reach the destinations. This approach combines the advantages of tunnel-based approach and routing-based approach and provides a suitable micro-mobility solution for wireless mesh networks. When integrating WANETs into cellular systems, the advantages of WANETs and cellular systems can potentially be both achieved. This type of networks, known as multi-hop cellular networks (MCNs), can potentially have better mobility support. For handoff performance, we propose an ad hoc networks-embedded handoff assisting (ANHOA) scheme, which utilizes the embedded ad hoc networks in MCNs to exchange information of different handoff options and assists multi-hop handoffs. The handoff reservation in each BS can be lowered by considering the handoff attempts to different cells. We proposed the information exchange among neighboring BSs and the algorithm
of finding the minimum handoff reservation. For location management, we proposed to utilize aggregative devices to aggregate location management signallings. For the high capacity transit (HCT) systems, we specially proposed a grouping method to reduce the location management signallings. Through these methods, the congestions caused by heavy location request arrivals can be greatly relieved.

The performance results have verified the contributions of these works.
CHAPTER 1
OVERVIEW OF RESEARCH TOPICS

1.1 Wireless Ad hoc Networks

Wireless networks usually refer to any type of communication networks which employ wireless communication techniques. There are two primary reasons that we want to introduce wireless networks to our lives. One is the wireless characteristic and the accompanied low cost, convenience and flexibility. The other is the mobility support which allows us to be connected as long as we are in the coverage of wireless signals.

Wireless ad hoc networks (WANETs) are one special type of wireless networks, which can be defined as decentralized wireless networks where there is no preexisting infrastructure such as BSs in cellular systems, or APs in WLANs. Unlike the infrastructure-based wireless networks, the nodes in WANETs are capable to relay the packets from the upstream nodes to the downstream nodes. Therefore, multi-hop transmission should be seen as one of the most important features of WANETs. This feature can create a lot of useful applications when the telecommunication infrastructures are not available. For example, in a disaster rescue, when all the infrastructures are destroyed, with the well-functioning hand-held devices, communications can still take place if each terminal can receive transmissions from others and relay them. These networks therefore seem very attractive for various military and civil purposes.

WANETs have been existing for over two decades. Mostly we see them in 3 major categories:

- Mobile Ad hoc Networks (MANETs)
- Wireless Sensor Networks
- Wireless Mesh Networks

In all three types, only MANETs are a pure ad hoc type of networks. Both wireless sensor networks and wireless mesh networks assume the traffic aggregate to a center point, i.e., sink in sensor networks and gateway in mesh networks. Among these three
types of networks, MANETs focus on mobility and connectivity; sensor networks focus on power consumption; mesh networks focus on throughput performance.

WANETs can be applied in many scenarios due to their low-cost, self-organizing, and flexible properties. However, there are a lot of issues in WANETs that need to be solved, such as the low throughput, routing, and mobility support. After all, this is a totally different type of networks. In order to utilize them, it is necessary to understand them first. Moreover, besides the aforementioned 3 major forms, WANETs have applications in other forms, such as vehicular ad hoc networks, cognitive ad hoc network, and delay tolerant networks. In fact, WANETs are potentially be capable to integrate with other types of networks in order to achieve certain applications. This variety creates more interesting research topics. Current research topics in WANETs include the following areas.

- Modeling of wireless networks and theoretical capacity bound
- PHY techniques
- Networking protocol layers, including MAC, routing, network coding, and cross-layer design
- Resource management and scheduling, including traffic-related topics, energy consumption, channel assignment and dynamic spectrum usage
- Mobility management
- Security and incentive schemes
- Issues in different types of wireless networks, including delay tolerant networks (DTNs), vehicular networks, cognitive radio ad hoc networks, and other hybrid networks

Solving issues and understanding problems in the existing forms of WANETs help us to utilize them effectively and efficiently. Thinking of the prospective benefits that offered by new forms of WANETs can also expand new ways of utilizing them. All the existing forms of WANETs were once new and now they flourish. However, before we go
to the variety of forms, we ought to take care of existing tough problems. Focusing on
the existing problems can not only help us improve the current systems' performance but
also enhance our understanding of the essence of WANETs.

1.2 Observations and Motivations

All types of WANETs suffer from low throughput performance because neighboring
nodes share the same spectral resource and interfere with each other. According
to the result of [22], with the scale of the networks increasing, due to the growing
interference and the number of intermediate hops of flows, the per node end-to-end
throughput performance will be decreasing to 0. This result depicts the trend of the
throughput performance with increasing network scale. However, for practical systems,
this should not be the excuse of the low throughput performance. The low spectral
efficiency of current protocols is instead the main reason of the currently low end-to-end
performance. When the throughput performance is satisfactory, a wireless network
should provide mobility support to its clients, which is one of the most important reasons
that attract people to utilize wireless networks. However, due to lack of infrastructure,
it is difficult for WANETs to support mobility. To employ WANETs, these are the two
primary problems requiring to be solved urgently.

As of my researches, I focus on the protocol design as well as the system
architecture in order to give solutions to these two problems. The current layered
approach of protocol design has its advantages in software engineering and standardization.
However, since every node in wireless medium interferes with each other in the mean
time of transmitting/receiving and the wireless channel conditions fluctuate all the time,
this approach cannot achieve a good performance in many scenarios. Incorporating
more information into the protocol entities can potentially decrease the spectrum
wastage and improve the throughput performance. Although end-to-end throughput is
directly controlled by transport layer, the spectral efficiency is decided by the design of
MAC layer, assuming that PHY techniques are given.
Unlike the protocol design for improving throughput performance, which focus on
the protocol layers or cross-layer designs, protocol design for mobility management of
WANETs requires more systematic concerns not only because currently there is no
standard existing but also because it may involve more network elements, such as the
counterparts of VLRs, HLRs in cellular systems.

There are a lot of pending issues in WANETs, such as routing, transport layer, and
security. However, I cannot stretch my research to all the areas although I wish I have
enough energy to do so. I select these two themes due to two reasons. First, these
two problems are the most crucial problems for the practicality of WANETs. Although
PHY techniques directly drive the data rate improvement, since it is less related to
networking areas, I chose MAC layer as my major focus. Second, there are two methods
to understand a system, micro view and macro view. Holding only one will lead to
incomplete and even distorted understanding. I select MAC layer as the micro view and
select mobility management as the macro view to WANETs. Indeed, I will not limit my
view of WANETs within these two problems. Employing WANETs is a long term work
and it includes more than just throughput and mobility.

1.3 My Research

1.3.1 Throughput Improvement

My efforts on throughput improvement for WANETs are mostly done in the MAC
layer, or combined with a cross-layer approach. The first effort is to improve the
throughput performance by introducing multiple channels. IEEE802.11 standards
specify multiple non-overlapping channels which are the foundation of this type of
researches. We proposed a self-adjustable multi-channel MAC (SAM-MAC) to efficiently
utilize the spectrum. We base our design on the observation that the overhead of the
multi-channel MAC mainly comes from the signaling of channel assignment. When the
number of radios of each node is limited, the channel assignment is critical in that the
unpredictable channel switch without effective channel usage information can easily
cause transmission collisions. We let each node memorize its neighbors’ channel usage so that channel assignment is not necessary for every transmission.

Redesigning MAC layer with only one-hop neighbors’ information cannot effectively boost the systems’ end-to-end throughput because allocating the spectral resource only according to MAC layer information cannot let heavy-traffic areas have more resource than light-traffic area, especially in the case when there are multiple channels to be assigned. Moreover, spatial reuse is critical to the throughput performance in multi-hop WANETs and it is difficult to design because the centralized control is either unavailable or hard to achieve. We proposed a spatial reuse paradigm for multi-channel multi-radio wireless mesh networks. To best efficiently utilize the available spectrum, the mesh networks are divided into the central part and the outside part. The central part uses a centralized algorithm to find the optimum spectrum allocation. The mesh routers outside the center can distributedly reuse the resource in the central part by sensing the spectrum availability. In this paper, we also proposed a concept named as portal capacity which indicates the capability a mesh network can provide with the assumption of a given traffic distribution. This capacity is of great help in the network constructions.

In WANETs, different areas may have different traffic load, as mentioned in my previous work. In neighborhood areas, different nodes may have different traffic demand. Moreover, the traffic of different nodes may have correlations due to multi-hop flows traversing across. Without these considerations, spectrum allocated can be easily wasted either in current hop or in following hops. In another work of mine, I investigate the spectral inefficiency of contention-based MAC for WANETs. Although scheduling-based standard is available to provide higher MAC efficiency, e.g. IEEE 802.16, it is not suitable for an unplanned ad hoc network. For a self-configured ad hoc networks, do we have to live with the low spectrum efficiency? We proposed to let each node in ad hoc networks allocate spectrum intelligently, instead of contending the channel blindly as the current IEEE 802.11. Furthermore, we let the neighboring nodes
exchange the traffic dependency information in a cross-layer manner. With the collected information from its one-hop upper stream nodes and its one-hop down stream nodes, which is the reason for the protocol’s name “two-hop MAC”, each node can allocate the spectrum usage with a locally optimized end-to-end throughput.

In order to adapt to different wireless networks, the customized MAC is highly needed and the cross-layer information is probably needed. I believe there is still a long way for us to fully understand the spectrum efficiency of wireless networks.

1.3.2 Mobility Support

Mobility management can partially be seen as the routing in mobile networks. It is a key topic in cellular systems which has been well-researched in the past decade, in a pure hierarchical structure. With the concern of mobility management, moving clients can experience the service without interruptions. However, in other types of wireless networks, this issue has not been well addressed yet.

In IP networks, the mobility is usually addressed by mobile IP protocol. However, mobile IP is only suitable for the inter-domain roaming, i.e., macro mobility. Changing IP address in a small area is undesired for a continuous service. If the same IP address is used in one domain, locating the moving node will consequently be an issue. We proposed to combine the tunnel-based approach and the routing-based approach to deal with the micro-mobility in wireless mesh networks. The mesh gateway stores the recent location of each node and tunnel the downlink packets to the corresponding mesh router. Within one period, the temporary routing entries in each router that the moving node passes help the intermediate router to forward the downlink packet from the stored location to the current location. Periodical update will remove these temporary routing entries and control the routing delay.

I am currently delving into the issues in multi-hop cellular networks (MCNs), such as handoff performance and location management scheme. MCNs are a new emerged type of networks. Introducing ad hoc mode into cellular networks can potentially let us
benefit from combining the advantage of different wireless networks and have a versatile platform which can connect different networks together.

Assuming handoff nodes can access adjacent cells via multi-hop connections, by letting neighboring Base Stations (BSs) exchange traffic information, we can reserve less resource for the handoff calls to meet the required handoff dropping rate. We can also let the stationary/semi-stationary nodes form small-scaled ad hoc networks and let them assist the multi-hop handoff. In this way, the spectrum in each cell can be more efficiently utilized and the handoff dropping rate can be greatly decreased. In my first MCN paper, we proposed ANHOA (ad-hoc networks embedded handoff assisting) scheme, which utilize the embedded ad hoc networks in cellular systems to assist handoffs, and an algorithm to reduce the resource reservation for handoff calls by considering that handoff attempts that fail in one cell may succeed in another through multi-hop connections. The extended framework of information exchange among BSs can fulfill the load balancing task, which is otherwise very difficult.

My second MCN paper focus on the benefits that the introduction of ad hoc to cellular systems can bring to the location management. We proposed a grouping method for high capacity transit systems to reduce the location management signaling overhead. We also proposed a generic aggregative scheme which assigns some aggregation devices to collect the location update requests from the arrival mobile terminals. These two scheme can utilize the ad hoc mode’s relay feature and significantly reduce the location management signaling overhead.

1.4 Outline

This dissertation is arranged with the following outline.

Chapter 2 presents my work on the multi-channel MAC scheme for WANETs. Chapter 3 presents my work on the throughput improvement effort on multi-channel multi-radio wireless mesh networks. Chapter 4 presents my work on improving the
overall spectral efficiency for WANETs with the consideration of traffic correlation. All the above 3 chapters address the throughput issues in WANETs.

Chapter 5 presents a micro-mobility scheme for wireless mesh networks. Chapter 6 presents two schemes for handoff performance improving in multi-hop cellular networks (MCNs), among which one utilizes embedded ad hoc networks to assist the handoff calls and the other aims at reducing the BSs’ handoff reservation considering multi-hop handoff attempts. Chapter 7 presents our work on using aggregative devices in MCNs to reduce location management signallings.

Chapter 8 concludes this dissertation.
CHAPTER 2
SAM-MAC: AN EFFICIENT CHANNEL ASSIGNMENT SCHEME FOR MULTIPLE CHANNEL AD HOC NETWORKS

2.1 Background

Multi-channel MAC schemes for mobile ad hoc networks (MANETs) are attracting more and more attention nowadays, although they are still quite unexplored. Both 802.11b and 802.11a support multiple channels in infrastructure mode, but mobile nodes in ad hoc mode could only use one single channel [1]. The benefits of adopting multiple channels in MAC layer are shown in many papers. The most apparent benefit is the throughput improvement.

Nasipuri et al. [48] showed that with the same total bandwidth, dividing a single channel to multiple ones under CSMA mechanism gains a certain throughput improvement. The reason is that the utilization of multiple channels can mitigate collisions and contentions. In the extreme case when channel assignment is perfect and each pair of nodes have a dedicated channel, the contention and collision disappear. Therefore, in this situation, the capacity of the networks can be fully used.

The conclusion above can be only partly true because the overhead of channel assignment in a real multi-channel system cannot be ignored. The carrier sensing coupled with an efficient channel assignment mechanism is always used to select the channel with the least interference for transmission.

The other benefit we will gain from multi-channel is the fairness. We know that in 802.11 protocols, due to hidden/exposed terminal problem, in some topology scenarios, some nodes may become more disadvantaged and get less opportunity to successfully transmit than other nodes. Sometimes this situation will cause more severe problem. By moving the disadvantaged nodes to another channel, the fairness problem of 802.11 system can be alleviated. In other words, the nodes have more choices of channels than in a single channel thus better fairness can be achieved in multi-channel systems.
IEEE 802.11 DCF mode is widely used in MANETs and becomes the de facto standard. Most multi-channel research works are focused on the 802.11-like protocols. However, the previous 802.11-like multi-channel schemes cannot efficiently utilize the frequency band by reducing the channel assignment overhead. Moreover, the control channel saturation problem has significant impact on all these schemes, which results in a limited number of channels.

In this paper, we propose a new 802.11-like multi-channel MAC protocol, called Self-Adjustable Multi-channel MAC (SAM-MAC). The novel part of this scheme is the channel assignment, where a self-adjustment mechanism is used to balance the traffic of multiple channels, thus a more efficient utilization of channels is achieved and the throughput performance is improved. It also mitigates the control channel saturation problem greatly and consequently reduces the overhead further.

The rest of this paper is organized as follows. Section 2.2 reviews the related works. Sections 2.3 and 2.4 present the basic ideas and protocol description of SAM-MAC. Section 2.5 provides a discussion of the problems SAM-MAC has solved and the improvements it has obtained. Finally the simulation result is presented and conclusions are drawn in section 2.6 and 2.7, respectively.

2.2 Related Works

There are a lot of schemes using multiple channels to realize the ad hoc MAC. Nasipuri’s scheme [48] is one of the first multi-channel CSMA protocols, which uses “soft” channel reservation. If there are N channels, the protocol assumes that each host can monitor all N channels simultaneously with N transceivers. A host ready to transmit a packet searches for an idle channel and transmits on that idle channel. Among the idle channels, the one that was used for the last successful transmission is preferred. The protocol is extended by others to select the best channel based on signal power observed at the sender. This multi-channel scheme is a simple extension from the single channel MAC (802.11). It requires each node have N transceivers with one for each
channel, which seems not feasible for a practical system. Despite the infeasibility, this paper gives a useful conclusion that even with the same total bandwidth, separating the channel can improve the throughput performance.

DBTMA [24] and DUCHA [72] also divide a channel into multiple sub-channels, specifically, one data channel and one control channel. Busy tones are transmitted to avoid hidden terminal problems. Through this way the spatial utilization is increased thus a better throughput performance than 802.11 can be achieved. These schemes aim at the hidden/exposed problems in multi-hop topologies.

To exploit the spectrum efficiency with multi-channel schemes, channel assignment is the focus of many other papers.

S. Wu et al. [64], proposed a protocol that assigns channels dynamically, in an on-demand style. This protocol, called Dynamic Channel Assignment (DCA), requires one dedicated channel for control messages and other channels are for data transmission. Each host has two transceivers, so that it can listen on both the control channel and the traffic channel simultaneously. RTS/CTS packets are exchanged on the control channel, and data packets are transmitted on the traffic channel. DCA follows an “on-demand” style to assign channels to mobile hosts, and does not require clock synchronization. This kind of schemes does not perform well when the number of channels is large because all the negotiations are fulfilled on the control channel and too much contention will cause the saturation problem over the control channel.

Similar ideas are used in [14] and [73]. Additionally, Zhang used two common channels to solve the hidden/exposed terminal problems in [73].

MMAC [56] uses a different way to assign the channels. This protocol does not need a separate control channel. Instead, it utilizes an ATIM-like window in the default channel to fulfil the channel negotiation. The ATIM (Ad hoc Traffic Indication Message) window is the synchronization phase when 802.11 Power Saving Mechanism (PSM) is applied. Each node decides to be either in doze mode or awake mode according to the
announcement messages heard in the synchronized ATIM window. Therefore, it has low overhead in channel assignment than DCA and need only one transceiver. However, the price of these benefits is the synchronization. It is known that the synchronization is difficult to realize in MANETs. Furthermore, how to solve the common channel saturation problem remains open.

Shi et al. [55] proposed AMCP scheme which is similar to DCA scheme except that it needs only one transceiver. This major feature comes from a direct timeout mechanism before nodes select the channels. This timeout mechanism solves the multi-channel hidden terminal problem. However, this scheme does not bring great improvement of throughput performance. The other major part of this paper is the fairness improvement. In Information Asymmetry (IA) scenario and Flow In the Middle (FIM) scenario, some disadvantaged nodes will starve due to their disadvantage in the topology to get the channel idle interval. Using multiple channels can mitigate such a starvation problem through allocating another idle channel to the disadvantaged flow timely. However, this paper only solved this problem by adjusting the disadvantaged flow to another orthogonal channel. In the case where the number of channels is much less than the number of flows, such a starvation becomes inevitable again.

A classification has been given recently in [46]. In this paper, the multi-channel schemes have been divided into four categories:

1. Dedicated Control Channel (DCC);
2. Common Hopping (CH);
3. Split Phase (SP);
4. Multiple Rendezvous using 1 radio (MR).

CH and MR use the idea of time division and frequency hopping. RICH-DP [58] is an example of CH and SSCH [8] is an example of MR. Though [46] shows MR has a better performance than DCC and SP, we exclude them (CH and MR) in our paper in that they
Figure 2-1. Comparison of DCC Type and SP Type

use a very different approach in which time synchronization is needed and a channel hopping sequence is followed.

It is clear that the DCA scheme belongs to DCC type and MMAC belongs to SP type, according to [46]. Fig. 2-1 illustrates the basic difference of channel assignment (CA) between DCC type and SP type of approaches.

2.3 Motivations and Basic Idea

2.3.1 Motivations

The direct way to increase the throughput is to reduce the overhead caused by channel assignment. Therefore the primary goal of this design is to use the available frequency bands as efficiently as possible, thus achieve greater throughput.

If each node has a dedicated transceiver for each channel, the channel assignment will have zero overhead. This is because every channel is “visible” to every node all the time. However, due to cost consideration, the transceivers are usually fewer than the available channels. Therefore, channel assignment needs to assign the available channel resource to limited transceivers when there are data transmission requests. It is important to know the channel usage information before actual channel assignment.
Otherwise, collisions may happen or the extra waiting time will be inevitable. DCC type schemes collect the channel usage information on common control channel and assign channels according to it. SP type schemes use time division to clear the past channel usage and use special phase (beginning of each time interval) to do the channel assignment on the default channel.

Both are feasible approaches to achieving throughput improvement. But using split phases needs time synchronization, which is difficult to realize in ad hoc networks. Also with split phases, how to divide time into different phases is still a two-fold problem. The first difficulty is that data packets have variable size which probably do not utilize the data phase efficiently. The second is how long the channel assignment phase should be. This is because traffic may be different and request load of channel assignment is also hard to predict. Without synchronization and time division problem, though a dedicated frequency band still means relatively high overhead, the DCC approach seems more appealing. (Here we assume the synchronization is not mandatory in the system).

Since the transceivers are fewer than channels, the transceivers need to switch among the available channels. Therefore, two important issues cannot be ignored in the scheme design. One is the channel switching delay problem and the other is the acquirement of the channel usage information. For the first problem, a per-packet-based channel assignment scheme is not preferred. Channel assignment should be valid for a longer period. For the second problem, when a single transceiver is used, the channel assignment is difficult after data transmission because the usage information of other channels will be “invisible” during the data transmission period. AMCP, which uses only one transceiver, bypasses this difficulty in time domain [55]. It requests the nodes to wait for one data transmission period before channel assignment if the preferred channel is not available, which is also an extra overhead of channel assignment. To address these two problems, using two transceivers is a better choice. This furthermore helps to reduce the overhead by allowing data transmission on the common channel.
2.3.2 Basic Idea

The basic idea of the proposed scheme can be summarized as follows.

- Channel 0 is used as the common channel for channel assignment and other usage. Other channels are used for data traffic, called traffic channels.
- Two half-duplex transceivers (Tx) are required for each node. Tx 0 stays on common channel and Tx 1 switches among all traffic channels dynamically.
- Every node maintains a table recording its neighbors’ traffic channel number and it updates this table according to the information heard from the common channel.
- Nodes keep listening on the same traffic channel unless the channel is too busy. Channel assignment is not needed before every transmission. If nodes know the receivers’ traffic channel numbers, they send RTS on the traffic channel. Otherwise, they first send query requests on common channel before RTS/CTS.
  - When nodes need to change their traffic channels, requesting frames are also required to be sent on the common channel. The transmission on the traffic channels always follow the back-off procedure with the NAV information indicated on the common channel.
- Nodes can get the information of channel usage status by listening on the common channel. Based on this information, nodes can choose another traffic channel when current channel is too busy. This channel re-assignment method is called self-adjustment and it is critical to the performance of the whole system.
- The common channel can also be used for data transmission when its traffic is relatively low.

The basic idea is based on such an observation that in all previous schemes, handshakes occur only at the common rendezvous. This mechanism makes the common rendezvous susceptible to be the whole system’s bottleneck. SAM-MAC distributes the handshakes to available channels and furthermore makes the data transmission on the common channel possible. In this way, our scheme achieves higher
bandwidth efficiency and removes the bottleneck from the system. According to the above, we can consider our scheme as an extension of DCC approach.

2.4 Protocol Description of SAM-MAC

The protocol is described as follows. We first introduce the basic messages and then describe detailed procedures.

2.4.1 Basic Messages

- **RTS/CTS**: Request to send/clear to send.
- **DATA**: Data.
- **ACK**: Acknowledgement to the data frame.
- **NCTS**: Negative clear to send.
  
  When the destination is not available, it can send a NCTS on channel 0 to the sender to notify the NAV (Network Allocation Vector) [11]. It is used to solve the receiver blocking problem.
- **RTF/ATF**: Request to find/Acknowledgement to find.
  
  If the destination node’s listening traffic channel number is unknown, the sender sends RTF on Channel 0 to find it out and the destination node answers it also on Channel 0;
- **RCT/ACT**: Request to change traffic channel/Acknowledgement to change traffic channel.
  
  If a sender feels the traffic channel situation unsatisfying, it will send RCT to the destination to change the sender’s traffic channel to another one, on Channel 0. Then the destination node will answer back ACT when it decides which traffic channel to choose, also on Channel 0. The old traffic channel information should be included in these messages to let other nodes know the busy status of this traffic channel.
- **NBC**: NAV Broadcast.
Every time a receiver wants to send the RTS/CTS in traffic channel, it will copy its NAV information and Channel Number to NBC's field and send NBC on Channel 0. This message is used to avoid the hidden terminal problem caused by Multi-channel. This mechanism makes the NAV be a vector instead of a scalar.

2.4.2 Basic Operational Procedures

We first describe the initial channel selection of each node. After that, how a node transmit to a node which the sender has no knowledge of the receiver’s operating channel is described. When this knowledge is acquired, the transmission procedure can be simplified to the basic IEEE802.11 procedure. Under multi-channel scenarios, dynamically adjusting the operating channel is needed for efficiently utilize the channel resource. Finally, different from other DCC schemes, our scheme can allow common channel to be used for data transmission.

2.4.2.1 Choosing traffic channel (Initial channel selection)

With two half-duplex transceivers, one node listens on both the common channel (Channel 0) and one of the traffic channels. At the very beginning, when nodes join the networks, each of them picks randomly one traffic channel as its listening traffic channel.

Nodes keep listening on the same traffic channel until they need to transmit to some nodes on other channels or this channel is greatly saturated. The change of one node's listening traffic channel should be published to all other neighbors via the common channel.

How the sending nodes find the receiving nodes and how they change channels are described in following procedures.

2.4.2.2 Transmitting to a node listening on an unknown traffic channel

If the channel information of the destination node is not known, as Fig. 2-2 shows, the transmitting node sends a RTF on the common channel first. Since each node listens on the common channel, the destination node answers an ATF on the common channel, which includes the traffic channel information. Hereafter, the transmitting
node switches its Tx 1 to the given traffic channel and starts the RTS/CTS/DATA/ACK procedure. This procedure is required when the sending nodes have no memory of the receiving nodes, e.g., when new nodes join the networks. After one transmission, the listening traffic channel number of the node is known to all its neighbors and the following transmission becomes a simpler case.

### 2.4.2.3 Transmitting to a node listening on a known traffic channel

Each node maintains a table (neighbors’ channel table) recording the traffic channel number of its neighbors. Every time an ATF/ACT/NBC is transmitted on the common channel, each node listening on it updates its table. When there are data to transmit, each node looks up this table first to find out the channel number of its destination node. Then it switches to the given traffic channel and sends data using RTS/CTS/DATA/ACK procedure, as shown in Fig. 2-3. If the sender node is listening on a different channel before RTS, it should send a NBC message to notify all its neighbors of the channel change after RTS is sent. The purpose is to broadcast the NAV to all its neighbors and to avoid the Missing Receiver Problem [55] caused by the unnotified channel change. The receivers are also required to broadcast NAV via NBC frames.

The neighbors’ channel table also includes the available channel lists of its neighbors. This information is stored for the purpose of channel adjustment.

![Figure 2-2. Illustration of SAM-MAC Procedures: Unaware Case](image-url)
2.4.2.4 Adjusting to a different traffic channel when busy/collision (Self-adjustment procedure)

Nodes can adjust their channels to another one for the purpose of system’s load balancing. When the transmitter wants the receiver to change to another traffic channel, it sends an RCT on the common channel. The current traffic channel number and the available channel list should also be included in it. The traffic channel adjustment is decided by the receiver. After the receiver decides which traffic channel is most suitable (by the channel reassignment algorithm), it sends back an ACT on the common channel and switches Tx 1 to the chosen traffic channel. Then a RTS/CTS/DATA/ACK procedure follows on the traffic channel. If the receiver cannot find a more suitable traffic channel, it sends an ACT requesting the sender to stay in the same traffic channel.

If the users keep on changing channels, the system throughput as well as the QoS of the users will be degraded due to the communication overhead and switching delay. Therefore, a better channel re-assignment algorithm is required to reduce the frequency of channel changing. We assume the traffic of the users is not extremely unbalanced and bursty thus the traffic load can be well balanced among the channels. We can easily observe that only when a channel contains a greatly varying traffic can the frequency of channel reassignment be big. The assumption above is reasonable because the aggregate traffic in a channel is relatively constant and the fluctuation of small number of nodes’ traffic will not cause great traffic load imbalance among different channels.

![Illustration of SAM-MAC Procedures: Aware Case](image)

Figure 2-3. Illustration of SAM-MAC Procedures: Aware Case
We use two metrics for the channel re-assignment algorithm: the number of neighbors and the channel busyness ratio [68], which can both be counted or calculated from the information heard on the common channel.

2.4.2.5 Transmitting data on the common channel

Unlike previous DCC schemes, in which the handshakes always happen on the common channel, SAM-MAC’s handshakes happen on traffic channels. This means that the common channel has much less traffic than traffic channels and data transmissions on traffic channels are less sensitive to the traffic load on common channel than previous DCC schemes. The traffic on common channel consists of RTF/ATF, RCT/ACT and NBC messages. RTF/ATF messages are needed only when a new node joins the network. NBC is needed for the nodes on a different channel to acquire the busy status of the operating channel. RCT/ACT are used for channel re-assignment. Consequently, RCT/ACT and NBC form the major traffic on the common channel. As we mentioned above, when the traffic of the system is assumed not to be extremely unbalanced, the traffic on the common channel only utilize a small amount of channel resource. When the traffic on each channel is constant, this type of traffic can be ignored. Therefore, data packets can also be transmitted on the common channel.

Channel assignment algorithm allows the common channel to be chosen for data transmission when the common channel is light-loaded and the channel adjustment is needed.
After this adjustment is done through RCT/ACT, both the sender and the receiver use Tx 0 for the data transmission. Tx 1 can be turned off until adjustment to a traffic channel is required again.

2.4.3 Other Issues

In multi-hop topologies, the exposed terminal may be blocked by other transmissions and the sender cannot get the response from it. In SAM-MAC, the exposed terminal can send back NCTS on the common channel instead of being silent as it does in 802.11 protocol. Through this way, the exposed terminals are not vulnerable to be blocked.

Since the channel reassignment is decided by the receivers, the asymmetric information of channels between the senders and the receivers may cause unreasonable channel reassignment. For an example, a receiver may choose a channel that the sender cannot transmit at all. This is a usual case in multi-hop topologies. To avoid this inconsistency, the channel busyness ratio of the sending nodes’ channels should also be considered during the channel reassignment. In the adjustment procedures, the senders should always include a channel list with a decreasing order of channel busyness ratio. The receivers store this information to the neighbors’ channel table. Before the decision of channel reassignment, the receivers choose a channel with the lowest channel busyness ratio from its own traffic channel status table which is also in the available channel list of the neighbors’ channel table. Therefore, the receivers would not choose a channel that cannot be accepted by the senders.

All the channels are assumed here to have the same wireless status and the channel busyness ratio can embody the real channel status of each channel. Without the consideration of frequency-selective interference, all the channel usage can be acknowledged by the NAV broadcasting, which is the virtual carrier sensing. If the physical wireless channel status needs to be considered, the second transceiver should scan around all the traffic channels to do the physical carrier sensing. In this case this overhead also needs to be considered.
2.5 Performance Analysis

Next we evaluate the performance of SAM-MAC. Before we present the numerical simulation result, we first discuss some well-known problems and the impacts to our solution.

2.5.1 Multi-channel Hidden Terminal Problem

Multi-channel hidden terminal problem is an inborn problem in the multi-channel schemes. The illustration below assumes the MAC is using RTS/CTS-like mechanism, typical in IEEE 802.11 standards. Another assumption is that each node has less transceivers than channels.

In Fig. 2-5, node A is communicating to node B in a common channel. Node C switches to this channel around time T1 when, unfortunately, it misses the CTS sent out by B. If node C, a hidden terminal to A, proceeds to transmit to some other node after sensing an idle channel, the signals could collide with signals sent by A at the receiver of B. We can see the reason of this problem is that node C misses the channel status, the NAV in CTS. If each node can have the same number of transceivers as channels, this problem is avoided.

For DCA [64] and MMAC [56], since handshakes occur only at the common rendezvous and each node can obtain the channel usage information by monitoring handshakes, multi-channel hidden terminal problem does not occur. For AMCP [55], with one dedicated channel and only one transceiver, multi-channel hidden terminal problem could not be easily avoided. With a direct timeout mechanism in AMCP, each node avoids this problem by acquiring the channel usage information during this period.
Though this timeout mechanism helps to solve the multi-channel hidden terminal problem, it introduces an extra requirement of packet size due to the setting of the timer length.

For SAM-MAC, since handshakes are distributed, nodes on different channels are blind about each other’s channel usage information if the control channel is absent. However, with one dedicated transceiver on the control channel, SAM-MAC has the capability of knowing all channels’ usage information. The message NBC is used for this purpose. With this message, every node knows the NAV of all the channels and avoid the multi-channel hidden terminal problem.

2.5.2 Control Channel Saturation Problem

Though the dedicated control channel is the main overhead of DCC schemes when the channels are fewer, when the number of channels becomes greater, this overhead becomes relatively small and in some scenarios the control channel’s capacity becomes the bottleneck of total throughput. In [64], this saturation problem of DCA scheme has been shown by the simulation results. Neither can MMAC overcome this problem due to the limit of ATIM window length. In [55], AMCP uses the same serial contention procedure on the control channel, so it cannot avoid such a problem, either.

For SAM-MAC, we observe that most of the handshakes before each transmission occur on different traffic channels. The major traffic on control channel, NBC messages, is not related to the handshake. Consequently the saturation problem can be overcome. Only when there are a lot of channel adjustments will the handshake traffic of control channel become heavier. This can be caused by extreme fluctuation of traffic aggregation on each channel, which is very rare.

2.5.3 Missing Receiver Problem

The nature of Missing Receiver Problem can be explained as follows. When a transmitter wants to send a packet to a receiver which happens to tune its transceiver on another channel, the Missing Receiver Problem occurs.
In DCA and MMAC, the dedicated control channel and the ATIM window, which are so-called common rendezvous, contains all the channel usage information. Hence this problem has no impact on them. For AMCP, after saving one dedicated transceiver on control channel, this problem becomes an issue. Since AMCP uses direct timeout mechanism when it loses channel usage information, during this delay period, the Missing Receiver Problem is hard to overcome.

For SAM-MAC, all the channel usage information can be obtained from control channel with a dedicated transceiver. Therefore, this problem is solved in SAM-MAC.

### 2.5.4 Overhead Comparison and Throughput Improvement

Due to the same RTS/CTS/DATA/ACK procedure in both our scheme and previous schemes, the throughput performance of the standard 802.11 protocols can be used as a reference of the performance comparison. Considering the overhead caused by channel assignment, the multi-channel schemes' maximum throughput can be expressed as follows:

\[ S_{mc} = n \cdot S_{standard} - S_{overhead} \]

where \( n \) is the number of channels, as shown in Fig. 2-1. The above equation shows that the maximum throughput with multiple channels and multi-channel schemes is the product of the number of channel times the maximum throughput of 802.11 protocols, minus the channel assignment overhead. Obviously, the channel assignment overhead is the major indicator of the performance of a multi-channel scheme.

Referring to Fig. 2-1, the ratio between channel assignment phase and total beacon interval in SP type schemes is denoted as \( \alpha \). Therefore, the channel assignment overhead of SP type is \( \alpha \) and the one for DCC type can be easily taken as \( \frac{1}{n} \). Once the system starts up, this part of overhead caused by channel assignment is fixed no matter this part of separate resource is redundant or deficient for the channel assignment.
Under the assumption of fully utilization of the control channel (DCC type) or CA phase (SP type), we compare the overhead of two types with different numbers of channels. We use the parameters set in [56] as an example. $\alpha = 20/100 = 0.2$. When the number of channels is less than 5, the SP type of schemes have less channel assignment overhead. If this $\alpha$ can be smaller than 0.2, this advantage of SP type can hold in the scenarios with more channels.

However, DCC type and SP type do not have the identical channel assignment capability. Although the whole control channel can be used for channel assignment, only $1/n$ of CA phase is useful in SP type because of the requirement of a common rendezvous. To achieve the same capability of channel assignment as DCC type, SP type should require $\frac{\alpha}{n} = \frac{1}{n}$, which is impossible because there is no room left for data transmission at all. We know from previous research, [55], [64], that there will be a saturation problem in control channel when $6 \sim 8$ channels are fully utilized in DCC systems. According to the previous reasoning, we conclude that SP systems cannot support so much channel assignment load. Therefore, the SP systems are obviously more vulnerable to the saturation problem than DCC systems.

The comparison above shows that in single-hop topology it is impossible for multi-channel schemes to get $n$ times throughput as in a single 802.11 channel because of the existence of channel assignment overhead. When the channel assignment load is heavier, even $n-1$ times throughput is impossible.

This limit is not the same in multi-hop topology. The multi-channel schemes allow transmission to be happened concurrently on different channels which may not be allowed in a single channel scheme. In other words, to some extent multi-channel schemes increase the spatial reuse. Even with these overheads aforementioned, the throughput improvement is more apparent when traffic is heavy or exposed/hidden terminal problems are greatly alleviated by multiple channels.
Following the overhead analysis of previous schemes, we can analyze SAM-MAC’s overhead. When there are no new neighbors joining the networks, no RTF/ATF procedures are needed for a transmission. The multiple channels’ traffic is assumed to be well balanced by the self-adjustment channel reassignment thus few channel adjustments (RCT/ACT) are needed. Therefore, before each data transmission only RTS/CTS are required for the handshake. Without channel assignment for each transmission, each traffic channel can obviously get a approximate throughput as a single 802.11 channel.

Furthermore, the difference from other DCC Schemes is the distribution of handshakes to traffic channels. We can model the contention as a M/M/1 queue. Assume each contender generates a packet with exponential inter-arrival rate \( \lambda \) and each resolution of contentions also follows the exponential distribution with the rate \( \mu \). The number of contenders in the system is \( m \). When using previous schemes’ handshake, the probability of \( k \) contenders \( p_k \) in the system and the average queue length \( \bar{N} \) are:

\[
p_k = (1 - \frac{m\lambda}{\mu})(\frac{m\lambda}{\mu})^k
\]

\[
\bar{N} = \frac{1}{(1 - \frac{m\lambda}{\mu})} - 1
\]

After using SAM-MAC, with \( m \) contenders dispersed to \( n \) channels, the result is changed to:

\[
p_k = (1 - \frac{m\lambda}{n\mu})(\frac{m\lambda}{n\mu})^k
\]

\[
\bar{N} = \frac{1}{(1 - \frac{m\lambda}{n\mu})} - 1
\]

It is shown from the result that \( p_0 \) increases and \( \bar{N} \) decreases. With the average contention queue length decreasing, the total duration of the handshakes decreases.
This change helps to increase the throughput because the mitigated contention is the inherent overhead of CSMA/CA protocols. From the overhead point of view, this solution mitigate the handshake overhead in 802.11 protocols. According to the result of [10], the saturated throughput is independent of \( m \) when \( m \) is large enough. This is because in saturated situation the difference of number \( m \) could not bring much difference of collision probability. When \( m \) is small or the system is non-saturated, \( m \) affects \( \bar{N} \) greatly and \( \bar{N} \) affects greatly the throughput performance as well as the delay performance.

For all the previous schemes in which handshakes occur only at the common rendezvous, the dedicated part (the common channel or the ATIM window) cannot be used for data transmission even if it is far from saturation because of the dependency of data transmission to handshakes. However, with distributed handshakes, our scheme allows the common channel to transmit data, which makes the common channel no longer “dedicated” and achieves greater throughput than any other schemes. Therefore, \( S_{\text{overhead}} \) is further reduced and the throughput performance is improved. Additionally, since our scheme greatly alleviates the saturation problem of control channel, this scheme can support much more channels than other schemes.

### 2.6 Simulation Results

NS-2 is used as the simulation platform. The simulation keeps other layer intact and only modifies the MAC layer. This simulation only cares about the throughput in MAC layer since this scheme focuses on the behavior of MAC layer. Therefore, the end-to-end throughput will not be concerned in the simulation. The throughput here means the number of all the packets successfully transmitted by MAC layer.

Shi’s work [55] is used as the reference of simulation result for single-hop topology. The reason is that this paper is the latest one about the multi-channel schemes. It has given a comprehensive generalization of previous work and performance comparison. However, in multi-hop scenarios, this work focuses mainly on the fairness issue.
Table 2-1. Simulation Parameters For Single-hop Topology

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>SIFS</td>
<td>10 $\mu$s</td>
</tr>
<tr>
<td>DIFS</td>
<td>50 $\mu$s</td>
</tr>
<tr>
<td>EIFS</td>
<td>364 $\mu$s</td>
</tr>
<tr>
<td>Time Slot</td>
<td>20 $\mu$s</td>
</tr>
<tr>
<td>PHY header</td>
<td>192 bits</td>
</tr>
<tr>
<td>MAC header</td>
<td>224 bits</td>
</tr>
<tr>
<td>RTS</td>
<td>160 bits + PHY header</td>
</tr>
<tr>
<td>CTS, ACK</td>
<td>112 bits + PHY header</td>
</tr>
<tr>
<td>DATA</td>
<td>8000 bits + PHY header + MAC header</td>
</tr>
<tr>
<td>RTF, ATF, RCT, ACT</td>
<td>160 bits + PHY header</td>
</tr>
<tr>
<td>Basic Rate</td>
<td>2Mbps</td>
</tr>
<tr>
<td>Data Rate</td>
<td>2Mbps</td>
</tr>
<tr>
<td>Switching Delay</td>
<td>0</td>
</tr>
<tr>
<td>Topology Range</td>
<td>100m*100m</td>
</tr>
<tr>
<td>Flow Number</td>
<td>15</td>
</tr>
<tr>
<td>Duration</td>
<td>25s</td>
</tr>
</tbody>
</table>

without addressing the throughput performance. Therefore, we evaluate our scheme in multi-hop scenarios through other method.

In the simulation, different nodes in the network start to deliver packets from different starting time. Channel re-assignment algorithm will distribute the traffic evenly to the available traffic channels.

After this period, the channel switching is not so frequent because the nodes keep staying on the same channel unless the traffic change or connectivity cannot be satisfied. Thus the switching delay plays a less important role in our scheme, which is ignored here.

In the following simulations, single-hop throughput performance is compared with AMCP [55] under the scenarios with different number of channels. Specifically the detailed throughput performance of 3+1 (3 traffic channels and 1 common channel) is shown. Multi-hop throughput performance is provided and evaluated after that.
Although the maximum aggregate throughput in single-hop scenarios can be formulated [10], it is necessary to address that the aggregate throughput in multi-hop scenarios depends much upon the topology and the coverage area. Simply saying, a bigger topology potentially has bigger aggregate throughput than a smaller one because of more nodes and more spatial reuse. For this reason, the value of aggregate throughput in multi-hop scenarios is meaningless without comparison. The benefit of our scheme in multi-hop scenarios is shown through comparing the throughput gain with different number of channels on a certain random multi-hop area.

2.6.1 Single-hop Topology

The simulation parameters are set as Table 5-1.

In single-hop topologies, each node can listen to all the other nodes. As mentioned before, throughput of multi-channel schemes cannot exceed \( n-1 \) times of the saturated throughput of a 802.11 single channel, which intuitively is the “upper bound” throughput of DCC type schemes. However, SAM-MAC can break this “upper bound” when common channel (“CC” in the figures) is used for data transmission.

![Figure 2-6. Throughput Gain in Single-hop Topology with 3+1 Channel, with and without common channel transmitting data.](image)

Fig. 2-6 shows the throughput comparison of 802.11, AMCP, and SAM-MAC with and without CC used for data transmission. Under saturated situation, SAM-MAC without CC used for data transmission can get a slightly better performance than AMCP,
with the gain very close to 3. However, with CC used for data transmission, this gain can be up to 3.7.

Next simulation shows the potential multiple channels’ throughput using our scheme. To get the potential multiple channels throughput, more traffic resources are needed. In this simulation, 100 nodes are used. A reference line is used for the gain comparison. We note that the common channel is not used for data transmission in this simulation in order to make the benefit of being free from saturation problem clearer.

![Throughput Gain in Single-hop Topology with Multiple Channels](image)

Figure 2-7. Throughput Gain in Single-hop Topology with Multiple Channels

From Fig. 2-7 we can see the following result. The “1 channel” point stands for the standard 802.11 protocols’ throughput. With the number of traffic channels being increased up to 11, the throughput of SAM-MAC is approximately linearly increasing, which shows this scheme is free from the control channel saturation problem. When 11+1 channels are used the throughput can achieve 9.5 times as a single 802.11 channel. The potential throughput can reach 1750 pkt/sec or even more. In this scenario, the overhead of a dedicated control channel can be almost ignored.

2.6.2 Multi-hop Topology

In multi-hop topology simulation, the topology range has been changed to 500m × 500m, the number of flows to 32, and basic rate to 1Mbps. To get a clear
comparison with the throughput in single-hop topology, CC is not used for data transmission in this simulation.

From Fig. 2-8, it is shown that in the topologies where the hidden/exposed terminal problems affect the throughput greatly, using multiple channels can achieve higher gains than single-hop topology case. The saturated throughput of one single channel is only 150 pkt/sec, which is much lower than the 185 pkt/sec in single-hop topology case. After using multiple channels with SAM-MAC protocol, the gain is more than in single-hop topologies. In 3+1 channels and 6+1 channels cases, the gains are more than 3 or 6 times of a single 802.11 channel's throughput, respectively. The reason that 10+1 cannot achieve more than 10 times gain is because the traffic cannot fully saturate the channels.

The result of our throughput performance simulation supports the overhead analysis aforementioned.

### 2.7 Summary

Using multi-channel scheme helps to improve the performance of ad hoc networks, including throughput and fairness. Finding a way to use the multiple channels more efficiently helps to gain more performance improvement.
SAM-MAC can help to achieve this goal when a single channel is not enough. When the number of channels is relatively large, this scheme is more attractive because the overhead of the control channel is relatively small and the control channel saturation problem has less impact. When the number of channels is limited and the dedicated control channel becomes the major overhead, the bandwidth of the dedicated control channel can be shared by data transmission.

This scheme has obtained a better throughput performance with one more transceiver than AMCP and MMAC. It is less restricted and easily designed because of this extra transceiver. Future work includes the detailed fairness analysis and fair algorithm in channel assignment. This part can be done as one single module of channel assignment. Therefore the software is easy to be upgraded.
CHAPTER 3
EXPLOITING THE CAPACITY OF MULTI-CHANNEL MULTI-RADIO WIRELESS MESH NETWORKS

3.1 Background

Wireless mesh networks (WMNs) can be deployed to extend the coverage of last mile access, which are seen as the extended version of WLANs aiming at larger coverage areas and higher bit rate. Compared to its counterpart, cellular networks, WMNs can provide mesh clients the ability to access the Internet with the advantages of low cost, easiness of deployment and configuration, and flexibility of construction. For example, WMNs can provide Internet services to unscheduled conferences, military deployments, or emergency rescues.

Akyildiz et al. have provided an excellent survey on WMNs [5]. Commonly, a WMN consists of a limited number of gateways to access the wired networks, many mesh routers each of which covers a certain area and a larger number of mesh clients. Fig. 3-1 gives an example of a WMN structure with one gateway. Gateways function as the portal devices, which provide access to the wired Internet. Mesh routers connect with gateways via wireless media through zero or more intermediate mesh routers. Mesh clients access the system through the local mesh routers. The local mesh routers help mesh clients to forward data packets to the gateways, with one or multiple intermediate hops.

Figure 3-1. The Common Structure of a WMN
Nowadays, two series of standards can be applied to WMNs. One is the contention-based IEEE 802.11 family, among which the ongoing one, 802.11s, is aiming at better support to mesh structure. The other is the IEEE 802.16 family with the scheduling-based feature. It is widely known that the current 802.16 family do not support multi-hop networks well. IEEE 802.16j, aiming at multi-hop applications, also known as Mobile Multi-hop Relay-based (MMR) WiMAX, is still under development.

Usually, it is deemed that mesh clients use WMNs mostly for accessing the Internet. Therefore, the ability that a WMN provides mesh clients to deliver/receive data across the gateways is the focus of the design. Due to limited number of gateways and irregular distribution of mesh clients with multi-hop distance from the gateways, it is difficult to construct a WMN which can satisfy both coverage and throughput requirement. Researchers have been delving into the design issues. Jiang et al., proposed a CDMA-based resource management scheme for the wireless mesh backbone [28]. In order to boost the throughput of WMNs, they design the scheme of resource allocation by using location and interference information from the receivers’ point of view. Attempting to find out what attribute to current low throughput performance of WMNs, we generalize two major features for WMNs: multi-hop and traffic aggregation.

Due to multi-hop feature of WMNs, a large part of resource is “wasted” in forwarding at the intermediate routers. This feature causes multiple channel contentions and flow discontinuity as well. The other major difficulty coming from multi-hop feature is the frequency reuse, or from another angle, concurrent transmission. When there is no frequency reuse, the system throughput is definitely low because so many links share limited amount of frequency resource, a large part of which is used for forwarding. In multi-hop scenarios, reusing faraway frequency resource with proper planning is a basic and effective way to improve the throughput performance.

Because of the traffic aggregation towards a gateway, central links always have more traffic load. When these links do not acquire more resource than outside links,
congestion is created and bottlenecks are formed. Generally, when a large amount of undeliverable traffic is generated in some area, the congestion creates bottlenecks. Several other reasons can create bottlenecks in WMNs as well. Due to the different sensing range among different routers, hidden/exposed terminal problems and blocking problem haunt WMNs. Hidden terminal problem can increase the collision of a link. Exposed terminal problem and blocking problem can create disadvantaged links. These abnormalities are all possible reasons for bottleneck formation.

Fortunately, with the aid of multiple channels and multiple radios (MC-MR), the WMNs’ performance can be improved with effective channel allocation [35]. IEEE 802.11b and IEEE 802.11a specify 3 and 12 non-overlapping channels, respectively. To this end, many previous works proposed architectures or algorithms with different techniques [35], [50], [6], [34], [51], [53], [65].

Unfortunately, up to now, no previous paper gives us the knowledge of the capability that WMNs provide the mesh clients to deliver/receive data across the portals. It will be very useful to know the maximum traffic a WMN can support to deliver across the gateways at the construction phase without the presence of mesh clients.

In this paper, we investigate the capacity that WMNs provide mesh clients to deliver/receive data across the gateways, which we term as “portal capacity”. It will be shown later that portal capacity varies with different fairness constraints. We assume that traffic distribution is known so that statistically each mesh router contributes a certain amount of traffic to the gateways. Under these assumptions, the objective function of optimal portal capacity is formulated without the requirement of the knowledge of the real-time traffic demand. Furthermore, we propose our solution of achieving the optimality in portal capacity through the centralized algorithm while retaining the optimality via a distributed mechanism.

The contributions of this paper are as follows.
• It proposes the concept of portal capacity which characterizes the wireless access capability of WMNs provided to mesh clients.

• It formulates a target problem for maximizing the portal capacity with the fairness constraint and reduces it to a much simpler and more soluble one via the frequency reuse.

• It gives two parts of solution among which the centralized part focuses on the realization of the optimum portal capacity and the distributed part renders flexibility and robustness to the system.

The rest of the paper is organized as follows. Section 6.2 discusses the related works. Section 3.3 formulates the target problem. Section 3.4 describes the proposed solution. Simulation and evaluation are provided in Section 4-4. Conclusions are given in the final section.

3.2 Related Works

Some of previous works proposed heuristic approaches of utilizing multi-channel and multi-radio to increase the throughput performance [51], [53], [65], [57]. In [53], an architecture called Hyacinth is proposed for the MC-MR WMNs. Each mesh router carries out load-balancing routing and load-aware channel assignment in a distributed fashion. In [51], the interference mitigation is introduced in the design. The multi-radio conflict graph and a breadth-first searching algorithm are used for the interference-aware channel assignment. In [57], the channel assignment is jointly designed with multi-path routing and scheduling. In [65], superimposed code is used in channel assignment because of its property of s-disjunct. By assigning superimposed code to each mesh router in advance, the communication channel can be determined by the manipulation of two superimposed codes. However, this method has a requirement for a relatively large number of channels to avoid the co-channel interference. Most of these heuristic works are distributed schemes, which can adapt the traffic variation and link failure.
quickly, thus the robustness can be achieved, though they lack of the consideration of
the optimality of throughput performance.

Some works apply optimization techniques instead to achieve the optimality
of the system [50], [6], [34], in a centralized manner. Both [6] and [34] use a scale
factor $\lambda$ to scale the flow rate under the constraints in WMNs, while they use different
interference models and different algorithms to solve the optimization problems. In
[50], the target problem is to directly maximize the summation of utility function of rate
for each flow. These works are mostly too complicated in algorithm because of the
need of decomposing the NP-hard optimization problem and also lack of robustness
for supporting a real system. Meanwhile, the above algorithms do not give us the
information about the capability that a WMN can provide its clients to deliver packets
across the gateways.

In addition, to achieve and retain the optimality in these centralized works, WMNs
are always assumed to have stationary mesh topology and static traffic demand
when calculating the optimal value. Although gateways and mesh routers are mostly
stationary, due to inevitable wireless link failures, occasional node failures and node
maintenance, the topology cannot be considered as immutable. Moreover, although
mesh routers usually have a more constant traffic load than mesh clients by aggregating
mesh clients’ traffic, the traffic loads of mesh routers can still vary greatly from time to
time due to a variety of group events just like city traffic. Consequently, with the above
algorithms, any of these locally minor changes in the network can cause the loss of
fairness and optimality, which requires further global channel assignment and resource
allocation. Therefore, the variation of topology and traffic load cannot be ignored in
WMNs’ design. Though the relatively stable and stationary information of the topology
and traffic can be the input for calculating optimal performance, a dynamic mechanism is
required as well to address the robustness issue.
3.3 Problem Formulation

3.3.1 Portal Capacity

Unlike other ad hoc networks, a WMN is usually deployed by placing a number of mesh routers around one or several gateways so that mesh clients are able to deliver/receive data across this infrastructure and access the Internet via the portal devices, i.e. gateways. (In this paper, we first investigate the scenario with one gateway.) Obviously, the throughput the gateways carry is the aggregate end-to-end throughput of the system. Previous works investigated the aggregate end-to-end throughput performance with the knowledge of real-time traffic demand from mesh clients. However, the wireless ability that a WMN provides its clients to deliver/receive data across the gateway/gateways has never been touched. Hereafter we name this wireless ability as the **portal capacity**. It is obvious that each mesh router in a WMN takes its capacity to/from the gateway as its individual portal capacity for its local mesh clients.

Similar to APs in WLANs, the maximum achievable throughput of gateways in WMNs defines the capacity of the whole system. The calculation of a WLAN's capacity is simply the capacity of the operating channel due to the fact of its one-hop nature. In multi-hop WMNs, the system capacity is not a constant as it is in WLANs because the spectrum is shared by multi-hop links. Different spectrum resource sharing among multi-hop links will cause different portal capacity. According to the traffic aggregation property, the system capacity can be learned via the summation of the capacity of the interfering last hop links connected to gateways. Without the consideration of fairness, a WMN's portal capacity reaches the maximum value when the last hop links take all the channel resource while all other mesh clients and mesh routers are starved. Therefore, portal capacity is meaningful only when fairness constraint is considered.

If the fairness can only be achieved when each link's traffic demand is satisfied with the predefined proportion as in [34], the retainment of this fairness will be too difficult because each time when the traffic of any mesh client changes, fairness will
change and the traffic should be re-scheduled. Due to this consideration, we consider
the fairness in another way. We assume that the traffic distribution follows a fixed PDF
(Probability Density Function). The resource allocation can be based on the knowledge
of this distribution so that the optimum portal capacity can be achieved. In addition, a
dynamic mechanism is necessary to adjust this resource allocation to efficiently use
up the spectrum in real time. In this paper, we assume that the traffic follows uniform
distribution and the coverage of each mesh router is identical. Consequently, each
mesh router is seen equivalent with respect to its traffic contribution to the gateway
(forwarded traffic from other mesh routers excluded). Obviously, this problem seems
much easier to solve without considering the real-time traffic demand of mesh clients.
The optimal portal capacity under this fairness constraint is then determined only by
the infrastructure consisting of the gateway, mesh routers and the links among them,
which are relatively stationary. The knowledge of portal capacity can thereupon be
obtained without the presence of mesh clients at the construction phase and provide the
engineers important information beforehand.

Although this model ignores the diversity of traffic demand among the clients, it can
give us useful knowledge of system’s performance and guide the design of the system.
Such information cannot be obtained if the real-time traffic demand is considered for
fairness. Furthermore, we can incorporate the diversity of traffic demand statistically
by giving different traffic weights to different mesh routers according to their geographic
characteristics. For example, in busy areas, such as shopping malls, or important areas,
such as hospitals, the mesh routers can have bigger traffic weights than others.

3.3.2 Problem Formulation

Apparently, the optimum portal capacity of the system and its realization are our
focus in this paper. We denote the set of final hop links to the gateway(s) as $L_1$, the
set of mesh routers as $M$, the pre-defined traffic demand weight for mesh router $k$ as
$w_k$. The portal capacity via the final link $i$ is denoted as $P_i$. The final links are those
ones converged to the shaded area (the gateway) in Fig. 3-2. The aggregate portal
capacity without consideration of fairness can be expressed as $\sum_{i \in L_1} P_i$. Mesh router $j$’s
proportion of portal capacity via final link $i$ is denoted as $u_{ij}$

$$\max \sum_{i \in L_1} \sum_{j \in M} P_i u_{ij}$$

subject to:

*link capacity constraint*

*channel resource and radio resource constraint*

*interference (concurrent transmission) constraint*

*fairness constraint:*

$$\sum_{i \in L_1} u_{ij} P_i = \frac{w_j}{\sum_{k \in M} w_k} \sum_{i \in L_1} P_i$$

This formulated problem searches in all the possible resource allocation and
channel assignment for the maximum portal capacity with the fairness constraint. The
fairness constraint requires each mesh router share proportional amount of portal
capacity corresponding to its traffic demand weight $w_k$. The objective function embeds
$u_{ij}$ into each $P_i$ to add the fairness constraint. This is still an NP-hard problem even after
it removes the requirement on traffic demand. However, we can further simplify this
problem.

Figure 3-2. An Example of Link Distribution in WMNs and the Proposed Frequency
Reuse Pattern
3.3.3 Problem Simplification

We know that frequency reuse is critical to system’s performance because it can actually bring in more resource for allocation, especially when the system covers multi-hop distance. However, in the backbone of WMNs, a good frequency reuse pattern is difficult to find.

Usually in WMNs, the gateways are taken as the centers and mesh routers surround the gateways with different distances, most of the time counted in the number of hops. Due to the traffic aggregation, the closer to the portal the links are, the heavier traffic they carry. It is easy to observe that if a virtual circle is drawn in a WMN and the circle is large enough, links outside the circle can totally reuse the resource of the inside links, as shown in Fig. 3-2. The first reason of the feasibility of this frequency reuse is that outside links have less traffic load in total because all traffic needs to be aggregated in the gateway. The second reason is that the distance between frequency reuse links can be guaranteed by the proper frequency reuse distance \( D \). Finally, because of larger area and less traffic, this reuse can be repeated when the distance to the gateway is larger. Therefore, we can reduce the original problem to a problem with smaller area of interest. We term this circle with radius \( D \) as Circle \( D \). We call the mesh routers outside Circle \( D \) as outside mesh routers.

The criterion to discover the minimum Circle \( D \) is that all the links outside the circle can reuse the frequency resource used inside the circle. Usually the sensing range of a node is about twice as long as its transmission range (Simulator ns-2 adopts the typical value 2.2). Therefore, in this paper, we set this distance \( D \) equal to 3-hop coverage. It is clear that links outside the Circle \( D \) are always able to reuse some frequency resource used by some inside links with proper assignment. This frequency reuse separation remains valid no matter what the real topology is because of the reasons mentioned above.
Therefore, the former transformed problem can be reduced to a problem with fewer constraints.

\[
\max \sum_{i \in L_1} P_i \sum_{j \in M} u_{ij}
\]

subject to:

*link capacity constraint within Circle D*

*channel and radio resource constraint within Circle D*

*interference constraint within Circle D*

*fairness constraints:* \( \sum_{i \in L_1} u_{ij} P_i = \frac{w_j}{\sum_{k \in M} w_k} \sum_{i \in L_1} P_i \)

Note that we do not loosen the fairness constraints since each link within Circle \( D \) is required to carry the traffic from downstream routers as well. To unify the concerned area, the traffic demand weights of outside mesh routers are to be merged to the inside mesh routers. We assume that traffic from one outside mesh routers is bound to a certain inside mesh router on the boundary of Circle \( D \). Although this assumption means that the optimality is based on a certain fixed routing, it does not change the optimality when routing is changed. This can be easily demonstrated as follows. We first assume that each router has enough radios that can support as much as resource it has been allocated. Each link is allocated resource according to its weight. As long as the total weight is not changed, e.g., each mesh router has a fixed hop distance to the gateway, then each mesh router can always obtain the same amount of portal capacity. Otherwise, the optimality might be changed while the system performance would not have great impact in that this would not bring significant weight change. In reality, the number of radios cannot be as large as required. An improper routing does make some links over-loaded, which would affect the portal capacity. However, this difference rarely makes the non-optimality happen because the limited number of channels makes it impossible for a link to use more than one radio. Therefore, small variation of links’ weights does not change the system’s optimal performance significantly. In our solution,
we use the simple tree structure routing. Though this routing suffers from lack of robustness, dynamic adjustment mechanism can make up to this defect.

Based on this tree structure, the last hop links are aware of their load, the total traffic they are carrying. Therefore, the proportions of all the $P_i$ can be acquired. Hence, $\mu_{ij}$ can be known if multi-path routing within Circle $D$ is ignored. With the given tree structure and without considering frequency reuse, the simplified problem becomes a linear programming (LP) problem within Circle $D$ since the objective function and constraints are all linear. If the accumulated weights of the final-hop links change, a new global resource allocation and channel assignment is needed to regain the fairness.

### 3.4 Proposed Solution

Unlike previous papers related to the optimality, our solution is based on the standard of IEEE 802.11 family because the frame structures in 802.16 standards cannot support multi-hop very well even though the inherent scheduling mechanism can greatly reduce the contention overhead in IEEE 802.11. Note that we assume that the links between mesh routers and their local mesh clients use a different set of frequency bands so that they can be ignored in this paper and the network planning about this part can refer to the cellular systems directly.

From the problem formulation, we have found a new way to achieve WMNs’ optimality. Assuming that each wireless link within Circle $D$ has the same wireless condition and each wireless router has at least two radios with one for up-link and the other for the down-link and ignoring the frequency reuse, the aforementioned LP problem can be easily solved by allocating available bandwidth of given multiple channels to links among mesh routers according to their weights. Therefore, the key to the optimal portal capacity is the resource allocation and channel assignment for the links within Circle $D$.

For a practical system scheme, allocating channel resource to links within Circle $D$ to achieve an optimal aggregate portal capacity is not enough. Firstly, each outside
mesh router should have a way to acquire a certain amount of resource which supports the optimal portal capacity. Secondly, as aforementioned, outside mesh routers can totally reuse the spectrum resource within Circle $D$. A scheme is required to realize this frequency reuse which prevent the links within Circle $D$ from being interfered by outside links.

Finally, the achieved optimality requires the condition that each mesh router generates the preplanned traffic load. Real-time traffic varies from the expected value from time to time, thus the calculated resource allocation may not result in the optimal portal capacity. With a distributed and dynamic mechanism to adjust the allocated resource, when traffic varies the optimal portal capacity can still be maintained. Robustness is also gained through the distributed and dynamic scheme.

The overall proposed solution consists of centralized part and distributed part, which are discussed next, respectively.

3.4.1 Part I: Centralized Part

3.4.1.1 Overview

The optimal portal capacity is mostly determined by the resource allocation and channel assignment for the links within Circle $D$. The resource allocation here only means the allocation of bandwidth resource of given channels to the available radios of mesh routers. Though the difference of allocated resource may be limited by the number of available radios of considered mesh routers, this probability is rather low because, due to the limited number of channels, one link can rarely occupy more than one channel.

We propose a centralized scheme to carry out the task of resource allocation and channel assignment.

As aforementioned, with the assumption of uniform traffic distribution and identical coverage area among mesh routers, each mesh router contributes the same amount of traffic to the gateway. Therefore, the weight of each link can be defined as the number of
downstream mesh routers. Without the consideration of frequency reuse, the resource allocation is straightforward with the knowledge of total resource and links’ weights. A weighted allocation is enough for this task. Afterwards we need to face the challenges from frequency reuse, exact total resource and deviation of resource allocation.

Some notations are introduced beforehand. The channel capacity is firstly assumed as a constant $B$ and there are $K$ available channels. We denote the total weight of each tier as $\omega_I$, $\omega_{II}$, $\omega_{III}$, respectively. The total weight is $\omega_{total}$. The numbers of each tier mesh routers are denoted as $M_I$, $M_{II}$ and $M_{III}$, respectively. The number of all mesh routers is denoted as $M$.

Conservatively, when there is no frequency reuse within Circle $D$, a lower bound of portal capacity can be achieved by dividing the whole spectrum by the whole weights of the links. The lower bound is $\frac{K \cdot B \cdot M}{M + (M - M_I) + (M - M_{II})} \approx \frac{K \cdot B}{3}$, when $M_I, M_{II} \ll M$.

An upper bound is achieved when the 3rd-hop links can all reuse the frequency allocated to the preceding links, as shown in Fig. 3-3. To achieve this upper bound, the network topology and links’ weights should be roughly symmetric to the gateway. The most ideal aggregate portal capacity is $\frac{K \cdot B \cdot M}{M + (M - M_I)} \approx \frac{K \cdot B}{2}$ when $M_I \ll M$. 

Figure 3-3. Ideal Case When Frequency Reuse is Perfect
Therefore, the achievable value range for problem (3–2) is known to vary from $\frac{K \cdot \mathcal{B}}{3}$ to $\frac{K \cdot \mathcal{B}}{2}$. However, if the scheme is not well designed, the real system’s aggregate portal capacity can be far lower than these values. From this section on, we describe the scheme of exploiting the portal capacity of WMNs.

For an IEEE 802.11 channel, the maximal throughput $\mathcal{B}$ that can be achieved is not a constant. It depends on the number of contending nodes, i.e., the collision probability. If the channel’s maximal throughput is not the same for different links, the weighted resource allocation is not accurate, which may lead to under-utilization of portal capacity or non-optimal portal capacity. The real channel capacity with different number of contending nodes should be figured out and the capacity difference needs to be considered in the resource allocation. A revised weighted allocation is used for this purpose.

We first present the algorithm for the centralized resource allocation and channel assignment with the consideration of frequency reuse.

### 3.4.1.2 Centralized resource allocation and channel assignment

It can be easily observed that the difference of upper bound and lower bound of aggregate portal capacity comes from the difference of frequency reuse. Obtaining the frequency reuse as much as possible leads to the optimal aggregate portal capacity.

The most common way to allocate resource is the weighted allocation. $R_i = \omega_i \frac{R_{\text{total}}}{\omega_{\text{total}}}$, where $R_i$ stands for the allocated resource to link $i$ and $R_{\text{total}}$ as total resource. When one link can reuse other link’s resource, its weight $\omega_i$ (assumed not greater than the reused one’s weight $\omega_i$) is not counted again, thus the share $\frac{R_{\text{total}}}{\omega_{\text{total}}}$ will be increased due to a smaller total weight.

Based on the rule that frequency cannot be reused by two-hop neighbors [50], we can derive another two rules: *frequency of first tier links cannot be reused by 3rd tier links, nor can the 2nd tier be reused by 2nd tier*. Therefore, it is clear that the frequency reuse for the links within 3 hops is only possible between faraway 2nd and 3rd tier links.
or between faraway links of 3rd tier. From the engineering point of view, it is not difficult to figure out all the reused link pairs within Circle $D$ during the network construction. This information can be configured into gateways for the centralized resource allocation.

After allocating the proportional resource to each link, the gateway is required to assign the channels to the radios of each mesh router according to the allocated resource to each link. Because the number of channel is an integer, the channel assignment can make resource allocation deviate from the target proportion due to rounding operations. A fittest allocation strategy is used to mitigate this problem. The frequency-reusing links do not need to be assigned the channel resource because of previous assignment for the frequency-reused links.

Due to the limited number of channels, most links in WMNs have to share a channel with other neighboring links. This fact imposes extra difficulty on the frequency reuse. Under a contention-based protocol, if the frequency reused links share the channel with other links, the frequency reuse is not always feasible. An example is that channel assignment makes a first-tier link share a channel with its downstream 3rd-tier link. In this case, its downstream 4th-tier links cannot reuse the frequency resource of this first-tier link because it will bring undesired interference between the 3rd-tier link and the 4th-tier links.

When the frequency reuse cannot be fulfilled under a certain channel assignment, the resource allocation is required to re-calculate due to the change of total weight $\omega_{\text{total}}$. Therefore, the resource allocation needs to be considered with channel assignment jointly due to the uncertainty of frequency reuse. Our strategy is to compare the gain from frequency reuse and the cost to exclude other links from the frequency reused channel.

Resource allocation is firstly done with the knowledge of total weight $\omega_{\text{total}}$ and total resource $R_{\text{total}}$. During the channel assignment, when a frequency reused link is met, we attempt to exclude all the links that cannot be frequency-reused from the current
<table>
<thead>
<tr>
<th>Step</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Resource Allocation($\omega_{total}$)</td>
</tr>
<tr>
<td>2</td>
<td>Channel Assignment start</td>
</tr>
<tr>
<td>3</td>
<td>while(END != LINK)</td>
</tr>
<tr>
<td>4</td>
<td>if LINK == REUSE</td>
</tr>
<tr>
<td>5</td>
<td>Find the frequency reuse link set</td>
</tr>
<tr>
<td>6</td>
<td>if gain &gt; price</td>
</tr>
<tr>
<td>7</td>
<td>$\omega_{total} = \omega_{l} - price$</td>
</tr>
<tr>
<td>8</td>
<td>CHANNEL = REUSED$\rightarrow$CHANNEL</td>
</tr>
<tr>
<td>9</td>
<td>Resource Allocation($\omega_{total}$)</td>
</tr>
<tr>
<td>10</td>
<td>continue</td>
</tr>
<tr>
<td>11</td>
<td>else</td>
</tr>
<tr>
<td>12</td>
<td>Abandon Frequency Reuse</td>
</tr>
<tr>
<td>13</td>
<td>Clear frequency reuse mark</td>
</tr>
<tr>
<td>14</td>
<td>Wait for channel assignment</td>
</tr>
<tr>
<td>15</td>
<td>if RESOURCE &gt; 1</td>
</tr>
<tr>
<td>16</td>
<td>check the number of radio</td>
</tr>
<tr>
<td>17</td>
<td>and split RESOURCE to radios</td>
</tr>
<tr>
<td>18</td>
<td>if RESOURCE &gt; radios</td>
</tr>
<tr>
<td>19</td>
<td>error report</td>
</tr>
<tr>
<td>20</td>
<td>search the fittest CHANNEL</td>
</tr>
<tr>
<td>21</td>
<td>if(CHANNEL == UPLINK CHANNEL)</td>
</tr>
<tr>
<td>22</td>
<td>search again</td>
</tr>
<tr>
<td>23</td>
<td>LINK = LINK$\rightarrow$next</td>
</tr>
<tr>
<td>24</td>
<td>end</td>
</tr>
<tr>
<td>25</td>
<td>Channel Assignment end</td>
</tr>
</tbody>
</table>
channel and quantify this resource as a virtual weight $\omega_{\text{virtual}}$. The gain of this frequency reuse is the saved resource for these reused links, quantified by their weights $\omega_{l'}$. If the cost is less than the gain, we add this virtual weight $\omega_{\text{virtual}}$ to and subtract the gain $\omega_{l'}$ from $\omega_{\text{total}}$ and redo the resource allocation. Otherwise, the frequency reuse for these links is abandoned. This procedure is repeated until all the links have been assigned channels. Note that if more than one link can reuse the same frequency reuse link set, the cost and gain comparison should consider all of them. Table 3-1 is the pseudo code of centralized resource allocation and channel assignment with the consideration of frequency reuse.

As shown in Line 21, an important rule for the assignment is that the channel frequency for the up-link and down-link of each mesh router should be different. The constraint of the radio resource is also considered in Line 15. When there are more than one radio for down-link or up-link, the strategy is to spread the allocated resource evenly to each radio. This strategy considers the prospective dynamical adjustment so that when unused resource can be reallocated an extra channel assignment is not necessary. The channel assignment should be continuous, which means the same channel is preferred to be allocated to a close neighborhood. The purpose for this strategy is to avoid the hidden terminal problem as well as to enhance the frequency reuse probability.

3.4.1.3 Actual total resource for allocation

The portal capacity being discussed does not mean the nominal channel bandwidth. This value is defined as the achievable throughput across the gateways. It is used by the parent nodes to indicate the child nodes their feasible traffic rate. Each child node should follow this quota otherwise the parent cannot guarantee the delivery and the optimal portal capacity cannot be guaranteed, either. The allocated resource for each link equals to the portal capacity the child node inherits from the parent node, which will be shared by all the child node's downstream mesh routers.
Due to the traffic aggregation property of WMNs, the central links' weights are always greater than outside links'. Therefore, when weighted resource allocation is used, the central links usually get more resource than outside links. Therefore, in contention-based MC-MR WMNs, the number of sharing links closer to the gateways is much different from the one further away from the gateways. For a contention-based protocol, the channel's capacity depends much on the contention probability. When the number of links sharing the same channel is different, the actual capacity of this channel is different. According to the result of [71], this difference can be up to 10%~20%. If we use the nominal channel capacity as the resource for allocation, the central links's allocated resource will greatly exceed their target proportion. Eventually the optimality will be lost. To exploit the realistic optimal portal capacity, the first task is to find out the actual resource for allocation.

The saturation throughput of IEEE 802.11 channels has been given out in [10]. However, this throughput cannot be used to indicate the child nodes' traffic rate. The saturation throughput means the real delivery rate in 802.11 systems when each node has packets to send all the time. The traffic rate (arrival rate) of the senders is usually much larger than the delivery rate in this case. We do not want to use too big arrival rate to achieve the target delivery rate under saturation because an overstocked system is not preferred. A mapping from arrival rate to delivery rate under non-saturation is required for our purpose.

For analysis purpose, we assume that the arrival of packets for one mesh router is modeled as a Poisson process. The packet delivery of a station can be modeled as a queue system and the MAC service time is the service time of this M/G/1/K system when system is non-saturated. When the arrival rate is greater than the service rate, this system is overstocked, i.e., non-ergodic. We can find the biggest non-overstocked arrival rate in order to get the knowledge of achievable delivery rate, which is the value used to
indicate the child nodes. The following is the procedure of getting such actual channel capacity.

We denote the number of contenders in one channel as \( n_c \) and each contender is assumed to have the same packet arrival rate \( \lambda_c \). According to the result of [71], the MAC service rate \( \mu \) can be acquired as the inverse of the average service time \( \frac{1}{E(T_s)} \).

To make the queue system non-overstocked, the arrival rate should be less than the service rate.

\[
\lambda_c < \mu = \frac{1}{E(T_s)} \quad (3-3)
\]

The maximum arrival rate can be achieved when the equation holds:

\[
\lambda_c = \frac{1}{E(T_s)}
\]

The maximum acceptable traffic load of this channel can be achieved through this process. Thus can the maximum throughput is

\[
S = n_c \lambda_c (1 - p_B) (1 - p_c^{\alpha+1}) \quad (3-4)
\]

where \( p_B \) is the packets’ dropping probability due to finite queue length, \( p_c \) is packet discard probability in one transmission and \( \alpha + 1 \) is the retransmission time. The derivation can also be found in [71].

However, when each contender has different arrival rate, these values may change.

We rewrite the relationship between \( p_c \) and the conditional transmission probability \( \tau \) as

\[
p_c(j) = 1 - \prod_{i \neq j} (1 - (1 - p_0(i)) \tau_i) \quad (3-5)
\]

where \( p_0(i) \) is the probability that the wireless station \( i \) has no packet to transmit, which can be derived from \( \lambda_c(i) \). Meanwhile, from the Markov chain model, the conditional transmission probability can be derived as

\[
\tau_i = \begin{cases} 
\frac{2(1-p_c(i)^{\alpha+1})}{1-p_c(i)^{\alpha+1}+(1-p_c(i))W\sum_{k=0}^{m-1}(2p_c(i))^k}, & \alpha \leq m \\
\frac{2(1-p_c(i)^{\alpha+1})}{1-p_c(i)^{\alpha+1}+p_c(i)W\sum_{k=0}^{m-1}(2p_c(i))^k+W(1-2^m p_c(i)^{\alpha+1})}, & \alpha > m
\end{cases}
\]

where \( m \) is the maximum number of the stages allowed in the back-off procedure.
From these two equations, $p_c(i)$ can be figured out with the input of $n_c$ and $\lambda_c(i)$.

When $\lambda_c(i)$ are provided $p_c(i)$ can be figured out. Average MAC service time for each station can also be calculated. The total throughput is the summation of each station’s throughput which can be derived from $\lambda_c(i)$ and $p_c(i)$.

An extreme case of different arrival rate is that among $L$ contenders only one has intense traffic and the others transmits sparsely. The maximal throughput in this case approximates the one in case of $n_c$ equals to 1.

This time-varying channel capacity adds more difficulty on the centralized resource allocation task. However, it is shown that the unequal traffic pattern leads the maximal throughput to approximating the value with fewer contenders. Therefore, the equal traffic case is reasoned to have the most conservative channel capacity and this value is used as the input of our resource allocation scheme.

### 3.4.1.4 Revision of weighted resource allocation

With different number of contending nodes $n_c$, the maximum arrival rate can be acquired from the above process. We can use a factor $\eta$ to indicate the actual channel capacity as $\eta B$. Since the number of contending nodes for the same tier links does not vary greatly, we can assume $\eta$ for the same tier links to be the same. Factors for different tiers can be denoted as $\eta_I, \eta_{II}, \eta_{III}$ for tier 1, 2, 3, respectively.

The derivation of each tier’s average number of contending nodes ($C_I, C_{II}$ and $C_{III}$) can be obtained easily as follows.

\[
C_I = \frac{M_I + 1}{K \frac{\omega_I}{\omega_{total}}} \\
C_{II} = \frac{M_{II} + M_I}{K \frac{\omega_{II}}{\omega_{total}}} \\
C_{III} = \frac{M_{III} + M_{II}}{K \frac{\omega_{III}}{\omega_{total}}} 
\]  

(3–6)

Through the weighted allocation process, the allocated resource for 3 tiers is $\frac{\eta_I \omega_I}{\eta_I \omega_I + \eta_{II} \omega_{II} + \eta_{III} \omega_{III}}$, $\frac{\eta_{II} \omega_{II}}{\eta_I \omega_I + \eta_{II} \omega_{II} + \eta_{III} \omega_{III}}$, and $\frac{\eta_{III} \omega_{III}}{\eta_I \omega_I + \eta_{II} \omega_{II} + \eta_{III} \omega_{III}}$, respectively. The difference among
and $\eta_{III}$ makes the allocation greatly deviated from the target, implying that the actual resource allocated to 3 tiers should have the proportion of $\omega_I$, $\omega_{II}$ and $\omega_{III}$, respectively.

The solution to correct this deviation is to use $\frac{1}{\eta_I}, \frac{1}{\eta_{II}}$ and $\frac{1}{\eta_{III}}$ to modify the weights of links for different tiers. For example, $\omega_I = \frac{\omega_I}{\eta_I}$. Given that a link $i$ is in tier 1, the resource allocated to it is

$$R_i = KB \eta_I \frac{\omega_I}{\omega_{total}} = KB \frac{\omega_I}{\omega_I/\eta_I + \omega_{II}/\eta_{II} + \omega_{III}/\eta_{III}}$$

(3–7)

The total allocated resource is then

$$R_{total} = KB \frac{\omega_{total}}{\omega_I/\eta_I + \omega_{II}/\eta_{II} + \omega_{III}/\eta_{III}}$$

(3–8)

Through this correction, the target proportion of allocation can be attained.

### 3.4.2 Part II: Distributed Part

#### 3.4.2.1 Overview

The centralized part of our solution provides a straightforward way to optimizing the portal capacity of a MC-MR WMN. To fully fulfil this optimum value, a distributed mechanism is necessarily needed as well. This includes not only the distributed mechanism for outside mesh routers to handle the channel assignment and the portal capacity fulfillment, but also a dynamic adjustment mechanism for each router to dynamically and fully utilize the portal capacity.

There are a couple of reasons that the distributed mechanism is preferred in addition to the centralized part. First reason is that a distributed scheme can fit the dynamic environment better than a whole centralized one. For a centralized scheme, whenever this network is changed either in topology or traffic dynamic, an updating of channel assignment is needed. A distributed scheme can update the channel assignment in response to local minor changes. The second reason is the difficulty of frequency reuse decision in multi-hop scenarios. For centralized approach, to decide
the whole frequency reuse, it requires the center gateway to know all the knowledge of topology, the connectivity graph, which needs significant overhead to collect and maintain. A distributed scheme does not need this dependency and can cope with the frequency reuse according to the local knowledge and real-time information. The task of outside mesh router reusing the frequency resource of the links within Circle $D$ is also included in the proposed distributed scheme.

We have already demonstrated that when the central links have been allocated the channel resource appropriately, the outside tiers should have enough channel resource to realize their share of portal capacity. Given a mesh router's portal capacity, the task of its distributed channel assignment is to allocate its children their shares of portal capacity and assign them channels. The distributed algorithm can bring the network the adaptivity to possible minor local changes in traffic and/or topology, which should not cause an update across the whole network. The difficulties come from the acquirement of status information of all the channels and the coordination of channel assignment for neighboring mesh routers.

In addition, a dynamic mechanism is required to adjust the channel resource among different links after each of the mesh routers obtains its share of portal capacity and the assigned channels. Moreover, this adjustment can help the mesh routers recover from the link failures and link disadvantages.

3.4.2.2 Distributed channel assignment

Even with all the links within Circle $D$ acquiring their channel resource and fulfilling their portal capacity, the network still needs to address channel allocation for outside mesh routers in order to gain their shares of the portal capacity. We propose a distributed scheme for the outside mesh routers to accomplish their portal capacity.

Since the portal capacity of each mesh router is always passed down from its parent, the distributed channel assignment is determined to follow a top-down sequence. This means that the outside mesh routers can execute the distributed channel
assignment only after the centralized resource allocation and channel assignment have completed.

The centralized resource allocation is embedded in the procedure of the global channel assignment. Each mesh router waits for its parent to allocate the share of portal capacity and assign the channel resource for it. A certain channel is defined as the starting channel. Before channel assignment, all mesh routers reside on the starting channel waiting for the channel assignment from their parents or the gateway.

The major difficulty for a channel assignment scheme with fewer radios than channels is the acquirement of status information of all channels. Previous works use NAV (network allocation vector) knowledge collected from each of the channels to indicate the channel usage status for MANETs (Mobile Ad-hoc networks) \[64\]. However, per-packet-based channel assignment does not fit WMNs because of the relatively stationary topology and relatively constant flow in the backbone links. Instead, the channel assignment in WMNs is almost static except that the current requirement cannot be satisfied. For the above reason, the NAV knowledge is not necessary and a long-term indicator of the channel busyness level is more preferred in the channel selection for WMN routers. We specify a certain duration at the beginning of every \(T\) beacons on the starting channel as the channel busyness indicator phase (CBIP), as shown in Fig. 3-4. The CBIP is divided into \(K\) parts where \(K\) is the number of available channels. Each mesh router sets busy tones on the corresponding parts to signal the busyness status of its operating channels, with the duration length as the busyness level. Each mesh router chooses the most idle and non-conflicting channel for the links to its children. The busy tone’s length is always determined by the neighbor which senses the busiest channel status of this channel because the result of the addition of busy tones is the one with the longest duration. When a mesh router senses that some channel’s busy tone is not set, it can tell that there is no neighbor using this channel.
The indicator of the channel busyness level is determined by the following formula:

$$\delta = \left\lceil L \frac{T_{busy}}{T_{total}} \right\rceil \quad (3-9)$$

where $L$ is the number of mini-slots used in CBIP for one channel, $T_{busy}$ denotes the busy time duration of the given channel and $T_{total}$ denotes the whole time duration. The granularity of this indicator is decided by $L$, which cannot be too large. However, in a WMN, the outside mesh routers greatly under-utilize the channels because they reuse the frequency resource in a greatly larger area with significantly less traffic demand. Therefore, $L$ does not need to be large. For the mesh routers within Circle $D$, although this indicator is not useful because every channel is fully allocated, this mechanism is also incorporated in the mesh routers within Circle $D$ because of the neighborhood requirement from outside mesh routers and the requirement of a globally uniform protocol. Consequently, the centralized resource allocation is required to subtract this part of resource.

The algorithm for the distributed channel assignment can be described as follows.

After a mesh router (3 hops or further) gets its portal capacity and channel assignment from its parent (or the gateway), it calculates the portal capacity that each of its children can get after it subtracts its own usage. It collects the busyness status of each channel via CBIP and decides the channel assignment for its children afterward. For the case of more than one radio for down-link, the mesh router spreads the portal...
capacity to each available down-link radios for the convenience of prospective dynamic adjustment. After deciding the channel assignment or re-assignment, both parent and child start to set the busy-tone for this channel.

When the neighboring mesh routers carry out the channel assignment at the same time, it is possible that a certain channel is assigned by a lot of mesh routers and other channels are spared. In our algorithm, each mesh router waits for all the higher tier routers to finish their channel assignment before it starts its own. Among the mesh routers with the same tier number, the mesh router carrying the largest weight starts first.

The overhead of this mechanism is the CBIP and one extra radio for the busy tone in the starting channel.

### 3.4.2.3 Dynamic adjustment mechanism

Dynamic adjustment mechanism handles two types of changes: infrastructure-involved change and non-infrastructure-involved change.

If a mesh router’s share of portal capacity cannot be used up, its parent can let its siblings to share this unused resource. Within Circle $D$, the unused portal capacity can only be shared by links using the same channel because the channel resource is already fully allocated. Outside mesh routers have no such limit as long as they have vacant channel resource to fulfill the unused portal capacity.

This resource usage change is temporarily caused by traffic imbalance. This type of change has no impact on the relationship among mesh routers and is thereby defined as non-infrastructure-involved change. It also includes the case that a mesh router changes to another channel while keeping its share of portal capacity. The triggering condition of this change can be link disadvantage caused by hidden terminal problem, or poor link quality due to frequency-selective fading and interference.

We can show that with only non-infrastructure-involved change, the optimality of current assignment is not changed assuming that all the link status is identical. The
proof is straightforward as follows: as long as there is unused portal capacity due to some under-utilized links, it can be used up by other over-utilized links. Therefore, the optimum portal capacity will not lead to under-utilization unless the total traffic demand is below the aggregate portal capacity.

A mesh router can monitor other mesh routers of higher tiers. It can change its parent when its portal capacity cannot be fulfilled via current parent and there is an alternate parent who can provide a certain better share of portal capacity. This change involves the change of the relationship of parent-and-child. It belongs to the infrastructure-involved change. The case of a new router’s joining the system and the case of a router’s detaching its parent are also included in this type. The dynamic adjustment for infrastructure-involved change makes the system self-recoverable and might cause some unfairness in return. However, unfairness is the price for the robustness. Another global channel assignment can regain the fairness of the system.

For the sake of dynamic adjustment, a mesh router should broadcast its portal capacity and current usage ratio. A mesh router is required as well to broadcast its tier number, weight and the number of downstream mesh routers.

3.4.2.4 Infrastructure formation

When the system is being initialized, all the mesh routers carry out the association with the gateway on the starting channel. Each mesh router has a counter recording the number of downstream mesh routers and use this counter as the initial weight to get the share of portal capacity. In this initializing phase, each mesh router listens to the neighbors’ broadcasting messages and find the neighbor with the fewest hop counts to the gateway and the strongest signal strength as its parent. In this way, the initial tree is formed. This tree structure can be adapted to link failures or load imbalance dynamically. The centralized resource allocation is always based on the WMN’s current tree structure.
3.4.3 Part III: Support of Multi-gateway

In this paper, we only give a simple discussion of multi-gateway support with our solution.

Unlike other centralized approaches, our solution requires centralized resource allocation and channel assignment within Circle \( D \). Due to the smaller centralized area, supporting multiple gateways within one WMN becomes possible with our solution. When there are several gateways in the field, as long as the Circle \( D \) of each router is not overlapped, links inside each circle can achieve their portal capacity from the gateways and links outside can acquire portal capacity from any gateway distributedly. To guarantee the channel assignment of each circle having no interference with each other, the separation distance between each circle should be at least two hops away. Although there are more issues to be considered when apply our solution to multi-gateway WMNs, our solution provides a good option which many previous works do not support at all.

With multiple gateways connected to the Internet, a WMN can provide bigger portal capacity and larger coverage area.

3.5 Performance Evaluation

Previous works have not proposed concepts similar to portal capacity so that no knowledge is provided about how much channel resource can be transferred to the portal capacity in WMNs. In this paper we propose a feasible solution leading to the optimum portal capacity. Since there is no previous result similar to the portal capacity, we use two simple schemes based on current IEEE802.11 standards for comparison. The first scheme for comparison is a distributed single-channel scheme (SC). With this scheme, each node in the system contends for the total channel resource fairly. The second scheme for comparison is a distributed multi-channel scheme, in which each node randomly picks a channel and contends for transmission (RM). Using C program, we implement our proposed centralized algorithm of weighted resource allocation and
channel assignment (CW) to compare the portal capacity performance with these two schemes, SC and RM. For the purpose of comparison, all three schemes are applied to topologies limited in Circle $D$. We also assume that there is no frequency reuse in all three schemes and there is no geographical disadvantage among all links. For both compared schemes, SC and RM, portal capacity comes from the aggregation of the maximum throughput over the last hop links. In both distributed schemes, since each node contends for the channel resource without considering the portal capacity share of its downstream nodes, the fairness has no guarantee. Thus a fairness metric is included in the simulation as well.

This simulation set $K$, the number of available channels to 12 for our solution and the distributed multi-channel scheme; $\eta_1$, $\eta_2$, $\eta_3$ to 1.0, 0.8, 0.6, respectively. The channel bandwidth $B$ is ignored in the simulation since it is a constant. Therefore, the allocated resource and portal capacity can be quantified by values ranging from 0 to 12. We compare the value of the realized portal capacity among our centralized weighted allocation approach (CW), and the compared schemes, SC and RM.

Fig. 3-5 shows that in SC case, only a small proportion of total channel resource is transferred to portal capacity, or end-to-end throughput. RM scheme can get a better
performance. With the channel assignment overhead ignored, the gain of RM is roughly the same as $K$, which is still relatively low. Our solution (CW) can transfer about $1/3$ of the total resource to portal capacity, which is roughly 200% more than RM.

As mentioned above, each mesh router is designed to be able to deliver the same amount traffic to the gateways generated by itself. If each link does not acquire the allocated resource according to the target proportion of weight, the portal capacity cannot be fairly shared by all mesh routers. Therefore, we use $X_i$, the ratio between the allocated resource and the corresponding weight to calculate the fairness index.

$$\zeta = \frac{(\sum X_i)^2}{n \cdot \sum (X_i)^2}$$  \hspace{1cm} (3–10)

If the resource allocated to each link matches the corresponding weight, with the existence of admission control and message exchange between mesh routers, the fair share of portal capacity can be easily reached. With the proper setting of the weight for each link, our solution achieves the best fairness as shown in Fig. 3-5.

If the three schemes are applied to larger topologies, due to less-planned resource allocation, SC and RM will have much poorer performance than our solution.

The second part of our simulation is focused on the frequency reuse issue. Frequency reuse can bring more actual channel resource and the improved portal capacity. This part of the simulation shows the portal capacity improvement owed to frequency reuse in our solution. In our solution, centralized resource allocation only considers frequency reuse within Circle $D$ because outside mesh routers can reuse the resource inside the circle. Therefore, frequency reuse inside Circle $D$ determines the optimum frequency reuse of the whole system, which can be easily figured out through the field measurement.

We test the portal capacity with ideal frequency reuse using our algorithm under the scenarios of different numbers of outside mesh routers. For comparison, the portal capacity without frequency reuse is also provided.
Fig. 3-6 shows the portal capacity improvement using ideal frequency reuse. From the first part of Fig. 3-6 we can see a significant increase (roughly 40%) of the proportion of the portal capacity, which means the ratio of the real portal capacity and the real allocated channel resource not including the overhead and the unallocated channel resource. When the number of outside mesh routers changes, there is a slight decrease of both portal capacity proportion and real portal capacity. The reason is that when the number of mesh routers increases, the weights of second- and third- tier links increases. Therefore, more channel resource will suffer from smaller $\eta$.

### 3.6 Summary

This paper attempts to investigate the capability that a WMN provides its mesh clients to deliver/receive packets across the gateways. This capability can be quantified by the defined portal capacity with a certain fairness constraint. By formulating and simplifying the target problem, we propose a new solution to exploiting the portal capacity of WMNs. The centralized part of the solution provides the optimality of portal capacity. The distributed part maintains this optimality when system has local minor changes and it brings robustness to the system as well. In addition, this solution makes the support of multi-gateway possible, which can significantly improve the system capacity.
4.1 Background

The surprisingly poor performance of multi-hop wireless networks has attracted more and more attentions in the literature. During recent years, new transmission techniques are sprouting quickly. However, the traffic rate in multi-hop wireless networks is not increasing accordingly. Usually, when the scale of the networks becomes large, due to the increasing interference and the increasing number of intermediate hops of flows, the end-to-end throughput performance starts to deteriorate. Gupta and Kumar theoretically characterize this in [22]. However, for practical multi-hop wireless networks, such as WMNs, WSNs and some battle-field ad hoc networks, there is plenty of room to improve the throughput performance since the current poor throughput performance of these networks is mainly due to the inefficiency of spectrum usage. Most of recent research works are focusing on the spectrum efficiency for this reason.

As the de facto standard of most of the multi-hop wireless networks, IEEE 802.11 was originally designed for the single-hop Wireless LANs. Its performance in multi-hop scenarios is much below our expectation due to its blindly-contending and mechanically forwarding properties [69] [72]. This random access property of IEEE802.11 is one of the major reasons for the inefficient resource usage. Scheduling-based protocols, like IEEE 802.16, can provide better spectrum efficiency because it does not require nodes to contend for the channel before each transmission with the assumption of relatively constant traffic flows. However, since this type of protocols needs a fixed frame format which is vulnerable to the scheduling conflict, it is not suitable for the multi-hop wireless networks.

Random access MAC provides a roughly fair mechanism for wireless nodes to access the medium. The effort of differentiating the uplink and downlink resource allocation has been first applied to WLANs in [33] because of the observation that as
the central point, APs should occupy more resource than other nodes. To achieve better performance in multi-hop networks, several previous schemes attempt to break the fairness by prioritization, [70], [66], [31]. These schemes heuristically search for better spectrum sharing mechanism among wireless nodes, by differentiating the forwarding priority according to the priority tags of packets or flows. However, when the traffic pattern is more complicated, these schemes cannot guarantee significant performance improvement.

On the other side, with centralized approaches, scheduling-based MAC can allocate the resource in a more efficient way. This approach can find the optimal solution with knowledge of the topology and traffic when the network is not large. However, for large-scale networks, this approach becomes infeasible due to the NP-hardness and the difficulty of information collection. Therefore, the distributed scheduling approaches are proposed to address this dilemma, [40], [47]. These two approaches give us the insight of how good performance the networks can achieve. However, they always require a perfect scheduling, a MAC with no collision and no hidden/exposed terminals, which is almost impossible in multi-hop wireless networks. Previous distributed scheduling schemes also ignore the multi-hop nature of flows in multi-hop networks, which causes a lot of wastage in spectrum allocation. Moreover, within a neighborhood range, different neighbors can sense different condition of channels, resulting in potential conflicts of distributed scheduling, which is difficult to solve in distributed scheduling schemes.

In this paper, we propose a new MAC with a different spectrum allocation mechanism based on IEEE802.11. Similar to previous distributed approaches, the efficient spectrum usage of this scheme comes from the collection of neighbors’ traffic information. Different from previous works, our scheme collects traffic dependency information from neighbors as well. By addressing “asymmetric neighborhood” and “traffic dependency” issues, this paper gives a comprehensive way to improve the throughput performance in multi-hop wireless networks. Upon the observation of efficiency of spectrum allocation, a
new metric called “Allocation Inefficiency Ratio (AIR)” is introduced for better evaluating the proposed scheme.

The rest of this paper is organized as follows. Section 6.2 introduces the related works. Section 4.3 discusses the spectrum usage issues in multi-hop wireless networks. Section 4.4 describes the proposed schemes. Section 4.5 provides the evaluation study for our scheme. Finally, conclusion is drawn in Section 4.6.

### 4.2 Related Works

In the literature, there are two types of research aiming at improving the performance of multi-hop wireless networks distributedly. One is the heuristic approach based on the existing random access MAC, such as IEEE802.11. The other assumes a perfect MAC and gives out distributed scheduling algorithm. The former approach proposes practical ways to differentiate the medium access probability, thus a better spectrum usage can be achieved. The latter one takes the traffic load of each neighbor as input and tries to optimize the overall throughput performance in a distributed manner. We start the survey with the first approach.

In [70], forwarded packets are given higher priority than upstream packets by a shorter IFS, thus the packet accumulation at the forwarding nodes is alleviated and the delivery failure due to forwarding congestion can be reduced. Yang and Vaidya tried to ensure medium access for high priority source stations in their priority scheduling scheme [66]. Two narrow band busy-tone signals are used to ensure medium access for high priority nodes. Kanodia et al. further proposed a distributed priority scheme to differentiate the different packets’ priority in transmission [31]. Accordingly, the forwarded packets have their priority increased for the same reason as addressed in [70]. All the above works attempt to improve the performance by introducing differentiation among different nodes, different packets, or different flows. However, such differentiations cannot improve the overall performance of the system remarkably when the traffic and topology become more complicated or when flows cannot be prioritized. IFA scheme
proposes another way of spectrum usage allocation [18] based on the random access MAC. The neighbors’ traffic information are collected and the spectrum resource is allocated accordingly. By throttling input traffic to its system wide fair time share, random access MAC can greatly improve the fairness and throughput performance. However, the traffic demand of each node simply uses the value of arrived traffic and only parking lot topologies are analyzed.

The distributed scheduling approach focuses on the overall throughput performance analytically since they are derived from centralized scheduling algorithms. The scheme in [29] requires each node to collect each neighbor’s queue-length information and use a probability $\alpha$ to transmit its packets. The probability $\alpha$ is derived from the relationship of each node’s queue-length with its neighbors’. This paper proves that when this policy of distributed scheduling is applied, the largest capacity region $\Omega$ can be achieved with an efficiency ratio $\gamma$. Lin’s work has taken multi-channel and routing into consideration besides each neighbors’ queue-length [40]. According to the calculated contention cost, radio cost and congestion level, which are derived from neighbors’ queue-length, channel condition and nodes’ other information, each node decides the assignment of packets to different channels and different slots. This paper also shows that the provided distributed algorithm is provably efficient, which means that a provable fraction of the maximum system capacity can be achieved.

Although these works provide theoretical results of overall throughput performance, the assumptions of a perfect MAC and the lack of a real protocol support prevent them from transiting the theoretical results to a practical protocol. Moreover, when all queues are full and the traffic is backlogged due to over-injection of traffic, these distributed scheduling algorithms cannot address the congestion problem.

Furthermore, the analysis of these works do not consider the multi-hop nature of flows. They consider the traffic load (queue length) independently while the inherent traffic correlation among neighboring nodes is ignored. Similar to [18], they take the
arrived traffic as the traffic demand. We will discuss this ignored issue in the later section of “traffic dependency”.

No matter which approach is concerned, the purpose is to efficiently utilize the spectrum resource. The reason of the collection of neighbors’ information is to let each node efficiently share the spectrum resource. For multi-hop wireless networks, only neighbors’ traffic load information is not enough for the ideal spectrum usage allocation.

4.3 Spectrum Usage Issues in Multi-hop Wireless Networks

4.3.1 Ideal Spectrum Usage

Assuming a random access MAC, we hereby give an example of the spectrum inefficiency caused by fair access to the medium in multi-hop wireless networks. Fig. 4-1 shows 3 flows in a simple 4-node 2-hop topology. Flow 1 and 2 share the same forwarding node, Node 3. With IEEE802.11 protocols, Node 3 can share $\frac{1}{3}$ of the channel capacity like Node 1 and Node 2, with Node 4 only receiving. In this case, a part of the channel capacity that Node 1 and Node 2 hold is wasted because of Node 3’s delivery limit. Consequently, we can expect that when the traffic consists of stable and continuous flows, adjusting each node’ medium access can bring significant improvement of the throughput performance.

Usually the performance of multi-hop wireless networks can be evaluated by two metrics: fairness and end-to-end throughput. Suppose the flow demand vector is \( \{F_1, F_2, F_3, \ldots, F_N\} \), where \( N \) denotes the number of existing flows. With a certain
scheme, each flow can achieve a flow rate according to the flow allocation vector
\( \{ f_1, f_2, f_3, \ldots, f_N \} \). A perfectly fair allocation means the proportion of each allocated flow rate to its demand is identical, as discussed in [18]. However, this does not guarantee an optimum allocation since the aggregate end-to-end throughput is not considered. Using max-min delivery ratio as the criteria as in [15] includes the end-to-end throughput into consideration. In this paper, the delivery ratio is defined as follows.

\[
\zeta_i = \frac{f_i}{F_i} \quad (4-1)
\]

The objective is:

\[
\max \{ \min_{i \in F} \{ \zeta_i \} \} \quad (4-2)
\]

Let us consider Fig. 4-1 again. When legacy IEEE802.11 is applied, each active node obtains the same share of wireless channel. The overall end-to-end throughput of this small network is roughly 1/3, with each flow taking 1/9 equally. When scheduling is used and no fairness is concerned, the maximum overall end-to-end throughput can reach 1 when Flow 3 takes all the spectrum resource. However, the minimum flow rate is 0, with flow 1 and 2 totally starved. If we apply the max-min criterion to this simple topology, we can obtain the ideal spectrum usage allocation simply by observation. The maximum overall end-to-end throughput is 3/5, with each flow taking 1/5 and the remaining 2/5 taken by the first hop of Flow 1 and Flow 2. Each flow obtains the same flow rate 1/5, thus the max-min goal is achieved.

For complicated topologies, it is not easy to find the ideal allocation by simple observation since more complicated interference and frequency reuse need to be considered. When the knowledge of the whole topology and flow information is given, there are many algorithms to find this ideal allocation while NP-hardness is a big obstacle for the solutions. The distributed scheduling proposes to utilize the queue-length or traffic load information of neighboring nodes as the input of distributed scheduling algorithm [40], [47], [29]. The distributed scheduling is in nature another
type of spectrum allocation. The intuition is straightforward in that each node can get its spectrum share based on the proportion of its traffic load to the total traffic load within the interference range.

However, most of the distributed algorithms have a hidden assumption that each node’s traffic load is independent, which does not hold when multi-hop flows exist.

4.3.2 Traffic Dependency

When distributed algorithms are applied in multi-hop ad hoc networks, as in previous papers [40], [47], [18], [29], the traffic demand of each node is assumed to be the local traffic input. However, this traffic input depends on not only the arrival traffic from local upper layer entities, but also the forwarding requirement from the upstream nodes, which depends on current scheduling or spectrum allocation. Therefore, using the local traffic as the input of the algorithm can make the spectrum allocation deviated from the ideal one. Previous papers ignore this traffic relationship in their algorithms, which we term it as “traffic dependency”.

We use Fig. 4-1 as the example to illustrate the problem of ignoring traffic dependency. Suppose Node 1, 2 and 3 have the same original traffic load, valued as 1, and Node 3 needs to forward the traffic from both Node 1 and 2. Suppose each node collects only the arrival traffic information of each neighbor and allocates the spectrum accordingly. The initial arrival traffic rate of each node is assumed to be the result of random access. The result turns out to be far deviated from the ideal result, with each of Node 1 and 2 occupying 3/11 of the channel capacity and Node 3 occupies 5/11. We can see that Node 3’s allocated resource cannot even cover its forwarding requirement, and thus a big wastage is created. This allocation deviation comes from the ignorance of traffic dependency of neighboring nodes.

Traffic dependency comes from flows’ multi-hop delivery property. Li has raised similar concerns of this “inherent correlation” of upstream and downstream sub-flows in [39]. For centralized algorithms, the central point is assumed to have all the
knowledge of existing flows. Therefore, all flow and topology information, thus the traffic dependency information, can be acquired and processed at the central point. For distributed algorithms, the information mentioned above is difficult to obtain for individual nodes. Fortunately, traffic dependency information can be obtained by the information exchange of forwarding request from upstream nodes (node-based information), which bypasses the flow-based approach.

### 4.3.3 Allocation Inefficiency Ratio

In multi-hop networks, end-to-end throughput can only indicate the absolute value of deliverable traffic given the traffic pattern and topology. It cannot indicate the performance of applied schemes because this value depends on the traffic pattern and topology heavily. We need a metric which can characterize not only a system’s absolute performance but also the efficiency of the scheme applied. To better evaluate how the spectrum usage allocation is, we propose a new metric, called allocation inefficiency ratio (AIR). It is defined as the ratio between the air-time of the dropped traffic amount due to inefficient spectrum allocation and the total air-time allocated
due to inefficient spectrum allocation and the total air-time allocated

\[ \text{AIR} = \frac{\sum_{i \in N} T_i \frac{T_e}{PHY_i} - \sum_{j \in F} \sum_{k \in K_j} T_e \frac{T_e}{PHY_{j,k}}}{\sum_{i \in N} T_i \frac{T_e}{PHY_i}}, \]

where \( N \) is the node set, \( T_i \) is the outflow rate of node \( i \), \( F \) is the flow set, \( K_j \) is the hop set of flow \( j \), \( T_e \) is the end-to-end rate of flow \( j \), and \( PHY_{j,k} \) is the physical layer rate of flow \( j \) at hop \( k \).

Apparently, this metric gives us the knowledge of how much allocated spectrum is wasted. This wastage mainly comes from allocation discrepancy at different hops for one flow. A good spectrum allocation scheme should have an AIR with the value of 0. We can verify this claim according to the AIR calculation in above examples. The AIR value in the schemes in [47, 29, 40, 18], which ignore traffic dependency, is \( \frac{8}{33} \). This means \( \frac{8}{33} \) of the allocated spectrum is wasted because the corresponding traffic amount originally transmitted is not finally delivered.
4.3.4 Asymmetric Neighborhood

Due to multi-hop topology, each node has overlapping and non-identical neighborhood which may make information asymmetric. This information asymmetry is sometimes undesirable to a distributed scheme. For example, some node senses a relatively idle channel and decides to use its demanded traffic rate. However, some of its neighbors sense the channel in busy status instead. Therefore, conflicts are introduced when distributed scheduling or distributed allocation is implemented. Previous works have not provided solutions for this problem. Incorporating 2-hop neighbors’ traffic information will not help to solve this problem because the sensed total traffic load is not exchanged and can be different from node to node.

4.4 Proposed Scheme

4.4.1 Overview of The Proposed Scheme

For practical ad hoc networks, like WMNs or WSNs or other military ad hoc networks, the traffic is not totally ad hoc. Usually traffic aggregates at some points or areas with certain patterns. In this paper, we assume that the multi-hop wireless networks which we concern have certain traffic patterns and the flows inside have relatively stable traffic load. For these ad hoc networks, we design an efficient MAC which can utilize the limited spectrum resource in a more efficient way.

The basic procedure can be described as follows. Each node is required to broadcast its traffic demand to its neighbors, which is the same as previous works. Meanwhile, each node is required to notify its neighbors about the traffic dependency between them, which differentiates our work from others. Afterwards, each node allocates the spectrum individually according to the information collected and apply the calculated traffic rate to its transmission.

Traffic dependency information from different neighbors affects the estimation of accurate traffic load in different ways. Three different roles of neighbors are defined in this paper. When one neighbor has traffic for the current node to forward, we name this
role as *upstream neighbor*. Similarly, when the current node has traffic for its neighbors to forward, these neighbors are called *downstream neighbors*. Other neighbors are called *uncorrelated neighbors*. It should be noted that only when traffic demand of neighbors are correlated with the current node, are these neighbors to be seen as upstream neighbors or downstream neighbors. Therefore, one node sees the neighbors who have traffic ending at itself as uncorrelated neighbors because the traffic ending at itself will not affect its traffic demand.

The traffic demand from each node consists of two parts: the traffic that requires to be forwarded from its upstream neighbors and the traffic originated from its upper layer locally. It can be expressed by the following formula:

$$TD_i = TDO_i + \sum_{j \in N_i} TD_{fwding,j,i}$$

(4–4)

where $TD_i$ is the traffic demand, $TDO_i$ is the traffic originated from local upper layer, $TD_{fwding,j,i}$ is the traffic that requires to be forwarded from its upstream neighbor $j$ and $N_i$ is the set of neighbors for node $i$. The latter part is dependent with its neighbors’ traffic demand and the former part is not. Therefore, an accurate traffic demand of one node should be based on the knowledge of all upstream neighbors’ traffic dependency information. In this scheme, upstream nodes should notify their downstream neighbors about their forwarding request. Consequently, the downstream nodes update their traffic demand accordingly. The knowledge of accurate local traffic demand is not enough for ideal spectrum allocation. It is also important to acquire the correct traffic demand of the neighbors, $TD_i$. When downstream nodes broadcast their new traffic demands, since the downstream neighbors’ traffic includes forwarding requirement from the upstream nodes, the upstream nodes should be able to extract the dependent traffic from the messages, thus the pure change of the original traffic of the downstream nodes can be known. This knowledge is important in obtaining the accurate traffic demand of neighbors. Since each node considers both one-hop upstream neighbors’ and one-hop
downstream neighbors’ traffic dependency information, we name this scheme “2-hop MAC”.

It is obvious that due to complicated traffic patterns, each neighbor can play different roles simultaneously. However, we keep these terms in this section to illustrate the scheme clearly.

In this scheme, one node’s traffic change will affect the traffic demand of correlated neighbors. The change of one node’s traffic demand is passed to other nodes as if there is no channel limit. In this way, the traffic demand in this scheme can reflect the true traffic demand of the neighborhood. Therefore, the spectrum allocated to one packet in the current hop will also be allocated to this packet in other hops, and thus the bandwidth waste due to allocation discrepancy is reduced from the beginning of the allocation phase. In the spectrum allocation phase, the channel limit comes into play to give each node the identical allocation ratio, and thus the fairness is well-addressed.

There exists another type of traffic dependency when CSMA/CA is applied, because the receiving nodes need to send CTS or ACK messages upon receiving packets from the sending nodes. In our spectrum allocation, we count this part into sending nodes’ spectrum usage.

If each node in the neighborhood has the same information as each other, the distributed spectrum allocations should be identical. In multi-hop scenarios, it is common that neighbors have different neighborhood, thus the distributed spectrum allocations are probably different. If the total spectrum allocation does not exceed the channel limit, the distributed algorithm can be acceptable even though some fairness is sacrificed. However, spectrum allocation conflicts usually exist with the asymmetric neighborhood. The proposed scheme uses a feedback mechanism to regulate the neighboring nodes’ traffic rate from over-injection.

When a single rate is used, fairness requires the real traffic rate of each node is proportional to the ratio between its traffic demand and the total traffic demand. For
multi-rate support, the fairness is based on air-time as previous papers, such as [30]. Each node’s traffic load is expressed by the traffic rate. Based on the links’ achievable PHY rate, this traffic load can be converted to air time, which is used for fair spectrum allocation in local area. We ignore how the rate adaptation mechanism is implemented in this paper. We also simply assume that during one period each node uses one constant PHY rate in one link to its neighbors. The proposed scheme requires the air-time fairness among neighbors, which means the air-time fraction of each node’s allocated traffic rate is proportional to the ratio between its air-time of traffic demand and the summation of the air-time of each node’s traffic demand. This can be expressed as Formula 4–5.

\[
\frac{R_i}{\text{PHY}_i} \leq \frac{\text{Channel Limit}}{\sum_{j \in N} \frac{T_{Dj}}{\text{PHY}_j}}
\]

(4–5)

\( R_i \) and \( \text{PHY}_i \) denote the determined traffic rate and physical layer rate for node \( i \), respectively. We assume the channel limit in multi-rate environment to be a constant in this paper.

4.4.2 Scheme Description

In a real network, one node can play different roles of neighbors, downstream, upstream or uncorrelated neighbors, for its neighbors, due to the co-existence of different flows. Therefore, the proposed scheme assumes each neighbor as all possible types of neighbors. The messages from each neighbor are processed with a uniform procedure.

In this scheme, each node maintains three tables which record its own traffic information, its neighbors’ traffic information and the traffic dependency information. Each node periodically gets knowledge of its original traffic load and forwarding traffic load from its upper layers and updates these tables. It also updates these tables when it receives/overhears messages from its neighbors. Each node calculates the achievable traffic rate for itself and its neighbors distributedly, with the algorithm which will be
presented later. Regulation indicators are used to eliminate the over-injection of traffic due to allocation conflicts.

We denote the maximum normalized IEEE802.11 one-hop MAC throughput as $\eta_C$, which also means the maximum busy air-time fraction ($\eta_C = \frac{R_{\text{PHY}}}{P}$). The overhead incurred by the rate adaptation is assumed to be negligible. Related works, such as [71], can provide ways on how to determine the value of $\eta_C$.

In the following subsections, we first present the supporting data structure and messages’ format. We then describe the detailed procedure and algorithm.

4.4.2.1 Data structure and message format

Each node should maintain a parameter set as described in Table 4-1 which includes its own traffic information:

Table 4-1. Local Parameter Set

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>TDO</td>
<td>The traffic demand originated from upper layer locally</td>
</tr>
<tr>
<td>$R$</td>
<td>The node’s allocated and broadcasted traffic rate</td>
</tr>
<tr>
<td>$R^*$</td>
<td>The node’s adjusted traffic rate after regulation</td>
</tr>
<tr>
<td>TD</td>
<td>The node’s traffic demand</td>
</tr>
<tr>
<td>Status</td>
<td>The status parameter used to indicate different message requests</td>
</tr>
</tbody>
</table>

Each node should also maintain a table which records its neighbors’ corresponding information as the potential input of distributed spectrum allocation. The detail information is listed in Table 4-2.

The traffic dependency information is stored in an $(n + 1) \times (n + 1)$ matrix $D$, with the local traffic demand included. $D_{i,j}$ means the forwarding requirement from node $i$ to node $j$. When $i$ equals to $j$, $D_{i,j}$ stores the original traffic excluding its forwarding demand from its neighbors in this neighbor set. Note that this original traffic demand may not purely be original. In node $i$’s storage, if neighbor $j$ has some forwarding request from its
neighbor k, which is not a neighbor of node i, neighbor j’s original traffic demand in the matrix includes this part of forwarding request. Obviously, we have the relationship as follows:

\[ TD_i = \sum_{j \in N} D_{i,j} \]  

\[ (4-6) \]

\( TD_i \) in Formula 4–6 means the broadcasted version of \( TD_i \). Throughout this paper, superscript \( \prime \) stands for the broadcasted version and superscript \( * \) stands for the regulated version. Apparently, matrix \( D \)’s storage has some overlapped information with the former two tables. The difference is that matrix \( D \) only stores the broadcasted version and the other two tables gather the latest information.

The broadcasted messages should contain the following information: the traffic demand of the node \( (TD) \), the traffic load/rate \( (R) \) and the adjusted traffic rate \( (R^*) \) of the node, a list of traffic that the host needs the neighbor to forward \( (TDtfw_i) \), an indicator that a traffic regulation is necessary or this regulation is to be deactivated. There is one more parameter, unit demand allocation \( \varepsilon \), is implied in the broadcast messages.

\[ \varepsilon = \frac{R}{TD} \]

\[ (4-7) \]

This value is used to eliminate the conflicts due to asymmetric neighborhood, whose usage will be illustrated in a later section.

Table 4-2. Neighbor Parameter Set

<table>
<thead>
<tr>
<th>ID</th>
<th>Neighbor Address, as the ID of each record.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Status</td>
<td>Neighbor Status: Active or inactive</td>
</tr>
<tr>
<td>( R_i )</td>
<td>Neighbor i’s traffic rate</td>
</tr>
<tr>
<td>( R_i^* )</td>
<td>Neighbor i’s regulated traffic rate</td>
</tr>
<tr>
<td>( PHY_i )</td>
<td>Neighbor i’s physic layer rate</td>
</tr>
<tr>
<td>( TDtfw_i )</td>
<td>Forwarding requirement to neighbor i, the latest value</td>
</tr>
</tbody>
</table>

neighbor k, which is not a neighbor of node i, neighbor j’s original traffic demand in the matrix includes this part of forwarding request. Obviously, we have the relationship as follows:

\[ TD_i = \sum_{j \in N} D_{i,j} \]  

\[ (4-6) \]

\( TD_i \) in Formula 4–6 means the broadcasted version of \( TD_i \). Throughout this paper, superscript \( \prime \) stands for the broadcasted version and superscript \( * \) stands for the regulated version. Apparently, matrix \( D \)’s storage has some overlapped information with the former two tables. The difference is that matrix \( D \) only stores the broadcasted version and the other two tables gather the latest information.

The broadcasted messages should contain the following information: the traffic demand of the node \( (TD) \), the traffic load/rate \( (R) \) and the adjusted traffic rate \( (R^*) \) of the node, a list of traffic that the host needs the neighbor to forward \( (TDtfw_i) \), an indicator that a traffic regulation is necessary or this regulation is to be deactivated. There is one more parameter, unit demand allocation \( \varepsilon \), is implied in the broadcast messages.

\[ \varepsilon = \frac{R}{TD} \]

\[ (4-7) \]

This value is used to eliminate the conflicts due to asymmetric neighborhood, whose usage will be illustrated in a later section.
4.4.2.2 Procedure

The tables are updated according to local cross-layer notification or neighbors notification as mentioned above. The accurate traffic load of each node in the neighborhood is calculated according to these tables. Spectrum allocation is executed with the knowledge of traffic load and channel limit. Meanwhile, a check for regulation indicator or deactivation indicator is necessary to avoid the allocation conflicts due to asymmetric neighborhood.

For upstream neighbors, traffic demand is as broadcasted since no other nodes in the neighborhood can affect this value. However, local node’s traffic demand is affected by its upstream neighbors. The local traffic demand calculation can be found in equation (4–4). For downstream neighbors, the traffic demand depends on local traffic change. It can be calculated as follows:

\[ TD_{\text{down}} = \sum_{j \in N} D_{j,i} - TD_{\text{tfw}}, + TD_{tfw} \]

(4–8)

where \( TD_{\text{tfw}} \) can be found in matrix \( D \) according to the current node index (If it is calculated by Node \( k \), \( TD_{\text{tfw}} = D_{k,i} \)). Since each node can play different roles simultaneously, the overall traffic demand can be calculated in this way:

\[ TD_i = TDO_i + \sum_{j \in N, j \neq i} D_{j,i} - TD_{\text{tfw}}, + TD_{tfw} \]

(4–9)

If local node index is not \( i \), \( TDO_i \) uses the value of \( D_{i,j} \). We can see that \( TD_{\text{tfw}} \) changes when upstream neighbors’ traffic changes, which embodies the traffic dependency. Through this way, the accurate traffic demand can be calculated by each node locally and the ideal spectrum allocation can be fulfilled.

The asymmetric neighborhood information can cause different spectrum allocation at different neighbors, leading to spectrum usage conflicts. When one node senses the total spectrum allocation by individual neighbors exceeding the channel limit, a regulation indicator is sent out. The regulation mechanism in this scheme uses the
parameter of $\varepsilon$. This parameter stands for the uniform allocatable traffic air-time per unit traffic demand in air-time form. Each individual node allocates the traffic rate based on the air-time fairness, which means $\frac{R_i}{TD_i}$ is identical for all nodes within its neighborhood from its point of view. When it senses that the conflicts happen and the total channel limit has been exceeded, if each of its neighbors can follow its $\varepsilon$ by applying $\varepsilon \times TD_i$ to their traffic rate, the over-injection can be regulated. When the node senses the total traffic demand decreasing and the regulation is no longer necessary, a regulation deactivation indicator is required to recover the neighbors' traffic rate. Fig. 4-2 shows the flow chart.

Before every data transmission, if parameters in the record have been updated, a distributed spectrum allocation algorithm is carried out. When the node starts to transmit, it piggybacks its broadcast information in RTS or DATA packets. After broadcasting, the newly calculated values update the old record. If $TD_i$, $R_i$ and $TD_{tfw}$ are not changed since last broadcasting, no information is attached to the data. The algorithm executed before broadcasting is shown in Fig. 4-3.

The gross spectrum allocation can be briefly described as follows. If the total traffic does not exceed the channel traffic limit $\eta_C$, the node uses its required traffic rate.

Figure 4-2. Flow Chart for Message Processing
Figure 4-3. Flow Chart for Spectrum Allocation

Otherwise, the node uses the traffic rate in proportion to the channel traffic limit. The basic allocation algorithm can be expressed in the following equation.

\[ R_i = \begin{cases} 
  TD_i, & \text{if } \eta_C \geq \frac{TD_i}{PHY_i} + \sum_{k \in N, k \neq i} \sum_{j \in N} \frac{D_{j,k} - TD_{tfw_k} + TD_{tfw_k}}{PHY_k} \\
  \frac{TD_i}{PHY_i} + \sum_{k \in N, k \neq i} \sum_{j \in N} \frac{D_{j,k} - TD_{tfw_k} + TD_{tfw_k}}{PHY_k} \cdot \eta_C, & \text{if } \eta_C < \frac{TD_i}{PHY_i} + \sum_{k \in N, k \neq i} \sum_{j \in N} \frac{D_{j,k} - TD_{tfw_k} + TD_{tfw_k}}{PHY_k}
\end{cases} \]

If no information has been changed since last broadcasting, the spectrum allocation is not needed. If the total traffic demand does not exceed the channel limit, the spectrum allocation is not needed, either, though the traffic information change should be broadcasted.
It is worthwhile to note that although this scheme requires the modification of legacy IEEE802.11 messages’ format and the regulation outgoing traffic, the algorithm is not complicated and it is feasible to upgrade the current 802.11 protocol to our scheme.

4.5 Performance Evaluation

There are two common metrics to judge the performance of a multi-hop MAC scheme, the end-to-end throughput and the fairness. In the proposed scheme, the spectrum sharing follows the proportions of each nodes’ traffic demand. Although this fairness is based on nodes’ traffic, the flows’ fairness can be achieved when each node allocates its spectrum share based upon the flows’ traffic proportions. We will evaluate the flows’ fairness index in the following simulation. The end-to-end throughput is an important performance metric of multi-hop networks. However, it is not a good metric to evaluate the design of a scheme for the reason described above. Additionally, we adopt the AIR evaluation in this study.

As we mentioned above, the end-to-end throughput performance depends on the topology and the traffic pattern heavily. The end-to-end throughput varies when the topology or the traffic pattern is slightly changed. In this simulation, we construct 4 different and typical 3-hop (4-hop in the last topology) topologies for evaluation, as shown in Fig. 4-4. Since the benefits of the proposed scheme is more prominent when the traffic pattern is less homogeneous, we abandon the commonly-used grid topology. The circle nodes have original traffic while the rectangular nodes only forward traffic for the neighbors. The elliptical nodes only receive traffic. Consequently, there are both long-hop (3-hop or 4-hop) flows and short-hop (2-hop) flows in the constructed networks. In the last topology (topology (d)), there is a case for bi-directional flows. The original traffic demand is marked under each node. To illustrate the spectrum allocation result clearly, the traffic demand is set to exceed the channel limit, which is assumed to be 11 Mbps. The traffic demand’s unit is in kbps.
For comparison purpose, we apply two other schemes to these topologies. One is random fair access MAC (RFA) and the other is a distributed scheduling scheme (DSS). RFA lets each node access the medium randomly, thus fairly (assumed). DSS requires each node collect the neighbors’ traffic demand, however, with no consideration of traffic dependency and allocate spectrum accordingly. The traffic demand in this scheme is simply taken as the traffic arrival rate. Since the traffic arrival rate depends on a predefined scheduling output, we set the initial scheduling as fair medium access. Although the first two schemes lack of mechanisms to overcome the asymmetric neighborhood problem, in this simulation, we assume there exists a regulation mechanism which can regulate the traffic within each neighborhood so that it will not exceed the channel limit. The first two schemes require each node to deliver the traffic from different sources according to the arrival proportions. The overhead of all 3 schemes is ignored. The fairness index in the simulation is calculated in the following...
way:

$$\zeta = \frac{(\sum X_i)^2}{n \cdot \sum (X_i)^2}$$  \hspace{1cm} (4–10)

where $X_i$ is the ratio between the realized flow rate and the corresponding flow rate requested for flow $i$.

Figure 4-5. Comparison of end-to-end throughput (a), fairness index (b) and AIR (c) for RFA, DSS and 2hopMAC schemes

From Fig. 4-5, we can see the performance comparison for RFA, DSS and 2-hop MAC. In this figure, the benefits of end-to-end throughput and fairness of 2-hop MAC are shown by the comparison with RFA and DSS. Part (a) compares the end-to-end throughput performance. The proposed scheme has achieved significant gain in every topology setting. It is interesting to note that the end-to-end throughput of RFA is better than DSS in Topology (a), (b) and (c). The reason is that DSS does not consider the multi-hop property of the traffic, thus the spectrum allocation can deviate far from the ideal case. For Topology (d), because there is a center-like point, DSS can allocate the spectrum better than RFA. Note that the regulation mechanism for asymmetric neighborhood problem is assumed for RFA and DSS, and thus the realistic performance should be worse. The other comment for this simulation is that the traffic demand is intentionally set so that the total traffic demand barely exceeds the channel limit and each flow has identical demand. When the traffic demand of each flow and the
difference of flow's traffic demand increase, the throughput performance gain of the proposed scheme can be higher because each node has more different traffic demand and a poor spectrum allocation can cause more significant throughput degradation.

Part (b) shows the fairness performance comparison. The fairness among different flows is guaranteed only with 2-hop MAC. Since each node is assumed to deliver the traffic according to the proportion of the traffic from different sources, nodes which do not know the real traffic demand from different sources but the arrival traffic amount, cannot allocate the spectrum usage fairly according to their knowledge.

End-to-end throughput is not a good indicator of system’s performance in the sense that it cannot tell how good a scheme is. It depends greatly on the traffic demand of each node as well as the flow distribution in the networks. AIR is introduced as the indicator of how good the spectrum is allocated using the given scheme. If AIR reaches 0, the scheme allocates the system’s spectrum to the serviced nodes without any wastage, which means the packets obtaining the spectrum to deliver at their first hop have obtained the virtually reserved spectrum for their following hops. There is no doubt that when the available spectrum is allocated without wastage, the efficiency of the spectrum allocation achieves its best performance. However, due to the complexity of topology and traffic distribution, non-zero AIR is usually inevitable. Part (c) shows the AIR comparison among 3 schemes, through which we can observe the spectrum allocation efficiency of each scheme. It can be seen that both RFA and DSS waste a lot of spectrum they allocate. Since DSS uses inaccurate traffic information to allocate spectrum, its AIR is higher than RFA’s in some scenarios. For 2-hop MAC, AIR is almost 0 in all four topologies. This result is obvious in that each neighbor incorporates the exact traffic information from its neighbors and pass this information to the 2-hop neighbors via these neighbors. Through this method, the knowledge of accumulated traffic demand for each node in the neighborhood is correctly acquired. Note that in Topology (d), the AIR has a non-zero value because of the regulation for asymmetric
neighborhood problem. This cannot be avoided when the distributed algorithm is applied since asymmetric neighborhood always exists in multi-hop wireless networks.

4.6 Summary

Throughput performance is always a key issue in multi-hop ad hoc networks. Distributed spectrum allocation/scheduling algorithms are commonly applied in the multi-hop networks to improve the efficiency of the spectrum usage and thus improve the poor throughput performance. However, without considering the multi-hop nature of flows, the spectrum allocation can have significant wastage especially for long hops flows. In this paper, we propose 2-hop MAC scheme. This scheme incorporates multi-hop consideration into spectrum allocation so that the spectrum allocated for one hop transmission will not be wasted due to lack of spectrum at the next hop. The asymmetric neighborhood, which is usually ignored in previous papers, is discussed and addressed in this paper. The end-to-end throughput performance and fairness improvements are investigated in several typical multi-hop scenarios.
CHAPTER 5
A MOBILITY MANAGEMENT SCHEME FOR WIRELESS MESH NETWORKS

5.1 Background

Over the past few years, wireless mesh networks (WMN) are gaining growing interest. This trend follows the popular needs for the inexpensive, continuous wireless wide-area coverage. A seamless wireless access is a common goal of the future communication.

Akyildiz et al. [5] have proposed a few models of WMNs. Usually, WMN consists of various types of entities: gateways, mesh routers, access points (AP) and mesh clients. Gateways are the connection points to the wire-line networks. Mesh clients are the terminal users which have no or limited routing function. Wireless APs are the entities in charge of the wireless access for the mesh clients. Stationary mesh routers form a wireless multihop backbone with long-range high-speed wireless techniques such as WiMAX. In different models, a mesh node can contain one or more functional entities, e.g., mesh routers usually implement AP functionalities.

When the mobile clients are stationary, with the support of backbone routing, the wireless access for them can be accomplished within a few hops. However, difficulty arises when there are needs for the mesh clients to move across the coverage area of different APs. How to maintain the ongoing connection and how to forward the downstream and upstream packets are not solved by the current standards. IEEE 802.16e adds amendments to the original standard to support mobility, but only specifies MAC and PHY layer [2]. IEEE 802.11s attempts to extend the WiFi to support the mesh mode and provide mobility support, which is still under development.

Mobility management is not a new topic in other existing networks. Akyildiz et al. [4] presented a survey on this topic. In cellular systems, this part has already been a critical part to the continuous service of the mobile clients. Handoff quality is one of the most indispensable testing items in each field trial test. However, wireless mesh networks,
which lack of infrastructure such as HLR and VLR, face more challenges in mobility management. Mobile IP is an approach which provides mobility support to mobile clients with IP identity [49]. The main idea is very similar to the HLR/VLR mechanism in cellular systems. Home Agent (HA) and Foreign Agent (FA) play the roles of home database and visiting database in the IP networks, respectively. Home address is used as the ID of a mobile client and the Care-of-Address (CoA) is used to locate the current position of the moving mobile clients.

Mobile IP can provide a solution to the inter-domain movement in WMNs. However, it is not suitable for the intra-domain movement, which is much more frequent than the inter-domain movement. The reason is that if FA is implemented in every AP, signaling cost and handoff latency become the major problems to the mobility support. Therefore, the solution to cope with the local movement is required. Protocols for IP micro-mobility have been proposed to solve the mobility dilemma in small-scale networks [23], [27], [52], [59]. Though these protocols can be applied to WMNs, heavier signaling cost and longer handoff latency due to more frequent local movement in WMNs still impede the practical mobility support.

In this paper, we propose a mobility management scheme in WMN, termed Mesh Mobility Management (M3). Some features of WMNs, such as multi-hop, mesh topology and continuous coverage, have been taken into consideration to better support the IP micro-mobility in WMNs.

The rest of this paper is organized as follows. Section II discusses some related works. Section III describes the proposed scheme. Performance analysis is carried out in Section IV. Conclusion is given at the end.

5.2 Related Works

In this paper, we focus on the mobility management within one WMN, which can be regarded as a micro-mobility issue. However, we search for a solution feasible with or without the Mobile IP support.
Not many related works of mobility management can be found in the literature of WMNs. Ganguly et al. [54] mentioned the mobility management issue in their comprehensive work. The experiment results confirm that handoff latency using a tunneling scheme is much longer than that using flat routing. However, since mobility management is not the focus of this paper, the authors discuss only the feasibility of mobility support and do not include detailed analysis. In SMesh [7], multiple APs monitor the moving mobile clients to achieve seamless handoff. This scheme eliminates the handoff latency at the price of high signaling cost.

However, previous works on IP micro-mobility are possible to be applied to WMNs, since WMN can be treated as one type of mobile IP networks. We now review some IP micro-mobility protocols.

In Cellular IP [59], mobile clients use the gateway’s IP address as their CoA and each router in this domain use the home addresses of the mobile clients to route the downstream packets. The default routes for each router to the gateway are used to direct the upstream packets.

HAWAII is another important framework of IP micro-mobility [52]. The CoA of each mobile client in HAWAII is a unique IP address allocated by the gateway of the domain. Different from the Cellular IP, HAWAII uses the CoA of each mobile client to route the downstream packets. This difference makes HAWAII less coupled with Mobile IP protocol and also enables the per-flow QoS support in the backbone network.

In both schemes, each domain is identified by a single gateway and the entire domain is constructed to a tree-like structure. Both schemes require each router to maintain a routing entry for each mobile client in the downstream APs’ coverage. When handoff occurs, the corresponding routing entries will be updated in all the routers involved from the new AP to the crossover router which is shared by the new AP and old AP. The invalid routing entries in the routers of the old path need to be removed. Due
to the major feature of per-host routing, this type of schemes is called **mobile-specific routing approach** [11].

Another important type of IP micro-mobility protocols is the **hierarchical tunneling approach** [11], an example of which is Mobile IP Regional Registration (MIP-RR) [23]. Hsieh et al. [27] proposed another scheme, namely, Hierarchical Mobile IPv6. This type of schemes replaces the mobile-specific routing by introducing the tunneling technique. Through the hierarchical registration procedure, the higher-level FA knows the location information (ID of the lower-level FA) of the mobile clients and encapsulates the data packets with the destination address of this lower-level FA. Per-host routing entry is not required for the routers in these schemes while per-host location information is still stored in FAs. Due to the extra processing of encapsulation and decapsulation as in Mobile IP [49], larger delay is introduced to each flow. Additional cost of this type of schemes is that two or more CoAs have to be used. When handoff takes place, the registration with a different CoA also adds extra delay. The intuitive idea of this approach is to extend the Mobile IP mechanism to local movements.

### 5.3 \( M^3 \) Description

#### 5.3.1 Model Description

In this paper we model the WMN with multiple mobile clients, one gateway, multiple routers with AP’s functionality (called “AP” hereafter) and their covering area (called “cell” hereafter). The case of more than one gateway can be easily derived from this paper. Each AP has the functionality of AP, router and database for the subscriber information.

The gateway is required to assign a unique IP address in its domain to a mobile client when it is powered up. This unique IP address of a mobile client can be the CoA when mobile IP is provided for the inter-domain roaming. The foreign agent (FA) and home agent (HA) can reside in the gateway. In the scenario where more than one
gateway present, our scheme can be easily extended by placing the FA/HA at the intersection of the gateways and using different IP address pools for each gateway.

We use a 3-level hierarchical structure to illustrate our scheme, as shown in Fig. 5-1. The three APs connecting to the gateway have superior status than their downstream nodes. They are required to collect the location information of the mobile clients in the cells of the subordinate APs. We name these APs “superior routers (SR)” hereafter. The rest of the APs have equivalent status. SRs act as the delegates of the gateway and share the signaling traffic. In a smaller mesh network, if the gateway is not the bottle neck, these superior routers can be removed which yields only 2-level hierarchical structure.

As discussed in [53], a WMN can be constructed in a tree-like structure. Each router has its only parent node and may have a number of children. This kind of modeling has its benefit for the routing where only the traffic flows between the gateway and each mobile client are considered. This model shows its limitation when the mobility management is taken into account. The tree structure is extracted from the real geographical topology based on the criterion of the shortest path from each AP to the gateway, which cannot be used to obtain the optimal path between any two geographical neighboring APs. The routing of previous schemes strictly follows the tree structure even when there exist shorter paths. Unlike other WMN models, our scheme allows the communication along the paths which are not in the tree. We assume that most of the time, geographically adjacent APs have shorter communication paths other than the only path along the tree. Therefore, this structure embodies a mesh topology.

We assume that the routing in the backbone (APs, superior routers and the gateway) has been set up. Since the backbone nodes in WMNs are mostly stationary, this assumption is reasonable. The remaining problem is on ensuring a mobile client to move around in this area without incurring high packet loss, long handoff latency and high signalling cost to the system.
5.3.2 Proposed Solution

5.3.2.1 Power-up

We know that for a mobile client, the subscriber information should include authentication, authorization and accounting (AAA) information and QoS profiles. If every AP in the domain maintains a copy of all the mobile clients' subscriber information from the gateway, the network will be less scalable and difficult to administrate. In our scheme, when a mobile client is powered up the authentication procedure should be fulfilled before an IP address is allocated to this client according to the subscriber information in the gateway. The gateway activates the record of this mobile client and records the location information hereafter. The serving AP keeps a copy of the subscriber information to avoid frequent visiting the database in the gateway.

Database of each AP contains only the subscriber information of the mobile client currently in the cell. Database of each superior router additionally contains the location information of all the mobile clients residing in the subordinate APs’ cells.
5.3.2.2 Handling downstream packets

The downstream packets, in which the destination address is not the AP’s address, cannot be routed by the intermediate superior router and APs without routing entries. In this scheme, tunneling technique is used to forward the downstream packets. These packets are attached with extra IP headers in which the destination address is the destination AP’s address. Upon receiving these tunneled packets, the destination APs decapsulate and forwards them to the addressed mobile clients in the cells.

In Fig. 5-1, the bold lines illustrate the downstream process, with the dashed lines and solid lines indicating the routing part and the tunneling part, respectively. From the gateway (GW) to the SRs, the packets are routed according to the location information. The other routing part in downstream forwarding will be discussed in section 4).

5.3.2.3 Handling upstream packets

For the upstream packets, the tunneling is not needed. The APs can use the default routes to forward packets to the gateway.

5.3.2.4 Handling handoff

Handoff occurs when the mobile client moves to a new AP’s cell. Upon receiving a handoff request message from the moving client indicating the former AP’s ID, the new AP sends a handoff request message to the former AP. The former AP sends back the corresponding subscriber information to the new AP after receiving the handoff request message. Meanwhile, it adds a temporary entry in its routing table with the destination address of this mobile client. A timer with length $T_r$ is started. After the timer expires, the routing entry and the corresponding subscriber information will be removed from the former AP. If the downstream packets are decapsulated by the former AP but the addressed mobile client is not found in the cell, these packets are routed to the new AP using the temporary routing entry. To guarantee that this routing can reach the new AP, each router on the path from the old AP to the new AP is required to add this routing entry. When the mobile client moves again, the chain of the downstream routes
continues to be concatenated. The similar idea of this chain-like structure for the location update has been used by HMIP [43] and POFLA [44] for different applications. In cellular system, this method is called “pointer forwarding” and we borrow this name [44].

Suppose mobile client A moves from position 3 (A(3)) to position 4 (A(4)) in Fig. 5-1, the downstream packets are first tunneled to AP3. AP3 forwards the decapsulated packets to AP5 according to the “pointer” of mobile client A. Upstream packets from A are routed to gateway by the default routes of AP5, AP7 and SR3, sequentially.

To prevent the encapsulated packets from passing the final destination AP to a former AP, the attached IP header can include the mobile client’s IP address in the option field. Therefore, when the encapsulated packets reach the final destination before the end of the tunnel, the AP of final destination can decapsulate them instead of simply passing them on.

5.3.2.5 Periodic location update

Adding the temporary routing entry is not the final solution to mobility management. Triangular routing problem is introduced by this method [49]. HMIP uses the idea that after a certain number of the hops the mobile client triggers a location update to the HA [43]. Our scheme uses the time interval to be the triggering condition. We assume that the mobile clients in the WMN are not so fast that during the period of $T_{lu}$ they cross less than $N_{hndf}$ APs. Thereafter, once every $T_{lu}$, the mobile clients can trigger the location update to control the triangular routing problem. To make this update more efficient, we let each AP, instead of each mobile client, be the initiator to trigger this update. Each AP reports the current set of mobile clients to the superior routers. The superior routers select another interval $T_{hu}$ to periodically update the set to the gateway. $T_{hu}$ is obviously required to be no less than $T_{lu}$.

After this periodic location update, downstream packets can be tunneled to the AP where the mobile client exactly locates without traversing all the APs the mobile client has visited.
It is necessary to consider the case that if the time instant when each AP involved in a mobile client’s handoff reports to its superior router is different, the downstream packets’ routing might not follow the exact shortest route. The solution is simple. If \( T_r \) is longer than the intervals of location update \( T_{lu} \) and \( T_{hu} \), the downstream packets can always find a path to the destination, which might not be the shortest. These extra hops can be expected less than \( N_{hndf} \).

5.3.3 Extended Discussion of The Protocol

Our scheme has combined the tunneling and per-host routing techniques which are the major features of previous two approaches [11].

Having compared these two types of approaches, we now discuss their pros and cons. By using the mobile-specific routing, the necessity of encapsulation/decapsulation is eliminated, and vice versa. The reason that mobile-specific routing cannot be applied to macro-mobility is the difficulty to find a crossover router which can maintain the mobile-specific routing entry. In other words, the scalability problem makes it infeasible. Moreover, this approach highly depends on the routing protocols. Another problem of this approach is pointed out in [11], which is, when update messages are lost due to physical reasons such as radio black-out, the routing entries in different routers might be inconsistent. Maintenance signaling might be an addition to guarantee the consistency. For the hierarchical tunneling approach, if the number of hierarchical levels are not small enough, the encapsulation/decapsulation will cause the delay performance intolerable. However, if the number of hierarchical levels are small enough, the signaling cost of handoff and handoff latency may be instead intolerable.

Our scheme achieves the advantages of both previous approaches. Tunneling the downstream packets in the backbone lower the routing requirement for each intermediate APs. Without the multiple-level registration procedure in the hierarchical tunneling approach, our scheme achieves shorter handoff latency. Consequently, the packet loss problem is greatly alleviated. A simple buffering technique can eliminate
the packet loss without the out-of-order problem of packet forwarding. On the other hand, applying the per-host routing only between geographical neighboring APs does not require each AP to maintain too many intermediate routing entries. This “pointer forwarding” method significantly reduces the location update to the gateway despite the extra periodic location update which is introduced to control the triangular routing problem. The delay of downstream packets is controllable due to the controllable triangular routing.

The features of $M^3$ can be related to some features of WMNs. Continuous coverage is the reason for applying the mobile-specific routing in the last few hops. The stationary characteristic of WMNs’ backbone yields simpler backbone routing. The hierarchical structure and the simple backbone routing render tunneling more appealing. Low speed of mobile clients limits the delay of downstream forwarding in this scheme. Moreover, due to the controllable delay, the periodic location update can be applied without side effects. Unlike the strict hierarchical structure of cellular system, mobility management under the physical mesh structure of WMNs can be realized in a more flexible way.

5.4 Performance Analysis

In what follows we focus on the benefits obtained related to the most important two factors in mobility management: signaling cost and handoff latency. We compare these two factors between our scheme and two previous approaches [11]. As we mentioned in the previous section, the encapsulation/decapsulation introduced by the tunneling brings us extra delay to downstream packets. This is the price when we want to use less per-host routing. However, this delay is not significant and should be tolerable because encapsulation and decapsulation only happen once for each downstream packet. Therefore, our performance analysis will not include this part.

The signaling cost is defined as the total amount of the extra signaling traffic due to the mobility. The handoff latency is defined as the service disruption time of the mobile clients due to handoff.
The signaling cost includes two major parts: update signaling cost during handoff and other maintenance signaling cost. The maintenance signaling in our scheme is the periodic location update signaling. To simplify the comparison, we calculate the signaling cost incurred during one period $T_{lu}$, with $T_{hu} = T_{lu}$ to further simplify the procedure.

In different mobility management schemes, the update procedure always starts from the new AP/BS to some anchor point, such as VLR/HLR or HA, then from the anchor point to the old AP/BS. If we assume the update signaling procedure is the same in different schemes, the cost depends only on the update path length, which is the focus of our analysis.

Let $n$, $m$, $\rho$, $C$, $C_u$, $\overline{PL}$ denote the number of mobile clients, the number of APs, the handoff times per mobile client during one period $T_{lu}$, the total signaling cost, the update signaling cost during handoff, and the average signaling path length during handoff, respectively.

$$ C = C_u + C_{maintenance} \tag{5-1} $$

$$ C_u = \rho \cdot n \cdot \overline{PL} \tag{5-2} $$

For our scheme, $\overline{PL}_{Ms}$ depends on the network planning and the geographical separation of neighbors. The worst case of this value occurs when all the available communication paths are those existing in the tree. However, usually in WMNs, the coverage is continuous and the geographical neighbors are interconnected to guarantee the connectivity so that each mobile client can reach the gateway in a limited number of hops. Therefore, $\overline{PL}_{Ms}$ can be assumed to be 1.

Next, we procure $\overline{PL}_{ps}$ of the previous schemes and evaluate the gain by using our scheme under different traffic situation with different $\rho$. 

109
In previous schemes, the communication always takes place along the tree paths. The average length from one node to its neighbors depends closely on the tree structure, such as the hierarchical level and the average number of children. Let \( k \) denote the average number of children and \( l + 1 \) the number of hierarchical levels as (including the level of gateway). We obtain the following equalities.

\[
\begin{align*}
    m &= \sum_{i=0}^{l} k^i \\
    &= \frac{k^{l+1} - 1}{k - 1} \\
\end{align*}
\]  

(5–3)

Excluding the links in the tree, the number of which is \( m - 1 \) (excluding the gateway), we define the handoff paths to be the sequential sibling visiting. The number of neighbor handoff paths with \( 2 \cdot j (j \in [1, l]) \) hops can be expressed as follows.

\[
\begin{align*}
    n_{2j} &= \sum_{i=1}^{l-j+1} (k - 1) \cdot k^{l-1} \\
    &= k^{l-j+1} - 1 \\
\end{align*}
\]  

(5–4)

Therefore, assuming that the handoff of each case is equal likely, the average path length can be expressed as:

\[
\overline{PL}_{ps} = \frac{2 \cdot \sum_{j=1}^{l} j \cdot (k^{l-j+1} - 1)}{(m - 1) - \frac{l(l+1)}{2}}
\]  

(5–5)

The gain of our scheme is defined as \( g = \frac{C_{ps}}{C_{m^3}} \cdot C_{maintenance} \) in our scheme is the signaling cost of periodic location update, which is in the order of \( m \). Assume \( n \) is much larger than \( m \), with a factor of \( a = n/m \). The gain \( g \) will be approximately the ratio of \( \frac{\overline{PL}_{ps}}{\overline{PL}_{M^3}} \) when \( \rho \) becomes large enough. When \( \rho \) is very small, the periodic update becomes the major signaling cost of handoff. The trend of \( g \) under different \( \rho \) can be shown in Fig. 5-2.
Figure 5-2. Gain of $M^3$ over previous schemes

We now discuss the average path length. In the literature, the maximum number of hops is not recommended to be large, with 4 or 5 preferred. The average number of children $k$ should be relatively small in order to avoid the performance bottleneck. Table I shows the typical value set of the average path length.

Table 5-1. Typical Value Set of Average Path Length

<table>
<thead>
<tr>
<th>parameters</th>
<th>$k=2, l=3$</th>
<th>$k=3, l=3$</th>
<th>$k=2, l=5$</th>
<th>$k=3, l=5$</th>
<th>$k=m, l=1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>average path length</td>
<td>4</td>
<td>2.9</td>
<td>4.21</td>
<td>3</td>
<td>2</td>
</tr>
</tbody>
</table>

If there are normal routers (without AP functionality) in the network, the handoff between parents and children becomes less possible. Therefore, the average path length for the previous schemes will be larger. The gain of $M^3$ will be higher.

This signaling reduction is mainly for the SRs and the gateway of the WMN, which are the bottleneck most of the time.

Handoff latency can be expressed as:

$$D = D_{hndf−detc} + D_{u−path−upd} \cdot \overline{PLi}$$

(5–6)

where $D_{hndf−detc}$ is the delay of handoff detection, $D_{u−path−estb}$ is the unit-step delivery delay in path update and $\overline{PLi}$ is the average length of path update.
Handoff detection is out of the scope of this paper. Therefore, we ignore the difference of $D_{\text{hndf-detc}}$ in different schemes. For the same reason as $\overline{PL}_{M^3}$, $\overline{PL}_{M^3}'$ can be assumed to be 1. For previous schemes, since the average position of the crossover routers is in the middle of the signaling path, we have $\overline{PL}_{ps}' = 0.5 \cdot \overline{PL}_{ps}$. Commonly, according to the typical value set of $\overline{PL}_{ps}$, our scheme performs better in handoff latency than previous schemes.

5.5 Summary

$M^3$ is proposed to meet the requirement of lower signaling cost and shorter handoff latency. This scheme utilizes some of the characteristics of the WMN to combine the two previous types of approaches. Consequently, it mitigates their shortcomings and achieves the advantages of both.
6.1 Background

Traditional cellular networks are typical one-hop wireless networks since only the last hop transmissions are wireless between mobile terminals and base stations. The existing infrastructures of cellular systems make seamless connections for mobile clients possible via handoff management. In the last few years, due to the intensive development in wireless ad hoc networks and the multi-interfaced mobile devices such as smart phones, multi-hop cellular networks (MCNs), the integration of ad hoc mode into cellular systems, have started to emerge. The origin of this new type of networks comes from the intention that the new cellular networks should provide an all-IP platform, where new types of wireless networks, e.g., WiFi and WiMAX, and the traditional cellular networks can interwork seamlessly with variety of services. 3GPP has been already engaged in the study of this trend [21] and the specification of the interworking between WCDMA networks and WLANs was standardized [20]. It is now well-known that these wireless ad hoc networks, usually operated in the ad hoc mode, can be configured more flexibly with much lower cost and potentially higher data rate due to much shorter transmission range, though the coverage and mobility may not be guaranteed. Consequently, it is expected that MCNs can take advantages of both traditional cellular networks and wireless ad hoc networks.

A MCN is commonly considered as a cellular-based integration of a cellular network and ad hoc networks because ad hoc mode is treated as an additive component to cellular systems, which does not impair the traditional cellular infrastructure, which means that mobile users can access BSs through either direct (one-hop) connections or multi-hop connections. With the cellular-based MCNs under consideration, the first advantage over the traditional cellular systems we observe is the low cost of extending service coverage. Multi-hop wireless relaying can easily provide faraway
wireless terminals with connections without building extra infrastructures (adding more BSs). Since the relay devices are usually much more portable than BSs, the flexibility in configuring more service connections is the second advantage we can expect. Moreover, this flexibility creates new ways, other than the traditional channel borrowing [61], to implement load balancing among different cells. Finally, with the ad hoc links bearing potentially higher data rate with extra spectrum, the system capacity can be expected to improve. Of course, we realize that the hidden assumption is that mobile terminals, in addition to the operator added ad hoc relaying stations, also participate in relaying and helping, which may not be practical if there is no incentive for the helping mobiles. Therefore, certain incentive protocols should be designed to stimulate mobile terminals to help the MCNs to operate. This topic will be investigated in a separate paper.

In this paper, we focus on the benefits that the ad hoc mode can bring to the handoff management. Handoff management is an important part of mobility management in Personal Communication Services (PCS) systems and any wireless cellular system in general in maintaining seamless connection [16]. Handoff dropping (HOD) rate is the major metric for assessing handoff performance and it usually requires the designed system to keep the HOD lower than certain threshold in order to meet the customers’ satisfaction. In wireless cellular networks, HOD rate is usually guaranteed by bandwidth reservation for handoff calls in cells a mobile user most likely visits during its call connection. More bandwidth reservation can indeed meet the HOD rate requirement, but at the price of blocking more new call connections because less resource is made available for new call connections. Thus, a good handoff management should not only keep the HOD rate low, but also tie up less bandwidth. As a supplement to cellular systems, ad hoc links (links via relaying by using harvested either inband or outband resource opportunistically) create more opportunities for MNs to connect to BSs. Therefore, HOD rate is expected to decrease with ad hoc mode introduced into the
cellular systems. Moreover, since handoff calls may be connected to adjacent BSs and thus use adjacent cells’ reservation, each cell can reserve relatively less bandwidth to achieve the same HOD rate, and hence will accommodate more new calls.

As we alluded earlier, the introduced ad hoc links create plenty opportunities for access to the cellular systems. By choosing good link connections, we can decrease the call dropping rate while inappropriate choice of links may lead to poor performance. Careless design of handoff decision process may increase the handoff frequency, increase the call dropping rate, or lead to the inefficiency of resource usage. How to design the handoff strategy in the MCNs becomes a more complicated task than purely comparing the received signal strength (RSS) because of the options offered by the multi-hop connections. This problem for MCNs has not been touched upon previously. Most previous works assume that the relaying is made by specific planted devices (relay stations) according to proper network planning [13, 63]. This assumption really limits the utilization of benefits offered by the self-organizing ad hoc networking and diminish the flexibility of the ad hoc mode introduced. In this paper, we propose a new scheme, called ANHOA (Ad-hoc-Network-Embedded Handoff Assisting Scheme), which utilizes the embedded, self-organized small-scale ad hoc networks to assist the handoffs. By exchanging information inside the embedded ad hoc networks, relay nodes can help handoff calls choose better handoff options and thus reduce the call dropping probability. Consequently, to meet a certain HOD rate, BSs can reserve less bandwidth for handoff calls than before. We also design an algorithm to enable adjacent BSs to lower the bandwidth reservation according to the traffic information of their cells to loosen up more bandwidth for new calls while meeting the HOD rate requirement. Furthermore, we have also propose a framework for adjacent BSs to exchange information through which the load balancing among the surrounding cells can be implemented. Although there are some works on different issues in MCNs, they rarely address the handoff
management and corresponding resource management by taking full advantage offered by the introduction of ad hoc mode.

The rest of the paper is organized as follows. Section 6.2 discusses the related works. In Section 6.3, we describe the system model and the basic ideas of this paper and in Section 6.4, we present the proposed ANHOA scheme. The algorithm for finding minimum reservation in MCNs and the framework of information exchange among adjacent BSs are given in Section 6.5. We carry out the performance evaluation in Section 6.6 and conclude the paper in the final section.

6.2 Related Works

The integration of multi-hop ad hoc mode into cellular systems has been conceived in late 1990s and early 2000s. However, one of the first thoughts of integration is to introduce relaying systems into cellular systems. Hsu and Lin might be the first presenting the potential idea called Multi-hop Cellular Networks (MCNs) [41]. In this paper, the authors suggested to use multi-hop communications from mobiles to a base station via possibly multiple hops with lower transmission power so that more simultaneous communications can be accommodated. In iCAR (Integrated Cellular and Ad Hoc Relaying Systems), Wu et al. proposed to proactively deploy a new set of relaying nodes in areas where traffic congestion start forming and use these relaying nodes, called ad hoc relaying stations (ARSs), to relay the traffic from the congested cell to the non-congested cells where traffic can be served [63]. To investigate how much ad hoc relaying could enhance the system capacity in MCNs, Law et al. studied the capacity by assuming that a cell is divided into two co-centered areas where direct communications are carried out within the near range between mobiles and the BS while multi-hop communications will be carried out only when mobiles are outside the near range [36] and found that for certain scenarios, the system capacity can indeed be increased.
To take more advantages of relaying capability and by harvesting potential outband resource, Luo et al. proposed a unified cellular and ad-hoc network architecture (UCAN) based on 1xEV-DO (HDR) and 802.11b [42]. This work was motivated by the observation that a higher downlink data rate may be needed for many applications. By allowing wireless clients to relay the downlink traffic in this scheme, the system can achieve better throughput performance. The authors developed the discovery algorithm of wireless proxy (relay clients), which plays very important role in implementing their proposed architecture. In view of the potential gain in using ad hoc mode, future generation wireless cellular standards have also considered this issue and proposed ODMA (Opportunity Driven Multiple Access) protocol [19], which is a similar scheme to UCAN [42], but focuses on the improvement of data rate by allowing relaying.

There are many proposals on ad-hoc/cellular integration architecture in the literature which address different aspects of various integrated networks. Cavalcanti et al. provided a survey of all these integrated networks [12]. Some of the features of the related works are compared. Le and Hossain have also given a survey on the existing MCNs [37], which focuses more on resource management. Cho et al. dealt with handoff issues of various types in MCNs [13]. In [9], Bhargava et al. took a different approach and intended to use cellular networks to help the management of ad hoc networks by utilizing the benefit of coverage offered by cellular systems. As we can observe, most integration research works are based on cellular systems, focusing more on enhancing the performance of cellular systems because cellular systems are deeply commercialized and they can provide seamless coverage and mobility service.

6.3 System Model and Basic Idea

6.3.1 System Model

In the MCNs, there are two operational modes for a link. One is the traditional cellular operational mode using the cellular spectrum and following the cellular standard. The other we call the ad hoc mode may use outband spectrum and follows the ad hoc
protocols such as IEEE802.11. The service coverage areas are divided into cells as in traditional cellular system while ad hoc links can cross the boundaries of adjacent cells when relaying traffic. Each BS is assumed to have the same transmission power as before which can cover the whole cell area.

Since the focus of this paper is the handoff issue in MCNs, in our model here, each MN (mobile node or mobile terminals or user equipment) is always bound to one certain BS, no matter whether via only cellular links or via multi-hop connections. We call the handoffs involving multi-hop connections as the multi-hop handoffs. When an MN moves from one cell to another and chooses the target BS as its serving BS, the handoff process is no different from that in the traditional cellular systems. Only when a multi-hop connection exists either before or after an MN alters its connection, do the handoffs face the new type of network challenges. Fig. 6-1 illustrates the multi-hop handoffs in a multi-hop cellular system.

![Figure 6-1. Model of Multi-hop Handoffs](image)

In Fig. 6-1, MN A moves from cell 1 to cell 0. Unfortunately, the BS in cell 0 has no spare spectrum. In the illustrated scenario, MN A has two options. It can either access the BS 5 through MN B or access BS 7 through MN C and MN D. Either option can enable MN to access to the cellular system.
Usually, a connection with more hops implies more potential disconnection vulnerability due to the mobility of intermediate nodes. There are also cases that the multi-hop connection may have high data rate due to less interference and shorter distance between nodes on the connection. How to choose between a single-hop connection and multi-hop connection is a challenging but important question. In this paper, to simplify the analysis, we assume that each MS attempts the direct connection first before sinking multi-hop connection unless being redirected. When a direct connection is not available, a disconnected MN, or an MN looking for a handoff, searches for the nearby ad hoc spectrum to find the alternative connections. Some connected MNs broadcast their corresponding BSs’ IDs and some other information via the ad hoc spectrum. For new calls, MNs search for all possible connections until they succeed. For handoff calls, due to the time limit of signaling process, MNs make several attempts to connect to different BSs until they succeed or the time limit is reached. The BSs’ IDs are used to prevent repeated attempts. The connected MNs are responsible for relaying the connecting MNs’ requests to the destination BS. The destination BS allocates cellular spectrum resource to the connecting MN in care of the last hop relay MNs. The BS keeps the information of the relay relationship of its serving MNs in order to deliver packets correctly. As a remark, the security may be a concern because the MNs on the connection may cause security leak. We will address the security issue elsewhere.

In traditional analysis of cellular networks, it is assumed that the call droppings caused by link failure are usually ignored and only the call dropping is caused mainly by handoff failure. In fact, a connection can be disconnected due to wireless link failure, especially when the connection is multi-hop. Thus call droppings caused by the failure of ad hoc links cannot be ignored in MCNs for our analysis. The call dropping rate can be expressed in such a formula as follows.

\[
P(dropping) = 1 - (1 - P(\text{con\_fail})) \times (1 - P(HOD))
\]  

(6–1)
\( P(HOD) \) is the probability of call dropping due to handoff failures. \( P(\text{con.fail}) \) is the probability of call dropping caused by connection failures. Multi-hop connections is disconnected more easily due to the smaller transmission range, mobility of relay nodes and nature of multiple-hop. Obviously, the call dropping rate is related to the hop counts of the multi-hop connection and the mobility of relay nodes as well as the roaming nodes.

To maintain a certain \( P(HOD) \) level, certain bandwidth should be reserved for handoff calls. If there is only one cell’s resource under consideration, according to Markov Chain model, such as [26], the HOD probability can be obtained when the traffic arrival and departure rate are given by

\[
P_0 = f(\lambda_n, \lambda_{ho}, \mu, Rsv)
\]

where \( f() \) is the derivation function of HOD rate which is described in detail in Appendix. We denote the HOD rate as \( P_0 \) when only one cell’s resource is considered. The arrival rates of new calls and handoff calls are denoted as \( \lambda_n \) and \( \lambda_{ho} \), respectively. The departure rate of all calls is denoted as \( \mu \) and \( Rsv \) is the number of reservation channels.

We know that the transmission range of an MN is much smaller than that of a BS and many MNs are highly mobile. It seems ad hoc mode could not significantly help the cellular system due to much shorter span of certain links. However, there are still many relatively stationary MNs existing in the system because low mobility can achieve higher data rate or purely because of users’ behaviors. Devices such as ARS in [63] can also be introduced to increase the stability of multi-hop connections. Therefore, due to the existence of plenty of potential relay nodes, we can rely on multi-hop connections to improve the performance of cellular systems.
6.3.2 Basic Idea

In MCNs, beside the BSs, the existing relay nodes can also assist the roaming MNs to complete handoff process via multi-hop connections. The multiple handoff options may connect the same roaming MN to different BSs, or to the same BS via different paths. Among multiple handoff options, how to find the best is very important but challenging. Connecting to inappropriate BSs through inappropriate relay nodes at inappropriate timing may not only increase the handoff frequency, increase the call dropping rate, but also lead to the inefficiency of resource. In order to make the right decision, it is important for roaming nodes to collect more information about the candidate choices. It is well known that self-organization is one of the major advantages of ad hoc networks, and thus relay nodes can easily exchange information in ad hoc mode. When the relay nodes form small-scale ad hoc networks, the resource information can be exchanged and maintained periodically within the networks, just as in some routing algorithms in IP networks. The roaming nodes can obtain the necessary information to enable the path selection for a connection with better QoS and lower the call dropping probability. Large-scale ad hoc networks are not suitable for this task because the information exchange will create too much signalling overhead.

The traditional way to control the HOD rate in cellular networks is to adjust the bandwidth reservation for the handoff calls [17]. Higher reservation means lower HOD rate. However, blindly increasing the bandwidth reservation also lead to the higher call blocking rate for new calls, thus decrease the spectrum efficiency. Therefore, searching for the optimal reservation which can also satisfy the constraint of the HOD rate should be the goal for each BS. For MCNs, due to the existence of multi-hop handoffs and multiple handoff attempts, each BS can practically have a smaller bandwidth reservation. The knowledge of adjacent cells’ traffic can help each BS making resource management more effective, furthermore, balance the load among different BSs.
Based on the insights above, in this paper we propose two schemes to improve the handoff performance. In the first scheme, we let a part of the stationary MNs form small-scale ad hoc networks to exchange and maintain necessary information for the potential handoffs. When handoffs occur, the roaming MNs can utilize the present information to choose the best choice for handoffs. The utilization of self-organizing characteristic of ad hoc networks in MCNs is manifested through the use of multi-hop handoffs. The second scheme deals with the bandwidth reservation at each BS. Each BS gathers the traffic information of its adjacent cells, takes the prospective multiple handoff attempts for roaming MNs into account and makes the least reservation. This mechanism is then further developed to a framework which can balance the load among different cells when traffic load is significantly unbalanced.

6.4 ANHOA: Ad-hoc-Network-Embedded Handoff Assisting Scheme

6.4.1 Overview

Multi-hop connections can reduce the HOD greatly in that more opportunities are provided than before. However, although potential paths exist in these nodes, it may not be easy to find the best choice. In MCNs, there are probably many connected MNs surrounding one disconnected MN. Among all these potentially connected nodes, some may have already been connected to the BSs that have no spare channels to accommodate new connections while others are highly mobile and not suitable for relaying. Although several attempts for a handoff process are generally acceptable, trial and error approach is not a good idea. Fortunately, as we all know, information can be effectively gathered and shared using the power of networking. To make information exchange more effective and more helpful to the handoffs, we may prefer to select stationary nodes to form mobile ad hoc networks (MANETs) to assist handoffs. Obviously, when the scale of these assisting networks becomes large, the coverage can be wider and the handoff can gather more help. On the other hand, the maintenance of the assisting networks will become more difficult and the information exchange will be
less effective. Thus, we need to employ stationary, small-scale ad hoc networks to more effectively assist the multi-hop handoffs in cellular systems. Based on this argument, we propose our scheme handling multi-hop handoffs, called “Ad-hoc-Network-Embedded Handoff Assisting Scheme (ANHOA)”, in which we address how to select nodes for the relaying and how to form the handoff assisting networks. In this paper, we focus on the handoff issues and simply assume that these networks can be formed by the instruction of operators.

6.4.2 Architecture and Roles

Figure 6-2. Embedded Ad hoc Network in Cellular System

We first describe the network architecture and roles of the assisting networks in Fig. 6-2. In this network architecture, each of these embedded ad hoc networks covers the area which may cover several adjacent cells. As Fig. 6-2 shows, several MNs form one embedded ad hoc network (EAN) in the cellular systems which spans across several cells. In Fig. 6-2, the circled nodes are the backbone nodes and the solid lines between them stand for the relatively stationary links. The triangular nodes stand for the roaming nodes which are searching for BS connections with the help of the embedded ad hoc networks. We call these nodes “attaching nodes (ANs)”. The nodes in the backbone connecting directly to the BSs are marked with “0” inside the circles. We call these
nodes “portal nodes (PNs)” because they have portal capability of connecting to the backbone cellular system. Other nodes in the backbone not directly connected to BSs use the number inside the circles to stand for the hop counts to portal nodes. We call them “relay nodes (RNs)”. RNs and PNs form the backbone of the embedded ad hoc networks which assist multi-hop handoffs in the cellular system. The ANs which cannot have the direct connection to BSs use the EANs to find alternative connections. The backbone node which have direct connection to the attaching nodes are their “dock nodes (DNs)”. ANs’ DNs do not have to be PNs. Each PN should also know the corresponding ANs which connect to a BS through it via different DNs. These ANs are called the “subordinate nodes (SNs)” of this PN.

6.4.3 Information Maintenance of EAN

The construction of EANs and the routing maintenance are omitted in this paper. We assume that the EANs have been already constructed, and the routing paths to potential PNs, the link capacity status and the interference link set of each link are all known, which can be used to estimate the so-called “portal capacity (PC)” and enable each AN to make the handoff decision. Notice that each backbone node in EANs should maintain a table storing the link capacities of all its neighboring links and to be updated when some interfered paths are taken or released.

Each node in the backbone of EANs is a potential DN for roaming MNs. When a roaming MN attaches to an EAN and searches for a multi-hop connection to a BS, its DN in the EAN provides connection information for it and then the handoff attempts follow. Portal capacity in this paper is defined as the maximum achievable bandwidth along the path in the EAN from DNs to the connected BSs.

For each PN, its own portal capacity is the average available bandwidth obtained from the Markov chain model. Given the arrival rate $\lambda$ and departure rate $\mu$ of a cell, according to Markov chain model as [17, 26], the expected unused channel/bandwidth can be derived. For each relay node, the portal capacity to different PNs also depends
on the path capacity of different paths. Since all links of one path share the same medium, they may tend to interfere with each other, hence the path capacity is not simply the minimum of link capacities on the path. It needs to take the link contention/scheduling of the channel into account. According to the result of [70], the path capacity is at most $1/4$ of minimum link capacity. Therefore, we use the following equation to estimate the portal capacity.

$$PC = \min\left(\text{EstCellBW}, \frac{1}{\min(4, HC)} \min_{i \in L}(L_i)\right)$$ \hspace{1cm} (6–3)

where $\text{EstCellBW}$ denotes the estimated available bandwidth in cellular spectrum, $HC$ is the hop count from DN to PN, $L_i$ denotes the link capacity of link $i$ and $L$ denotes the link set of the corresponding path.

This information is used by each AN to determine which PN to connect. The information about the connection to different BSs via different PNs is maintained in every backbone node of the EAN. Besides the link capacity status table aforementioned, each backbone node in EANs should maintain two tables which store the information of available connection resource and the information of its ANs, respectively.

Table 6-1. Portal Nodes' Information Table

<table>
<thead>
<tr>
<th>PNID</th>
<th>The ID for the candidate of portal nodes</th>
</tr>
</thead>
<tbody>
<tr>
<td>BSID</td>
<td>The ID of the BS which connects to the corresponding portal node</td>
</tr>
<tr>
<td>EstCellBW</td>
<td>Estimated available bandwidth of corresponding BS</td>
</tr>
<tr>
<td>HC</td>
<td>Hop counts to the corresponding portal node</td>
</tr>
<tr>
<td>PC</td>
<td>The available portal capacity to the corresponding BS</td>
</tr>
</tbody>
</table>

The portal capacity entries in the table are used by ANs to choose PNs with better dropping probability. These entries are correlated with each other. In other words, if one
Table 6-2. Attaching Nodes’ Information Table

<table>
<thead>
<tr>
<th>ANID</th>
<th>The ID of the attaching node</th>
</tr>
</thead>
<tbody>
<tr>
<td>PNID</td>
<td>The ID of the attaching node’s corresponding portal node</td>
</tr>
<tr>
<td>HC</td>
<td>Hop counts to the corresponding attaching node</td>
</tr>
<tr>
<td>OPC</td>
<td>The occupied portal capacity by the corresponding attaching node</td>
</tr>
</tbody>
</table>

PN is chosen by some AN, the portal capacity of other paths as well as this path will probably also change. The reason is that different paths may share same portal node, or share same certain link, or, links from different paths interfere with each other with high probability.

For each PN, it is necessary to collect and update the information of its SNs. Therefore, the following table is required to be maintained in every PN besides the above two tables.

Table 6-3. Subordinate Nodes’ Information Table

<table>
<thead>
<tr>
<th>SNID</th>
<th>The ID of the subordinate node</th>
</tr>
</thead>
<tbody>
<tr>
<td>DNID</td>
<td>The ID of the subordinate node’s corresponding dock node</td>
</tr>
<tr>
<td>HC</td>
<td>Hop counts to the corresponding docking node</td>
</tr>
<tr>
<td>OPC</td>
<td>The occupied portal capacity by the corresponding subordinate node</td>
</tr>
</tbody>
</table>

When the link capacity of each link changes, the portal capacity of each path is required to adapt accordingly in order to reflect the BS connection capability in real time. The change of the arrival rate $\lambda$ or the departure rate $\mu$ causes the change of $EstCellBW$, and thus the change of $PC$. These changes should be broadcasted in the EAN and each backbone node should update its entries accordingly. When one AN acquires a certain amount of bandwidth from one PN, all the intermediate links subtract the same amount of bandwidth from each link capacity. Moreover, all the interfered links
in the EAN should also subtract the summation of the occupied bandwidth they sense. The above information related to links is not stored in the above 3 tables, but in the link table we assumed before. \( PC \) in Table 6-1 and \( OPC \) in Table 6-2 and Table 6-3 are calculated based on the link capacity information via Formula 6–3.

In addition to these \( PC \) related updates, when an MN enters or leaves the EAN, the corresponding relationship should also be updated. These updates will be referred to as the “relationship update” in the following section. Upon the arrival (bandwidth granting) or departure (bandwidth releasing) of one SN, a PN notifies all the intermediate backbone nodes about this event hop by hop till the DN. The PN adds/clears the corresponding SN entry. The PN and all other corresponding backbone nodes update the \( PC \) information, and the DN adds/clears the corresponding AN entry. Besides the update of these 3 tables, the change of \( PC \) should trigger the change of link status table of the corresponding link, and thus the link status table of interfered links.

6.4.4 Handoffs via ANHOA

Apparently, introducing the ad hoc mode gives handoff decision more flexibility in cellular systems. The first possible change of handoff procedures is that with the existence of EANs, handoffs might involve multiple nodes in the EANs. Previously, the handoffs are always from one cell to another, with direct connections to BSs. With the existence of EANs, handoffs can also possibly take place from one EAN to another, from one EAN to a direct connection or vice versa, or within the same EAN. Procedures for these types of handoffs should be specified to avoid the possible resource waste and service interruption.

Moreover, the timing for a handoff also becomes complicated. In the traditional cellular systems, the timing of handoffs is when the current BS’s signal strength is below a certain threshold and the alternative BS’s signal strength is above a certain threshold (or more complicated strategy). When multi-hop connections are allowed in cellular systems, handoff decision are given more options and can be more flexible than before.
When an MN travels in the MCNs, it uses its ad hoc mode radio searching for the existing EAN signals while using its cellular radio for cellular services. It measures the signal strength of the detected signals (including direct BSs signals and EAN signals) and maintain its set of potential handoff candidates. Handoffs without EAN involved are no more different from the traditional cellular handoffs. Multi-hop handoffs happen when certain conditions are met. When multi-hop handoffs are needed and there is an embedded ad hoc network accessible, ANHOA can be executed.

![Multi-hop handoff diagram](image)

Figure 6-3. EAN-involved Handoffs

The basic procedure of EAN-supported handoffs has several elements just as the traditional handoffs as shown in Fig. 6-3. Update requests are sent to the new BSs. New BSs need to retrieve users' information from the old BSs. The connections are then established between the new BSs and MNs. Finally the old BSs are required to clear the old record. Small difference from the traditional cellular handoff procedures is the extra information that the update request should carry, the Portal Node ID. This information is used for the release of the old multi-hop connection through the same EAN or another one.

Although the extra signallings within EANs bring extra cost to the handoff procedures, it can be seen as the price to pay for alternative handoff options, which is intuitively better than the traditional handoffs only. The signallings within EANs simply deal with
the updates of 4 aforementioned tables. These updates change the nodes' relationship corresponding to the multi-hop connections. If a handoff happens within a single EAN, the procedure can be simplified because the only thing needs to do is the “relationship update” as mentioned before.

ANHOA is proposed to provide mechanisms to support multi-hop handoffs in cellular systems with the aid of EAN. Although there are a lot of issues in handoff timing and handoff procedures, we focus on how the MNs choose handoff options so that the call dropping can be greatly reduced.

First, let us describe the handoff procedures via ANHOA. Within the transmission range of any EAN node, the AN can acquire the knowledge of all PNs in the EAN and the $EstCellBW$ of each connected BS. The AN uses the “handoff decision algorithm” to choose a suitable PN and requests the DN to forward the update to the chosen PN. If the new EAN is different from the old EAN, which can be learned from the EAN ID information in the update request message, the “relationship update” of the old EAN is separated from the one of the new EAN, even if they connect to the same BS. The DN forwards the update request to the PN, with the requested QoS information and the old BS and old PN (null-valued if there is no). After QoS negotiation, the PN sends the update request to the new BS with the negotiated QoS. Before granting the update request, the new BS needs to request the subscriber information from the old BS and allocate corresponding resource to the AN, care of the PN. Especially, if BS does not change, the handoff only requires BS to update the MN's PN information. On the other hand, when EAN does not change, the adding and removing of the “relation update” can be processed within the same EAN. When the BS does not change and the PN changes, the connected BS only needs to update the MN's PN information.

When neither BS nor PN changes, BS does not need to do anything and all the handoff procedure is carried out by the EAN. Therefore, ANHOA can relieve a lot of signaling pressure off BSs by taking care of a lot of handoffs within EANs.
Beside the signaling, for the packets to be forwarded from the ANs to the PNs or the opposite direction, it is both feasible to use routing and tunneling approach. Considering we assume EANs to be small-scale ad hoc networks, routing is more efficient because tunneling may cause more delay due to encapsulation and decapsulation.

Notice that when a multi-hop handoff happens within an EAN, the procedure can be very simple and convenient. The clients can get the seamless service at very low price. For the continuity of the service, within one EAN, the MN can choose the same PN as before. In this way, the handoff procedure only involves the path information update in the corresponding nodes in the EAN.

ANHOA can also help new calls. This mechanism does not discriminate new calls from handoff calls although we should let handoff calls have higher priority as done in the traditional cellular systems.

### 6.4.5 Handoff Decision Algorithm

In this section, we present the handoff decision algorithm used by MNs in the environment along with EANs. The input of the algorithm is the candidates of handoff destinations, either BSs or EAN nodes, and their corresponding information. The output is the choice of the handoff destination. The criterion is to have lower call dropping and more QoS satisfaction. In this paper, we assume QoS can be always satisfied when a connection is available and we ignore the QoS satisfaction. As we formulate in Eq. (6–1), call dropping can be caused by handoff failure or by physical layer communication failure. For each connection candidate, i.e., one DN's connection to one PN, the corresponding call dropping rate is calculated. The AN compares the call dropping rates of all candidates and find the one with the lowest call dropping rate. Note that one DN can have more than one candidate because it can connect to different PNs.

There are several factors that affect the handoff performance. The first is the expected bandwidth of the destination BS. With the knowledge of arrival rate and departure rate of the corresponding BS, the steady probability of each state can be
derived and thus the expected available bandwidth can be derived as well.

\[ P_i(BS\_HOD) = f(\lambda_i, \mu_i, Rsv) \]  \hspace{1cm} (6–4)

\[ EstCellBW_i = g(\lambda_i, \mu_i, Rsv) \]  \hspace{1cm} (6–5)

where \( P_i(BS\_HOD) \) is the HOD rate derived based on the knowledge of reservation \( Rsv \), the initial/handoff call arrival rate \( \lambda_i \), and departure rate \( \mu_i \) of BS \( i \), \( f() \) and \( g() \) denote functions of \( P(BS\_HOD) \) and \( EstCellBW \), respectively.

The second factor is the handoff cost in terms of signaling traffic. If a handoff takes place within the same EAN without changing BS, the handoff cost is in fact very small. It is more appealing than those handoffs that changes BS, PN, or even EAN. We denote the HOD probability when only BS changes, only PN changes, and only DN changes as \( P(BS\_HOD) \), \( P(PN\_HOD) \), and \( P(DN\_HOD) \), respectively. With \( P(PN\_HOD) \) and \( P(DN\_HOD) \) of each candidate obtained from measurements, the one-time multi-hop HOD rate of each candidate can be derived via the following formula.

\[ P_k(one\_HOD) = 1 - (1 - P_k(BS\_HOD)) \cdot (1 - P_k(PN\_HOD)) \cdot (1 - P_k(DN\_HOD)) \]  \hspace{1cm} (6–6)

If the BS or the PN does not change, the corresponding HOD rate takes 0 value.

The third factor is the hop-count of the multi-hop connection. If the multi-hop connection has higher hop-count, it will be more vulnerable to disconnect. In this paper, we assume all EAN nodes have similar mobility, thus we can derive the connection failure probability via the following formula.

\[ P_k(con\_fail) = 1 - (1 - P_k(one\_hop\_fail))^{hop\_counts_k} \]  \hspace{1cm} (6–7)

The last factor is the relative position of destination BS to the MN’s movement. If the BS locates along the direction of the MN’s movement, the MN can avoid some unnecessary handoffs in the future, thus the call dropping probability can be further
reduced. Given that the MN keeps the moving history, it can have the conditional probability for each candidate of the handoff destination \( P(\text{BS}_i|\text{BS}_h) \), where \( \text{BS}_i \) stands for the BS \( i \) and \( \text{BS}_h \) stands for the previously-visited BSs. The extra HOD probability can be expressed via the following equation.

\[
P_i(\text{extr. HOD}) = (1 - P(\text{BS}_i|\text{BS}_h)) \cdot P(\text{HOD}) \tag{6-8}
\]

where \( P(\text{HOD}) \) denotes the averaged HOD rate derived by averaging the overall chosen dropping rates of the MN’s recent handoffs. Therefore, the overall HOD rate for each candidate can be derived as below. The overall call dropping rate can be derived via formula (6-1).

\[
P_k(\text{HOD}) = 1 - (1 - P_k(\text{extr. HOD}) \cdot (1 - P_k(\text{one. HOD}))) \tag{6-9}
\]

After we calculate the \( P_k(\text{call. dropping}) \) for each candidate \( k \), we simply choose the handoff option (the combination of BS and PN) with the smallest call dropping rate. Although there exist other factors that can affect the handoff performance, like QoS satisfaction, we will not present the details here because we think we can easily extend our framework to accommodate more factors.

### 6.5 Resource Management for Multi-hop Handoff

Multi-hop connection demands good resource management along the connection and in adjacent cells, which is addressed as follows.

#### 6.5.1 Minimum Reservation

To maintain a certain level of HOD rate, each BS should reserve certain amount of bandwidth for handoff calls. If the reservation is not enough, the expected HOD rate will be exceeded. However, since the reserved resource is separated from the shared resource, if the reservation is unnecessarily high, the blocking rate for new calls will increase, leading to dissatisfaction in new call blocking. Therefore, the focus of resource management is to find the minimum resource reservation which meets the
HOD requirement. In MCNs, since we can make more than one attempt for handoff, we can expect lower HOD rate than in one-hop cellular networks with the same traffic load and same reservation even though multi-hop connections bring in extra traffic from the adjacent cells. From the MNs’ point of view, one rejected handoff request in one cell in traditional cellular systems has always certain probability to access the other cells through ad hoc links. From the BSs’ point of view, we can reserve smaller amount of bandwidth to meet the same HOD rate requirement for the same traffic load but with multi-hop handoff connections so that we can gain larger amount of shared bandwidth resource to achieve higher trunking efficiency.

In MCNs, a given HOD rate is the input of a reservation scheme design. This input specifies the expected HOD rate for each MN when it is physically within a certain cell. We denote the required HOD rate for each MN in certain cell as $P_r(HOD)$. Our goal is to find the minimum feasible reservation $Rsv$. In other words, we need to find the relationship between $P_r(HOD)$ and $P_0$ in Eq. (6–2). Different from what is described in the ANHOA scheme, $P_r(HOD)$ cannot be derived from certain candidates’ information because it is an expected value for the whole cell. We assume each handoff MN first attempts to access its current cell and then try the neighboring cells one by one with a certain time limit. Therefore, besides $P_0$, the HOD rates of the adjacent cells are also needed in finding $P_r(HOD)$.

In the proposed algorithm of finding the minimum reservation, each BS is required to collect the HOD rate information of the adjacent cells. To simplify the analysis, we assume that each BS only considers the multi-hop handoffs to its one-hop neighbors. The periodic message exchange among the neighboring BSs can provide a mechanism to acquire this information. BSs use the past arrival rate $\lambda$, departure rate $\mu$, and the reservation $Rsv$ of the neighboring cells as the predictive value of next time interval. According to the previously measured value $\lambda$, $\mu$ and the current reservation $Rsv$, each BS calculates its own one-attempt HOD rate and broadcast this value. Note that
the measured arrival traffic $\lambda_{ho}$ consists of the original handoff traffic and the handoff traffic rejected in adjacent cells. The broadcasted value of one-attempt HOD rate $\tilde{P}_i$ is taken as the input of calculating the minimum reservation. Besides the HOD rate of neighboring cells, each BS also needs to know the probability that multi-hop handoffs can access the adjacent cell. This access probability, corresponding to a surrounding BS $i$, denoted as $q_i$, can be measured by dividing the number of calls in current cell (BS 0) which have access to BS $i$ to the total number of handoff calls in the current cell. Obviously, this measurement requires the handoff calls to report extra information to BS, the accessibility to other BSs.

$$
P_r(HOD) = P_0 \cdot ( \\
\prod_{i=1}^{M} (1 - q_i) + \\
\sum_{i=1}^{M} \frac{\tilde{P}_i}{M} \cdot q_i \prod_{l \neq i}^{M} (1 - q_l) + \\
\sum_{i=1}^{M} \sum_{j \neq i}^{M} \frac{\tilde{P}_i \cdot \tilde{P}_j}{2! \cdot C_2^M} \cdot q_i \cdot q_j \prod_{l \neq i, j}^{M} (1 - q_l) + \\
\ldots + \\
\sum_{i_1=1}^{M} \ldots \sum_{i_N \neq i_1, \ldots, i_{N-1}}^{M} \frac{1}{N! \cdot C_N^M} \prod_{k=1}^{N} \tilde{P}_{i_k} \cdot q_{i_k} \prod_{l \neq i_1, \ldots, i_N}^{M} (1 - q_l) 
)$$

(6–10)

With these broadcasted $\tilde{P}_i$ and measured $q_i$, we can find the expression of the final HOD rate as in Eq. (6–10). While we take the required HOD rate as $P_r(HOD)$, we can derive the minimum reservation by incorporate formula (6–2). We use $N$ to denote the maximum number of attempts after the direct handoff attempt fails and $M$ to denote the number of neighbors of the current cell.
Note that in Eq. (6–10), the required HOD rate is simply the summation of the probabilities that after a handoff fails in the current cell, it fails in different numbers of other cells.

In actual MCNs with ANHOA scheme, roaming MNs attempt to connect to the proper BSs in a more intelligent way rather than connecting to the adjacent cells by using trial and error approach. Therefore, the overall HOD rate can be reduced further. Moreover, a rejected direct handoff call can attempt to connect to not only the adjacent BSs but also BSs further away, as long as the multi-hop paths exist. Eq. (6–10) considers only the adjacent BSs in order to simplify the derivation of the minimum reservation.

Let us revisit Eq. (6–10). When each $q_i$ has value 1, which means MNs in the current cell can access the resource of all the adjacent cells, this system can be seen as a larger cell consisting of all adjacent cells with aggregated resource from all adjacent cells. Obviously, this system has higher trunking efficiency, which can lower the call dropping rate and the requirement for resource reservation. When more than one hop neighboring cells are considered, even higher trunking efficiency and lower call dropping rate can be expected.

### 6.5.2 Load Sharing among Adjacent Cells

With the handoff options via multi-hop connection to access other BSs, an MN can achieve a much lower overall HOD rate $P_r(HOD)$ than the HOD rate derived from the arrival/departure rate and reservation, $P_0$, based on the one-hop connection. However, we should not only focus on the requirement of $P_r(HOD)$, but also need to maintain a certain resource to control $P_0$. High $P_0$ implies too much rejected traffic from current cell, which will overflow to other cells with $P_0 \cdot \lambda_{ho}$, even though $P_r(HOD)$ can still be satisfied. This part of traffic, called “permeated traffic” in this paper, will deteriorate other cells’ HOD rates, as a chain effect, due to repeated handoff attempts. This means if one BS keeps a significantly insufficient bandwidth reservation and mostly relies on other cells’
help, it can end up deteriorating surrounding cells’ HOD performance. The mathematical explanation for this situation is that the derivations of $P_r(HOD)$ and $P_0$ are based on the assumption for an ergodic system. If the reservation cannot support the arriving handoff calls, this system is not ergodic anymore. Consequently, Eq. (6–10) is not valid any more. When one of the cells in the neighborhood has been overloaded with traffic, this problem might be inevitable and even more protuberant. Therefore, a certain threshold should be set to ensure that none of the cell is overloaded. Furthermore, load balancing mechanism is necessary in MCNs to utilize the whole resource more efficiently. When one cell is heavily loaded, load balancing mechanism can utilize the adjacent cells’ reservation to maintain the required HOD rate.

In this paper, we build a framework for this purpose. In this framework, the traffic information in adjacent cells is exchanged and shared and the minimum reservation is calculated in a distributed fashion. More importantly, when some cell is heavily loaded, this framework provides a mechanism for BSs from other cells to take over part of the traffic so that each cell can maintain a relatively low HOD rate, $\tilde{P}_i$. This framework is called “Traffic Information Exchange for BSs’ Reservation Calculation Procedure”. The reservation calculation is done periodically by each BS. Between two consecutive periods of calculation, there are three phases: information collection, load balancing, and reservation calculation/information broadcasting.

In the first phase, each BS collects its neighbors’ traffic load in the previous period, the neighbors’ $\tilde{P}_i$, and the corresponding access probability $q_i$. Each BS can predict the traffic load according to the previously measured traffic, which consists of local traffic and permeated traffic from adjacent cells. The traffic load is measured in Erlang, i.e., traffic intensity. If the traffic load does not exceed a predefined threshold, $\lambda_{th}$, the second phase can be skipped. Otherwise, the overloaded BS requires its neighbors to take over part of its traffic load. The requests to different neighboring BSs will be based on the access probability, $q_i$. When requests have been granted, indication will be
broadcasted in the broadcast channel (BCCH) to redirect part of the traffic to other cells. The redirection operation for partial traffic can be implemented by a simple modulus operation by each MN. Thus the accurate traffic load of each BS is known, which can be expressed as follows.

\[
\lambda_n(i) = \lambda_{nLocal}(i) + \lambda_{nTkov}(i) + \lambda_{nPerm}(i)
\]

\[
\lambda_{ho}(i) = \lambda_{hoLocal}(i) + \lambda_{hoTkov}(i) + \lambda_{hoPerm}(i)
\]

The equations mean that the traffic load of handoff calls (with subscript \(ho\)) or new calls (with subscript \(n\)) consists of the local traffic (\(\lambda_{Local}\)), permeated traffic from neighboring cells (\(\lambda_{Perm}\)) and the overflow traffic from neighboring cells (\(\lambda_{Tkov}\)).

During reservation calculation phase, each BS calculates its own \(P_0\) beforehand, according to the latest traffic load and the reservation. If any BS is requested for taking over partial traffic from current BS, the calculation of corresponding \(P_0\) should include this partial traffic. With \(\tilde{P}_i\) and \(q_i\) known, the minimum reservation can be calculated via Eq. (6–10) and Eq. (6–2).

### 6.6 Performance Evaluation

In this section, we study the benefits that multi-hop connections can bring to the cellular systems in terms of handoff performance.

Firstly, we look into the HOD rate when the reservation of each BS remains unchanged. We study a cellular system with seven-cell frequency reuse. For a single cell, we calculate the HOD rate according to Eq. (6–2). The detailed derivation is briefly introduced in the Appendix, based on the model in [26]. The basic setting of resource parameters is listed in Table 6-4.

Each BS calculates the overall HOD rate according to Eq. (6–10) with the knowledge of its adjacent BSs’ single-cell HOD rates. The access probability to the neighboring cells are set as the same value. This calculation gives the statistical HOD rate in a cell. The choices of multi-hop handoffs are assumed to be chosen randomly.
Table 6-4. Simulation Setting for Each Cell

<table>
<thead>
<tr>
<th>Number of Channels</th>
<th>30</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Reserved Channels</td>
<td>5</td>
</tr>
<tr>
<td>Average Number of Initial Calls</td>
<td>40</td>
</tr>
<tr>
<td>Average Number of Handoff Calls</td>
<td>120 ∼ 270</td>
</tr>
<tr>
<td>Access Probability</td>
<td>0 ∼ 1</td>
</tr>
<tr>
<td>Average Call Intensity per User</td>
<td>0.1erl</td>
</tr>
</tbody>
</table>

With ANHOA scheme, MNs are expected to achieve better performance because more handoff opportunity and the intelligent handoff decision algorithm should mitigate the ongoing connection call drop.

![Figure 6-4. HOD rate](image)

In Fig. 6-4, the overall HOD rate of the center cell has been shown. Part (a) shows the relationship between traffic load and HOD rate. From Part (a), we can see that when traffic load increases, the HOD rate increases accordingly. With the help of adjacent cells, HOD rate can be reduced greatly. From Part (a), we can see that even with a small access probability, such as \( q = 0.25 \), the HOD rate can be improved with 20dB. We can also see from Part (a), when traffic load increases to a certain level, the HOD rate will deteriorate badly. The reason is that the current reservation cannot support the
traffic load and the rejected traffic also forms a big burden on the adjacent cells. Part (b) shows the relationship between HOD rate and different access probabilities under a certain traffic load. We observe that when access probability increases, the HOD rate decreases dramatically.

Figure 6-5. Channel Reservation

Secondly, we evaluate how the channel reservation is relieved with the help of adjacent cells. Lower reservation can also greatly reduce the call blocking rate of new calls. In this part of simulation, we set the required HOD rate to $0.5 \times 10^{-3}$. Under this constraint, each BS finds the minimum channel reservation. In Part (a) of Fig. 6-5, the numbers of reserved channels are shown with different traffic loads. We can see that without the help of adjacent cells (“Single Cell Case” in the graph), more channels need to be reserved to achieve the required HOD rate. When traffic load becomes larger, reservation does not work for the required HOD rate. In Part (a) of Fig. 6-5, 30 reserved channels stand for this case. With the increasing access probability, the channel reservation can be greatly decreased. In the case of $q = 0.35$, under most of the traffic load, there is no need for channel reservation to meet the HOD rate requirement. Part (b) of Fig. 6-5 shows the corresponding call blocking rate when the minimum reservation
is applied. We can easily observe the great improvement especially when traffic load becomes heavier. Since we are using the model from [26], the reservation in this simulation means that even there is no channel reserved for the handoff calls, the shared channels can still satisfy the required HOD rate.

6.7 Summary

Ad hoc links introduced to the cellular systems can bring great improvements in terms of handoff performance because they provide additional options for handoffs via multi-hop connections to BSs, leading to better service for roaming MNs by connecting to cells with more channel resource. The proposed scheme in this paper, ANHOA can assist MNs’ handoffs by utilizing the self-organizing small-scale ad hoc networks. Better handoff choice can be made when there are multi-hop handoff alternatives. Moreover, with multi-hop connections, multiple cells can balance the traffic load and collaboratively serve users with better performance. In light of this, we have propose a load balancing framework to enable resource sharing without specifically tying resource up for handoff calls. In so doing, we could mitigate the handoff call dropping while not affecting too much to the new calls. Simulation results verify the benefits of our schemes.

6.8 Appendix

According to [26], HOD rate in a cutoff prioritized reservation system can be modeled as a finite-state Markov chain.

![Markov Chain Model for Handoff Reservation System](image)

Figure 6-6. Markov Chain Model for Handoff Reservation System

In Fig. 6-6, $S_i$ denotes the state of $i$ channels being occupied. $\lambda_N$ and $\lambda_{HO}$ stand for the arrival rate of new calls and handoff calls, respectively. $C$ is the total number of channels and $R$ is the number of reserved channels. $\mu$ is the departure rate for both types of calls. In this model, handoff calls and new calls start to use the shared channels
when there are still spare channels in the shared channel pool. When the shared
channels are used up, only handoff calls can use the reserved channels.

We can write the state equations as the follows.

\[
P(i) = \begin{cases} 
\frac{\lambda_N + \lambda_{HO}}{i \mu} \cdot P(i-1), & \text{for } i = 1, 2, \ldots, C - R \\
\frac{\lambda_{HO}}{i \mu} \cdot P(i-1), & \text{for } i = C - R + 1, \ldots, C 
\end{cases}
\]

With the normalization condition, we can derive the probability of state 0.

\[
P(0) = \sum_{k=0}^{C-R} \frac{(\lambda_N + \lambda_{HO})^k}{k! \mu^k} + \sum_{k=C-R+1}^{C} \frac{(\lambda_N + \lambda_{HO})^{C-R} \lambda_{HO}^k}{k! \mu^k} \cdot P(0)^{-1}
\]

The HOD happen when all the channels are occupied. Therefore, the HOD rate is
the probability of state \(C\).

\[
P_C = \frac{(\lambda_N + \lambda_{HO})^{C-R} \lambda_{HO}^R}{C! \mu^C} \cdot P(0) 
\]

The call blocking probability is equal to the summation of probabilities that the
states occupy state \(C - R\) to state \(C\).

\[
P_B = \sum_{k=C-R}^{C} P(k) 
\]
CHAPTER 7
RELIEVING THE LOCATION MANAGEMENT CONGESTION VIA AGGREGATION IN
MULTI-HOP CELLULAR NETWORKS

7.1 Background

In PCS networks, the timely and exact location information of each mobile user is critical to the system in the sense that one call or one packet will fail in delivery otherwise. Location management (LM) is the function to fulfill this requirement. Each mobile user is required to update its latest location to the system whenever it reaches a new location area. Meanwhile paging is the complementary process since a location area is usually bigger than a cell and the exact location can be retrieved only via the response of the destination user. Reducing the location area size can lower the paging cost, however, at the price of an increasing signaling burden of location update (LU). Contrarily, bigger location area brings bigger paging signaling. Clearly, there is a tradeoff between LU and paging. Moreover, although location update and paging both occupy wireless resource, the costs of them are not equivalent. Paging costs the paging channel resource while LU costs the access capability of the system. If LU requests overwhelm the RACH (random access channel), new calls and handoff calls will be blocked. Therefore, careful design is required to balance the two costs as well as to reduce the overall costs. Many researchers have done a lot of works for this purpose.

In modern cities, population becomes denser and denser and so do the PCS devices. People tend to move together due to their similar working and living patterns. Additionally, occasional big events attract large numbers of people to gather to certain locations at a certain time. To reach a location, people tend to take public transportation, especially in the metropolitan areas, because of energy and traffic concerns. The moving masses easily create big fluctuation of the population of certain locations. Consequently, people live in a chain of crowded environments, such as office buildings, stores, subway stations, stadiums, theaters, etc. We can easily fail to reach a contact when the other end is in a crowded environment. For the cells along the location
area boundaries, new/handoff calls are usually found to be rejected due to the RACH congestions while other resource, such as DCCH, has not been fully occupied. This is because the moving masses of mobile terminals (MTs) launch signaling processes at the same time and congest the access of PCS systems in an outburst. The LU congestion problem gradually becomes the biggest issue of LM in current systems.

Researchers have noticed this problem and some methods are proposed to address it. F. Wang et al. propose a group LU scheme for cellular systems in [60]. Y. Zhang et al. propose a buffering scheme to relieve the congestion caused by batch arrival of LU requests [74]. However, although these schemes are inspirational, they cannot effectively solve this problem. For the indispensable LU requests, the most effective way to mitigate the potential wireless congestion is to aggregate these messages and reduce the channel contentions. Unfortunately, observing the architecture of the LM signaling in traditional cellular networks, we can see that the rigorous one-hop wireless structure forbids the possible aggregation of wireless signaling.

Fortunately, cellular systems evolve themselves in the meantime. The introduction of ad hoc mode to cellular systems sheds light on the possible solution of this problem.

The idea of integrating ad hoc into cellular systems comes from the motivation that cellular systems should become an all-IP platform so that all the service of new wireless networks (WiFi, WiMAX) can be incorporated together without difficulties [21]. The integration can be either based on ad hoc systems, such as CAMA in [9], or based on cellular systems. Since our focus of this paper is on the LM issues, we consider only the cellular-based integration. The cellular-based integrated systems, usually called Multi-hop Cellular Networks (MCNs) [41] [42] [12] [36], are expected to bring the advantage of ad hoc networks into cellular systems, such as flexibility, robustness, self-organizing, and low cost.

In MCNs, each MT can be others’ relay node and they forward signaling and data packets for each other. Since the ad hoc links are allowed to use extra spectrum
different from the cellular spectrum, the signaling burden of cellular wireless is possible to be moved to ad hoc spectrum and thus we can expect the relief of wireless congestion. Firstly, we propose a grouping scheme to reduce the wireless cost of LM for the scenarios in high-capacity transit (HCT) systems. With our scheme, the LU requests from the moving-in-together MTs can be grouped together by a special MT in vehicle so that the group maintenance can avoid the big cost of the cellular spectrum and the LM can be fulfilled in an efficient way. Secondly, we propose a generic aggregative scheme, which also takes advantage of the ad hoc links in MCNs, for other scenarios where MTs do not follow the same movement but arrive at the same spot around the same time and cause wireless congestions.

Our research in LM includes the cost analysis after the proposals of two schemes. We formulate the cost function consisting of LU part and paging part. By changing the LU part of the cost with our proposed scheme, we can find the new optimum location area size in MCNs with our analytical model.

The rest of the paper is organized as follows. Section 7.2 discusses the related works. Section 7.3 gives out the underlying system model. Section 7.4 presents the grouping scheme. Section 7.5 proposes the generic aggregation scheme. Section 7.6 shows the simple way to formulate LM cost and find the optimum location area size. Section 7.7 provides the performance evaluation. Conclusions are given in the final section.

7.2 Related Works

In the literature, there are several approaches aiming at reducing the LM cost. Static approach is one of the most common methods. During network planning, the entire coverage is divided into location areas (LAs) which usually consist of several cells. When a certain MT reaches a new LA, it launches LU procedure. Paging happens at all the cells within the current LA when an incoming call arrive and the destination is in IDLE state. Ma and Fang provided a pointer forwarding scheme to reduce the
unnecessary location update[44] and alleviate the signaling burden of HLR (home location register). Dynamic schemes can report the location update more flexibly in more various ways. Distance-based approach requires the system to remember the location of one MT’s last LU and when the moving distance exceeds a certain threshold a new LU is triggered. This approach has been studied in [62] extensively. Movement-based approach is another dynamic LU approach [3], where the LU is triggered when the movement of a certain MT reaches a threshold. Besides these, Misra et al. proposed an information theory-based LM for vertical roaming users in [45].

For the signaling cost of paging, Zang et al. proposed to delay the paging which leads to a reduced paging cost [67].

However, for the potential LU congestion problem, there are not so many works existing in cellular literature. F. Wang et al. propose a group location update scheme for cellular systems in [60]. By letting the group head report the location of its members, wireless cost can be greatly reduced. However, even though utilizing grouping method can disperse the location update requests, the incurred extra grouping maintenance signaling makes this dispersion less attractive. Additionally, the base station (BS) is not a good choice to take charge of the grouping maintenance since it is difficult and inefficient for BSs to group the mobile users with same movement. Y. Zhang et al. propose a buffering scheme to relieve the congestion caused by heavy arrivals of LU requests [74]. By letting BSs buffer pending LU requests due to the exhaustion of DCCH (Dedicated Control Channel), LU retry attempts can be reduced and hence the congestion can be relieved to some extent. This method attempted to spread the LU requests along the time line by storing the arrived LU requests in BSs’ buffers. However, the congestion in RACH cannot be effectively relieved with this buffering approach. Han et al. similarly propose a group LM scheme for one-dimensional networks (transportation systems) in [25]. By specifying virtual visitor location register (VVLR) as the register of group information and an additional tier between VLR and HLR, the
wireless cost and the signaling burden of HLR can be both reduced. However, this scheme can only adapt to a dedicated system and cannot be applied to PLMN cellular systems. GSM-R approach is another work based on dedicated systems [32]. The focus is on accelerating handoff and cell re-selection procedure by predicting the locations of MTs moving along the railways.

### 7.3 System Model

Before we start to describe the schemes, let us refresh some preliminaries about the LM. As shown in Fig. 7-1, in LTE (long term evolution) architecture, when each MT detects the change of location area, it reports the location to HSS (home subscriber system, HLR in LTE). It first contends to access the RACH and then asks the eNB (enhanced node B, BS in LTE) for the channel assignment. Besides LU traffic $\lambda_{LU}$, new calls $\lambda_N$ and handoff calls $\lambda_{HO}$ also contends for the RACH. The total traffic in RACH $\lambda_{RACH}$ can be expressed as Eq. (7–1).

$$\lambda_{RACH} = \lambda_{LU} + \lambda_N + \lambda_{HO}$$  \hspace{1cm} (7–1)
The eNB assigns a DCCH for this request and the MT uses this assigned channel for the following signaling process. The signaling is then going on wire-line until it reaches HSS. HSS responds LU request to MME (mobility management entity, switch center in LTE), eNB and then the MT. The DCCH is released when the process is over. Paging process starts when there is an incoming call or a downlink packet and the intended MT is in IDLE status. Not knowing the exact cell location of the mobile user, MME asks all eNBs within the location area where the mobile user lastly resides to page on the PCCH (paging channel).

As aforementioned, the cost of LM is usually regarded as only the wireless cost. Therefore, we ignore the cost of wireline signaling in our analysis. Among the wireless cost of LU and paging, since they consume the resource of different channels, they are not equivalent in the overall cost.

LU is not happening in each cell since the location area consists of several cells and only when the location area changes, is the location update necessary. However, paging process consumes the paging channel of all the cells within the location area. In the proposed schemes, the paging process in MCNs in not changed. The effort we are making is to reduce the LU’s consumption on RACH.

7.4 Grouping To Save The Signaling Cost

7.4.1 Overview

Among the wireless congestion scenarios, many are caused by the arrivals of HCT vehicles. In HCT vehicles, many MTs move together along certain fixed routes.

When a HCT vehicle arrives at a new location, the MTs inside start to update their locations together. These LU requests can easily over-inject the RACH and cause the congestions. Apparently, the most direct way to solve this problem is to group these MTs and relieve the contentions of RACH. Previous work, GLU [60], proposed to use grouping method in 3G and 3G beyond systems to reduce the LM cost. Different from it, we propose a new group LM scheme for MCNs which utilizes special MTs for group
maintenance. We assume these special devices, which we name it group head (GH), in each HCT vehicle, control the group maintenance and interact with each eNB they traverse. The reason that MTs are willing to deposit LM to GHs is they can have higher probability to complete the LU process and will not miss prospective calls. Grouping can reduce the contending LU requests and increase the access probability. Moreover, higher access priority for the group LU can be designed to encourage MTs to use group LM. The automatical depositing can be done by software setting easily. By moving most of the group maintenance messages to the ad hoc links, LM cost in MCNs can be greatly reduced. However, to achieve this gain, the procedures need to be carefully designed.

In grouping approaches, there are two hard problems which require the designers to think over. One is how the groups are formed and the other is how to reduce the group maintenance signaling cost. Moreover, during the LU process, the systems will allocate new temporary identities to replace the old temporary identities. The grouping scheme should have the capability to fulfill this task as well.

7.4.2 Procedures of Group Location Management

In MCNs, the GH, which has both the ad hoc interface and the cellular interface, does not need to be specially fabricated except that it needs to install extra software, which takes care of the grouping tasks and group message assembling. In addition, the core-networks need to support the group information storage and other corresponding operations. People can interact with the mobile devices in order to decide whether or not to associate with the GHs and which GH to associate with. Doing this manually can easily challenge people’s patience. Automatic association can be configured via software and make people less concerned about how their cell phones are working. We describe the group LM procedures based on automatic association. The accompanied problem with automatic association will be mentioned along with its solution.
The group LM procedures include the following basic operations. An MT needs to execute the association if it is to joint a group. It need to be dissociated with the GH when it leaves the group, although this procedure is usually executed by the GH unilaterally in our scheme. During the existing period of the groups, GHs need to update the group information with the systems involved periodically. When a new LA is reached, the GH is required to launch the group LU. However, before these basic operations are described, we first discuss how a group is formed.

### 7.4.2.1 Group formation

Due to the existence of GHs, the group formation seems to be an easy task. Each GH built in the vehicle can easily take over the MTs’ LM task whenever they are close to or reside in the vehicle. However, if the group members change rapidly, the benefits coming from grouping will be diminished. In traditional cellular networks, as we know, it is difficult for any BS to form a relatively stable group since the members’ movement is purely unpredictable. In MCNs, with the GHs built in the vehicles, MTs with the same movement can be easily grouped together via signal strength comparison. Unfortunately, GHs cannot distinguish the MTs inside the vehicle from the ones outside. Moreover, one MT can associate with a GH located in any other vehicle nearby. Considering this potential mess, we need to find a way to make the group more stable and reasonable.

We propose to use a guarding period as the condition to start the association with any new group member. We denote the length of this period as $T_{ag}$. Each GH uses its ad hoc interface to communicate with its current members and at the meantime probes the existence of prospective members. A timer with the length of $T_{ag}$ will be started when a new MT is emerging. If this MT is still in range after $T_{ag}$ expires, the GH will confirm it to deposit its LM task. Due to the imperfect of wireless channels, the links between GHs and the group members can be interfered or even broken. Therefore,
another guarding timer $T_{dg}$ is used to prevent the members from being dissociated with the GHs inadvertently.

These two guarding timers are used to make the group constituent stable. The value of $T_{dg}$ depends on the channel condition. Therefore, it is better to use a practical value for this timer or allow the software configuration. On the other hand, the value of $T_{ag}$ decides the gain of using grouping approach. If $T_{ag}$ is too small, many passing-by MTs can easily associate with the GHs assuming that everyone configures the association as automatic mode. The unnecessary associations waste the channel resource and the gain of using groups decreases. If $T_{ag}$ is too big, it is difficult for MTs to become group members which leads to a smaller group size. With a smaller group size, the gain of grouping is not fully exploited. We will leave the process of finding proper value of $T_{ag}$ in the following section.

Besides the guarding timers, scenarios of co-existing signals of multiple GHs should be also considered. The strategy we propose is the simplest one. If the MT is previously associated with one GH, it will maintain the association even if some other GH has stronger signal strength. If the MT is not associated with any GH when multiple GHs are available, it chooses the one with the strongest signal strength. If the link outage duration reaches $T_{dg}$, both the GH and the MT deem themselves dissociated. Although in some scenarios this strategy will lead to wrong GH selection, such as when a passenger changes to the connecting line nearby, the association will be corrected after two vehicles separate and the delayed association will not impact the system performance.

### 7.4.2.2 GH LM association

When a passenger board the HCT vehicle, the MT he/she carries will receive an indication from the GH residing in the vehicle instructing the MT to deposit the LM task. Following the indication, the MT starts the association procedure as shown in Fig. 7-2. First it sends a deposit request message to the GH using its ad hoc interface.
Figure 7-2. Group Association Procedure

After the guarding duration $T_{ag}$, the GH responds with its identity which is used by the MT afterwards to verify the GH and report the association relationship to the eNB. With this process completed, the GH takes over the LM task of the corresponding MT, including the regular location update and periodic location update, and the system (eNB and MME) has the knowledge of the group and the associating MT. Note that the communications between MTs and GHs are via ad hoc links and the communications between GHs/MTs and eNBs are via cellular links.

7.4.2.3 Group information update

GHs are required to detect the existence of the members constantly. If some MT tries to join the group, it follows the procedure of association. If some group member is lost, the GH is responsible to update the group information periodically to the network.

Periodically, the updated group information in one MME will be broadcasted to its neighboring MMEs. Since HCT vehicles always follow a certain path, the broadcast can be replaced with multi-cast to reduce the unnecessary communications. These information updates do not involve the HLR (or HSS).
Through this way, the group information can be synchronized to the MMEs along the traveling path. With this information ready beforehand, the wireless transmissions can be greatly saved. This idea is similar to the shadow cluster in [38]. However, with the group under consideration, the signalling overhead in our scheme is far less.

### 7.4.2.4 Group location update

Upon the LM depositing, the security issues cannot be ignored. In this paper, we simply assume each MT sends its authentication keys along with its ID to the GH in a safe way. Other security issues are not within the scope of this paper and thus skipped.

Upon arriving at the new location area, the GH launches the location update procedure with the group ID. If the constituent of the group is changed since last report, a list of current members is sent out with the LU request. If there is no change of the group’s constituent, a group ID is enough for the group location update.

The new MME checks out the MTs requiring LU according to the previously received group information and then sends these LU requests to the corresponding HSS.

Since all the group members stay in the same old LA before entering the new LA, their process of temporary ID (TMSI in GSM) reallocation can be grouped as well. If some MTs are required with some special process, such as IMSI verification, the following procedures can be carried out in an individual way. As of the system performance, the group LU requests have mitigated the RACH congestion problem no matter the following procedures are carried out in a grouped way or not.

### 7.4.2.5 Dissociation

Upon leaving the GH, MTs take back the LM task. This is triggered by the timeout of timer \(T_{dg}\).

When the HCT vehicle arrives at the destination, the GH fulfils the dissociation task with all the group members to clear the information storage in MMEs.
7.4.3 Gain Analysis

7.4.3.1 Signaling cost reduction

In MCNs, the cellular spectrum is more critical than ad hoc spectrum in that ad hoc mode is usually seen as an addition to cellular systems in MCNs while lack of available cellular bandwidth can directly cause call blockings. As shown in Fig. 7-1, if we utilize the grouping method to report the location, multiple location update requests can be combined as one. The utilization of grouping can first enhance the success rate of LU. As we know, LU requests contend the RACH with new/handoff calls, following slotted-ALOHA MAC mechanism. We denote the LU traffic rate as $R_{LU}$ and call traffic as $R_c$, respectively. The averaged group size is denoted by $K$. The improved LU success ratio (we assume LU fail only because of RACH congestion) can be expressed as follows.

$$\eta = 1 - \frac{e^{-(R_c+R_{LU})}}{1 - e^{-(R_c+\frac{R_{LU}}{K})}}$$  \hspace{1cm} (7–2)

We can see from Eq. (7–2), that when the LU traffic overwhelms the RACH, grouping can greatly enhance the LU success rate as well as the success rate of new/handoff calls.

In Eq. (7–2), the traffic of LU is approximately reduced to $K$ times when grouping is executed. However, the exact wireless cost of location update should be analyzed in order to evaluate the performance quantitatively. We first investigate the wireless cost of regular LU without grouping in one location area with $M$ MTs in unit time duration. The periodical LU cost is not included in this analysis for the simplicity purpose. The average speed of the HCT vehicles is assumed as $\bar{v}_{HCT}$ in number of cells per unit time. We assume the location area consists of a cluster of cells which form a series of concentric hexagons with the radius of $d$ cells. The expected regular LU signaling cost can be expressed as follows.

$$C_{LU} = \frac{M \cdot \bar{v}_{HCT}}{2d - 1}$$  \hspace{1cm} (7–3)
The LU cost means the average times of LU during unit time among \( M \) MTs, assuming each MT update their location after they move a \( 2d - 1 \) distance.

With our grouping method, location update traffic is reduced \( K \) times along with the incurred MT-eNB registration traffic and the GH-eNB group information update traffic. Since ad hoc spectrum is not critical for mobile users to access the system, we ignore the wireless cost in ad hoc spectrum. For analysis, we assume that all \( M \) MTs are moving in \( K \) groups together, creating the \( K \) times-reduced location update traffic, the group information update for each group once every interval, and the association traffic for each MT once every trip. \( \bar{L} \) denotes the average trip length of passengers in the unit of number of cells. \( T_{gr} \) is the GHs’ interval of reporting group information update. \( \omega_{gm} \) and \( \omega_{ass} \) stand for the weights of group maintenance messages and individual association messages, respectively. Therefore, we can have the wireless cost of group LM as follows.

\[
C_{GLM} = \frac{M}{K} \cdot \frac{\bar{v}_{HCT}}{2d - 1} + \omega_{gm} \cdot \frac{M}{K} \cdot \frac{1}{T_{gr}} + \omega_{ass} \cdot \frac{M \cdot \bar{v}_{HCT}}{\bar{L}}
\]  

(7–4)

The wireless gain through grouping is hereby defined as follows.

\[
\zeta_{GLM} = \frac{C_{LU}}{C_{GLM}}
\]

(7–5)

The first item in Eq. (7–4) stands for the group LU traffic cost, which is reduced \( K \) times compared to the traditional location update. Since different messages in wireless have different impact on system performance, we put different weights for them. This part has unit weight in order to be consistent with Eq. (7–3). The second item stands for the incurred traffic of updating group information by GHs. The last item is the MTs’ association cost, which needs only once during the whole trip. Obviously, with the aim of mitigating congestion, \( \omega_{gm} \) and \( \omega_{ass} \) should be less than 1. It can be observed from Eq. (7–4) that when group size \( K \) and \( \bar{L} \) are big, the total cost is greatly reduced. We can easily conclude that when \( K \) is relatively bigger, the wireless gain is mainly determined
by the ratio of \( \frac{\bar{L}}{\omega_{\text{ass}}(2d-1)} \). When \( \bar{L} \) is much bigger than \( 2d - 1 \), the diameter of a location area, the wireless gain is determined by the value of \( K \).

Besides this gain, grouping greatly disperses the signaling along the timeline, leading to the mitigation of RACH congestion, which is a more important benefit of grouping method.

**7.4.3.2 Finding \( T_{ag} \)**

As aforementioned, the wireless gain of the proposed grouping scheme depends on the value of \( T_{ag} \). If \( T_{ag} \) is too small, the group size \( K \) can be big while the average trip length of the passengers \( \bar{L} \) is shortened since it is easy to join in the group and many non-passengers are recruited in the group. On the contrary, when \( T_{ag} \) becomes big, \( K \) will be reduced and \( \bar{L} \) is lengthened. In this case, many short-term passengers might be excluded from the group.

There is no available current model which can define the functions of \( \bar{L} \) and \( K \) in terms of \( T_{ag} \). This is a task which necessitates many real data. Suppose we acquire the relationship, i.e. \( \bar{L} = f(T_{ag}) \) and \( K = g(T_{ag}) \), we can derive the optimum \( T_{ag} \) with Eq. (7–3), (7–4) and (7–5).

**7.4.3.3 Other prospective benefits**

Beside group LU, grouping mechanism in MCNs can potentially improve the system’s performance in many ways. For example, GHs can help the adjacent cells to reserve resource for better QoS provision. GHs might provide an alternative way for a better handoff. GHs can even provide in-vehicle data service. Group concepts in cellular systems may quickly go beyond what we can picture for the time being. Using GHs in the HCT vehicles provides a necessary hardware preparation.

### 7.5 Generic Aggregative Location Management in MCNs

#### 7.5.1 Overview

When big events are held, a lot of people gather at the same location around the same time which usually brings wireless congestion as well as the traffic jams. We
cannot use grouping technique to solve the LU congestion because the MTs are not following the same path. However, we still can take advantage of the ad hoc feature of MCNs to aggregate the LU requests.

In this paper, we propose to utilize the aggregation devices (ADs) along the location area boundaries to collect the location update requests via ad hoc links periodically. After each time interval $T_{aggr}$, these ADs combine the collected location update requests into one message and send to the eNBs so that the RACH congestion can be greatly mitigated. Each eNB is required to decide the $T_{aggr}$ for the ADs connecting to itself according to the traffic load in the cell. To encourage mobile users to delegate their LUs to the ADs, $T_{aggr}$ should be set smaller than the expected delay caused by prospective collisions and retransmissions. Shorter delay of LU means smaller probability of missing incoming calls. Therefore, ADs which can provide smaller delay with a proper $T_{aggr}$ can attract MTs to delegate their location updates and thus mitigate the usage of RACH.

Since these ADs periodically access the system to do the LU for multiple MTs, eNBs can further allocate fixed channel resource for them and then the repeated contention for RACH is unnecessary. Next we describe the generic scheme of aggregative LM in MCNs.

7.5.2 Scheme Description

The ADs in this scheme can be any device which has the aggregation software built in. The ADs and the delegating MTs should have both ad hoc interface, usually a WiFi card, and cellular interface. The ADs should be authorized to be delegated with the LU task by the systems. Moreover, the ADs cannot be the ones in high mobility for fulfilling their aggregation tasks. When the ADs are in the positions where the LM aggregation is needed, they collect the delegation of location update requests through ad hoc links and launch LU procedure periodically, acting as the agents of corresponding MTs. To indicate the arriving MTs to delegate the location update task to ADs, each eNB sends out indications in the broadcast channel. The indication messages should also include
the list of the authorized ADs. Although security concerns should be included in this process, we do not cover them in our scheme since they are out of the scope of this paper. Each MT searches for the nearby ADs and decide whether or not to delegate LU according to the broadcasted information from both the AD and eNB. How to indicate the LU congestion probability \( P_{\text{col}} \) in RACH and what the interval length of ADs’ aggregation period \( T_{\text{aggr}} \) should be are the two major design issues in this scheme.

\[
P_{\text{col}} = 1 - e^{-\sigma \lambda_{\text{RACH}}}
\] (7–6)

Figure 7-3. Generic Aggregative Location Management
With the knowledge of timeout length of LU $t_{LU}$ and the transmission delay $D_t$, the expected delay of LU via RACH can be known as well, as shown in Eq. (7–7). The second item of $D_{RACH}$ is simply the product of the expected retransmission times and $t_{LU}$.

$$D_{RACH} = D_t + \frac{t_{LU} \cdot P_{col}}{1 - P_{col}}$$  
(7–7)

The total LU delay $D_{LU}$ is the summation of $D_{RACH}$ and wireline signaling delay $D_{wire}$, which can be expressed as $D_{LU} = D_{RACH} + D_{wire}$. Furthermore, with the knowledge of call-to-mobile ratio $\gamma$, we can know the expected number of missed calls.

$$N_{missed} = D_{LU} \cdot \gamma$$  
(7–8)

On deciding the aggregation period $T_{aggr}$, we first need to find the constraints. $D_{ALU}$ denotes the wireless delay of the proposed aggregative scheme which is mainly due to the aggregation interval. $D_{CSMA}$, the CSMA access delay, also contributing to $D_{ALU}$, can be ignored when the traffic is light.

$$D_{ALU} = D_{CSMA} + D_{aggr}$$  
(7–9)

$D_{CSMA}$ can be expressed by Eq. (7–10), where $P_f$ is the failure rate of one transmission in CSMA, $W_i$ is the backoff window size of $(i + 1)$th transmission, and $EIFS$ is the timeout length for one failed transmission.

$$D_{CSMA} = D_t + \frac{W_0}{2} + \sum_{i=1}^{Retrans} \left({EIFS} + \frac{W_i}{2}\right) \cdot P_f^i$$  
(7–10)

Obviously, when $D_{ALU}$ has smaller value than $D_{RACH}$, i.e. $D_{ALU} < D_{RACH}$, MTs prefer to delegating LUs to ADs.

When the access of ADs is not congested, the expected value of $D_{aggr}$ is $T_{aggr}/2$. Therefore, we have the first constraint of $T_{aggr}$ as Ineq. (7–11).

$$T_{aggr} \leq 2 \times \left(\frac{t_{LU} \cdot P_{col}}{1 - P_{col}}\right)$$  
(7–11)
If one AD gathers too much LU delegation, the delay of accessing AD via ad hoc links, $D_{CSMA}$, can be so big that $D_{ALU}$ will be expected to be bigger than $D_{RACH}$ even if $T_{aggr}$ is set with a small value. The MTs can choose other ADs or contend for the RACH instead.

There is another constraint of the aggregation interval, which is that the aggregated traffic should be no more beyond the capacity. Although the periodical aggregated location update via ADs can bypass the contention of RACH, there is still an upper-bound of acceptable LU traffic, denoted as $\hat{\lambda}_{LU}$, which is related to the traffic of new calls and handoff calls. Regarding the aggregative LU traffic as the Poisson process, we have the second constraint expressed as follows. $N_{AD}$ denotes the number of ADs, $\lambda_{LU}$ is the LU traffic, and $\delta$ is the aggregation rate of the LU traffic.

$$\frac{N_{AD}}{T_{aggr}} + (1 - \delta) \cdot \lambda_{LU} \leq \hat{\lambda}_{LU}$$

$$\Rightarrow T_{aggr} \geq \frac{N_{AD}}{\hat{\lambda}_{LU} - (1 - \delta) \cdot \lambda_{LU}} \tag{7–12}$$

Ineq. (7–12) indicates that, when all the LU traffic goes through ADs, i.e. $\delta = 1$, larger aggregation interval can gather more LU traffic and reduce the wireless cost more.

With these two constraints, we can roughly decide the aggregation interval. It should be no bigger than a certain value as in Ineq. (7–11) in order to stimulate mobile users to use ADs. It also needs to be no smaller than another certain value as in Ineq. (7–12) so that the aggregative traffic would not exceed the handling capability of the eNB. Although some certain ADs may not collect any LU delegation during some periods due to skewed LU traffic distribution, we ignore in the analysis the possibility that these ADs would release the assigned channels and re-contend the RACH when new LU traffic arrives. In the case that the value of $T_{aggr}$ cannot satisfy both inequalities at the same time, we simply remove the first constraints (Ineq. (7–11)). This case means that the LU traffic has not congested RACH and the delay via RACH is acceptably small. Therefore, there is no need to stimulate the aggregation by blindly shortening the
aggregation interval. It is also worthy to note that only when there is RACH congestion is
Ineq. (7–11) required.

According to the value decided based on the above two inequalities, we can have
the wireless cost gain as follows.

\[
\zeta_{ALU} = \frac{\lambda_{LU} \cdot T_{aggr}}{N_{AD} + (1 - \delta) \cdot \lambda_{LU}}
\] (7–13)

7.6 Location Management Cost Analysis and Application in MCNs

Although we only strive to reduce the wireless cost of location update, the proposed
scheme can potentially change the overall design of LM. Before we can achieve this
goal, we need to deepen our understanding of the LM. Previously there are several
papers formulating the cost function of LM. Unfortunately, none of them shows the
inversely proportional relationship between LU and paging in their formulation.

To formulate the cost function of LM, we need to have the knowledge of two
parameters. One is the call-to-mobile ratio, \( \gamma \), which shows the probability that a mobile
user is to be called. The other is the averaged moving speed, \( \nu \), which shows how
soon an MT is to change the location area in average. We assume that the location
area consists of a cluster of cells which form a series of concentric hexagons as
aforementioned. We denote the radius as \( d \) in number of cells. For the convenience of
understanding we also define the average number of users in the location area, \( M \), and
the time interval for the cost measurement, \( \tau \).

The LM cost consists of the wireless cost of location update and the wireless cost of
paging, as shown in Eq. 7–14.

\[
C_{total} = \omega_{LU} \cdot C_{LU} + \omega_{paging} \cdot C_{paging}
\] (7–14)

To minimize the overall cost, we need to put different weights on these two parts. \( \omega_{LU} \)
and \( \omega_{paging} \) stand for the weights for LU and paging, respectively. Apparently, since
the RACH resource is more critical than PCCH as we mentioned before, we have the following relationship, $\omega_{LU} >> \omega_{paging}$.

LUs happen when the MTs detect the location area changed. Assuming each MT moves straightly with the average speed $v$, we can expect an LU to occur at most after it moves across the diameter length of the location area, which is $2d - 1$. Therefore, for $M$ MTs in $\tau$ duration, the wireless cost of LU can be expressed as follows.

$$C_{LU} = \frac{v \cdot M \cdot \tau}{2d - 1} \quad (7-15)$$

Different from LU, paging will happen at each cell within the location area, whenever there is an incoming call or packet. Note that $M$ and $\tau$ are in both equations and hence can be removed before further analysis.

$$C_{paging} = \gamma \cdot M \cdot \tau \cdot (3d^2 - 3d + 1) \quad (7-16)$$

By finding the first and second derivative of $C_{total}$, we can obtain the minimum value of $C_{total}$ and the corresponding $d$.

$$\frac{\partial C_{total}}{\partial d} = -\frac{2\omega_{LU} \cdot v}{(2d - 1)^2} + \omega_{paging} \cdot \gamma \cdot (6d - 3) \quad (7-17)$$

$$\frac{\partial C_{total}}{\partial d^2} = \frac{8\omega_{LU} \cdot v}{(2d - 1)^3} + 6\omega_{paging} \cdot \gamma \geq 0 \quad (7-18)$$

Since the second derivative is positive, by letting the first derivative equal to 0, we can know $d$ which is making $C_{total}$ minimum.

$$d_{opt} = \sqrt{\frac{\omega_{paging} \cdot \gamma}{\omega_{LU} \cdot v}} + 1 \quad (7-19)$$

### 7.7 Performance Evaluation

We investigate the benefits of our proposed schemes in terms of wireless cost in this section.
For the grouping scheme, we look into the grouping gains in terms of wireless signaling cost with different group sizes, $K$, and different average trip length, $\bar{L}$. The signaling cost is defined as Eq. (7–4) and Eq. (7–3). The signaling gain is defined in Eq. (7–5). The basic settings are listed in Table 7-1.

Table 7-1. Parameters’ Settings For Evaluating The Grouping Scheme

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$M$</td>
<td>800</td>
</tr>
<tr>
<td>$d$</td>
<td>4cell</td>
</tr>
<tr>
<td>$\omega_{gm}$</td>
<td>0.2</td>
</tr>
<tr>
<td>$\omega_{ass}$</td>
<td>0.02</td>
</tr>
<tr>
<td>$T_{gr}$</td>
<td>0.5hour</td>
</tr>
<tr>
<td>$\bar{v}_{HCT}$ (if it is not the variable)</td>
<td>6.0cell/hour</td>
</tr>
<tr>
<td>$L$ (if it is not the variable)</td>
<td>7.3cell</td>
</tr>
</tbody>
</table>

By setting $\bar{v}_{HCT}$ as 1.5, 3.0 and 6.0, we calculate the gains of the signaling cost. Fig. 7-4 shows the result. We can see that the gains increase with the group size $K$, however, not linearly. Bigger $\bar{v}_{HCT}$ brings bigger gain since the periodical group reporting occupies less and less signaling cost compared with other $\bar{v}_{HCT}$ related signaling. Because the individual association’s cost is not related to $K$, the gains do not have a linear relationship with $K$.

Figure 7-4. Grouping Gains of Signaling Cost With Different Group Sizes ($K$)

Fig. 7-5 shows the comparison of the gains with different average trip length $\bar{L}$. We set $K$ as 10, 50 and 200 for the calculations. Apparently, bigger $\bar{L}$ along with bigger $K$ brings bigger signaling cost’s gains. When $K$ is as big as 200, the gains almost increase
linearly with \( \bar{L} \). This means when the group size is large enough, the cost of location update and group reporting is relatively negligible so that bigger average trip length directly increases the total gain.

Figure 7-5. Grouping Gains of Signaling Cost With Different Trip Length (\( \bar{L} \))

Although big \( K \) and big \( \bar{L} \) together bring great gains, they usually do not come together. We have mentioned earlier that in a certain scenario, when we set the association timer with a big value, the big average trip length can be acquired whereas with the group size sacrificed. Short association timer is on the contrary. Therefore, if we want to optimize the signaling cost’s gain, we need first gather real data and find the relationship between \( K \), \( \bar{L} \) and the timer.

For the generic aggregation scheme, we need to first look into the interval \( T_{aggr} \)’s range. Based on the longest allowed \( T_{aggr} \), we then observe the gain of signaling cost. Before we can evaluate the scheme, we list the parameters’ setting as Table 7-2.

Table 7-2. Parameters’ Settings For Evaluating The Generic Aggregation Scheme

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \lambda_c )</td>
<td>2.5( p/s )</td>
</tr>
<tr>
<td>( \lambda_{LU} )</td>
<td>6.0( p/s )</td>
</tr>
<tr>
<td>( \sigma )</td>
<td>0.1( s/p )</td>
</tr>
<tr>
<td>( t_{LU} )</td>
<td>5( s )</td>
</tr>
<tr>
<td>( \delta )</td>
<td>0.8</td>
</tr>
<tr>
<td>( N_{AD} ) (if it is not the variable)</td>
<td>20</td>
</tr>
</tbody>
</table>
The upper bound of $T_{aggr}$ is defined in Ineq. (7–11) and the lower bound is defined in Ineq. (7–12). By changing the LU traffic load, we can see the change of the range of the aggregation interval.

![Figure 7-6. The Aggregation Interval's Ranges with Different LU Traffic](image)

Fig. 7-6 shows the calculation results with different LU traffic load. We can see that the lower bound of the aggregation interval does not change apparently with the LU traffic. This is because only non-aggregated LU traffic, which is only a small proportion, affects the lower bound of the aggregation interval. The upper bound of the aggregation interval increases exponentially with the LU traffic in that it depends on the collision probability in a slotted-ALOHA channel.

Fig. 7-7 shows the signaling cost's gain of the generic aggregation scheme. This scheme does not bring other extra traffic and aggregates the LU traffic only. By setting the aggregation timer with its biggest allowed value, we calculate the gains. It shows that fewer ADs can apparently bring bigger gains. This is because fewer ADs are required to aggregate more LU traffic. However, if ADs are not enough, high aggregation rate $\delta$ cannot be guaranteed and more congestions will happen in AD access.

It is worthy to note that this performance evaluation does not include periodical LU for the purpose of simplicity. However, periodical LU can be easily aggregated with our aggregative approach and thus higher signaling cost gain will be achieved actually.
7.8 Summary

Based on MCN structure, in this paper, we proposed two schemes to aggregate the LU traffic and thus mitigate the potential LU congestions. The grouping scheme is designed for the HCT scenarios where many MTs move together in HCT vehicles. By forming a big group of moving-in-together MTs, this scheme combine the regular LU and periodical LU into group LU messages, which lead to less wireless cost even with the incurred association messages and other group maintenance messages. For the scenarios where MTs gather without a uniform path, the proposed generic aggregation scheme provide a mechanism to aggregate the arrived LU requests and deliver them to the network in batch which is apparently a more efficient way. The feasible value of the aggregation interval has been extensively discussed. Further in this paper, the overall LM cost has been analyzed and an simple algorithm to find the optimum LM has been proposed. For MCNs, due to the existence of ad hoc links, the LM can have less wireless cost with the help of extra wireless spectrum. The aggregative approach in MCNs will provide a substantial help for various aspects of future networks.
CHAPTER 8
CONCLUSIONS

Wireless ad hoc networks (WANETs) have a variety of applications. They can appear in stand-alone forms such as MANETs, wireless sensor networks, or wireless mesh networks. They can also appear as one embedded component in hybrid systems, such as MCNs. No matter in which forms, WANETs have multi-hop transmissions and self-organization as their outstanding characteristics. These characteristics bring us the benefits of low-cost, convenience and flexibility and in the meantime they are the reasons of severe interference, bad spatial reuse and low throughput performance. It needs a lot of efforts to employ them efficiently.

This dissertation presents the efforts on utilizing WANETs from two aspects, throughput improvement and mobility support.

Although PHY layer techniques are the driving force of raising data rate, it is MAC layer that controls spectrum efficiency in multi-user scenarios. Due to the multi-hop property, schedule-based MAC will finds it limitation in WANETs and contention-based MAC appears to be a better choice. In multi-channel systems, channel assignment always cost a substantial part of spectrum. Reducing this part of signaling will greatly increase the MAC throughput. However, high MAC throughput does not always mean high end-to-end throughput. A flow's end-to-end rate is decided by the hop with the lowest rate. Therefore, it is important to allocate more spectral resource to the area with heavy traffic load. Moreover, in multi-hop WANETs, spatial reuse is crucial to resource allocation since if more spectrum can be reused then larger throughput can be expected. In the special case of WANETs, wireless mesh networks, where all the traffic goes to the gateways finally, resource allocation can be done in a semi-centralized semi-distributed way so that the central areas can easily acquire more spectrum and the outer areas can easily reuse the spectrum. A step closer, nodes with heavy traffic in a neighborhood should acquire more resource than the others for the same
reason. Although distributed scheduling provides an approach to achieve this goal, it is less helpful for contention-based networks. Distributed resource allocation among neighboring nodes appears to be the suitable approach. Using this approach, we should not overlook the fact that the traffic of neighboring nodes is correlated with each other since forwarding neighbors’ packets is a major task for nodes in WANETs. Incorporating the consideration of the traffic dependency into resource allocation in WANETs will help us to find the locally optimized end-to-end throughput. Obviously, a cross-layer design should be included in this incorporation. Wireless MAC layer has been well researched for many years. However, for different scenarios, there are still a lot of white space before we can reach the optimum throughput performance.

Due to the lack of infrastructure, it is difficult for WANETs to support mobility, which is usually managed in a hierarchical structure. To solve this, we can add some functional modules into the systems to build the hierarchical structure. For example, we can add foreign agents (FA) and home agents (HA) into wireless mesh networks so that mobile IP can be supported. However, mobile IP is not a suitable solution for intra-domain roaming because frequently changing IP brings frequent service interruption. We can store the locations (serving mesh routers) of mesh nodes in the gateway as the anchors and use temporary routing entries between the anchors and the exact locations to reach the destinations. This approach combines the advantages of tunnel-based approach and routing-based approach and provides a suitable micro-mobility solution for wireless mesh networks. When we integrate WANETs into cellular systems, we can achieve WANETs’ advantages in cellular systems. It is interesting to look into the mobility issues in such systems. With multi-hop connections existing, roaming clients can access the resource of the adjacent cells and better handoff performance can be thus expected. If we form embedded ad hoc networks in the systems, the information of handoff options can be collected and maintained before handoff takes place. We can choose better handoff options to reduce the handoff dropping rate. We can adjust the handoff
reservations in BSs according to the information exchanged among neighboring BSs so that the spectral resource will be used more efficiently. Moreover, we can utilize the ad hoc links in cellular system to aggregate the location management signallings and relieve the potential congestions caused by heavy arrivals of location update requests. This method is especially helpful in high capacity transit systems.

Surely, many more issues need to be solved in order to make use of WANETs well. I am currently looking into the routing algorithms in WANETs and thinking of the ad hoc routing issues when embedding WANETs into cellular systems.
REFERENCES


BIOGRAPHICAL SKETCH

Rongsheng Huang received his BS and MS degrees in electrical engineering from Xi’an Jiaotong University, Xi’an, China, in 1996 and 1999, respectively. From 1999 to 2001, he was working with Huawei Technologies Co. Ltd. as an R&D engineer on GPRS and 3G projects. From 2002 to 2005, he was working with UTStarcom Research Center, Shenzhen, China, as a senior engineer and team leader on 3G project. Since 2005, he has been working on the PhD degree in the Department of Electrical and Computer Engineering at University of Florida. His research interests cover the area of media access control, mobility management, protocol and architecture design for wireless networks. He is now a student member of the IEEE.