AN ABSTRACT OF THE DISSERTATION OF

Jong bum Ryou for the degree of Doctor of Philosophy in
Electrical and Computer Engineering presented on June 3, 2011.

Title: Adaptive Load Balancing Metric for WLANs

Abstract approved: _____

Ben Lee

As the number of mobile devices accessing large-scale WLANs such as campus and metropolitan area networks increases, the need for load balancing among the cells becomes crucial. In addition, the network must also support some minimum handoff tolerance defined by an application.

A number of load balancing techniques have been proposed in the literature that focuses on formulating new load metrics rather than using *Received Signal Strength Indicator* (RSSI) as the association metric. These schemes consider a variety of factors such as number of STAs, enhanced RSSI, channel utilization, queue length, bandwidth, and throughput to achieve balanced load. However, some of these techniques require protocol modifications to both APs and STAs or need special agents such as admission control server, extra software, and switches. Others do not consider *Quality of Service* (QoS) requirements of applications, which vary from one application to another, and thus do not satisfy

users requiring minimized handoff latency and real-time services. Moreover, most techniques ignored the *hidden node problem*, which causes packet collisions and thus the presence of such nodes can severely affect the performance of WLANs.

This dissertation proposes a new metric that provides load balance as well as timely handoffs for WLANs by taking into account both direct and hidden node collisions as well as the types of traffics in order to support QoS. Another novel feature of the proposed method is the use of probe requests during the discovery phase to monitor the states of the channels to determine the best *Access Point* (AP) for association. Our simulation results show that the proposed method is significantly better than relying only on signal strength in term of utilization, end-to-end delay, collision rate, and packet loss.

©Copyright by Jong bum Ryou
June 3, 2011
All Rights Reserved

Adaptive Load Balancing Metric for WLANs

by

Jong bum Ryou

A DISSERTATION

submitted to

Oregon State University

in partial fulfillment of
the requirements for the
degree of

Doctor of Philosophy

Presented June 3, 2011
Commencement June 2011

Doctor of Philosophy dissertation of Jong bum Ryou presented on June 3, 2011.

APPROVED:

Major Professor, representing Electrical and Computer Engineering

---

Director of the School of Electrical Engineering and Computer Science

---

Dean of the Graduate School

I understand that my dissertation will become part of the permanent collection of Oregon State University libraries. My signature below authorizes release of my dissertation to any reader upon request.

Jong bum Ryou, Author

ACKNOWLEDGEMENTS

I would like to express my immeasurable gratitude to my advisor Prof. Ben Lee. It has been an honor for me to have worked under his supervision. He has been a great source of motivation and encouragement not only for my Ph.D. degree but also for human life through the entire process. This dissertation would have not been possible without his constant special guidance and professional expertise. I look forward to his continued friendship.

I also want to express my appreciation to Prof. Mario E. Magaña, Prof. Huaping Liu, Prof. Thinh Nquyen, Prof. Luca Lucchesse, and Prof. Abdollah Tavakoli for serving in my graduate committee and helping my study with discussions and comments. In addition, I would like to thank Prof. Chansu Yu at Cleveland State University and Mohammed Sinky for their comments and proof reading my dissertation.

I would also like to express my thanks to Republic of Korea Air Force and military officers of seniors, juniors, and colleagues for giving me an opportunity to study at Oregon State University with full support.

Lastly, I would like to thank my family all most sincerely for their support in everything. Specially thank my mother, Minja Jeong and mother-in-law, Kisun Kim for their caring and encouragement for my whole life. Sincerely thank my

wife, Yunjeong Paeng for her tireless support, patience, love and friendship, and thank my lovely children, Daehyun (Danny) and Hyunmin (Harry) for their love and great pleasure. In addition, thank my father, Moonho Ryou who is now heaven and proud of me for not giving up and completing this work. I forever long for his voice giving me the strength to keep up with my study.

TABLE OF CONTENTS

TABLE OF CONTENTS (Continued)

LIST OF FIGURES

LIST OF FIGURES (Continued)

LIST OF TABLES

LIST OF APPENDIX FIGURES

Chapter 1 – Introduction

Large-scale WLAN deployment is popular in locations such as conferences, university campuses, and airports, as well as metropolitan areas due to their low cost and ease of deployment [1]. A WLAN consists of multiple Access Points (APs) with overlapping cells to provide a wide coverage and offers high transmission rates. In current implementations, each user associates with an AP with the strongest signal strength. However, recent studies have shown that this simple approach leads to inefficient association of mobile stations (STAs) to available APs [2–7]. In particular, multimedia applications, such as streaming video and audio, video conferencing, VoIP, and interactive games, require a certain level of guaranteed *Quality of Service* (QoS) in terms of bandwidth, delay, jitter, and packet loss. Uneven distribution of user loads among APs increases congestion and packet loss, and thus reduces throughput. This results in inefficient medium utilization, reduced performance, and occasionally, network collapse. Therefore, proper *AP selection* is an important issue in WLANs.

A number of load balancing techniques have been proposed in the literature that focuses on formulating new load metrics rather than using *Received Signal Strength Indicator* (RSSI) as the association metric. These schemes consider a variety of factors such as number of STAs [8], enhanced RSSI [7, 9], channel utilization [10], queue length [11–13], bandwidth [14–16], and throughput [17, 18] to

achieve balanced load.

However, there are several issues with these methods. First, most of these methods require protocol modifications to both APs and STAs [8, 9, 19] or need special agents such as admission control server [7], extra software [10, 16, 18], and switches [20]. In addition, these approaches require APs to monitor the state of the channels and the details of the entire network by using a lot of extra information frames, which waste bandwidth and decrease overall throughput. Second, most of the STA-based methods [11, 12, 14, 15, 17], which function without assistance from APs, monitor the load after associating with an AP. Thus, a STA can only evaluate the load balancing metric of the cell it is associated with, which results in extra delay when the STA searches for a better channel and reassociates with another AP. In addition, a ping-pong effect occurs when the STA frequently changes its association, which in turn degrades the performance of WLANs. These limitations do not satisfy users requiring minimized handoff latency and real-time services. Third, prior techniques ignored the *hidden node problem*, which causes packet collisions and thus the presence of such nodes can severely affect the performance of WLANs. Fourth, most of the prior techniques do not consider QoS requirements of applications, which vary from one application to another. For example, in load balancing schemes that consider bandwidth as the main metric [14–16], all application types are treated equally even though some applications do not have tight constraints on bandwidth, such as email, file sharing, and web surfing. Although there have been numerous research efforts that separately consider load balancing with or without QoS and the hidden node problem for WLANs, unfortunately no

work exists that considers both issues at the same time.

Therefore, this dissertation proposes a new *Probe Request based Adaptive Load Balancing Metric* (PR-ALBM), which has several unique features. First, probe requests are utilized during the discovery phase to unobtrusively observe the states of all the adjacent channels to decide on the best AP to achieve a balanced load. This is done by observing the frequency of DCF (Distributed Coordination Function) Interframe Spaces (DIFSs), probing delay, and different types of traffic to evaluate not only contention rate but also end-to-end delay for each channel. Second, a fixed, *optimized backoff time*, instead of random backoff time, is used for probe request frames. The choice of backoff time, which is represented in terms of number of slots, is optimized for each application to provide a sufficient amount of time to properly observe the channel and yet lead to timely handoff. Third, a load balancing metric is developed based on the probabilities of carrier sense and hidden node collisions, where the former is based on the number of DIFS and probing delay and the latter is modeled using an M/M/1/K queue. Finally, PR-ALBM can be applied to all types of IEEE 802.11 networks, including IEEE 802.11e, without requiring any modification to the APs or new agents, such as load balancing servers and extra software.

Chapter 2 – Background

The basic *Medium Access Control* (MAC) mechanism of 802.11 is the *Distributed Coordination Function* (DCF). Although IEEE 802.11 has become more and more popular, it does not support QoS in terms of throughput, end-to-end delay, jitter and packet loss.

IEEE 802.11e [21] is an enhanced version of IEEE 802.11 with improved DCF and mechanisms similar to DiffServ, which is used in wired networks in order to support QoS requirements. In IEEE 802.11e, each traffic is assigned one of several user priorities according to the type of application. In addition, service differentiation is performed using a different set of medium access parameters for each user priority.

2.1 Distributed Coordination Function (DCF)

2.1.1 Introduction

The IEEE 802.11 standard provides MAC and *Physical* (PHY) layer functionality for wireless connectivity of different types of STAs, which can be fixed or mobile at pedestrian and vehicular speeds within a local area [22]. Table 2.1 shows the basic comparison of the different 802.11 standards.

In June 1997, the *Institute of Electrical and Electronics Engineers* (IEEE) ini-

tially released the 802.11 WLAN standard, which defines MAC and PHY layer specifications [23]. The initial IEEE 802.11 standard defines an *Infrared* (IR) layer and two different spread spectrum radio layers such as *Frequency Hopping Spread Spectrum* (FHSS) and *Direct Sequence Spread Spectrum* (DSSS) with data rates of 1 Mbps and 2 Mbps. The IR layer has not been popular, but FHSS and DSSS are widely used within the 2.4 GHz band.

In 1999, the IEEE standardized two enhanced versions of 802.11a [24] and 802.11b [25]. 802.11b also operates in the 2.4 GHz band based on DSSS and at speeds of up to 11Mbps. Most WLANs today comply with the 802.11b version [22]. 802.11a operates in the 5 GHz band and at up to 54Mbps using *Orthogonal Frequency Division Multiplexing* (OFDM).

In 2003, the IEEE released 802.11g [26], the third modulation standard based on 802.11b in the 2.4 GHz band. It can support transmission rates higher than 802.11b (up to 54 Mbps) due to OFDM.

In 2009, the IEEE published 802.11n, which improves the previous 802.11 standards by using *Multiple Input Multiple Output* (MIMO) antennas. It operates not only in the 2.4 GHz band but also in the less crowded 5 GHz bands.

Table 2.1: Basic Comparison of Different 802.11 Standards.

Types	802.11b	802.11a	802.11g	802.11n
Release	Sep 1999	Sep 1999	Jun 2003	Oct 2009
Frequency	$2.4GHz$	$5GHz$	$2.4GHz$	$2.4GHz$ or $5GHz$
Range	100-150ft	25-35ft	100-150ft	100-165ft
Speed (max)	$11Mbps$	$54Mbps$	$54Mbps$	$540Mbps$
Modulation	DSSS	OFDM	OFDM	OFDM

2.1.2 IEEE 802.11 Design

The basic building block of the 802.11 standard is the *Basic Service Set* (BSS), which is a set of all STAs that can communicate with an AP using the same channel. There are two types of BSS: Infrastructure BSS and *Independent BSS* (IBSS). Fig. 2.1 illustrates infrastructure BSS and IBSS, which are distinguished by either the presence or the absence of an AP, respectively.

In an infrastructure BSS, all communications take place through the AP (called a *cell* in cellular communications), including communications between STAs within the same BSS. Therefore, BSS represents a basic service area covering a set of STAs within the range of the AP. In addition, the infrastructure BSS adopts a *BSS identifier* (BSSID) defined by the MAC address of the AP in order to distinguish one BSS from an another.

In contrast, STAs in an IBSS can communicate directly with each other without a backbone infrastructure IBSS is often referred to as an ad hoc network due to quick network establishment and termination without a backbone infrastructure.

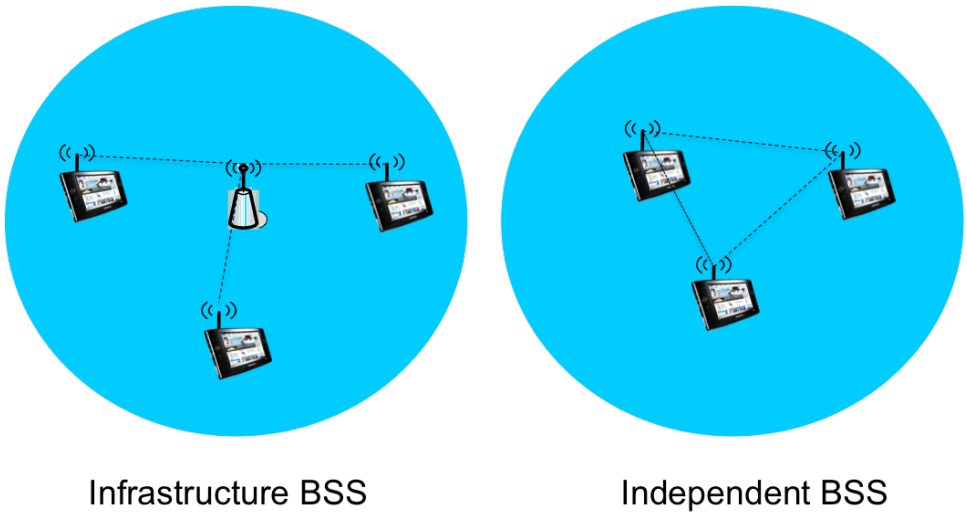

Figure 2.1: Infrastructure BSS and IBSS.

Several BSSs can be connected together via a *Distribution System* (DS) in order to extend network coverage to a larger area, which is called an *Extended Service Set* (ESS). The IEEE 802.11 standard does not restrict the composition of the DS, i.e., whether it is a IEEE 802 compliant network or not. A DS connects all the APs and acts as a bridge, and thus forwards network traffic to a STA and supports mobility of STA within an ESS. Fig. 2.2 shows an example of BSS and ESS.

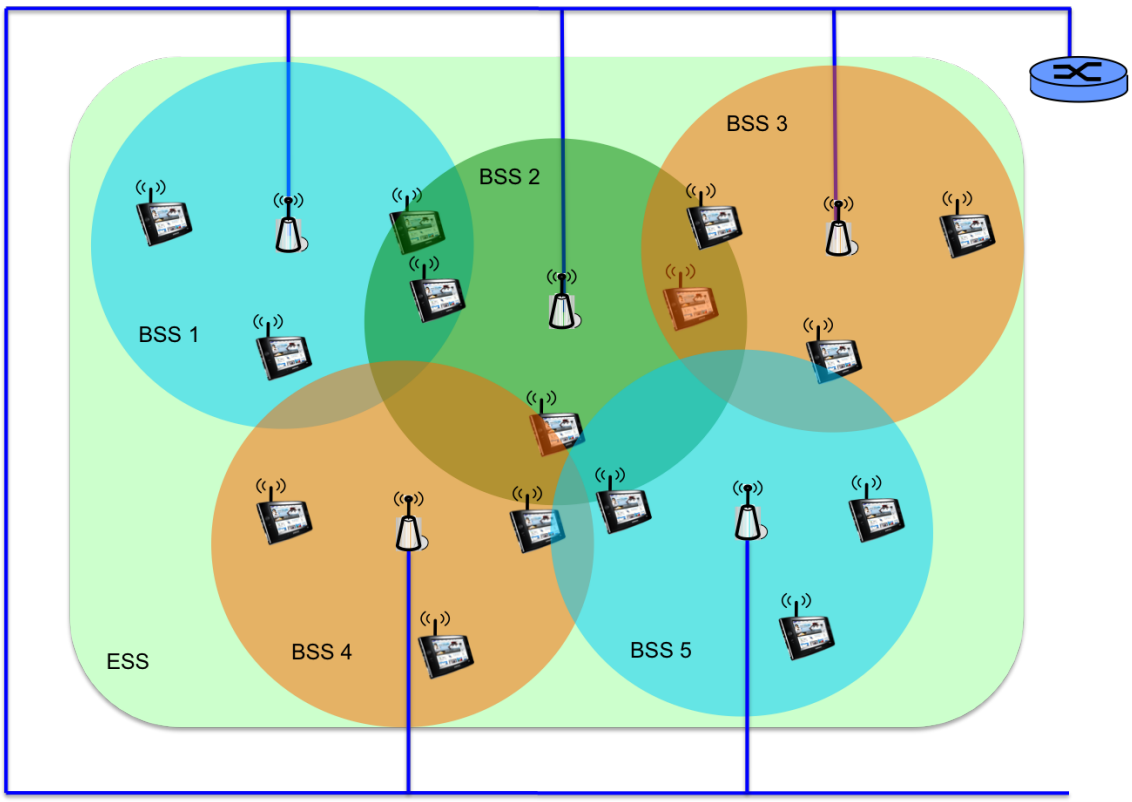

Figure 2.2: Figure of BSS and ESS.

2.1.3 Distributed Coordination Function (DCF)

The IEEE 802.11 MAC defines two different access mechanisms, *Distributed Coordination Function* (DCF) and *Point Coordination Function* (PCF). DCF is mandatory and the primary access protocol for sharing of the wireless medium between STAs and APs. Therefore, most traffic uses DCF since multiple STAs can transmit their data frame without central control regardless of whether infrastructure BSS or IBSS is employed. The optional PCF is priority based and provides a contention-free channel access mechanism. In PCF, STAs do not contend for the wireless medium and instead access to the medium is controlled by point coordinators that reside in APs. This dissertation discusses the operations of DCF in detail since our proposed metric is related to DCF.

DCF is the basis of *Carrier Sense Multiple Access and Collision Avoidance* (CSMA/CA). A STA first checks that the medium is available before transmitting a frame, which is called carrier sensing. In DCF, there are two types of carrier sensing functions: physical carrier sensing and virtual carrier sensing. Both functions are concurrently used to sense the medium and enable the MAC coordination to decide the status of the medium. Physical carrier sensing is provided by the physical layer and its result is sent to the MAC layer. On the other hand, virtual carrier sensing is provided by the *Network Allocation Vector* (NAV) performed by the MAC coordination. NAV is a timer for indicating how long the medium is busy and is set in the *duration* field of 802.11 frames. This information is used to reserve the channel for a certain time period as shown in Fig. 2.3. A STA that acquires

the medium reserves it by setting the NAV to the expected time to complete the current transmission. Meanwhile, other STAs sensing the medium read the duration field and count down from NAV to zero. When the NAV becomes zero, it indicates the medium is idle.

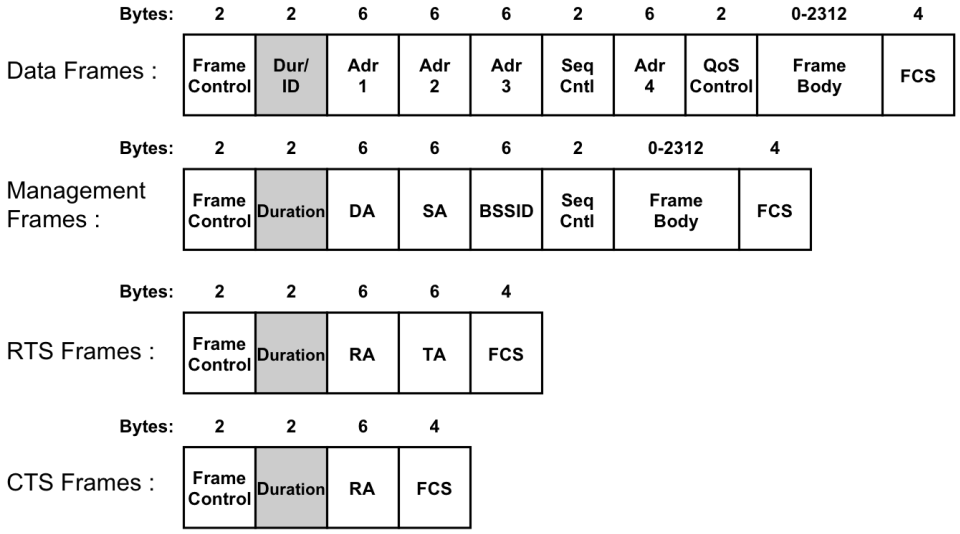

Figure 2.3: MAC Frame Format and Duration Field.

If the medium is idle for a time period equal to that of *DCF Inter-Frame Space* (DIFS) and *Random Backoff* (RBO), the STA starts transmission, but other STAs wait until the medium becomes idle again during the DIFS and RBO time period. When the receiver successfully receives the frame, it replies by sending an *Acknowledge* (ACK) frame after a *Short Inter-Frame Space* (SIFS) time period.

The IEEE 802.11 defines four different interframe spaces in order to coordinate access to the medium. Fig. 2.4 illustrates the relationship of these spaces. Technically, these spaces provide different levels of priority for different types of frames, such as ACK, RTS/CTS, management, and data frames.

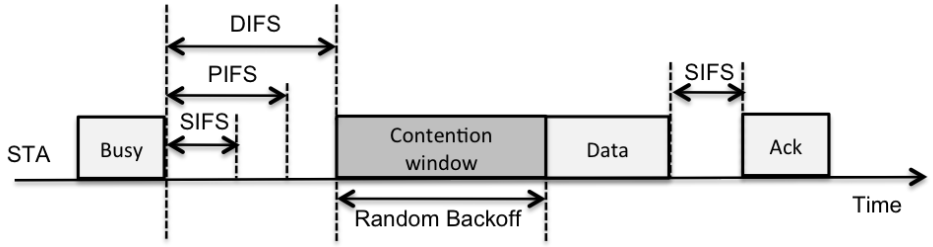

Figure 2.4: Relationship of Interframe Space.

SIFS is the shortest of the interframe spaces and is used by ACK and RTS/CTS frames. Since SIFS is shorter than PIFS or DIFS, the ACK and RTS/CTS frames have priority over other types of frames that can be transmitted from STAs. The second shortest interframe space, *PCF Inter-Frame Space* (PIFS), is used by AP in the PCF mode for contention free operation. For example, STAs controlled by PCF can send a frame after the PIFS has elapsed. Therefore, they have priority over another data frame transmitted from a STA in the DCF mode. DIFS represents the

longest interframe space and is used prior to sending data frames and management frames.

2.1.3.1 Random Backoff Procedure with DCF

IEEE 802.11 adopts a *Collision Avoidance* (CA) scheme to avoid collisions. This is because, unlike wired networks, collisions are hard to detect in wireless networks and waste valuable transmission capacity. In a wired ethernet environment, transceivers can transmit and receive at the same time and thus collisions can be detected. However, it is not possible to transmit and receive frames at the same time on the same channel using radio transceivers [22]. Therefore, WLANs adopt the CA scheme instead of *collision detection* (CD) schemes used in Ethernet to reduce collisions.

In order to perform CA in WLANs and thus reduce the probability of collisions among STAs in the same channel, a STA waits an additional time period, called a *Random Backoff* (RBO) time, following DIFS and before transmitting a frame. The RBO period is determined by the *Contention Window* (CW) consisting of CW slots, where the length of a slot depends on the underlying physical layer. RBO can be represented by the equation

$$RBO = Random() \cdot SlotTime, \tag{2.1}$$

where $Random()$ is a pseudo-random integer drawn from a uniform distribution

over the interval [0, CW-1] and SlotTime is a constant value that depends on the physical layer.

The STA starts decrementing its assigned RBO time slots by one slot during the RBO process as the medium is sensed to be idle for DIFS. When the medium is busy by other STAs during the RBO period, its counter is not decremented. The RBO process will resume when the medium is sensed again to be idle for DIFS. Finally, the STA can transmit when the RBO timer reach zero.

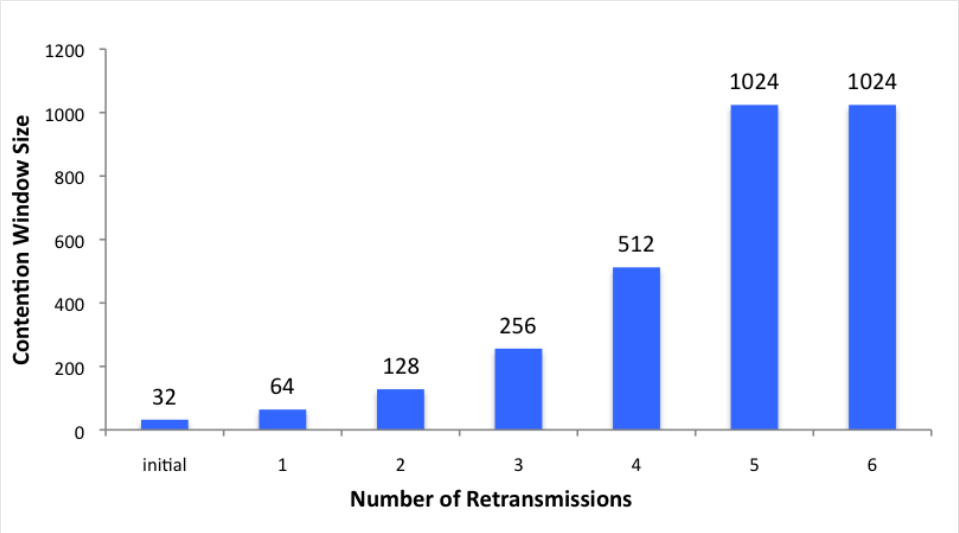

Figure 2.5: Exponential Random Backoff.

Fig. 2.5 illustrates the exponential Random Backoff process. Initially, the CW size is set to a CWmin value, which is the minimum size of CW when the STA first attempts to send a frame. However, when a collision occurs, CW increases exponentially up to a maximum value, CWmax. The size of CW is limited by the underlying PHY layer. For example, in the case of DSSS, the size of CWmin

and CWmax are 32 and 1024, respectively. Therefore, the size of CW increases exponentially (i.e., 32, 64, 128, 256, 512, and 1024) as shown in Fig. 2.5. When CW reaches its maximum size, this value is used until it can be reset. The CW is reset when transmission is successful or the number of retransmissions for a frame reaches a limit, referred to as the Maximum Retry Limit.

2.1.3.2 Hidden node problem and RTS/CTS Mechanism

A *hidden node* collision occurs when multiple nodes that cannot sense each other transmit at the same time. Thus, the presence of such nodes can severely affect the performance of WLANs. Moreover, the importance of handling the hidden node problem has increased with the increase of uplink data for applications such as VoIP, video gaming, and video conferencing. The hidden node problem is well known in ad hoc networks; however, relatively little work has been done to consider the problem in wireless infrastructure networks.

The IEEE 802.11 standard defines *Request To Send* (RTS) and *Clear To Send* (CTS) mechanisms to prevent *hidden node* collisions. Fig. 2.6 illustrates the RTS/CTS handshake process by exchanging RTS and CTS control frames.

When STA 1 has a frame to transmit, it starts the process by transmitting an RTS control frame and the receiver, STA 2 responds with a CTS frame after waiting for a SIFS time period. Then, STA 1 can transmit the frame. RTS/CTS can reserve the medium for the required duration of the complete frame exchange sequence including data frame and ACK using the NAV mechanism. In addition,

collisions are prevented because hidden nodes for STA 1 are noticed by the CTS from STA 2, which significantly reduces hidden node collisions.

Although RTS/CTS is designed to mitigate the hidden node problem, a significant amount of overhead is incurred and thus is typically not used in infrastructure mode [18]. Moreover, efficiency will be seriously degraded when small data frames are retransmitted.

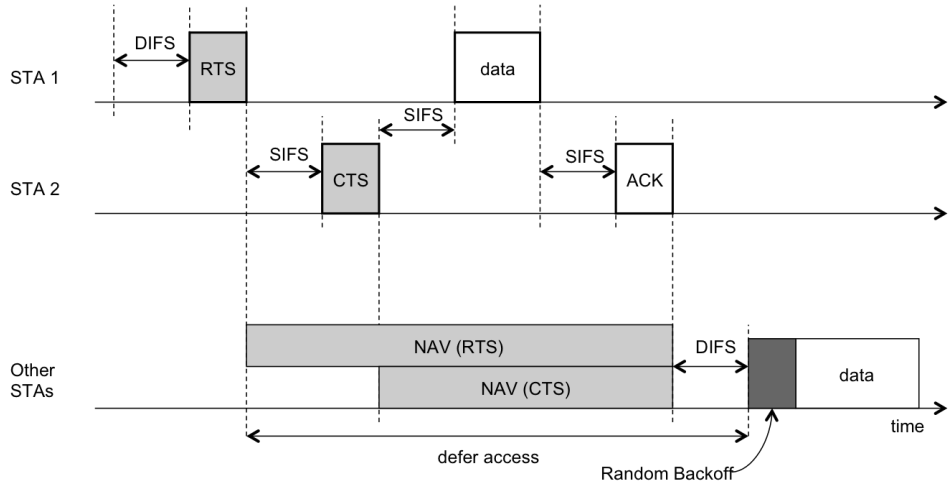

Figure 2.6: RTS/CTS Handshake Process.

2.1.4 Summary

DCF is based on *Carrier Sense Multiple Access with Collision Avoidance* (CSMA/CA). When STAs need to access a WLAN in either infrastructure or ad hoc mode, each STA checks whether the medium is idle before attempting to transmit a frame. Two types of carrier sensing functions, i.e., physical and virtual, are performed to

determine if the medium is available. *Virtual carrier sensing* is provided by the *Network Allocation Vector* (NAV), which is a timer for indicating how long the medium is busy and is set in the *duration* field of 802.11 frames. A STA that acquires the medium reserves it by setting the NAV to the expected time to complete the current transmission. Other STAs sensing the medium read the duration field and count down from NAV to zero. When the NAV becomes zero, it indicates the medium is idle. The STA then waits for the channel to become idle for the *Distributed Coordination Function Interframe Space* (DIFS) and performs the exponential *Random Backoff* procedure.

The random backoff period is determined by the *Contention Window* (CW) consisting of CW slots, where the length of a slot depends on the underlying physical layer. The backoff period is determined by a random integer drawn from a uniform distribution over the interval [0, CW-1]. The STA transmits a frame after waiting for its assigned random backoff time. When the medium is busy by other STAs during the random backoff, its counter is not decremented. This way, random backoff time is the only factor that determines priorities among contending transmissions.

When the receiver (i.e., AP) receives the frame, it sends back an *Acknowledgement* (ACK) frame after waiting for *Short Interframe Space* (SIFS). Since SIFS is shorter than DIFS, the ACK frame has priority over data frames that can be transmitted from other STAs. When the STA does not receive an ACK, it considers the frame to be lost and retransmits the frame. The size of CW is initially assigned to CW_{min}. However, the size of CW is doubled each time a frame is lost

with an upper bound of CW_{max}. When the transmission succeeds, CW is reset to CW_{min}.

2.2 Enhanced Distributed Coordination Function (EDCF)

2.2.1 Introduction

IEEE 802.11 DCF only provides best effort service and does not differentiate between different types of data traffic. However, there has been a significant increase in the use of mobile devices, not only various notebooks and netbooks, but also smart phones (e.g., iphone, Galaxy S) and tablet devices (e.g., iPad and Galaxy-tab) that can access WLANs using various applications.

Some of these applications, such as multimedia applications, require a certain level of guaranteed service, *Quality of Service* in terms of throughput, end-to-end delay, and packet loss, which vary from application to application. For example, VoIP has tight constraints on end-to-end delay, while FTP has tight constraints on data loss. Technically, all application types are treated equally by the IEEE 802.11 DCF even though some applications have tighter constraints on throughput, delay, and packet loss.

Therefore, IEEE 802.11e was developed to provide QoS for delay sensitive WLAN applications, such as streaming video and VoIP. IEEE 802.11e introduces *Enhanced Distributed Coordination Function* (EDCF), which is designed to provide prioritized QoS by supporting differentiated, distributed access to the medium

using different priorities for different types of data traffic similar to Diffserv in a wired network.

2.2.2 Enhanced Distributed Coordination Function (EDCF)

EDCF specifies enhanced interframe space, random backoff, and reservation mechanisms for the channel to support QoS.

Table 2.2: Priority and Access Category in 802.11e.

User Priority (UP)	Access Category (AC)	Types of Traffic
1 (lowest)	0 (BK)	Background
2	0 (BK)	Background
0	1 (BE)	Best Effort
3	1 (BE)	Best Effort
4	2 (VI)	Video
5	2 (VI)	Video
6	3 (VO)	Voice
7 (highest)	3 (VO)	Voice

IEEE 802.11e defines eight *User Priorities* (UPs) and four *Access Categories* (ACs) for different types of data traffic as shown in Table 2.2. Each frame from

the higher layer is assigned an UP ranging from 0 to 7 and is mapped to an AC according to the type of data traffic. Note that UP0 has higher priority than UP1 and UP2. The ACs are classified into *Background* (BK), *Best Effort* (BE), *Video* (VI), and *Voice over IP* (VO), where BK has the lowest priority and VO has the highest priority.

Technically, each AC uses a different set of parameters shown in Table 2.3 in order to obtain differentiated service, which results in contention differentiation.

Table 2.3: EDCF Parameters.

AC	CW_{min}	CW_{max}	AIFSN	TXOP (DSSS)
BK	CW_{min}	CW_{max}	7	0
BE	CW_{min}	CW_{max}	3	0
VI	$\frac{(CW_{min}+1)}{2} - 1$	CW_{min}	2	3.008 ms
VO	$\frac{(CW_{min}+1)}{4} - 1$	$\frac{(CW_{min}+1)}{2} - 1$	2	1.504 ms

Fig. 2.7 shows the EDCF mechanism, which provides differentiated access based on special parameters called *Arbitration Interframe Space* (AIFS) and *Transmission Opportunity* (TXOP) as shown in Table 2.3.

AIFS is the time period the medium has to be idle before a STA performs a random backoff and is similar to DIFS in DCF. However, AIFS values are determined

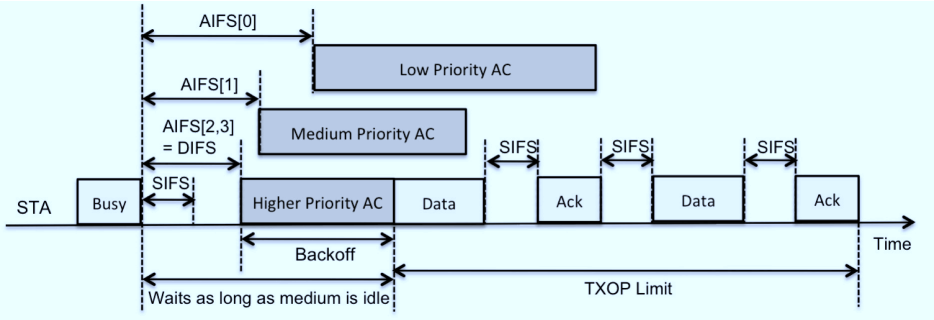

Figure 2.7: AIFS Relationships and TXOP Limit.

by ACs based on the following equation:

$$AIFS[AC] = SIFS + AIFSN[AC] \times SlotTime, \tag{2.2}$$

where $SIFS$ and $SlotTime$ are the same as those in DCF, and $AIFSN[AC]$ is the AIFS number, which is differentiated by the ACs. As can be seen by the default AIFSN values in Table 2.3, traffic with a high priority AC has a lower AIFSN than a low priority AC. Therefore, high priority ACs ensure that delay sensitive applications do not suffer from long delays. The duration of different AIFSs are shown in Fig. 2.7.

The CW_{min} and CW_{max} parameters in EDCF are applied based on Table 2.3. The higher the AC priority, the lower the values for both CW_{min} and CW_{max}. Therefore, ACs with lower CW values result in shorter backoff times and thus shorter medium access delays.

TXOP is the time period that a STA is allowed to transmit consecutive frames of the same AC separated by SIFS and ACK as shown in Fig. 2.7. This significantly

reduces delay for applications with high priority ACs (i.e., VO and VI). As shown in Table 2.3, the high priority ACs obtain the medium for longer durations of TXOP than do low priority ACs (i.e., BK and BE). The TXOP value of zero for BK and BE indicates that consecutive frame transmissions are not allowed for these low priority ACs. Each STA has four transmit queues that are maintained by EDCF as shown Fig. 2.8.

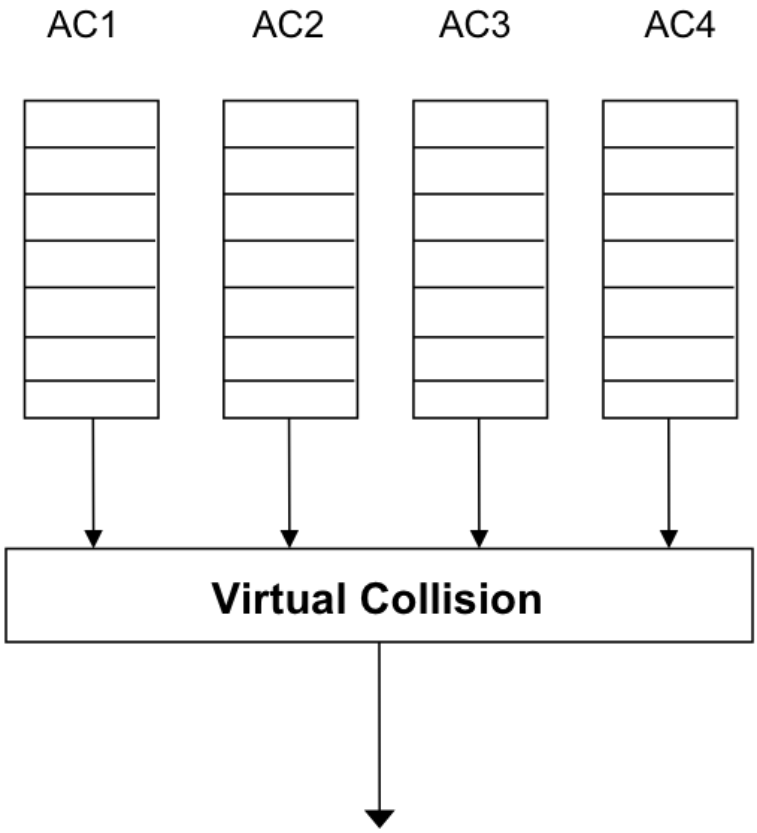

Figure 2.8: EDCF and Four Queues.

2.3 Association and Handoff in WLANs

In an infrastructure BSS, a STA must associate with an AP to obtain network service. Therefore, association is the crucial process for STAs to join a WLAN. The association process includes discovery, authentication, and association phase. The discovery phase, which involves identifying an AP for a network service in a STA's area, starts with probing for an available AP using either passive or active scanning.

In passive scanning, a STA switches its transceiver to a channel and waits for a beacon signal from an AP, or waits for a predefined maximum duration as defined by the channel time parameter, which is longer than the beacon interval. The beacon information gathered from APs on all the channels is used choose the best AP to associate with. However, a typical beacon interval is $100ms$ and the predefined maximum duration is longer than $100ms$. Thus, STAs must wait for a very long time in order to search all possible channels, which results in high packet losses during handoff.

For these reasons, active scanning is implemented as shown in Fig. 2.9. In active scanning, a STA broadcasts a probe request frame specifying a particular *Service Set IdentiFier* (SSID) and waits for either an indication of an incoming frame or waits for *Minimum Channel Time* (MinChannelTime) to expire. If the STA receives a probe response, it means one or more APs exist in the channel. Therefore, the STA waits until the *Maximum Channel Time* (MaxChannelTime), which has a typical value of 11ms. MinChannelTime is shorter than MaxChan-

nelTime to keep the overall handoff delay low, but it should be long enough for a STA to receive a possible response. This process is repeated for every channel. Then, the STA selects the best AP based on the information obtained from the probe response frames.

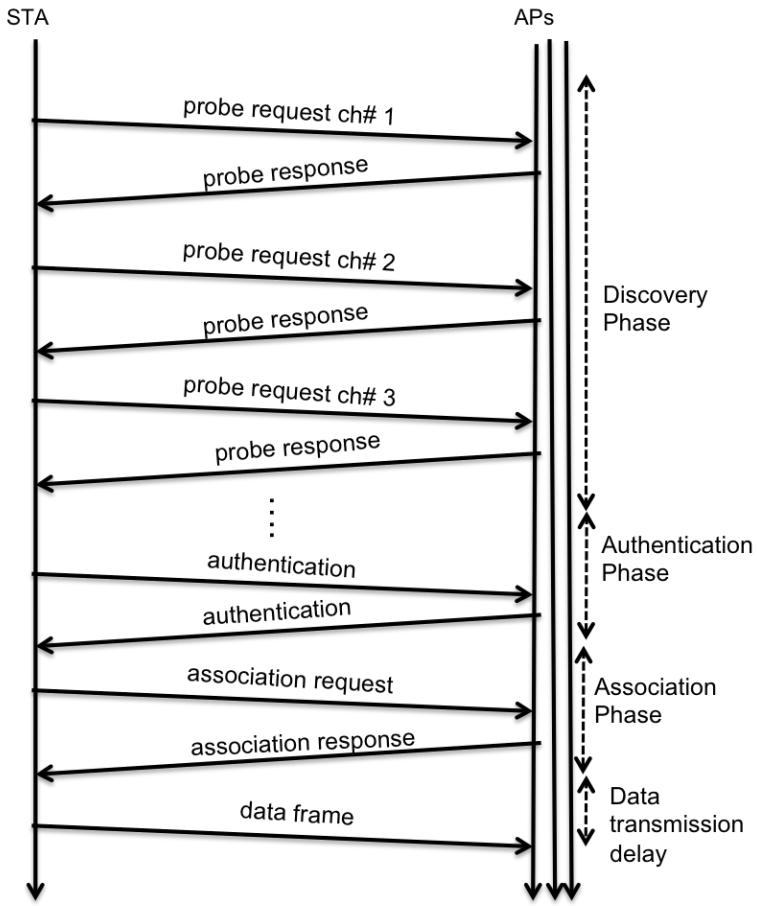

Figure 2.9: Active Scanning.

After scanning, the last two steps of the association process involve authentication and reassociation. Authentication is the process that a STA uses to announce its identity to the selected AP. In the IEEE 802.11 standard, authentication is performed using open system or shared key. Open system authentication is the default method for IEEE 802.11.

The STA initiates the association processes with a unicast management frame and an association request frame, then the AP replies with an association response frame to the STA. Finally, the STA can obtain network service.

The handoff process in WLANs is same as that of association. In the IEEE 802.11 standard, a STA senses the degradation of signal quality in the current channel as a STA moves from one BSS to another. The reduced signal strength of the current channel causes the STA to start the handoff process to another BSS.

2.4 QualNet

The work in this dissertation makes use of the QualNet tool for network modeling and simulation. QualNet [27] is a network simulator based on the C++ programing language for both wireless and wired communication networks. It is derived from *Global Mobile Information System Simulator* (GloMoSim) that was first released in 2000 by *Scalable Network Technologies* (SNT). However, QualNet is supervised by both SNT and GloMoSim, and it is developed and maintained by the UCLA Parallel Computer Lab. QualNet Developer is a discrete event simulator and has models and libraries for common network protocols that are provided

in source form. QualNet is organized by the *Open Systems Interconnection* (OSI) stack, which provides easy implementation and the building of new protocols in addition to network modeling at different layers. Therefore, QualNet provides a comprehensive set of tools with all the components for custom network modeling and simulation projects [27].

QualNet provides three different graphical modes such as design mode, visualization mode, and analyzer statistical graphic mode. The design mode is used to create and design simulations, the visualize mode is used to execute and animate experiments created in the design mode, and the analyzer statistical graphic mode displays statistical results generated from a QualNet simulation.

QualNet uses a layered architecture, which is similar to that of the TCP/IP network protocol stack. Within that architecture, where data moves between adjacent layers. Fig. 2.10 shows the QualNet protocol stack and the general functionality of each layer. For example, the link (MAC) layer provides link by link transmission as in the TCP/IP network protocol. At the source node, the link (MAC) layer receives data from the network layer and passes the data to the PHY layer for transmission over the wired or wireless channel. At the receiving node, the link (MAC) layer receives data from the PHY layer and forwards the data to the network layer.

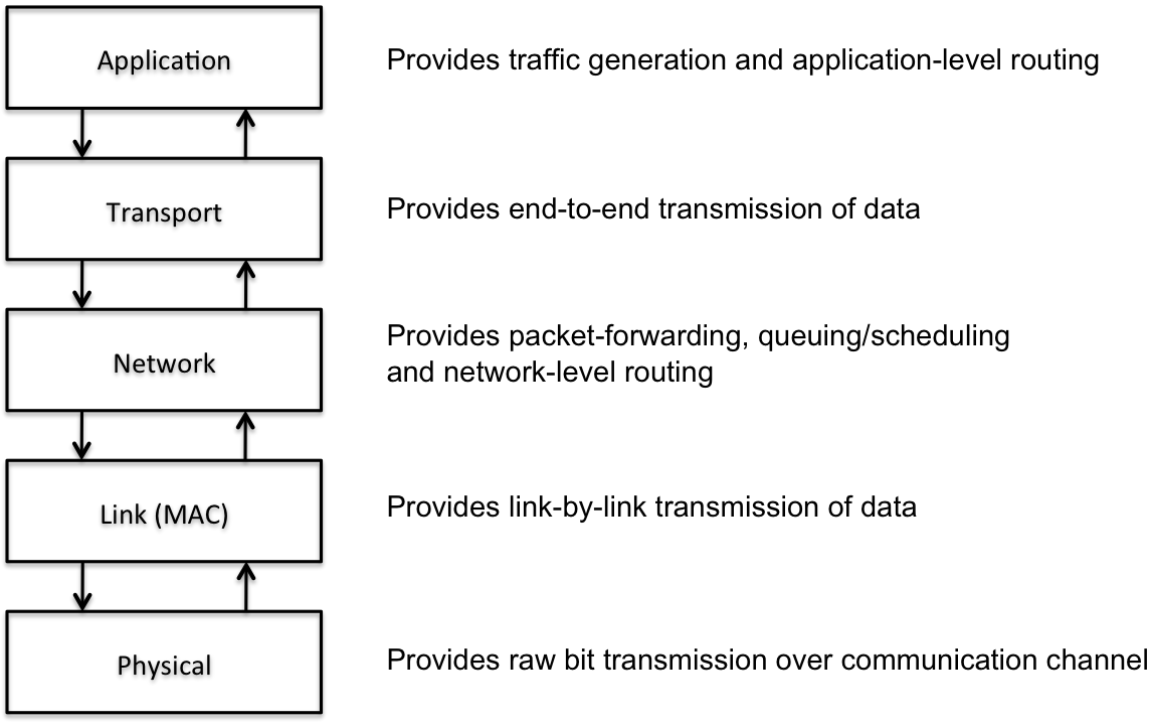

Figure 2.10: Qualnet Protocol Stack.

Chapter 3 – Related Work

3.1 Load Balancing Methods

A number of techniques have been proposed that considers network load rather than just RSSI as the association metric for WLANs.

The load balancing metrics proposed in [7–9] consider factors such as the number of users currently associated with an AP, the mean RSSI value of users currently associated with an AP, or the RSSI value and the bandwidth that can be obtained when a new user associates with an AP. However, these methods require protocol modifications on the AP side, and the number of users and RSSI alone cannot be used to predict the probability of collisions and the available bandwidth in the network.

In [10, 14], a STA observes the skewed time period of beacon frame receptions to estimate channel utilization or available bandwidth. However, the observation time increases because beacon frames are usually transmitted every 100 ms. The authors in [16] proposed bandwidth as the main metric, which is the percentage of time the AP is busy transmitting or receiving data during some time interval. However, each STA needs to be equipped with extra client software and wastes bandwidth by introducing additional test traffic.

Traffic Shaping [12] limits the maximum number of packets that can exist in

the queue of an AP. Selective dropping [11] prevents a new user from entering the system when the maximum number of users has been reached, so that offered throughput per each user remains above the specified threshold. These proposals are beneficial since overloading often results in queue overflow, which then increases the frame drop rate. However, Selective Dropping may worsen the starvation of some users, while Traffic Shaping restricts the throughput of individual connections in order to accommodate all users.

The authors in [13] proposed IQU, a practical queue-based user association management method for heavily loaded WLANs. IQU maintains a queue of users requesting network accesses, and only the STAs that can be simultaneously accommodated are permitted to access the network. Although each user is granted a fair opportunity to access the network while maintaining high overall throughput, admitted users are limited to assigned work periods and those not permitted need to wait in a queue for admission. These limitations can not accommodate users requiring minimized handoff latency and real time services, such as multimedia applications, for an extended period with the AP.

In [18,20], requests are accepted if the predicted load level after the association does not exceed some threshold, and a heavily loaded AP can disassociate a STA from its *Basic Service Set* (BSS) by sending an unsolicited disassociation frame. These approaches need modification to APs, and incur additional overhead since the AP needs to know not only the neighboring APs' load information but also the details of the entire network. 802.11k [19, 28] is expected to improve IEEE 802.11 by allowing a dynamic adaptation to the radio environment. For example,

a STA may request various information from other STAs or APs by exchanging extra measurement requests and report frames. However, most architectural and station service lists specified by the IEEE 802.11 standard, such as frame formats and procedures, need to be changed and a lot of extra measurement frames are needed not only in a BSS but also in an *Extended Service Set* (ESS). These incur wasted bandwidth by introducing a lot of extra measurement frames and decrease the overall throughput. Moreover, it needs extra time in order to converge towards an appropriate result in the case of burst pattern traffic since it may have different reports from different STAs.

The proposed PR-ALBM eliminates the shortcomings of these existing methods. First, the use of probe requests eliminates the need for extra information frames and long delays waiting for beacon frames from APs or searching for a better channel. Second, PR-ALBM is a STA-based method and thus it avoids issues suffered by AP-based methods, such as starvation of some users and throughput restriction of individual connections. Third, prior techniques ignored the *hidden node problem*, which is becoming more important with the increase in uplink data for applications such as VoIP, video gaming, and video conferencing. Finally, PR-ALBM considers QoS requirements of applications without any modification to the APs or without requiring any types of agents.

3.2 Handoff Delay

A number of techniques have been proposed to reduce handoff delay. These techniques focus on optimizing the probing process, since the probing delay represents more than 90 percent of overall handoff delay [29, 30].

In [31], handoff delay is reduced by using extra hardware in the form of additional radios. One radio is used for packet transmission, while the other radio is used for background scanning and pre-associating with alternate APs. Selective Active Scanning [32] uses an overlay sensor network to detect APs and the quality of their transmission channels. For example, a STA broadcasts an AP list request to the nearby sensor nodes, and starts the active scanning process based on this list.

In SyncScan [33], APs send staggered periodic beacons that allow a STA to scan for additional APs while still being associated with its current AP. In contrast, a STA actively probes for APs in [34]. Both techniques are achieved by either passively or actively scanning for available APs in the background [33, 34].

In [35], the authors limit the number of channels to be probed by defining the topological placement of APs and the mobility patterns of STAs using what is called the Neighbor Graph. A proposal in [36] predicts the next point-of-attachment based on signal strength, which minimizes the number of probes during handoffs. Mhatre and Papagiannaki [37] have shown that the use of long-term trends in signal strength measurements allow for better handoff decisions. However, they can frequently fail to provide correct Next-AP predictions because

STA movement was not always identical to the last path.

Unfortunately, these proposals only consider handoff delay and do not take into effect load balancing, which causes serious imbalance of user loads among APs and thus substantially reduces the network performance.

Chapter 4 – The Proposed Method: PR-ALBM

Although there have been numerous research efforts that separately consider load balancing with or without QoS, the hidden node problem, and handoff issues for WLANs, unfortunately no work exists that considers all these issues at the same time.

The proposed PR-ALBM eliminates the shortcomings of these existing methods. First, the use of probe requests eliminates the need for extra information frames and long delays waiting for beacon frames from APs or searching for a better channel in order to timely handoff. Second, PR-ALBM is a STA-based method and thus it avoids issues suffered by AP-based methods, such as starvation of some users and throughput restriction of individual connections. Third, prior techniques ignored the *hidden node problem*, which is becoming more important with the increase in uplink data for applications, such as VoIP, video gaming, and video conferencing. Finally, PR-ALBM considers QoS requirements of applications without any modification to the APs or without requiring any types of agents.

Note that the proposed PR-ALBM utilizes probe requests to observe the state of the channel and determines the best AP for association to balance the network load. Therefore, it can be applied to any application; however, VoIP is considered as a case study due to its delay sensitivity and has one of the shortest handoff

requirements. In addition, PR-ALBM is presented based on 802.11b, which has the slowest transmission rates and longest $SlotTime$, $DIFS$, and $SIFS$ among all the 2.4 GHz IEEE 802.11 networks (i.e., 802.11 b/g/n).

4.1 Timely Handoff Using Optimized Backoff Time

PR-ALBM uses a fixed, *optimized* backoff time, instead of random backoff time for probe request frames. The choice of backoff time, which is represented in terms of number of slots, is optimized for each application to provide sufficient amount of time to observe the channel and yet result in timely handoff.

Fig. 4.1 illustrates the timing for the discovery, authentication, association, and a successful data frame transmission. In order to obtain the most appropriate backoff time and thus provide timely handoff, the total delay D_{total} between when the first probe request is transmitted and when the association response is received needs to be known, which is given as

$$D_{total} = D_{probe} + D_{auth} + D_{assoc}, \tag{4.1}$$

where D_{probe} represents the probing delay, D_{auth} is the authentication delay, and D_{assoc} is the association delay.

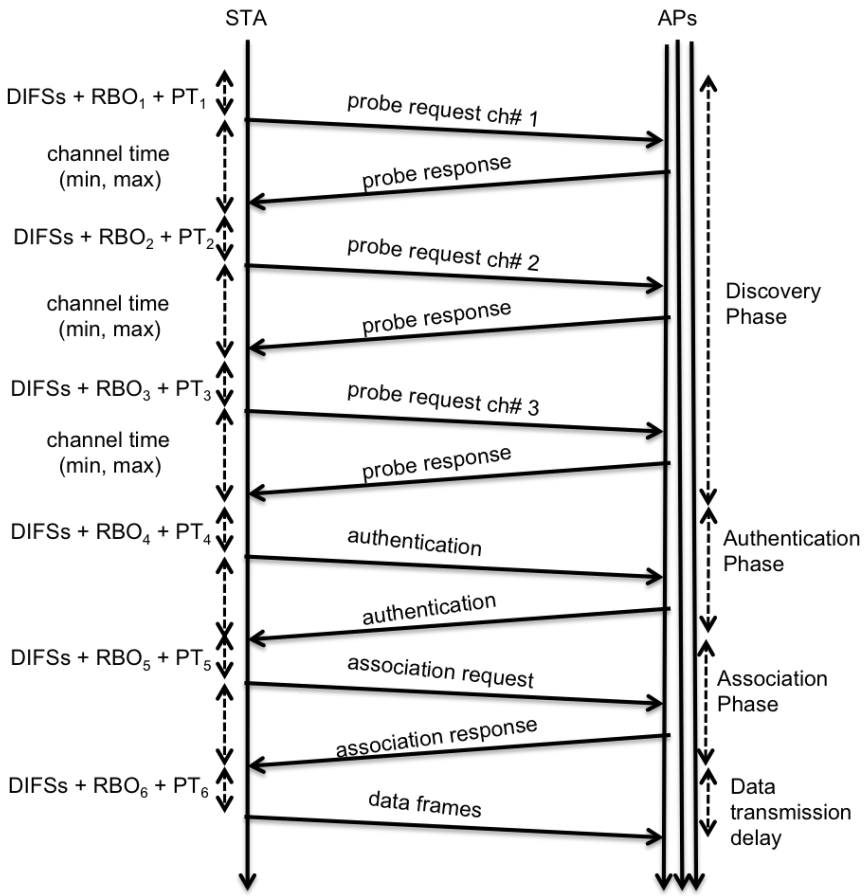

Figure 4.1: Total Delay between Probe Request Frame and Data Frame.

D_{probe} is defined by the following equation:

$$D_{probe} = n \cdot (DIFS \cdot d + RBO + PT) + n \cdot (t_{tx} + t_{prop} + t_{ch} + t_{switch}), \quad (4.2)$$

where n is the number of channels, d is the average number of $DIFS$s, RBO is the average random backoff time, PT is the average *pause time* during the discovery phase, t_{tx} and t_{prop} are transmission and propagation time of the probe frame, respectively, t_{ch} is the channel time, and t_{switch} is the channel switch time.

PT represents the additional pause time required when the medium is busy by other STAs or AP, and thus RBO is not decremented. Fig. 4.2 illustrates an example scenario consisting of four contending STAs, where STA1 observes the traffic of the other three STAs. The RBO values in terms of number of slots for STAs 1-4 are 9, 2, 5, and 7, respectively. The groups of slots indicated by (A), (C), (E), and (G) represent parts or sections of RBO or CW slots, while the other groups of slots indicated by (B), (D), and (F) are parts of PT. Note that the slot time, i.e., $SlotTime$, is a constant value found in the STA's Management Information Base (MIB). The $SlotTime$ for 802.11b is 20 μs, which results in a PT value of 200 μs for STA1.

The value t_{ch} can be either *minimum channel time* (t_{min}) or *maximum channel time* (t_{max}). t_{min} is the minimum amount of time a STA has to wait on an empty channel. On the other hand, t_{max} is the maximum amount of time a STA has to wait to collect all the probe responses, which is used when a response is received

within t_{min}. Note that t_{ch} is always less than or equal to t_{max}. Finally, t_{switch} is the average time required to switch from one channel to another.

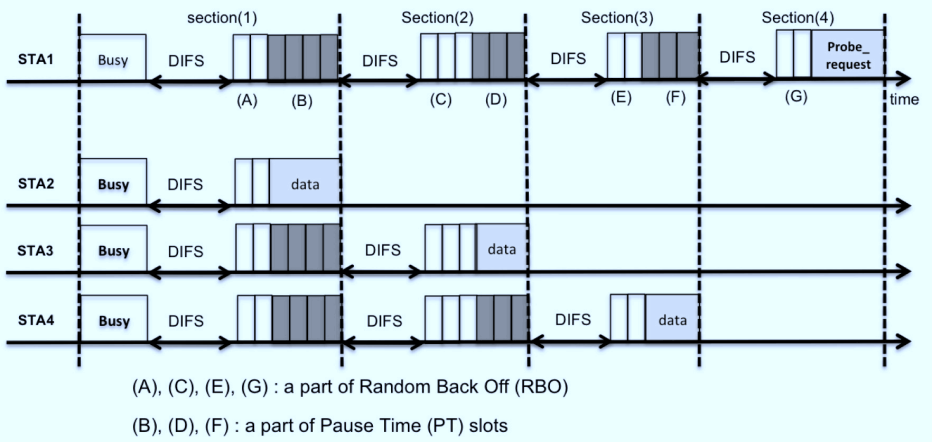

Figure 4.2: Differentiation between RBO and PT.

In order to provide timely handoff, D_{total} should not be greater than some minimum tolerance level defined by a multimedia application. For example, the handoff delay for VoIP is recommended to be no greater than 80 ms [38–42]. In our previous work [43, 44], D_{auth} and D_{assoc} were measured to be 6 ms and 4 ms, respectively, which are similar to the experimental results in [29] and ns-2 simulation results in [45]. Therefore, D_{probe} for VoIP should be less than 70 ms. Since APs in the adjacent cells use only non-overlapped channels 1, 6, and 11 to reduce interference among the cells [1, 44, 46], i.e., n=3, the time required to probe each channel should be lower than 23.3ms. Suppose that t_{ch} is 11 ms (i.e., $t_{ch} = t_{max}$), t_{tx} and t_{prop} are both 1 μs, and t_{switch} is 5 ms [45]. Then, the following inequality can be obtained from Eq. 4.2:

$$DIFS \cdot d + RBO + PT \leq 7.3ms. \tag{4.3}$$

SIFS is equal to 10 μs in 802.11b, which leads to DIFS of $50\mu s = SIFS + 2 \cdot SlotTime$. In 802.11e, the management frames have the highest priority and are thus classified into AC of 3 [23], which leads to AIFSN[3]=2. Therefore, AIFS for probe requests is 50 μs from Eq. 2.2, and is the same as DIFS. Therefore, determining optimum backoff time for non-QoS and QoS WLANs is identical.

As can be seen from Fig. 4.2, PT is proportional to d, and thus, can be written as

$$PT = d \cdot m \cdot SlotTime, \tag{4.4}$$

where m is the average number of pause slots in a section of PT, and thus term $m \cdot SlotTime$ represents the average pause time of a section. For example, STA1 observes 4 DIFSs and 10 pause slots in Fig. 4.2. Therefore, the average pause time in a section is 50 μs. Therefore, solving for RBO in Eq. 4.3 leads to the following equation:

$$RBO \leq 50\mu s \cdot \left(146 - \left(1 + \frac{2}{5}m\right) \cdot d\right). \tag{4.5}$$

RBO can also be represented by the equation:

$$RBO = optBO \cdot SlotTime, \tag{4.6}$$

where *optBO* is the *optimized backoff time* that provides a sufficient amount of time to observe the channel and leads to a timely handoff. Based on Eqs. 4.6 and

4.5, $optBO$ can be written as

$$optBO \leq \frac{5}{2} \cdot \left(146 - \left(1 + \frac{2}{5}m\right) \cdot d\right). \tag{4.7}$$

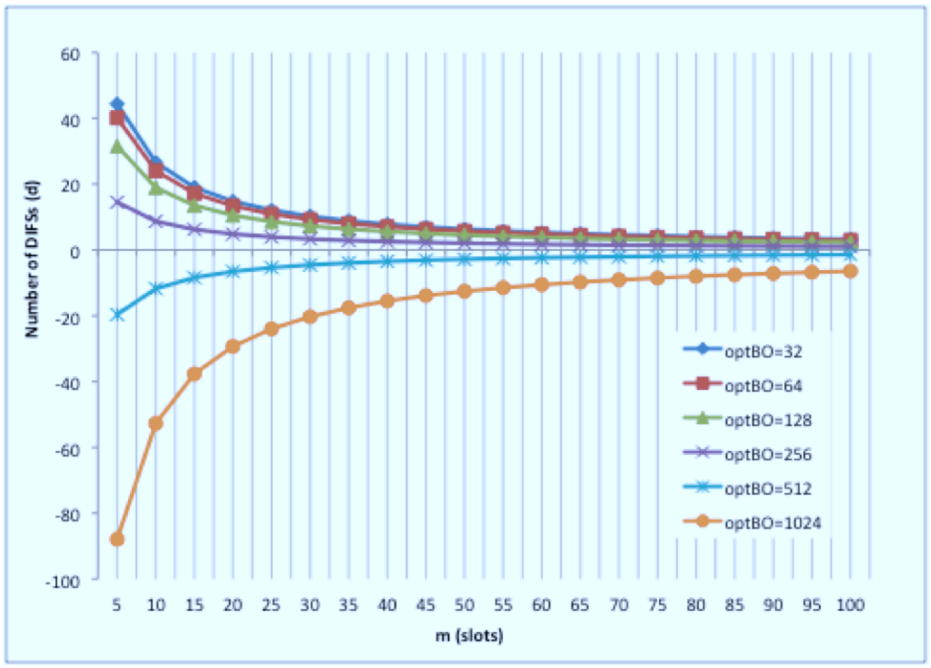

Figure 4.3: $optBO$ as function number of d.

Fig. 4.3 shows $optBO$ as a function of d and m based on Eq. 4.7. The figure shows that d is positive as a function of m for different $optBO$ values, except for the cases of 512 slots and 1024 slots, where d becomes negative. These two cases are impossible and indicate that $optBO$ values of 512 and 1024 cannot satisfy the delay requirement. This figure also shows that the best candidate for $optBO$ is 256 slots since it provides the longest duration for observing the state of the channel.

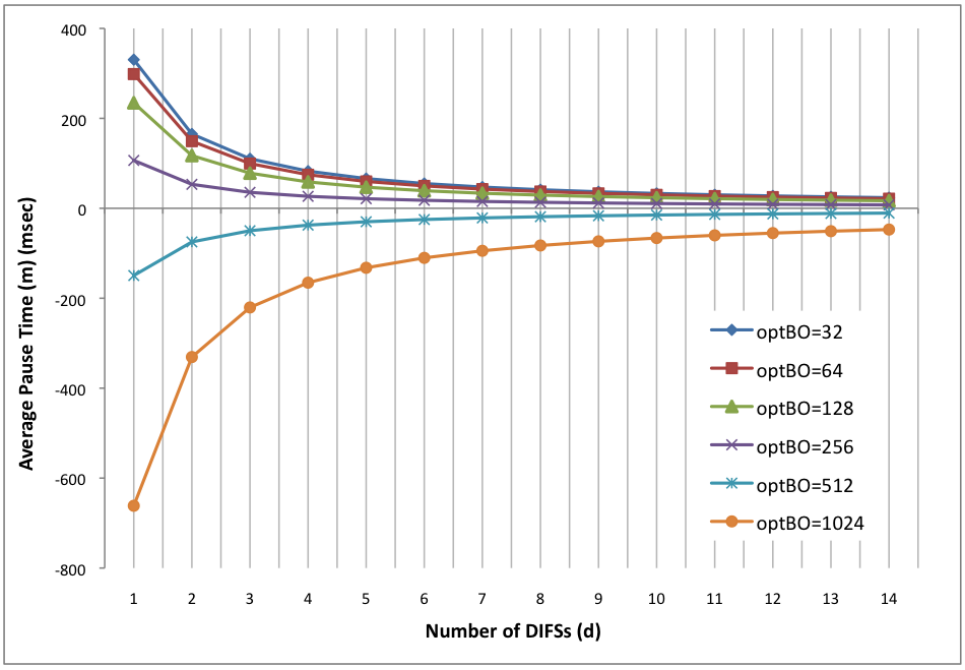

Figure 4.4: *optBO* as function number of m.

Fig. 4.4 also shows *optBO* as a function of m and d based on Eq. 4.7. The figure shows that m is positive as a function of d for different *optBO* values, except for the cases of 512 slots and 1024 slots, where m become negative. This result is very similar to that of Fig. 4.3. Thus, these two cases are impossible and indicate that *optBO* values of 512 and 1024 cannot satisfy the delay requirement. This figure also shows that the best candidate for *optBO* is 256 slots since it provides the longest duration for observing the state of the channel, which is same result of Fig. 4.3.

Both figures, Fig. 4.3 and Fig. 4.4 show that the best candidate for *optBO* is 256 slots since it provides the longest duration for observing the state of the channel.

However, m and d, and thus $optBO$, may not exactly follow the analytical results shown in Fig. 4.3 due to variations in a real environment. Therefore, the $optBO$ value of 128 slots will provide some cushion for these errors. Our simulation results also show that the best value of $optBO$ for VoIP is 128 slots (see Sec. 5.3).

4.2 PR-ALBM

Based on the $optBO$ discussed in Sec. 4.1, PR-ALBM considers QoS requirements of applications, which vary from one application to another. Therefore, it adopts end-to-end delay and throughput as main factors for VI and VO, and considers collision rate as the main factor for BK and BE. PR-ALBM observes the number of $DIFS$s and D_{probe} during the discovery phase to estimate the contention rate of all the adjacent channels in non-QoS WLANs. In addition, the types of traffic loads of all the adjacent channels are observed for QoS WLANs. The details of PR-ALBM are presented in the flow chart shown in Fig. 4.5.

In order to achieve balanced load in a non-QoS WLAN, PR-ALBM is defined by the probability that a STA x successfully transmits a data frame on channel i, P_x^i, which is represented as

$$P_x^i = (1 - P_{DC}^i) \cdot (1 - P_{HC}^i), \tag{4.8}$$

where P_{DC}^i and P_{HC}^i represent the probabilities that a data frame experiences a direct collision and a hidden node collision in a channel, respectively. A *direct*

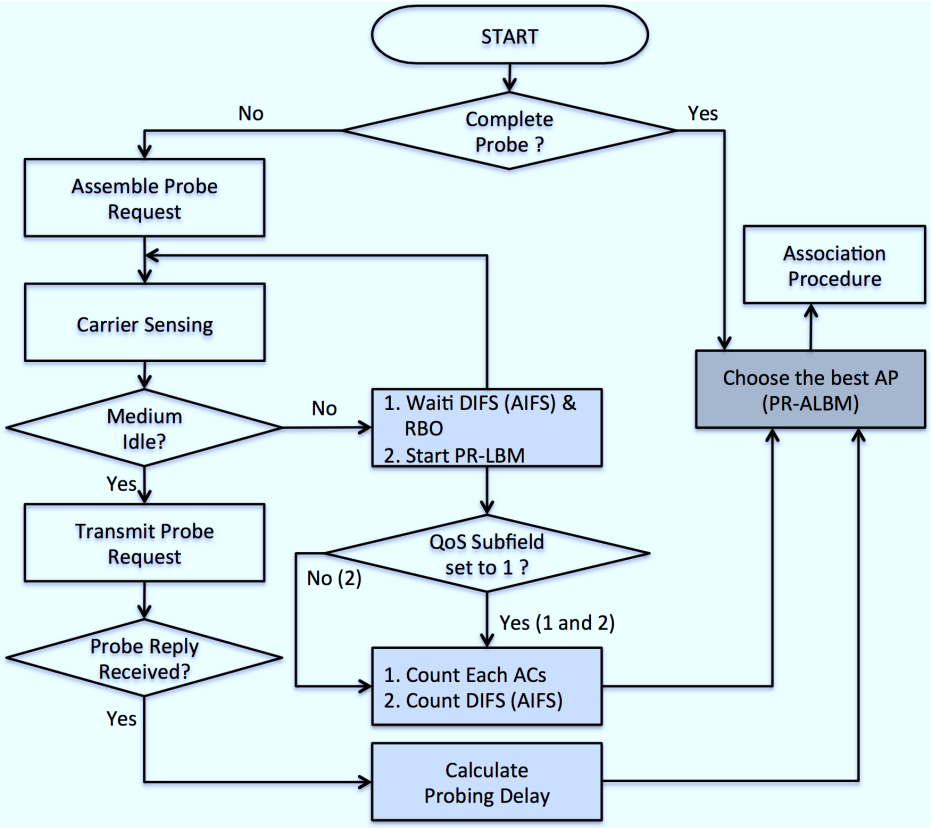

Figure 4.5: Flow Chart of PR-ALBM.

collision occurs when two STAs that can sense each other start transmitting packets at the same time, while a *hidden node* collision occurs when multiple far away nodes that cannot sense each other transmit at the same time.

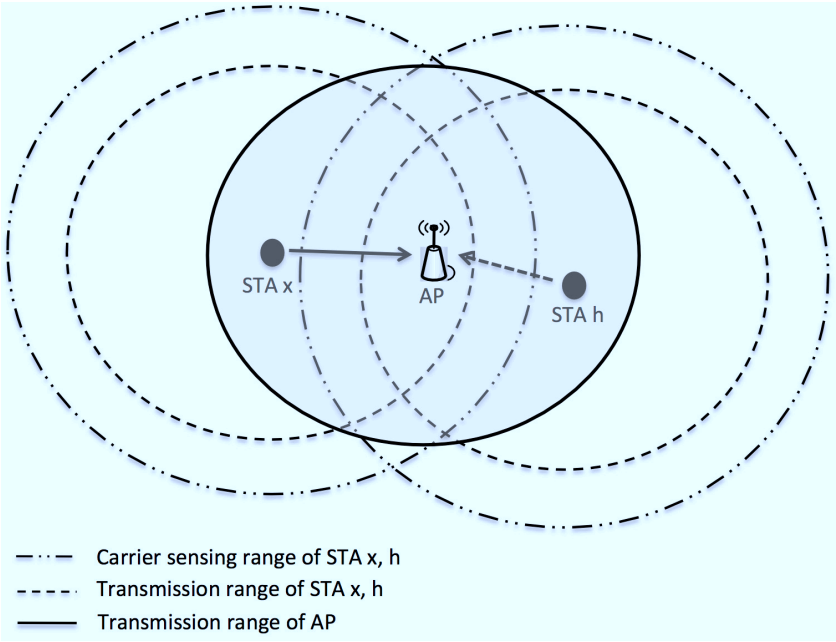

Figure 4.6: Hidden Node Collision in the Three-entity Topology.

Fig. 4.6 depicts the conditions under which hidden node collisions occur. The dashed and dotted circles around STAs x and h represent their carrier sense and transmission ranges, respectively, while the circle around the AP represents its transmission range. This figure shows that STAs x and h cannot hear each other, but AP hears both. Therefore, the transmission from STA x will collide with the transmission from STA h.

To estimate P^i_{DC}, a time-slot based observation method is used as was shown in Fig. 4.2. As can be seen in Fig. 4.2, the number of DIFSs for STA 1-4 is 4, 1, 2, and 3, respectively, and the number of contending frames observed by STA 1-4 is 3, 0, 1, and 2, respectively. Therefore, the number of DIFSs increases in proportion to the number of contending frames and thus, the contention rate can be estimated by observing the number of DIFSs. Probe delay represents the total time delay between the time attempt to transmit a probe request frame and receive a probe reply frame, and includes all types of delays during one channel probing phase, such as contention delay to seize the channel, transmission delay, propagation delay, and queuing delay at AP. Therefore, as contention increases, the probe delay increases.

Moreover, the ratio of d and the probe delay increase, as the level of contention increases in a channel. Therefore, both factors are used to estimate the contention rate, P^i_{DC} as follows:

$$P^i_{DC} = \alpha \cdot d^i \cdot D^i_{probe}, \tag{4.9}$$

where d^i and D^i_{probe} are the total number of DIFSs (or AIFSs) and the probing delay in channel i. On the other hand, α represents a weight that normalizes P^i_{DC} to be a fraction. Thus, $0 < \alpha \leq \frac{1}{d^i \cdot D^i_{probe}}$.

Unlike P^i_{DC}, P^i_{HC} cannot be evaluated by observing the channel since hidden nodes cannot be heard. Instead P^i_{HC} is modeled using an M/M/1/K queue at each STA, which follows the Poisson process with packet arrival rate of λ, and service

rate of μ. Thus, the probability distribution of the service time (T) is exponential with a mean of $1/\mu$. Since the queue can hold at most K packets, any additional packet will be refused entry into the system. Therefore, λ and μ based on the Birth-Death process shown in Fig. 4.7 are given as follows:

$$\lambda_k = \begin{cases} \lambda & k \leq K \\ 0 & k > K \end{cases}$$

$$\mu_k = \mu \quad k = 1, 2, \cdots, K \tag{4.10}$$

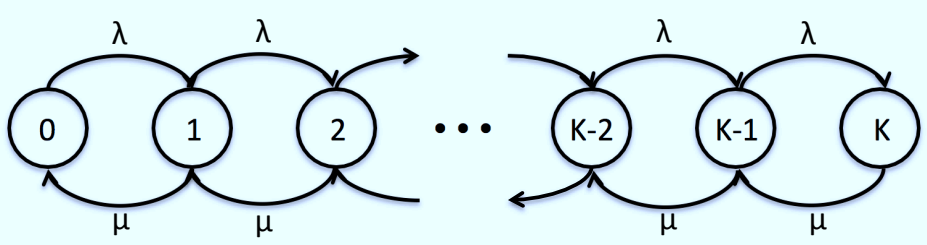

Figure 4.7: State-transition Diagram for M/M/1/K.

P^i_{HC} is divided into two conditional probabilities representing when the queue length is greater than zero and when it is zero. Let q_h denote the queue length at a hidden STA h. Then, the probability that a *hidden node collision* (HC) will occur with STA h is given as

$$P^i_{HC} = \sum_{k=0}^{K} P\{HC, q_h = k\} \tag{4.11}$$

$$P^i_{HC} = P(HC \mid q_h > 0) \cdot P(q_h > 0) + P(HC \mid q_h = 0) \cdot P(q_h = 0). \tag{4.12}$$

If $q_h > 0$, then STA h has packets in its queue and they will be transmitted. Therefore, the conditional probability that packets from STAs x and h will collide, i.e., $P(HC \mid q_h > 0)$, is one. In addition, even if the queue in STA h is empty at $t = 0$, a collision will occur only if a packet arrives at queue of STA h before $t = T$. Since STA h follows the Poisson process with the arrival rate of either λ_h or 0 and service rate of μ, $P(HC \mid q_h(0) = 0)$ follows an exponential distribution given by

$$P(HC \mid q_h = 0) = 1 - e^{-\rho_h}, \rho_h = \lambda/\mu. \tag{4.13}$$

Therefore, Eq. 4.12 can be rewritten as

$$P^i_{HC} = 1 \cdot P(q_h > 0) + (1 - e^{-\rho_h}) \cdot P(q_h = 0) \tag{4.14}$$

The probability that the queue length will be zero, $P(q_h = 0) = P_0$, can be obtained from the Birth-Death process and is given as (see Sec. A)

$$P_0 = \frac{1}{1 + \sum_{k=1}^{K} \prod_{j=0}^{k-1} \frac{\lambda_j}{\mu_{j+1}}}. \tag{4.15}$$

The probability that the queue length will be k, P_k, is obtained based on the

finite Markov chain diagram shown in Fig. 4.7, which leads to

$$P_k = \begin{cases} P_0 \left(\frac{\lambda}{\mu}\right)^k & k \leq K \\ 0 & k > K \end{cases} \tag{4.16}$$

Thus, $P(q_h = 0)$ is given by

$$P(q_h = 0) = P_0 = \frac{1 - \frac{\lambda}{\mu}}{1 - \left(\frac{\lambda}{\mu}\right)^{K+1}} = \frac{1 - \rho}{1 - \rho^{K+1}}. \tag{4.17}$$

Therefore, the following equation can be obtained for P_{HC}:

$$P_{HC}^i = \left(1 - \frac{1 - \rho}{1 - \rho^{K+1}}\right) + \left(1 - e^{-\rho_h}\right) \cdot \left(\frac{1 - \rho}{1 - \rho^{K+1}}\right). \tag{4.18}$$

For QoS WLANs, PR-ALBM considers CW_{max} and $AIFS$ as shown in Table 2.3 since their lengths differentiate an application's priority. Moreover, high ACs such as VO and VI have much more opportunity to seize the channel than BE and BK, which results in minimum delay and high throughput. As can be seen in Fig. 4.8, there is significant difference in maximum waiting time composed of AIFS and CW_{max} between high and low ACs. For example, VO and VI have 18 and 34 slots respectively, for maximum waiting time. On the other hand, BE and BK have 1027 and 1031 slots, respectively. The difference between VO and VI is significantly smaller than the difference between VI and BE or BK. The same phenomenon can be observed in Fig. 4.9. Therefore, PR-ALBM takes into account the

number of higher value ACs obtained during the discovery phase and then avoids channels with large volume of such traffic. This is done by counting the number of VO and VI types in each channel. Then, the AP with the smallest number of VOs and VIs is chosen as the best AP. If there is a tie, then PR-ALBM considers the number of DIFS and the D_{probe}.

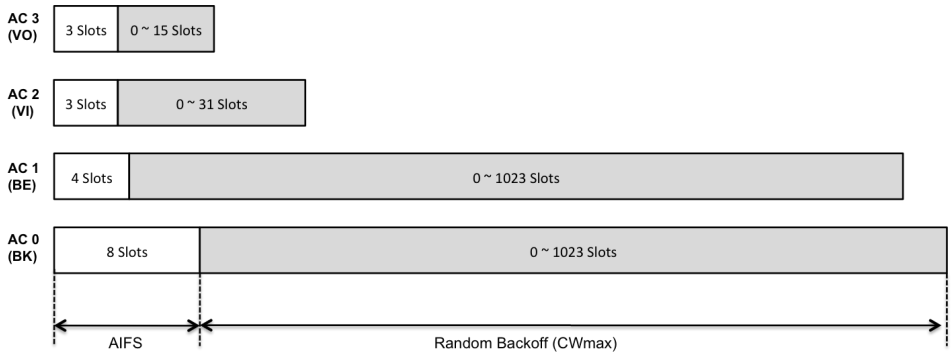

Figure 4.8: Maximum Waiting Time for Accessing a Channel.

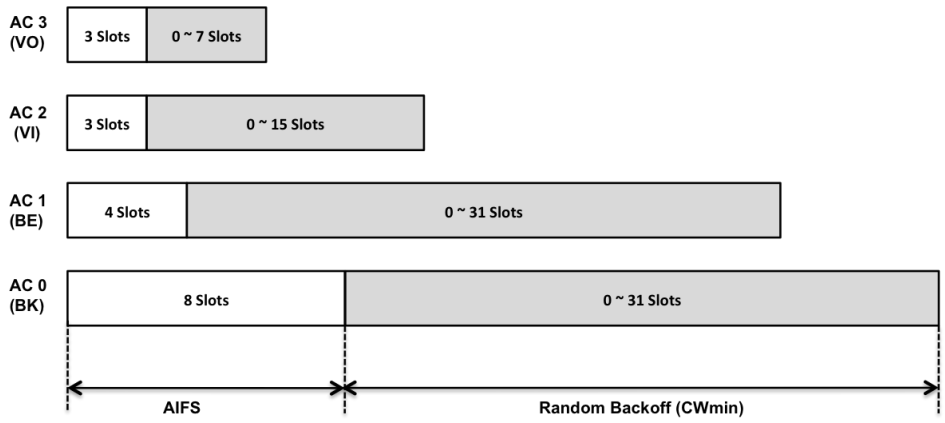

Figure 4.9: Minimum Waiting Time for Accessing a Channel.

The types of traffic loads are obtained from the QoS subfield and *Traffic Identifier* (TID) in the MAC header shown in Fig. 4.10 [21]. The TID field supports 8 different user priorities and the QoS subfield indicates whether a STA wants to use QoS. For example, if the QoS subfield is set to 1 and the TID field is set to 6, the traffic type is voice classified into AC [3] as indicated in Table 2.2.

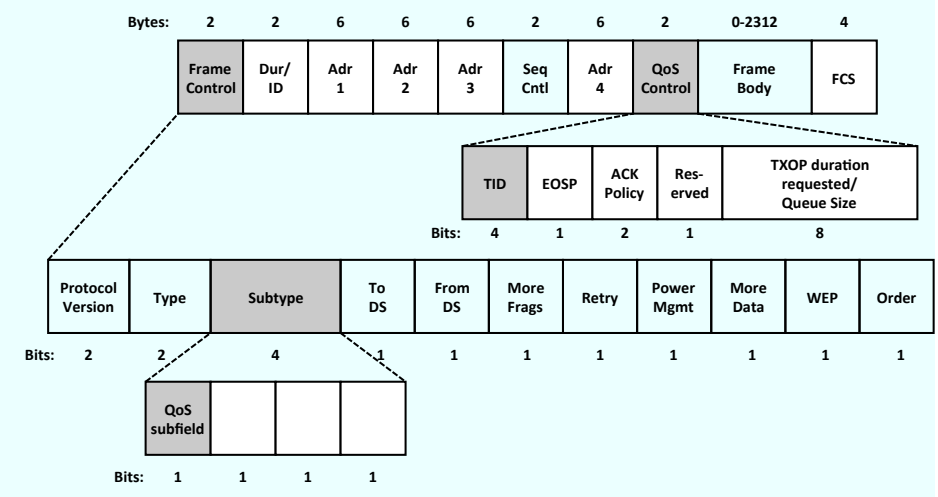

Figure 4.10: QoS Subfield and TID Field in the MAC Header.

Chapter 5 – Simulation and Analysis

5.1 Simulation Environment

This section evaluates the accuracy of our model with QualNet [27] based on the simulator parameters shown in Table 5.1, Table 5.2 and two scenarios discussed below. Our simulation is based on an infrastructure WLAN with three overlapped cells without Request-to-Send (RTS) and Clear-to-Send (CTS) handshake. Although RTS/CTS is designed to mitigate the hidden node problem, a significant amount of overhead is incurred and thus is typically not used in infrastructure mode [18]. There are two different types of scenarios for simulation and analysis. One evaluates the accuracy of PR-ALBM in non-QoS WLANs and another evaluates that in QoS WLANs. For both scenarios, the simulation results are the average of 100 simulation runs and the positions of STAs are different for each run.

5.1.1 Scenario 1

Fig. 5.1 shows the simulation topology for Scenario 1, which is implemented in order to verify $optBO$ for timely handoff and evaluate the accuracy of PR-ALBM for non-QoS WLANs. The accuracy of our metric is measured by varying the number of contending STAs and the number of backoff slots. For each simulation run, a STA (i.e., STA x) using VoIP, based on G.711 codec with packetization

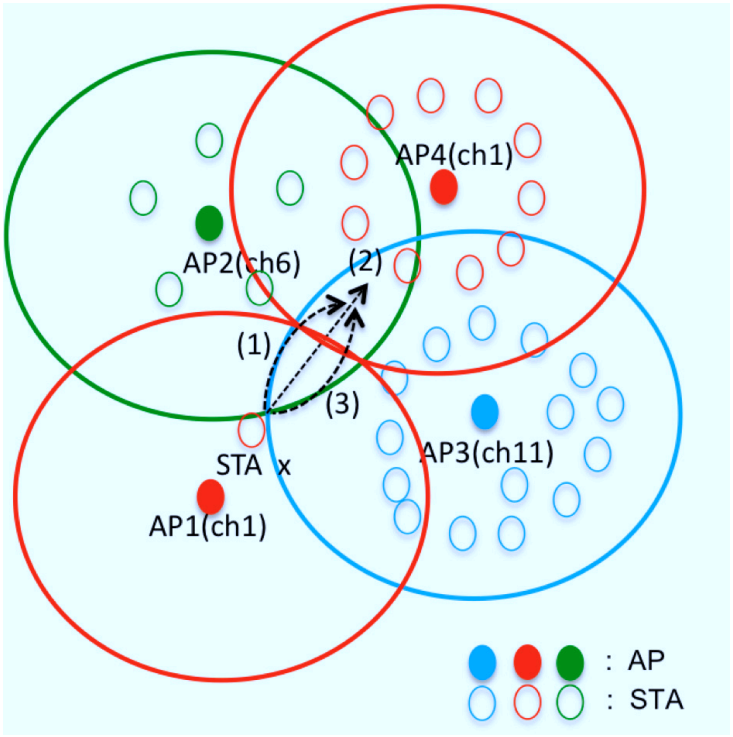

Figure 5.1: Simulation Topology (*Scenario 1*).

interval of 20 ms and packet size of 120 bytes, enters the area overlapped by AP2, AP3, and AP4 with random speeds of 1∼4 $miles/h$ and random routes of 1, 2, and 3 representing its relative proximity and thus RSSI to AP2, AP4, and AP3, respectively. In addition, the number of STAs in each cell randomly varies from 3 to 20, and these STAs are positioned within a cell in such a way that they can all sense each other, except one STA that is positioned to act as the hidden node for STA x. Each STA generates UDP packets at a constant interval of 200 ms with packet size of 256 bytes.

Table 5.1: Simulation Parameters (Scenario 1).

Parameters	Scenario 1	Parameters	Scenario 1
Simulation Time	1000 seconds	DIFS	$50\mu s$
Radio Types	802.11b	CWmin	31
Number of APs	4	CWmax	1023
Channels	1, 6, & 11	D_{auth}	$6ms$
Number of STAs	1∼50	D_{assoc}	$4ms$
Applications Types	CBR & VoIP	t_{switch}	$5ms$
Data Rates	64Kbps∼2Mbps	t_{ch}	$11ms$
Slot time	$20\mu s$	Propagation delay	$1\mu s$
SIFS	$10\mu s$		

5.1.2 Scenario 2

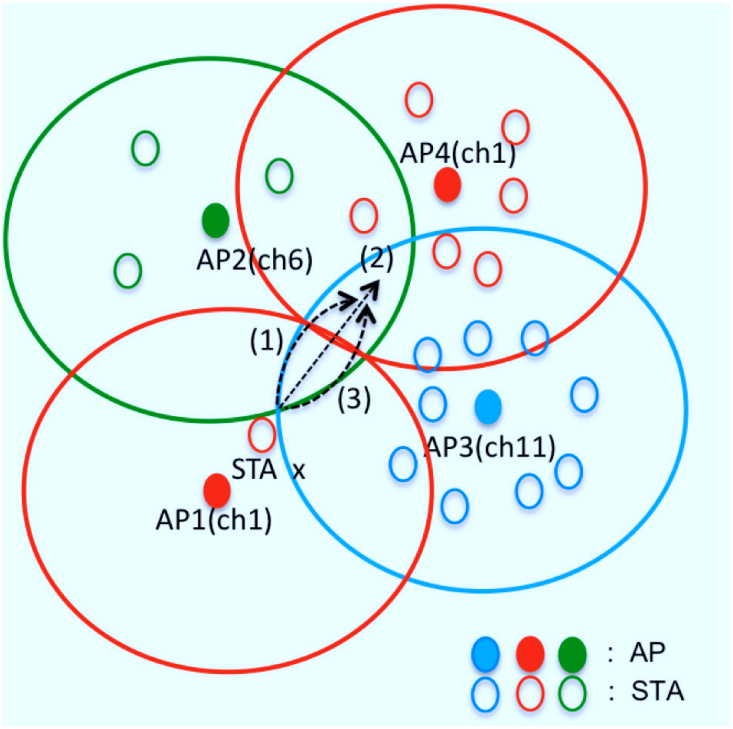

Figure 5.2: Simulation Topology (*Scenario 2*).

This scenario shown in Fig. 5.2 is implemented to evaluate the accuracy of PR-ALBM for QoS WLANs. STA x using VoIP enters the area overlapped by AP2, AP3, and AP4 with random speeds between 1∼4 $miles/h$ and random routes of 1, 2, and 3, again representing its relative proximity to APs, while it monitors the number of high priority AC traffics for all nearby channels. STAs are distributed across the three cells in such a way to represent different network loads not only in terms of the number of STAs but also traffic type. AP2 has 3 STAs composed of

1 BK STA, 1 VI STA, and 1 VO STA. AP3 has 6 STAs consisting of 2 BK STAs, 2 VI STAs, and 2 VO STAs. AP4 has 9 STAs with 3 BK STAs, 3 VI STAs, and 3 VO STAs. As in Scenario 1, each voice flow uses G.711 codec with a packetization interval of 20 ms and packet size of 120 bytes. BK traffic is generated at a constant interval of 200 ms with a constant packet size of 256 bytes. VI traffic is generated the same way as BK traffic because it leads to more reasonable results than using different traffic sizes and intervals [47–49]. In addition, the simulation results are obtained the same way as in Scenario 1.

Table 5.2: Simulation Parameters (Scenario 2).

Parameters	Scenario 2	Parameters	Scenario 2
Simulation Time	1000 seconds	DIFS	$50\mu s$
Radio Types	802.11b	CWmin	BK(31),VI(15),VO(7)
Number of APs	4	CWmax	BK(1023),VI (31),VO(15)
Channels	1, 6, & 11	D_{auth}	$6ms$
Number of STAs	1∼25	D_{assoc}	$4ms$
Type of App.	BK, VI, & VO	t_{switch}	$5ms$
Data Rates	64Kbps∼2Mbps	t_{ch}	$11ms$
Slot time	$20\mu s$	Prop. delay	$1\mu s$
SIFS	$10\mu s$	AIFSN	BK(7),VI&VO(2)

5.2 Evaluation Parameter

Some applications, such as multimedia applications require a certain level of guaranteed QoS in terms of throughput, delay, and packet loss. These requirements significantly vary from application to application in WLANs. For example, VoIP has tight constraints on end-to-end delay while FTP is prone to data loss. Therefore, the parameter used to analyze a certain application is very important to conduct a fair evaluation.

For the proposed metric, our evaluation is performed based on the following parameters : utilization, end-to-end delay, and data loss. This is because the value of utilization indicates performance rate, which is represented as successful frame delivery. In addition, end-to-end delay represents the total delay between source STA and destination STA and longer delays strongly restricts the performance of application. Data loss indicates the busy state of a channel.

5.2.1 Utilization

Utilization is one of the most important parameters for measuring the performance of wireless LANs and is defined as the achieved throughput related to the net bitrate, which is the transmission rate of the physical layer, in bit/s of a communication channel. For example, if the throughput is $1.5Mbit/s$ with a $2Mbit/s$ transmission rate in an IEEE 802.11 WLAN, the channel utilization is 0.75. Therefore, a large utilization value indicates high performance. Utilization is given as

$$utilization = \frac{achieved\ throughput\ (bits/s)}{transmission\ rate\ (bits/s)}. \tag{5.1}$$

where the achieved throughput is the average rate of successful frame delivery over a communication channel. The achieved throughput is usually measured in bits per second and is given as

$$throughput = \frac{total\ number\ of\ delivered\ packet \cdot packet\ size\ (bits)}{total\ time\ duration\ of\ delivery\ (sec)}. \tag{5.2}$$

Utilization is one of the most important parameters for measuring the performance of a wireless LAN. This is because utilization also indicates consumed bandwidth, corresponding to achieved throughput. Most multimedia applications, such as streaming media, VoIP, and video conferencing can be classified as utilization sensitive applications. These applications require constant bandwidth and may seriously suffer when they connect with higher utilization channels. Therefore, STAs using multimedia applications need to avoid associating with channels of higher utilization and select channels which have the smallest value of utilization.

5.2.2 End-to-end Delay

End-to-end delay is another important measurement, which represents the total delay between the time of generating a frame at the sender and the time of receiving the frame at the receiver. Therefore, end-to-end delay includes all types of delays

during the whole transmission time period : transmission delay, propagation delay, processing delay, and queueing delay and is given as

$$D_{endtoend} = D_{trans} + D_{prop} + D_{queuing} + D_{proc}, \tag{5.3}$$

where D_{trans} represents the transmission delay, D_{prop} is the propagation delay, $D_{queuing}$ is the queuing delay in routers, and D_{proc} is the processing delay.

Transmission delay is the amount of time required to transmit data into the wireless channel, which is caused by transmission rate or link bandwidth of the wireless link. It is given as the following formula:

$$D_{trans} = \frac{frame\ size\ (bits)}{transmission\ rate\ (bits/s)}. \tag{5.4}$$

For example, the transmission delay is $4ms$ when a frame with 512 bytes transmits over the wireless link of 2 Mbps.

The propagation delay is the amount of time for a frame required to travel in the wireless link which is located between sender and receiver and is given as

$$D_{prop} = \frac{d}{prop_{speed}}. \tag{5.5}$$

where d is the distance of the link, which is the distance between sender and receiver, whereas $prop_{speed}$ is the propagation speed over the wireless medium. In wireless communication, it is the speed of light.

Queueing delay is the amount of time for a frame waits in the queue before it is transmitted. The delay depends on the queue size and the number of arriving frames already waiting in the queue. As the line of frames waiting in the queue increases, the queuing delay is longer.

Processing delay is the amount of time required to examine the packet's header, check for bit level errors, and determine the direction of the packet. However, the delay value is typically on the order of ms or less, thus usually ignored in simulation study [42].

Longer end-to-end delay strongly restricts the performance of interactive real-time applications such as VoIP, video conferencing, and multiplayer network games. This is because longer delays severely decrease the performance of these applications. For example, end-to-end delays of smaller than $150ms$ are not perceived by a human; $150ms$ to $400ms$ may be acceptable but is not ideal; more than $400ms$ is often unacceptable in VoIP [42].

5.2.3 Data Loss

Data loss usually increases as traffic increases, thus it indicates the busy status of a channel. Although a lost packet may be retransmitted from a sender to a receiver, a lot of retransmissions seriously decrease the performance of the channel. Therefore, our performance measurement for a STA includes not only utilization and end-to-end delay, but also packet loss in terms of collision rate and drop rate.

5.2.3.1 Collision rate

In order to calculate collision rate, the formula is given as

$$Col_{rate} = \frac{N_{col}}{N_{frame}}. \tag{5.6}$$

where Col_{rate} represents the collision rate, N_{col} is the total number of collisions and N_{frame} represents the total number of frames sent from a STA.

5.2.3.2 Drop rate

The drop rate is given as

$$Drop_{rate} = \frac{N_{drop}}{N_{frame}}. \tag{5.7}$$

where $Drop_{rate}$ represents the drop rate, N_{col} is the total number of drop frames and N_{frame} represents the total number of frames sent from a STA.

Elastic applications such as email, FTP, and web surfing are generally data oriented. They generally have longer end-to-end delay and low utilization tolerant but are strictly loss sensitive and require reliable data transfer [42].

5.3 Simulation Results - Scenario 1

5.3.1 Number of DIFSs

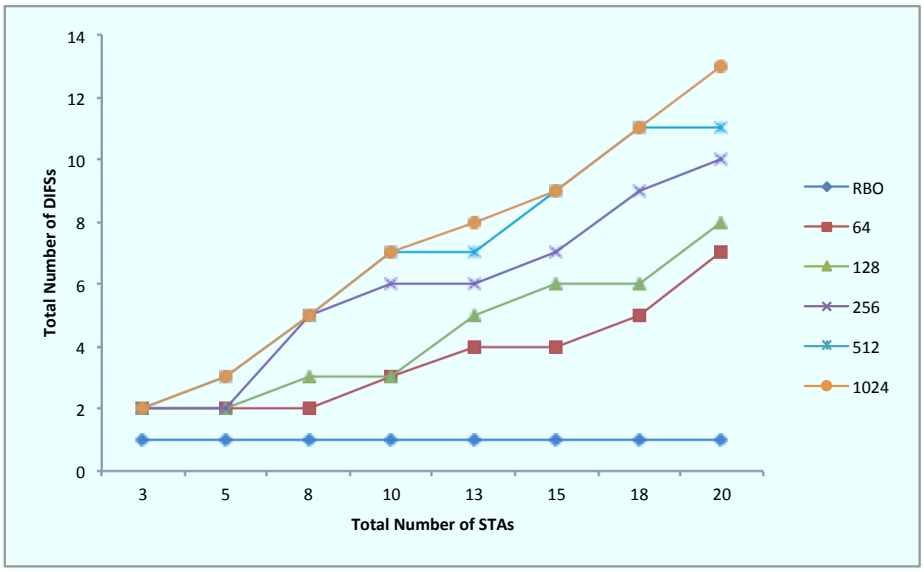

Figure 5.3: Number of DIFSs vs. Number of STAs for Different Number of Backoff Slots (Scenario 1).

Fig. 5.3 shows the number of DIFSs as function of number of STAs and backoff slots. The average number of DIFSs is around 2∼9 slots. With the exception of the default random backoff (RBO), which is 32 slots, the number of DIFSs increases almost linearly for different values of backoff slots. For example, when a STA chooses a fixed number of backoff slots of 1024 and there are three contending STAs, it observes on average two DIFSs before it can transmit a probe request frame. Moreover, the number of DIFSs increases linearly from 3 to 13 as the

number of STAs increases from 5 to 20. In contrast, when a STA adheres to the default random backoff mechanism, it only observes one DIFS regardless of the number of STAs. Therefore, the default random backoff does not provide a sufficient amount of time to properly observe the state of the channel.

5.3.2 Probing Delay

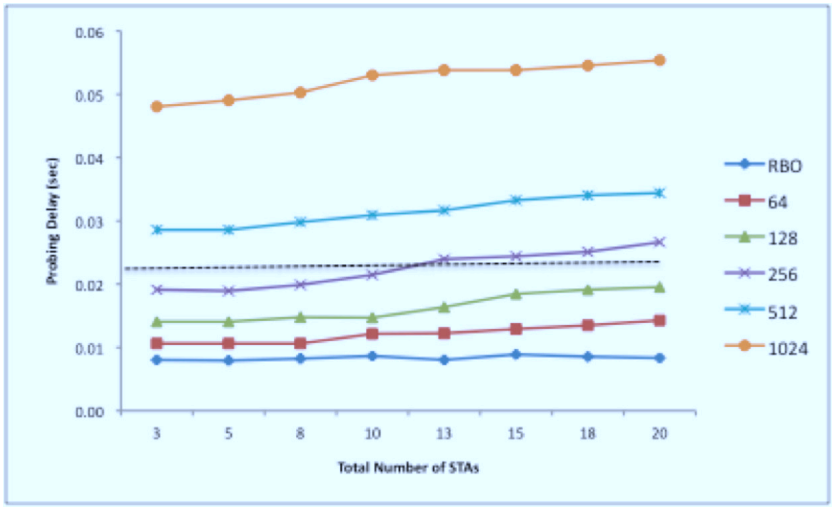

Figure 5.4: Probing Delay vs. Number of STAs for Different Number Backoff Slots (Scenario 1).

Fig. 5.4 shows the probing delay as a function of the number STAs and backoff slots. The probing delay increases slightly as the number of STAs increases. In contrast, the delay increases significantly as the number of backoff slots increases. Based on the VoIP probing delay requirement of less than 23.3 ms (see Sec. 4.1), which is indicated by the dotted line in Fig. 5.4, backoff slots of 512 and 1024 do not satisfy the delay requirement. The result for 256 backoff slots indicates that when the number of STAs is between 3 and 8, the delay requirement can be satisfied. However, when the number STAs exceeds 8, the delay requirement cannot be satisfied. Therefore, choosing a backoff slot of less than 256 will provide timely handoff.

5.3.3 Probability of Direct Collision

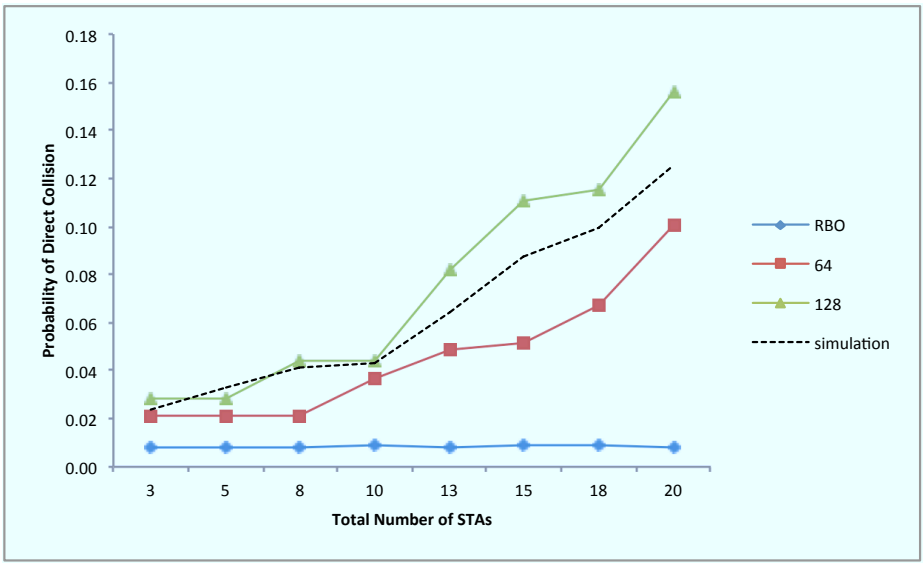

Figure 5.5: Probability of Direct Collision vs. Number of STAs for Different Number Backoff Slots (Scenario 1).

Fig. 5.5 shows P_{DC} as a function of the number of STAs with different numbers of backoff slots that would satisfy the delay requirement. It is important to note that these results, with the exception of the dotted line indicated as 'simulation', were generated using Eq. 4.9 based on the number of DIFSs shown in Fig. 5.3 and probing delays shown in Fig. 5.4. For this simulation, α is given as $1 \cdot \frac{1}{sec}$ for Eq. 4.9 because the values of d^i are between 2 and 8 and the values of D^i_{probe} are between 0.008 $msec$ and 0.018 $msec$, and thus, the values of $d^i \cdot D^i_{probe}$ are much less than 1. In addition, the purpose of estimating P_{DC} is to relatively select the best AP among candidates. If the values of $d^i \cdot D^i_{probe}$ are greater than 1, α should be 0

$< \alpha \leq \frac{1}{d^i \cdot D^i_{probe}}$. On the other hand, the result indicated by the dotted line was generated by keeping track of the actual number of collisions that occured during simulation. As can be seen from the figure, the result for the backoff time of 128 slots is the closest to the simulation result, which indicates that the best value for $optBO$ is 128 slots.

5.3.4 Probability of Hidden Node Collision

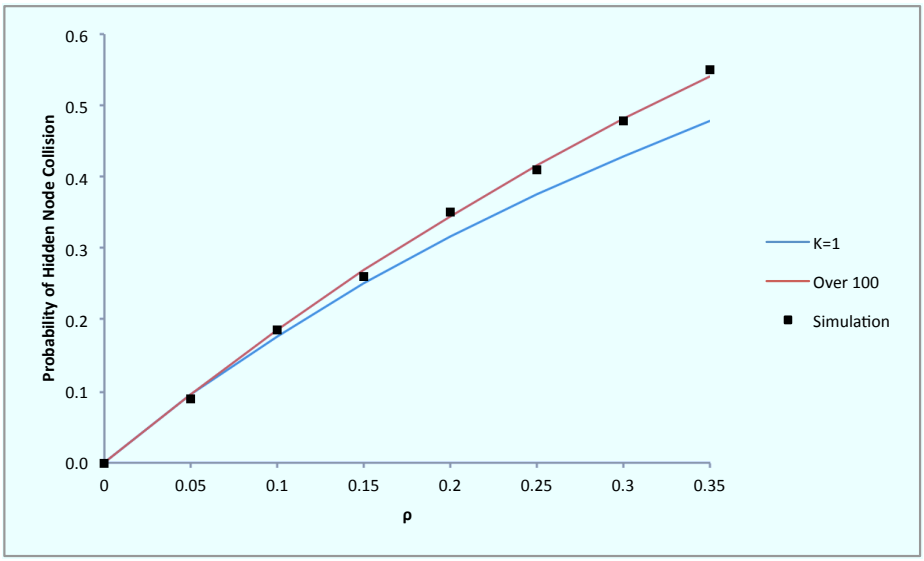

Figure 5.6: Probability of Hidden Node Collision vs. Number of STAs (Scenario 1).

Fig. 5.6 compares P_{HC} from Eq. 4.18 and the simulation results as a function of ρ with queue size of $K = 1$ and $K > 100$. The figure shows that the analytical result matches the simulation result when the queue size is over 100. Note that $K = 292$ for simulation, which is the default queue size in Qualnet [27]. Therefore, the simulation results refer the plot of $K = 292$. As can be seen from the figure, the analytical results for $K = 292$ matches with the simulation results, which show that the probability of hidden node collisions is strongly related to ρ. For example, when the ρ is 0.1, the collision probability is around 20%.

5.3.5 PR-ALBM vs. RSSI

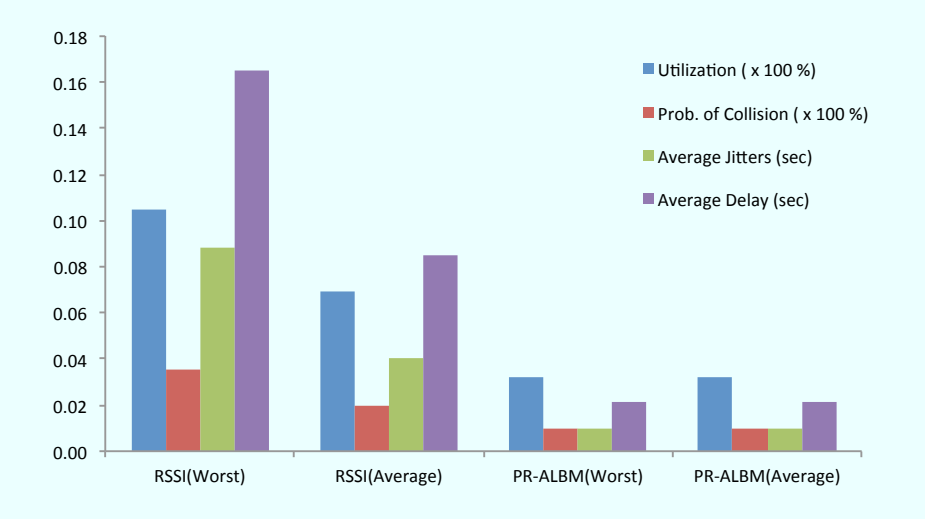

Figure 5.7: PR-ALBM vs. RSSI for non-QoS WLAN (Scenario 1).

Fig. 5.7 compares the performance of PR-ALBM against using simple RSSI for the cell that the STA associates with. Results are given in terms of utilization (i.e., the proportion of throughput utilized by the traffic in a cell), collision rate, average jitter, and average end-to-end delay, with each method reporting both worst case and average values. The PR-ALBM results in choosing an AP with lower utilization, which means that the available utilization for STA is higher. Furthermore, it chooses the AP with lower collision rate, lower jitter, and shorter delay than RSSI in both worst case and average values. This shows that the simple RSSI approach leads to inefficient association of STAs to available APs, which is similar to the results found in [2–5, 7].

5.4 Simulation Results - Scenario 2

5.4.1 Characteristics of ACs

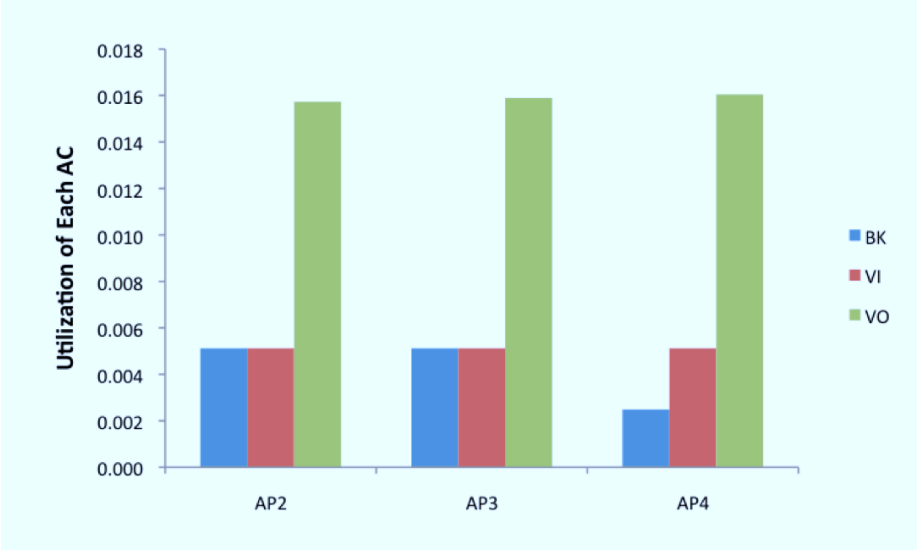

Figure 5.8: Average Utilization of ACs (Scenario 2).

Fig. 5.8 compares the average utilization of each AC traffic in different APs. As can be seen, VO traffic has the highest utilization and is stable across all APs. In contrast, the utilization of BK traffic is significantly lower in AP4, which has the largest number of high ACs (6 versus 2 for AP2 and 4 for AP3), while that of VI traffic is similar across all three APs. This is because high priority ACs restrict low priority ACs and their effect increases as their numbers increase in a cell.

Fig. 5.9 compares the average end-to-end delay of each AC traffic in different APs. These results are based on a configuration where the four APs are connected

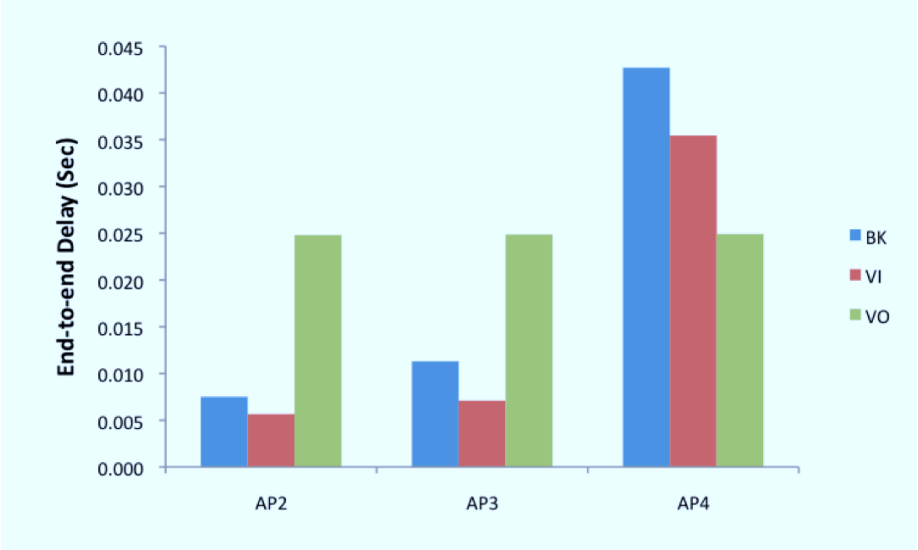

Figure 5.9: End-to-end Delays of ACs (Scenario 2).

to a switch which in turn is connected to a server that also acts as the destination server. Therefore, end-to-end delay represents the time duration between data frame transmissions from a STA to the destination server. Fig. 5.9 shows that VO traffic has the most stable end-to-end delay across all three APs. BK traffic in AP4 experiences the largest end-to-end delay because it has the largest number of high priority ACs. In addition, the delays for VO traffic are significantly less than BK and slightly less than VI in AP4. This is strongly related to the EDCF parameters shown in Table 2.3. That is, although VO traffic has the shortest AIFS, which is the same as VI, but it has shorter CW than VI. Therefore, VO traffic experiences the shortest waiting time to access the channel and thus much more opportunity to transmit frames than VI and BK resulting in shorter delay.

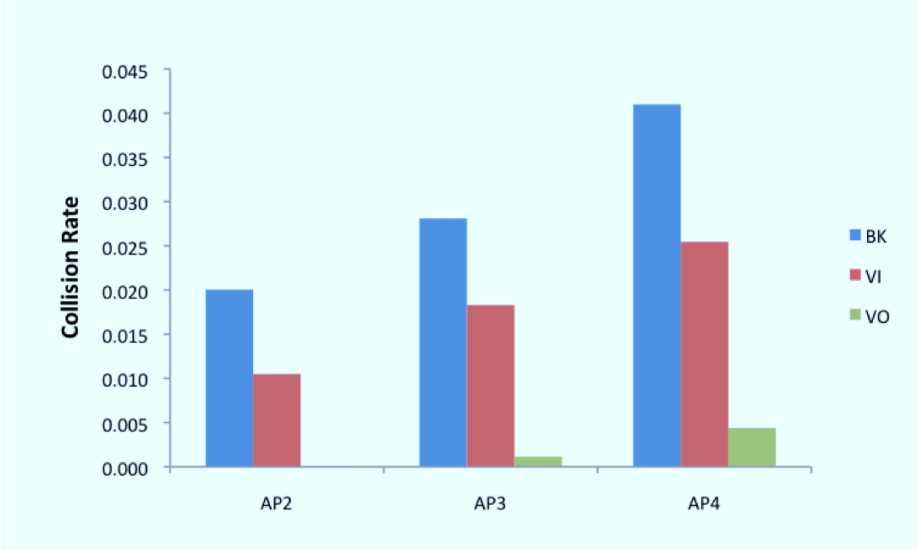

Figure 5.10: Collision Rates of ACs (Scenario 2).

The collision rate results in Fig. 5.10 are directly related to Fig. 5.9. In this figure, VO traffic has a significantly lower collision rate than VI or BK traffic and is relatively constant across all three APs. On the other hand, the collision rate of BK traffic significantly increases as the number of high priority ACs increases. This is because BK traffic has the largest CW and AIFS, and thus results in a higher level of contention among several BK flows. It is interesting to note that VI traffic experiences higher collision rates than expected. This is because VI traffic constantly generates packets every 200 ms similar to BK traffic. However, the rate of increase is much more gradual and much smaller than BK traffic.

The results in Fig. 5.11 are also directly related to Fig. 5.10. In this figure, BK traffic has the largest drop rate in AP4 while VO and VI traffics experience

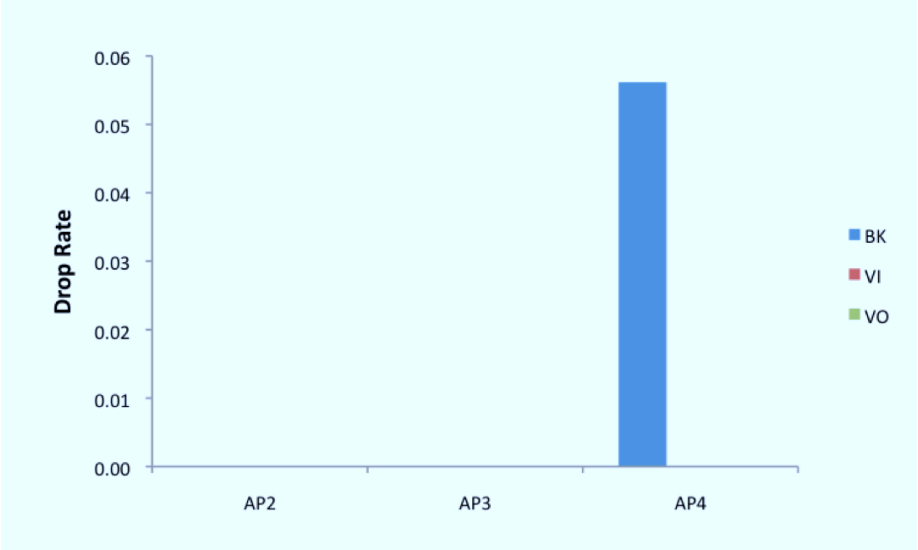

Figure 5.11: Drop Rates of ACs (Scenario 2).

no packet drops. This result is related to Fig. 5.10 in the sense that BK traffic has the largest CW and AIFS and results in higher level of contention among several BK flows. Although it does not experience packet drops with small numbers of high priority ACs, it significantly increases as the number of high priority ACs increases.

In summary, the number of high priority ACs ultimately determines the load of a cell and the cell with the least number of high priority ACs should be chosen as the best AP for association. These results are similar to those found in [50–55].

5.4.2 PR-ALBM vs. RSSI

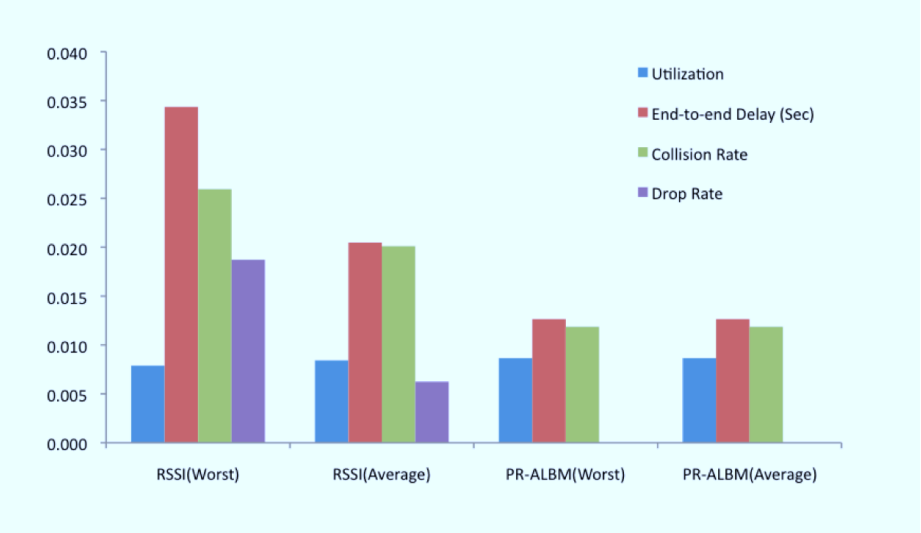

Figure 5.12: PR-ALBM vs. RSSI for QoS WLAN (Scenario 2).

Fig. 5.12 compares the overall performance of PR-ALBM against RSSI for QoS WLANs in terms of utilization, average end-to-end delay, collision rate, and drop rate. The simulation results are similar to those of non-QoS WLAN in Fig. 5.7. The PR-ALBM method chooses the AP that has shorter end-to-end delays, lower drop rates, and lower collision rates than RSSI in both worst and average cases. This shows that PR-ALBM is applicable for both non-QoS and QoS WLANs.

Chapter 6 – Conclusion

This dissertation proposed a new load balancing metric called PR-ALBM. The unique features of PR-ALBM are the use of probe requests to observe the channel and analytical models to estimate the characteristic of traffics and probability of collision for all surrounding channels. Our simulation results show that the proposed metric is accurate and thus very applicable in all types of IEEE 802.11 WLANs for applications requiring timely handoff and load balance.

Chapter 7 – Future Work

As part of future research, we plan to investigate couple of issues not only for load balance in infrastructured network but also for building a new routing protocol in ad-hoc network.

First, an enhanced probability metric of hidden node collisions will be developed and integrated into PR-ALBM. This can be achieved from not only the probability of hidden node collision metric based on traffic loads introduced in this dissertation but also the difference between the estimated number of nodes in a BSS based on the number of DIFS during the probe request period and the real number of nodes informed from AP during the probe reply period.

Second, we plan to develop a new routing algorithm for military *Mobile Ad-hoc NETworks* (MANETs). MANETs are considered as one of the network technologies for *Network Centric Warfare* (NCW) because they require no infrastructure or pre-configuration, and thus they can be used to dynamically create communication networks for groups of wireless users in any situation, such as open areas, mountain regions, and urban area (see Sec. B.1). Therefore, routing protocols are crucial for maintaining some degree of connectivity, even as network topology frequently changes based on the movement of military forces. In order to develop a new algorithm, we plan to consider the mobility models of not only ground forces such as infantry, tank, and artillery but also *Unmanned Aerial Vehicles* (UAVs) for provid-

ing communication support to troops on the ground (see Sec. B.1). These models consider real military tactical mobility and situation. We proposed a new ground mobility model, called *Urban Military Operation Mobility Model* (UMOMM) which will be set for ground mobility model and UAV mobility model is planned to be developed with several ground tactical issues (see Sec. B.1). This is because one of the missions of UAV is to provide communication support to troops on the ground. Then, we will develop a new routing algorithm for MANETs. This will be done by considering several military tactical environments. First, military command structure needs to be reflected in the routing protocol since traffics from each unit commander must be transferred within a given amount of bandwidth. Second, traffics from different emergency situation such as medical, chemical, and warfare dangerous situation will be taken into account. Third, commanders need to monitor everything from their solders, such as operation situation and medical state, and report the combined information to his directive commander. Therefore, understanding how PR-ALBM will perform under these conditions is crucial for properly adjusting some of the parameters such as utilization, end-to-end delays, and packet loss.

Bibliography

[1] Metro-Area and Campus Wi-Fi. Available at http://www.connect802.com/metro_wifi.htm

[2] G. Athanasiou *et al.*, "Dynamic Cross-Layer Association in 802.11-based Mesh Networks," *INFOCOM 2007*, May 2007, pp. 2090-2098.

[3] O. Ekici and A. Yongacoglu, "A Novel Association Algorithm for Congestion Relief in IEEE 802.11 WLANs,"' *Proc. of the 2006 International Conference on Wireless Communications and Mobile Computing (IWCMC06)*, July 2006, pp. 725-730.

[4] T. Korakis *et al.*, "Link Quality based Association Mechanism in IEEE 802.11h Compliant Wireless LANs," *1st Workshop on Resource Allocation in Wireless NETworks (RAWNET06*), Apr. 2006, pp. 725-730.

[5] Y. Bejerano *et al.*, "Fairness and Load Balancing in Wireless LANs Using Association Control," *Proc. of the 10th Annual International Conference on Mobile Computing and Networking (MobiCom04)*, USA, Sept. 2004, pp. 315-329.

[6] L. Yen *et al.*, "Load Balancing in IEEE 802.11 Networks," *Internet Computing*, Feb. 2009, pp. 56-64.

[7] A. Balachandran *et al.*, "Hot-Spot Congestion Relief in Public-area Wireless Networks," *Proc. of the 4th IEEE Workshop on Mobile Computing Systems and Applications (MCSA02)*, June 2002, pp. 70-80.

[8] I. Papanikos and M. Logothetis, "A Study on Dynamic Load Balance for IEEE 802.11b Wireless LAN," *8th Intl Conf. Advances in Communication and Control*, 2001, pp. 83-89.

[9] T-C. Tsai and C.-F. Lien, "IEEE 802.11 hot spot load balance and QoS-maintained seamless roaming," *Proc. National Computer Symposium (NCS)*, 2003.

[10] M. Lee *et al.*, "Enhanced Algorithm for Initial AP selection and Roaming," *US patent 0039817, Patent and Trademark Office*, Feb. 2004.

[11] A. Barbaresi *et al.*, "Admission Control Policy for WLAN Systems based on the Capacity Region," *IST Mobile Summit*, Jun 2005.

[12] M. Portoles *et al.*, "EEE 802.11 Downlink Traffic Shaping Scheme for Multi-User Service Enhancement," *14th IEEE Proceedings on Personal, Indoor and Mobile Radio Communications (PIMRC 2003)*, Sept. 2003, pp. 1712-1716.

[13] A. P. Jardosh *et al.*, "IQU: Practical Queue-Based User Association Management for WLANs," *Proc. of the 12th Annual International Conference on Mobile Computing and Networking (MobiCom06)*, Sept. 2006, pp. 158- 169.

[14] S. Vasudevan *et al.*, "Facilitating Access Point Selection in IEEE 802.11 Wireless Networks," *Proc. of the 5th ACM SIGCOMM Conference on Internet Measurement*, 2005, pp. 293-298.

[15] A. Veres *et al.*, "Supporting Service Differentiation in Wireless Packet Networks Using Distributed Control," *IEEE Journal on Selected Areas in Communications*, vol.19, no.10, Oct. 2002, pp.2081-2093.

[16] A. Nicholson *et al.*, "Improved Access Point Selection," *Proc. IEEE Intl Conf. Mobile Systems, Applications and Services (MobiSys06).*, 2006, pp. 233-245.

[17] L. Yen *et al.*, "SNMP-Based Approach to Load Distribution in IEEE 802.11 Networks," *Proc. of Vehicular Technology Conference (VTC2006-Spring)*, 2006, pp. 1196-1200.

[18] H. Velayos, V. Aleo, and G. Karlsson, "Load Balancing in Overlapping Wireless LAN Cells," *Proc. IEEE Intl Conf. Comm.*, 2004, pp. 3833-3836.

[19] E. Villegas *et al.*, "Load Balancing WLANs through IEEE 802.11k Mechanisms," *Proc. of ISCC'06*,2006.

[20] P. Iyer *et al.*, "System and Method for Centralized Station Management," *US patent 0213579, Patent and Trademark Office*, Sept. 2005.

[21] IEEE Std 802.11e-2005, Amendment 1 : Medium Access Control (MAC) Quality of Service Enhancements.

[22] Jim Geier, Wireless LANs : Implementing High Performance IEEE 802.11 Networks, Sams, second edition, 2001.

[23] IEEE Std 802.11-2007, Part 11 : Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications, 2007.

[24] IEEE Std 802.11a, Part 11 : Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications, 1999.

[25] IEEE Std 802.11b, Part 11 : Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications, 1999.

[26] IEEE Std 802.11b, Part 11 : Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications, 2003.

[27] QualNet Simulator. Available at http://www.scalable-networks.com/.

[28] IEEE Std 802.11k-2008, Amendment 1 : Radio Resource Measurement of Wireless LANS.

[29] A. Mishra *et al.*, "An empirical analysis of the IEEE 802.11 MAC layer handoff process," *ACM SIGCOMM Computer Communication Review*, 2003, pp. 93-102.

[30] H. Velayos and G. Karlsson, "Techniques to reduce IEEE 802.11b MAC layer handover time," in IEEE International Conference on Communications (ICC), Jun. 2004, pp. 3844-3848.

[31] V. Brik, A. Mishra, and S. Banerjee, "Eliminating handoff latencies in 802.11 WLANs using multiple radios: applications, experience, and evaluation," in Internet Measurement Conference (IMC), Oct. 2005, pp. 27-32.

[32] S. Waharte, K. Ritzenthaler, and R. Boutaba, "Selective active scanning for fast handoff in WLAN using sensor networks," in Mobile and Wireless Communications Networks (MWCN), Oct. 2004, pp. 59-70.

[33] I. Ramani and S. Savage, "Syncscan: practical fast handoff for 802.11 infrastructure networks," in IEEE INFOCOM, Mar. 2005, pp. 675-684.

[34] S. Pal, S. Kundu, and K. Basu. Handoff : Ensuring seamless mobility in IEEE 802.11 wireless networks. [Online]. Available: http://crewman.uta.edu/corenetworking/projects/handoff/newhandoff.html

[35] M. Shin, A. Mishra, and W. A. Arbaugh, "Improving the latency of 802.11 hand-offs using neighbor graphs," in The International Conference on Mobile Systems, Applications, and Services (MOBISYS), Jun. 2004, pp. 70-83.

[36] S. Shin, A. G. Forte, A. S. Rawat, and H. Schulzrinne, "Reducing mac layer handoff latency in IEEE 802.11 wireless LANs," in ACM International Workshop on Mobility Management and Wireless Access (MOBIWAC), Sep. 2004, pp. 19-26.

[37] V. Mhatre and K. Papagiannaki, Using smart triggers for improved user performance in 802.11 wireless networks, in Proceedings of MobiSys, June 2006, pp. 246-259.

[38] International Telecommunication Union, "Recommendation G.1020" July 2006

[39] International Telecommunication Union, "Recommendation Y.1541" July 2006

[40] F. Guo and T. Chiueh, "Device-transparent network-layer handoff for micro-mobility," *IEEE International Symposium on Modeling, Analysis & Simulation of Computer and Telecommunication Systems (MASCOTS '09)*, Sept. 2009, pp. 1-10.

[41] "Deploying VoWiFi: Handoff delays and QoS of voice over WiFi". Available at http://www.codealias.info/-technotes/network_performace_requirements_for_voice _over_ip.

[42] J. Kurose and K. Ross, "Computer Networking : a top-down approach featuring the Internet," *Addison-Wesley*, fifth edition. 2010.

[43] W. Wanalertlak*et al.*, "Global Path-Cache Technique for Fast Handoff in WLANs," *Proc. of 16th International Conference on Computer Communications and Networks (ICCCN07)*, Aug. 2007, pp. 45-50.

[44] W. Wanalertlak *et al.*, "Behavior-based Mobility Prediction for Seamless Handoffs in Mobile Wireless Networks," accepted for publication in *Wireless Networks, Journal of Mobile Communication, Computation, and Information (WiNET)*, 2010.

[45] I. Purushothaman and S. Roy, "FastScan: a handoff scheme for voice over IEEE 802.11 WLANs," *Wireless Networks*, March 2010, pp. 2049-2063.

[46] A. Jardosh *et al.*, "Understanding Congestion in IEEE 802.11b Wireless Networks," *Proc. of the 2005 Internet Measurement Conference*, 2005, pp 279-292.

[47] Tourrilhes. J, "Packet frame grouping : improving IP multimedia performance over CSMA/CA," *IEEE International Conference on Universal Personal Communications, 1998*, Oct. 1998, pp. 1345-1349.

[48] Y. Fakhri and B. Nsiri *et al.*, "Throughput Optimization Via the Packet Length and Transmission Rate For Wireless OFDM System in Downlink Transmission," *International Journal of Computer Science and Network Security*, Vol.6 No.3B, March. 2006, pp. 41-46.

[49] M. Mansor and J. Abdullah, "Evaluating the Communication performance of an Ad Hoc Wireless Network using Mesh Connectivity Layer (MCL) Protocol," *2009 Conference on Innovative Technologies in Intelligent Systems and Industrial Applications*, July. 2009, pp. 192-197.

[50] I. Inan *et al.*, "An Adaptive Multimedia QoS Scheduler for 802.11e Wireless LANs," *Proc. of 2006 IEEE International Conference on Communications (ICC06)*, Dec. 2006, pp. 5263-5270.

[51] G. Hwang *et al.*, "New access scheme for VoIP packets in IEEE 802.11e wireless LANs," *IEEE Communications Letters*, July. 2005, pp. 667-669.

[52] P. Engelstad and O. Osterbo, "An analytical model of the virtual collision handler of 802.11e," *Proc. of the 8th ACM international symposium on Modeling, analysis and simulation of wireless and mobile systems (MSWiM05)*, 2005, pp. 255-259.

[53] R. Acharya *et al.*, "WLAN QoS Issues and IEEE 802.11e QoS Enhancement," *International Journal of Computer Theory and Engineering*, Vol.2, No.1, Feb. 2010, pp. 143-149.

[54] J. Sengupta and G. Grewal *et al.*, "Performance evaluation of IEEE 802.11 MAC layer in supporting delay sensitive services," *International Journal of Wireless and Mobile Networks (IJWMN)*, Vol.2, No.1, Feb. 2010, pp. 42-53.

[55] D. He and C. Shen *et al.*, "Simulation study of IEEE 802.11e EDCF," *The 57th IEEE Semiannual Vehicular Technology Conference*, April. 2003, pp. 685-689.

[56] Leonard Kleinrock, "QUEUEING SYSTEMS. Volume 1: Computer Applications," Wiley-Interscience, Jan. 1975.

APPENDICES

Appendix A – Birth-Death Queuing Systems

The Markov process, named after the creator Andrey Markov, is a stochastic process holding a Markov property in terms of memorylessness. That is, its past states are irrelevant and its future state is determined only by its most recent state. A set of random variables X forms a Markov chain if the probability that the next state is x_{n+1} depends only upon the current state x_n, regardless of the state's previous history [56]. Therefore, Markov property may be written as

$$P[X(t_{n+1}) = x_{n+1} \mid X(t_n) = x_n, X(t_{n-1}) = x_{n-1}, \cdots, X(t_1) = x_1] = P[X(t_{n+1}) = x_{n+1} \mid X(t_n) = x_n] \tag{A.1}$$

where $t_1 < t_2 < \cdots\ t_n < t_{n+1}$ and x_i are included in some discrete state space.

Birth-Death process is a special case of a Markov process in which transitions from state K are permitted only to neighboring states $K+1$, K, and $K-1$. The state transition rate diagram for the Birth-Death process is shown in Fig. A.1.

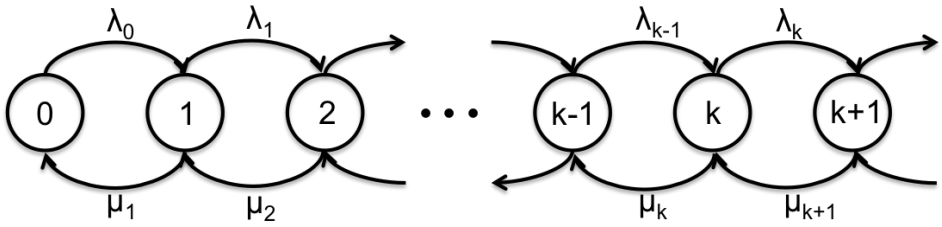

Figure A.1: State Transition Rate Diagram for the Birth-Death Process.

The transitions are permitted only from state K to neighboring states $K+1$ or $K-1$. A transition from K to $K+1$ denotes a birth and a transition from K to $K-1$ is referred to as a death in the state transition rate diagram for the Birth-Death process. In Fig. A.1, the state K in equilibrium, which clearly depend on time, can be observed as following a set of differential difference two equations:

$$\frac{dP_k(t)}{dt} = -(\lambda_k + \mu_k) \cdot P_k(t) + \lambda_{k-1} \cdot P_{k-1}(t) + \mu_{k+1} \cdot P_{k+1}(t), (k \geq 1) \tag{A.2}$$

$$\frac{dP_0(t)}{dt} = -\lambda_0 \cdot P_0(t) + \mu_1 \cdot P_1(t), (k = 0) \tag{A.3}$$

A Birth-Death system becomes more complicated and unmanageable when a complex time dependent birth-death situation is consider due to a lot of transient behaviors. A queue is in equilibrium if the following equation exists where the probabilities $P_k(t)$ keep constant as t goes to infinity and show no more transient behavior.

$$P_k = \lim_{t \to \infty} P_k(t) \tag{A.4}$$

If Eq. A.4 exists, a differential-difference equation, $\frac{dP_k(t)}{dt}$, becomes zero as t increases to infinity. Therefore, Eq. A.2 and Eq. A.3 will be given by

$$0 = -(\lambda_k + \mu_k) \cdot P_k + \lambda_{k-1} \cdot P_{k-1} + \mu_{k+1} \cdot P_{k+1}, (k \geq 1) \tag{A.5}$$

$$0 = -\lambda_0 \cdot P_0 + \mu_1 \cdot P_1, (k = 0) \tag{A.6}$$

Eq. A.5 and Eq. A.6 can be reformulated as given below since negative number of population is not allowed:

$$0 = -(\lambda_k + \mu_k) \cdot P_k + \lambda_{k-1} \cdot P_{k-1} + \mu_{k+1} \cdot P_{k+1} \tag{A.7}$$

Eq. A.7 can also obtained by inspecting the state K in equilibrium as shown in Fig. A.1 and expressed as follows:

$$\lambda_{k-1} \cdot P_{k-1} + \mu_{k+1} \cdot P_{k+1} = (\lambda_k + \mu_k) \cdot P_k \tag{A.8}$$

Therefore, the probability of state $K = 1$ is given as

$$P_1 = \frac{\lambda_0}{\mu_1} \cdot P_0 \tag{A.9}$$

Then, the probability of state $K = 2$ with Eq. A.8 and Eq. A.9 is represented by the following equation:

$$P_2 = \frac{\lambda_0 \lambda_1}{\mu_1 \mu_2} \cdot P_0 \tag{A.10}$$

In such way, the general probability of state K will be expressed using the following two equations:

$$P_k = \frac{\lambda_0 \lambda_1 \cdots \lambda_{k-1}}{\mu_1 \mu_2 \cdots \mu_k} \cdot P_0 \tag{A.11}$$

$$P_k = P_0 \prod_{i=0}^{k-1} \frac{\lambda_i}{\mu_{i+1}}. \tag{A.12}$$

Since the summation probability of all state is one, the probability of state zero can be obtained from Eq. A.14:

$$\sum_{k=0}^{\infty} P_k = 1 \tag{A.13}$$

$$P_0 = \frac{1}{1 + \sum_{k=1}^{\infty} \prod_{i=0}^{k-1} \frac{\lambda_i}{\mu_{i+1}}}. \tag{A.14}$$

Appendix B – Routing Protocol for NCW

B.1 Network Centric Warfare and MANETs

Network Centric Warfare (NCW) is a new theory that focuses on increasing combat effectiveness by linking or networking resources among military forces, and its fundamental goal is to provide accurate, detailed, real-time information to all levels of command and control. In NCW, future soldiers will be equipped with an integrated set of high-technology devices such as GPS, wireless communications, and sensors. These will be linked to an array of real-time and archived battlefield information resources. Moreover, these soldiers will form *Mobile Ad-hoc NETworks* (MANETs) to provide crucial information, such as videos and pictures of combat situations as well as location and vital signs of soldiers, back to the command structure.

MANETs are considered as one of the network technologies for NCW because they require no infrastructure or pre-configuration, and thus they can be used to dynamically create communication networks for groups of wireless users in any situation, such as open areas, mountainous regions, and urban areas. In a MANET, the network topology frequently changes based on the movement of mobile nodes, e.g., soldiers and Humvees. Therefore, routing protocols are crucial for maintaining some degree of connectivity, even as nodes move. There are a number of routing

protocols for MANETs, which include proactive and reactive routing protocols. However, the performance of routing protocols has been difficult to properly evaluate because of lack of mobility models that realistically represent the behavior of mobile nodes in military situations.

Although there have been many research efforts on synthetic entity and group mobility models, they cannot be applied in realistic combat operation scenarios. This is because mobile nodes in military situations are not independent but typically related to each other and have more complex mobility scenarios depending on tactical situations and military units. For example, one typical characteristic of military operations is that a group can dynamically partitioned into subgroups or merge with another group. For instance, in urban areas, a number of army units will first mobilize outside the urban area. When operation orders are given, the units will move toward their destinations within the urban environment. During the operation, a group may be divided into several subgroups where some of the subgroups are assigned new tasks while the rest of the subgroups continue towards their original objectives. After completing their new missions, subgroups will rejoin its main force.

Therefore, we proposed Urban Military Operation Mobility Model (UMOMM) with several unique features considering real military tactical environment as follows:

First, UMOMM is a military mobility model for infantry with restrictions on movements within an urban area which has become increasingly important for simulating modern battlefield situations encountered by troops. It adopts a platoon,

typically the smallest military unit, as the basic group model. A platoon can be dynamically partitioned into smaller squads each with a new task and they can be merged with its main force at an arbitrary point on the platoons. Moreover, mobility of platoons can be extended to model operations of all military units such as company, battalion, and division. The initial model of squad unit is as shown in Fig. B.1.

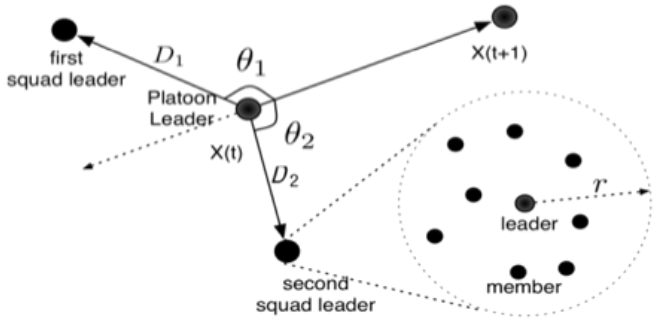

Figure B.1: Model Initialization.

Second, mobility patterns differ depending on the involved mission of each platoon, such as surveillance, gaining control of maneuver routes and intermediate locations, and occupying assigned locations on the map. In addition, UMOMM can model situations where each group encounters various obstacles constructed by hostile forces and must overcome them to reach the destination. The movement model including group partitioning, merging, and blocking area for UMOMM is as shown in Fig. B.2.

Finally, the UMMOM is used for military urban operation in urban. Fig. B.3 shows downtown of Portland, an example of urban area and the abstracted topol-

ogy of the area.

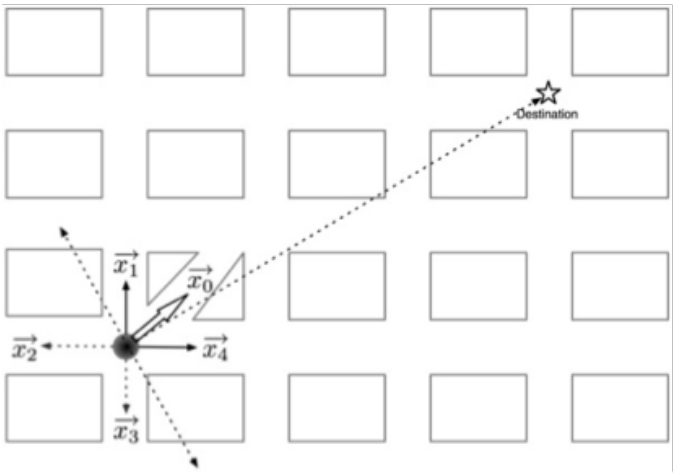

Figure B.2: Group Partitioning, Merging, and Blocking Area.

Figure B.3: Downtown of Portland and Abstracted Topology.

We plan to develop an UAV mobility model taking into account several issues as follows: First, UAVs can operate autonomously and maintain network connectivity among UAVs and provide coverage to its tactical operation area. Second, movement characteristics include speed, altitude, relative direction angle of UAVs.

Therefore, UAVs act as prominent radio nodes that connect disconnected ground radio nodes, thus they increase node connectivity and operational range for ground forces as shown in Fig. B.4. In addition, UAV mobility model, referred to as UAV-UMOMM, needs to be integrated with Ground-UMOMM to better understand the viability of the NCW concept.

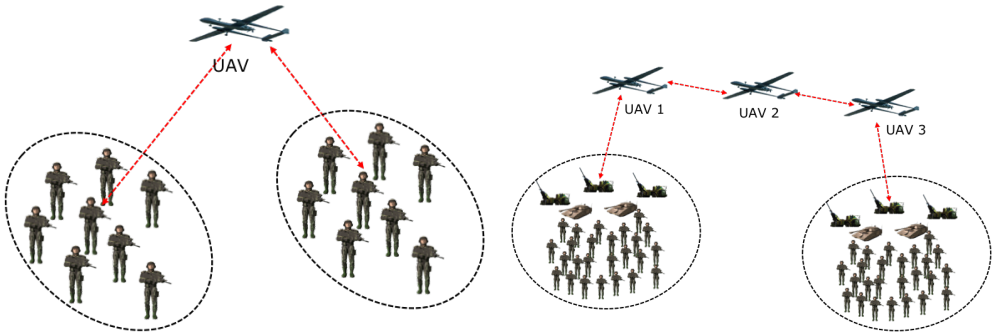

Figure B.4: Increase Node Connectivity and Operation Range.

B.2 Routing Protocol for UAV/Ground-UMOMM

The objective of building an UAV mobility model is to develop UAV/Ground-UMOMM to study the effectiveness of UAVs for providing communication support to troops on the ground. Obviously, a new routing protocol considering military tactical environment discussed above is required to implement NCW. In order to build a new protocol, we adopt MANET communication between nodes based on military command structure because MANET has been evaluated as an efficient network for military tactical requirements, such as quick termination and estab-

lishment of network. In addition, we also consider military command structure and emergency situation. Therefore, the protocol should be a hierarchical routing protocol.

Fig. B.5 and Fig. B.6 show the topology for our future study. Fig. B.5 also shows an example of a NCW conceptual model for UAVs consisting of several types of ground forces, i.e., infantry, tank, and artillery, and *Tactical Operation Center* (TOC). A military unit, e.g., division, consists of a set of sub-units that cover a tactical operation area. Each sub-unit forms a cluster so that all of its members can communicate with each other, and it has a commander (head) that can communicate with other sub-unit commanders. Sub-unit commanders can communicate with low-altitude UAVa, and these UAVs form a MANET to communicate amongst each other. Low-altitude UAVs can communicates with a high-altitude UAV and high-altitude UAVs may communicate with *Tactical Operation Center Head Quater* (TOC HQ) or TOC Division via satellite.

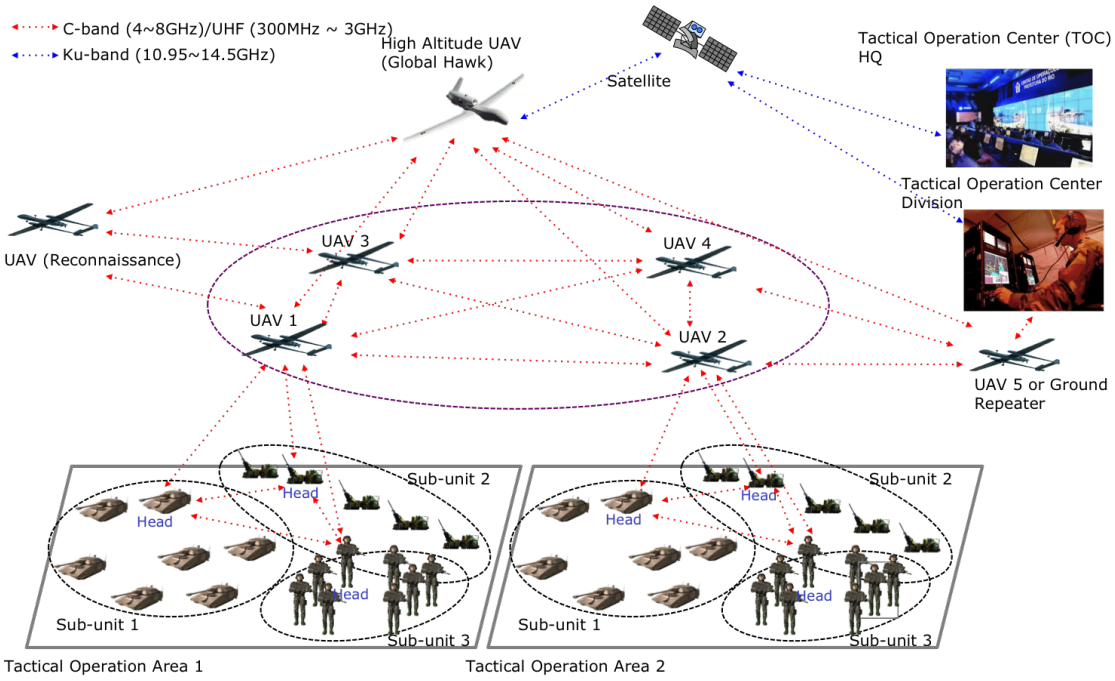

Figure B.5: NCW Conceptual Model.

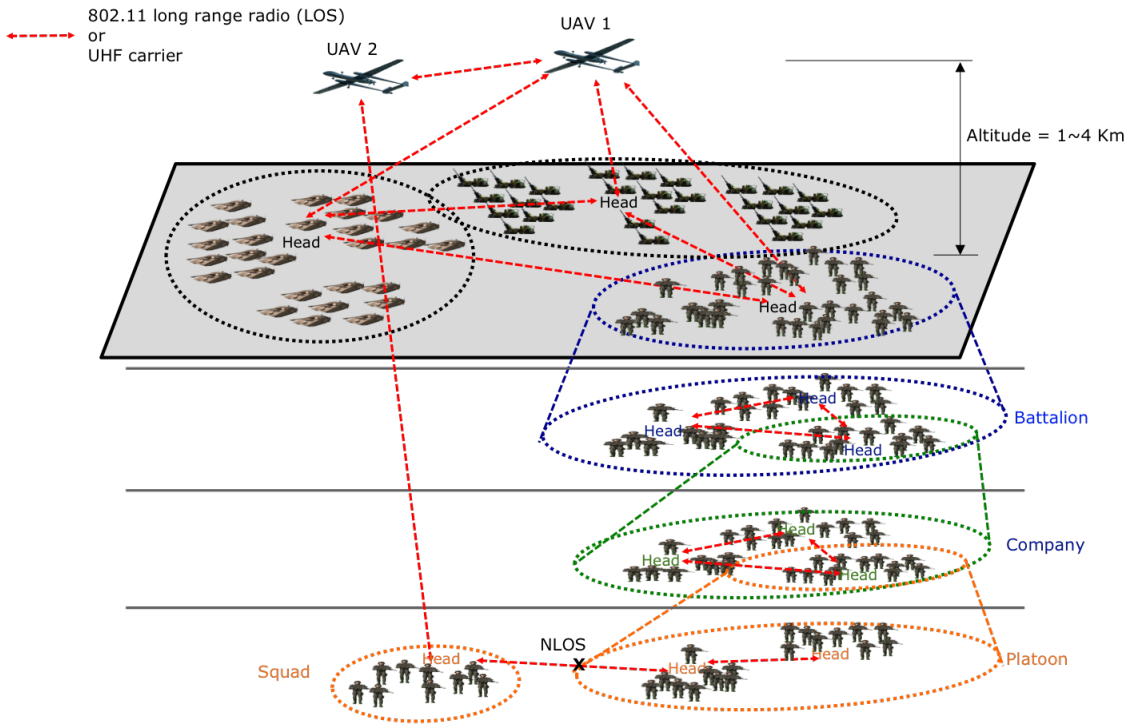

Figure B.6: UAV and Ground Network Topology.