INFRASTRUCTURE POWER SAVING AND QUALITY-OF-SERVICE PROVISIONING FRAMEWORK FOR WIRELESS LAN MESH NETWORKS

INFRASTRUCTURE POWER SAVING AND QUALITY-OF-SERVICE PROVISIONING FRAMEWORK FOR WIRELESS LAN MESH NETWORKS

BY

AHMAD M. KHOLAIF, M.SC., B.SC. (CAIRO UNIVERSITY)

A THESIS

SUBMITTED TO THE DEPARTMENT OF ELECTRICAL & COMPUTER ENGINEERING

AND THE SCHOOL OF GRADUATE STUDIES

OF MCMASTER UNIVERSITY

IN PARTIAL FULFILMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

Doctor of Philosophy (2008) McMaster University

(Electrical & Computer Engineering) Hamilton, Ontario, Canada

TITLE:	Infrastructure Power Saving and Quality-of-Service Provisioning Framework for Wireless LAN Mesh Networks
AUTHOR:	Ahmad Mohammad M. Kholaif, M.Sc., B.Sc. (Cairo University)
SUPERVISOR:	Dr. Terence D. Todd, Professor of Electrical and Computer Engineering, B.A.Sc., M.A.Sc., Ph.D. (University of Waterloo)
NUMBER OF PAGES:	xv, 166

Abstract

Internet access using IEEE 802.11 wireless local area networks has become very common. In home and office networks where voice, video and audio will be delivered, quality of service (QoS) support is essential so that customers can be offered video on demand, audio on demand, voice over IP and high-speed Internet access. In addition to the proliferation of WLAN hotspots, WLAN mesh networks are now being used as a cost-effective means for coverage extension and backhaul relaying between IEEE 802.11 access points (APs).

In conventional IEEE 802.11, APs are always continuously powered using fixed wired connections. In future WLAN mesh networks however, wired power connections may not always be readily available, especially in Wi-Fi hotzone installations which cover expansive outdoor areas. In such cases, fixed power connections can often be replaced by a battery operated or solar powered design. For this reason, power saving on the AP is highly desirable for this type of application. Unfortunately, this is not possible since the existing IEEE 802.11 standard requires that APs remain active at all times.

In this thesis, we propose and investigate a comprehensive framework for a power saving QoS-enabled access point (PSQAP), intended for use in solar and low power IEEE 802.11 infrastructure applications. An energy-efficient media access control protocol is proposed using the contention-based channel access mode for IEEE 802.11. When real-time flows are present, a PSQAP schedules its awakening/sleeping pattern in a manner that satisfies the delay and packet loss requirements for the admitted real-time flows. A dynamic connection-admission control algorithm is proposed for efficient management of wireless resources. We show that both background traffic and the synchronization of stations' transmissions due to AP's alternating between awake and sleep states can cause excess queuing and packet collision rate. These effects result in an increase in packet delay and power consumption at the mobile stations in contention-based channel access mode. We propose and investigate several scheduling methods for mitigating these effects. It is also shown that voice over IP over WLAN (VoWLAN) suffers a low capacity problem and high handset/AP power consumption. A novel adaptive voice packetization scheme is proposed which improves VoIP capacity and reduces power consumption. The work in this thesis is characterized by

analytical models and evaluated through extensive network simulations to study and analyze the key performance aspects of the proposed framework and the associated protocols.

Acknowledgements

I would like to acknowledge the advice and guidance of my supervisor, Dr. Terence D. Todd, for he was the person who has given me this unique opportunity to learn and explore knowledge. I am grateful for him allowing me to work on interesting projects and for his deep insight and constructive discussions from which I have learned much. Without his help throughout my research, this study would have not been successful. I would also like to thank my co-supervisor, Dr. Ted Szymanski, and the members of my supervisory committee Dr. Douglas Down and Dr. Dongmei Zhao for their support and for the time they spent discussing queuing and resource management problems even when I dropped by their offices with no prior notice. Special thanks to my unique research collaborator, Dr. Polychronis Koutsakis for his support. The many helpful discussions on various modeling, scheduling and admission control issues with Dr. Koutsakis contributed to the enrichment of a large portion of this work.

I would like to thank all the administrative and technical staff at the ECE department. Cheryl, Helen, Cosmin and Terry, thank you for your help that made my life easier during my period of study at Mac.

I would like to express my appreciation to my dear team-mates in the ECE department in general and in the Wireless Networking Laboratory in particular. Special thanks are to Mohammad, Waheed, Amir and Wesam Mesbah. My PhD study was a uniquely enjoyable experience just because of you and I wish you all the best in your life.

To my wife, Wesam Fandari, I honor you for all the home sickness, loneliness and for all what you did just to help me accomplish my goals. To my son, Yusuf, your innocent smile has been very inspiring for me.

I am as ever, especially indebted to my family and in particular to my dearest parents, Mohammad and Nadyah. It is because of all your constant care, encouragement and teachings I am able to stand here today. May ALLAH reward you the best for that.

List of Acronyms

3G	Third Generation Cellular Networks
AC	Access Category
ACK	Acknowledgment
AIFS	Arbitrary Inter – Frame Spacing
AP	Access Point
APSD	Automatic Power Save Delivery
AVP	Adaptive Voice Packetization
BSS	Basic Service Set
CAC	Connection Admission Control
CBR	Constant Bit Rate
CP	Contention Period
CSMA/CA	Carrier Sense Multiple Access with Collision Avoidance
CTS	Clear To Send
CW_{max}	Maximum Backoff Window Size
CW_{min}	Minimum Backoff Window Size
DAR	Discrete Autoregressive Model
DAR(1)	Discrete Autoregressive Model With Order One
DCF	Distributed Coordination Function
DL	Downlink
DIFS	Distributed Interframe Spacing
DS	Distribution System
DTIM	Delivery Traffic Indication Message
DTT	Downlink Transmission Time
EBSS	Extended Basic Service Set
EDCA	Enhanced Distributed Channel Access
EDF	Earliest Deadline First

GSM	Global System for Mobile Communications
GW	Gateway
HCCA	Hybrid Coordination Function Controlled Channel Access
HCF	Hybrid Coordination Function
HTTP	Hypertext Transfer Protocol
iBSS	infrastructure Basic Service Set
IBSS	Independent Basic Service Set
IP	Internet Protocol
IEEE	Institute of Electrical and Electronics Engineers, Inc.
ISM	Industrial Scientific and Medical Radio Bands
LAN	Local Area Network
LID	Logical Identifier
LM	Latency Margin
MAC	Media Access Control
MAP	Mesh Access Point
MP	Mesh Point
MPEG – 1	Moving Picture Experts Group Layer 1 Encoding
MPEG – 2	Moving Picture Experts Group Layer 2 Encoding
MS	Mobile Station
MSDU	Media Access Control Service Data Unit
NAM	Network Allocation Map
NAV	Network Allocation Vector
NIC	Network Interface Card
PB	Predicted – Bidirectionalframe
PBX	Private Branch Exchange
PC	Point Coordinator
PCF	Point Coordination Function
PHY	Physical Layer
PIFS	Point Coordinator Interframe Spacing
PLMN	Public Land Mobile Network
PoE	Power over Ethernet
PSM	Power Save Mode
PSMP	Power Save Multi – Poll
PSQAP	Power Saving QoS – Enabled Access Point
Ptime	Packetization Interval Time
QAP	Quality of Service Capable Access Point
QCIF	Quarter Common Intermediate Format
QoS	Quality of Service
QSTA	Quality of Service Capable Station

RF	Radio Frequency
RSSI	Received Signal Strength Indicator
RTP	Real – Time Protocol
RTS	Real – Time Server
S – APSD	Scheduled Automatic Power Save Delivery
S – PSMP	Scheduled Power Save Multi – Poll
SDP	Session Description Protocol
SF	Surplus Factor
SI	Service Interval
SIFS	Short Inter – Frame Spacing
SIP	Session Initiation Protocol
SP	Service Period
SSID	ServiceSetIdentifier
TB	TransmissionBudget
TCP	Transport Control Protocol
TIM	TrafficIndication Map
TS	Traffic Stream
TSPEC	TrafficSpecifications
TXOP	Transmission Opportunity
UDP	User Datagram Protocol
UL	Uplink
UP	User Priority
US – APSD	Un – Scheduled Automatic Power Save Delivery
US – PSMP	Un – Scheduled Power Save Multi – Poll
UTT	Uplink Transmission Time
VBR	Variable Bit Rate
VoIP	Voice over Internet Protocol
VoWLAN	Voice over Internet Protocol over Wireless Local Area Networks
WCETT	Weighted Cumulative Expected Transmission Time
WCETT – LB	Weighted Cumulative Expected Transmission Time with Load Balancing
Wi – Fi	Wireless – Fidelity
WiMax	World Interoperability for Microwave Access
WLAN	Wireless Local Area Network
WMNs	Wireless Mesh Networks
WWAN	Wireless Wide Area Network

Contents

List of Figures

Chapter 1

Introduction

1.1 Overview

Wireless multimedia transmission across Wireless Local Area Networks (WLANs) has been gaining attention in recent years because of the proliferation and advances in the integration of wireless technologies like Bluetooth, IEEE 802.11, 3G and IEEE 802.16 (WiMAX). Driven by huge demands for portable access, the WLAN market, in particular, is taking off quickly. Due to its convenience, mobility support, and high-speed access, WLAN represents an important future trend for "last-mile" Internet access and has emerged as a prevailing technology for (indoor) broadband wireless communications. In November 2005, the IEEE 802.11e amendment of the IEEE 802.11 standard was ratified. IEEE 802.11e provides adequate Quality-of-Service (QoS) mechanisms and handset power saving capabilities that can be used to provide different levels of QoS for multimedia applications such as video and voice. These features are described under the two general frameworks for: contention based, namely Enhanced Distributed Channel Access (EDCA) and controlled based access, namely

Hybrid Coordination Function Controlled Channel Access (HCCA). However, it was shown in the literature that the simple scheduler defined in IEEE 802.11e is inefficient for multimedia traffic streams with bursty instantaneous bandwidth requirements [ANT06, FGTB04, GMN03].

While still under development, the new IEEE 802.11s amendment for wireless mesh networks extends the IEEE 802.11 standard by defining an architecture that supports both broadcast/multicast and unicast delivery using radio-aware metrics over self-configuring multi-hop mesh topologies. Wireless mesh networks (WMNs) are now being extensively used as a cost-effective means for coverage extension and backhaul relaying between IEEE 802.11 access points (APs). In WLAN mesh networks, wired power connections may not always be readily available, especially in installations which cover expansive outdoor areas. In such cases, fixed power connections can often be replaced by a solar powered design. In the solar powered case, since AP power saving is not permitted in conventional IEEE 802.11, solar panel/battery cost is often much higher than would be possible otherwise. Unfortunately, IEEE 802.11 APs are required to be continuously active, and therefore there is no reasonable possibility for AP power saving. Additionally, when considering the issue of QoS provisioning for real-time multimedia applications, infrastructure power saving is extremely challenging due to the excess delay and jitter experienced by multimedia content served by a power-saving AP that schedules unavailability periods to extract power-conservation. On the other hand, client station power saving is strongly supported in the standard to provide standby and call times similar to those of cellular systems, e.g., Power-Save Mode (PSM) in IEEE 802.11b [std99], Automatic Power Save Delivery (APSD) in IEEE 802.11e [std05a] and Power-Save Multi-Poll (PSMP)

in IEEE 802.11n [std06a].

In order to support QoS provisioning for multimedia applications across heterogeneous wireless mesh networks that span WLANs and cellular systems, specific problems such as the very low capacity of VoIP over WLAN networks should be overcome. The typical low VoWLAN capacity is mainly due to the short audio payload used, combined with the large per packet overheads associated with various protocol layers. Many performance studies have shown that given the choice of small frame sizes in current IP phones (typical framing values are 10-30 ms), the maximum number of simultaneous VoIP calls that can be placed in a single IEEE 802.11 AP are far below theoretical capacities [GK02, GK03, AEF+03]. It was also shown that the low VoWLAN capacity is highly correlated with high handset power consumption.

1.2 Motivation

Wireless mesh networks are now being extensively used as a cost-effective means for coverage extension and backhaul relaying between IEEE 802.11 access points (APs). Recently, there has been an increasing proliferation of applications that require rapid and low-cost solar-powered wireless mesh network deployments. Examples include but are not limited to: emergency relief, mobile/military, temporary deployment and outdoor metro-WiFi. It has been also shown that reducing mesh node power consumption can lead to significant cost savings in terms of solar-panel/battery size [FT07], and that for this reason, power saving at the AP would be highly desirable. Additionally, WMNs are usually provisioned for worse-case traffic conditions, which are always far heavier than the time averaged traffic load values. Therefore, a huge potential for infrastructure power saving is inherent in the operation of these types

of networks. Since the existing set of IEEE 802.11 standards requires that APs remain active at all times, this restriction is a major impediment to the development of practical power saving infrastructure. Furthermore, when considering the issues of network capacity, bandwidth allocation and QoS provisioning for real-time multimedia applications in contention-based distributed WLANs, the problem is extremely challenging due to the time-varying bandwidth requirements for multimedia content, contention constraint, the packet-based network, and most importantly, intense competition for the channel. Additionally, delay-aware QoS schedulers are needed to guarantee delay requirements for voice/video streams served by a power saving AP that operates at a less than 100% duty-cycle.

Using the mandatory contention-based channel access mode of IEEE 802.11, the focus of this thesis is to devise a framework for a power saving QoS-enabled AP (PSQAP) intended for use in low power infrastructure applications. The objective of this framework is to allow a PSQAP to power save while maintaining user QoS requirements and without significantly affecting station power saving. In addition to cost savings gains possible through reducing mesh node power consumption, this work allows for a thorough understanding of the various effects of having a PSQAP alternate between sleep and awake states on the delay/loss performance of real-time applications and handset power consumption. Findings from this study are used to devise novel connection admission control schemes and scheduling algorithms for optimizing the power performance at both the AP and handsets. The focus of this study is restricted to VoIP and multiplexed videoconference traffic as two of the most predominant network applications with real-time requirements. The proposed framework pays a great deal of consideration to the low capacity and high handset

power consumption problems of VoIP over WLAN. This stems from the increasing demand for cost-efficient VoWLAN deployments.

1.3 Scope

A framework for a power saving quality-of-service (QoS) enabled access point (PSQAP) is introduced. The proposed scheme introduces infrastructure based power saving, not supported by the IEEE 802.11 family of standards/amendments, while preserving the QoS requirements for delay and loss intolerant real-time applications. Therefore, the work in this thesis assumes that it is possible to modify conventional IEEE 802.11 in order to allow for infrastructure power conservation. The proposed framework uses the mandatory contention-based channel access mode, rather than polling-based channel access mechanisms to accommodate the traffic backlog at mobile stations. This facilitates the application of the framework to all existing WLAN networks. In this type of setup, stations conserve energy by spending most of their time in a low-power state (doze) mode and only switching to active mode whenever they have packets to send or receive. From a handset point of view, backward compatibility is achieved through the use of network allocation maps (NAMs) which are a generalization of NAV setting defined in the legacy IEEE 802.11.

A scheme for dynamic activation and de-activation of scheduled service intervals is introduced to allow a PSQAP to maximize its power saving while dynamically adjusting the offered capacity according to time-varying bandwidth demands of the real-time multimedia content. The problem of QoS provisioning is considered only from the uplink traffic point of view since packets may experience violation of their QoS requirements when served by a power-saving AP which operates at a shorter

duty-cycle to conserve energy.

A dynamic connection admission control algorithm (CAC) is proposed for efficient resource management and bandwidth allocation for real-time multimedia streams. Admission decisions for video streams are based on an accurate prediction of the required capacity using a discrete auto-regressive model. The proposed CAC also prevents best-effort traffic starvation by reserving a minimum bandwidth quota for low-priority traffic flows. The correctness and viability of the proposed framework, including the energy-efficient MAC and dynamic CAC schemes, are verified using extensive simulation experiments.

Several scheduling schemes are proposed to solve the problem of increased queuing delay and high contention level experienced by handsets served by a power saving AP. A mobile station served by a power saving PSQAP is likely to consume more power than would be possible if served by a conventional, i.e., always awake AP. This is due to the fact that handsets which generate packets during an unavailability (i.e., sleep) period would have to awaken and contend for channel access once the PSQAP switches to the active state and the channel becomes available. It is shown that by combining careful scheduling with an energy-efficient MAC protocol, a near optimal power performance can be achieved for both the handsets and the PSQAP.

Another solution is also proposed to overcome the problem of low VoIP capacity in WLAN deployments. The solution is based on an enterprise-based adaptive voice packetization server (AVP-RTS) that splits the RTP VoIP connection into two legs. In this way, voice coding parameters at each end of the call are negotiated separately and the server can allocate the available latency margin (and the ensuing capacity gain) asymmetrically across the call. The efficiency of the proposed AVP scheme is

verified using simulations. It is shown that the improved VoIP over WLAN capacity is associated with a similar improvement in AP and mobile stations' power performance. The call blocking probability of the adaptive voice packetization scheme was analyzed using a Markov model. This model also confirmed the efficiency of using the AVP-RTS server in minimizing new call blocking probabilities for a typical VoWLAN installation.

1.4 Organization of Thesis

In Chapter 2 we provide an overview of the concepts relevant to low-power operation and Quality-of-Service (QoS) provisioning for QoS-enabled power saving access points in IEEE 802.11-compliant deployments. First, the chapter introduces wireless local area networks (WLANs), followed by a description of the building blocks, the different operational modes and power management mechanisms of WLANs. The chapter also reviews the QoS provisioning capabilities in the IEEE 802.11e standard and the relevant literature. Then, a description of wireless mesh networks (WMNs) and infrastructure power saving in solar-powered WMNs is provided with a comprehensive literature review. Finally, the chapter provides an overview of Voice over IP technology and discusses the problems of low capacity and poor power performance of VoIP over WLAN.

Chapter 3 describes an energy-efficient framework for a power saving quality-of-service (QoS) enabled access point (PSQAP). First, the frame structure and signaling for a low-energy MAC protocol are introduced. An adaptive connection admission control scheme, which uses the TSPEC information of admitted flows to calculate the service interval length of a PSQAP, is also described. The chapter provides various

performance results for the framework obtained from a simulation-based study.

In Chapter 4, the technical challenges associated with the transmission of video applications over solar/battery powered wireless mesh networks are discussed. The idea of using a discrete autoregressive model to make admission decisions for multiplexed videoconferencing applications in WLANs is introduced. The chapter also identifies a dilemma resulting from the conflicting PSQAP and handset power saving requirements. New bandwidth allocation and scheduling schemes are proposed and evaluated for the contention-based US-APSD and PSMP power management mechanisms. Finally, the chapter presents a performance evaluation using analytical models and simulation experiments for the various algorithms considered.

Chapter 5 is devoted to discussing the problems of low capacity and poor power performance of VoIP over WLAN. The chapter presents the use of an adaptive voice packetization server (AVP-RTS) which splits the RTP VoIP connection into two separate call legs. The chapter also provides algorithms for performing a sub-optimal assignment of the packetization intervals, followed by an analytic characterization of the system performance. The performance analysis section describes the simulation environment and presents various performance results.

Chapter 6 presents conclusions drawn from this work. Areas for future work in QoS provisioning in power saving infrastructure WMNs research are outlined.

Chapter 2

Background

2.1 Overview

This chapter provides an overview of the concepts relevant to low-power operation and Quality-of-Service (QoS) provisioning for QoS-enabled power saving access points in IEEE 802.11-compliant deployments. First, wireless local area networks (WLANs) are introduced. The building blocks of WLANs are described and the different operational modes and power management mechanisms of WLANs are explained. QoS provisioning capabilities in the IEEE 802.11e standard and the relevant literature are reviewed. Then, a description of wireless mesh networks (WMNs) and infrastructure power saving in solar-powered WMNs is provided with a comprehensive literature review. Finally, Voice over IP technology and the problems of low capacity and poor power performance of VoIP over WLAN are described.

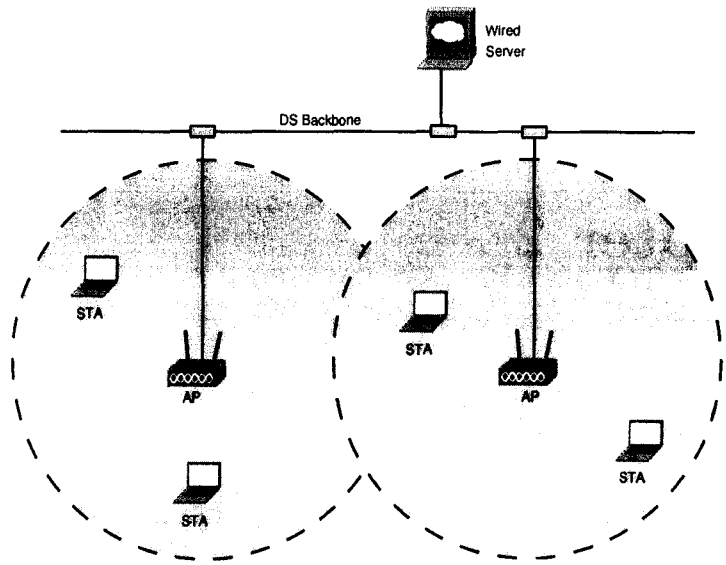


Figure 2.1: *WLAN with Two Basic Service Sets*

2.2 Wireless Local Area Networks

Wireless Local Area Networks (WLANs) have emerged as a prevailing technology for indoor broadband wireless access during the past decade. WLAN products have become commodity items used in professional and consumer products alike. Recently, the proliferation of WLANs as extensions of wired networks has been increasing dramatically, thereby giving stations equipped with wireless interfaces a higher degree of mobility. The two most common WLAN standards are the North American IEEE 802.11 (commonly branded as Wi-Fi) and the European HIPER (HIgh PErformance Radio) LAN. In the remainder of this section, IEEE 802.11 terminology is used to illustrate relevant concepts.

IEEE 802.11 defines two types of configurations, infrastructure Basic Service Set (iBSS) and Independent BSS (IBSS). Figure 2.1 shows the basic components of an

infrastructure-based IEEE 802.11 wireless network. The AP is the central entity of each coverage area with coordination functionality. Additionally, the AP acts as a bi-directional bridge between the wireless medium and the wired infrastructure (i.e., typically Ethernet). Stations (STA) are mobile devices equipped with IEEE 802.11 wireless network interfaces.

The first WLAN standard was ratified in 1997 by Working Group (WG) 11 of the IEEE LAN/MAN Standards Committee and is known as the IEEE 802.11 standard. The IEEE standards define the physical (PHY) and medium access control (MAC) layers specifications for short range, high data rate wireless communications [Gas05]. The PHY layer can use either the 2.4GHz unlicensed Industrial Scientific and Medical (ISM) radio frequency (RF) band or infra-red (IR) for communication. The 1997 IEEE 802.11 standard provided a basic data rate of 1Mbps and a maximum bitrate of 2Mbps.

Communication between the AP and the associated stations occurs over the shared wireless medium that carries the data. A station must associate with an AP in order for it to transmit and receive data from the wired infrastructure and to communicate with other stations in the same WLAN. A Basic Service Set (BSS) is the term used to refer to an AP and its associated stations.

In large WLANs, multiple BSSs can be joined together using a distribution system (DS), thereby providing sufficient coverage for a larger number of handsets. This setup of having two or more BSSs is referred to as an Extended Service Set (ESS). The DS is the wired backbone connecting APs in one network and allowing the associated stations to access services available on the wired infrastructure. Figure 2.1 shows an ESS consisting of two BSSs. The connecting DS in Figure 2.1 allows for

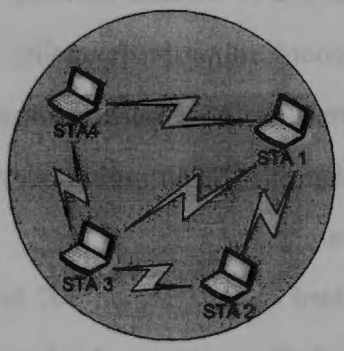


Figure 2.2: *WLAN Operating in Ad Hoc Mode*

inter-AP communication where both APs can bridge packets between any of their associated stations and the server connected to the wired infrastructure or any two stations residing in different BSSs. As mobile stations (MSs) cross IEEE 802.11 BSS boundaries in an ESS, horizontal handoff events are triggered.

Figure 2.2 shows the configuration of a less common mode for WLAN operation, independent BSS, in which four MSs form an ad hoc network. In this mode, stations discover and directly communicate with each other in a peer-to-peer fashion without involving a central AP, provided that communicating nodes are within direct radio range. To form an ad hoc network, the four MSs shown in Figure 2.2 must have their wireless adapters configured for ad hoc mode and must use the same Service Set Identifier (SSID) and the same channel number. An iBSS can be used in situations where a temporary deployment is desired without the need to access information that is external to the network, a function that is essential to most WLANs. Performance suffers as the number of devices grows, and a large ad hoc network quickly becomes difficult to manage. Ad hoc networks cannot bridge to wired LANs or to the Internet without installing a special-purpose gateway.

Table 2.1: Parameter Values of IEEE 802.11a/b/g

	IEEE 802.11a	**IEEE** 802.11g	**IEEE** 802.11b
DIFS	$34\mu s$	$28\mu s$	$50\mu s$
SIFS	$16\mu s$	$10\mu s$	$10\mu s$
Slot Time	$9\mu s$	$9\mu s$	$20\mu s$
CWmin	16	16	32
Supported Data Rates	6, 9, 12, 18, 24, 36, 48, 54 Mbps	6, 9, 12, 18, 24, 36, 48, 54 Mbps	1, 2, 5.5, 11 Mbps
Basic Rate	N/A	N/A	2 Mbps
PHY header	$20\mu s$	$20\mu s$	$192\mu s$
ACK frame	$24\mu s$	$24\mu s$	$248\mu s$

IEEE 802.11 was amended by the IEEE 802.11b standard released in 1999, which operates in the 2.4GHz ISM and introduced higher bit rates of 5.5 and 11Mbps. IEEE 802.11b operates in the same frequency band as many other short-range wireless devices such as microwave ovens, Bluetooth and cordless phones, and can experience interference from these devices. In addition, the 2.4GHz ISM band only allows for three non-overlapping frequency channels, which limits the capacity of an IEEE 802.11b-based WLAN deployment. As a result, the IEEE 802.11a standard was also ratified in 1999 which operates in the 5GHz unlicensed ISM frequency band. An IEEE 802.11a compliant device supports data rates up to 54Mbps, additionally, the 5GHz ISM band allows for 12 non-overlapping frequency channels. These two features, in addition to a shorter PHY preamble and contention time slot, allow for data rates that are 5 times higher than those of IEEE 802.11b WLAN deployments. However, to overcome the incompatibility issue of IEEE 802.11a with the widely deployed IEEE 802.11b Wi-Fi cards, the IEEE 802.11g standard was ratified in June 2003 which operates in the 2.4GHz band. IEEE 802.11g has two different operation modes. In the IEEE 802.11g only mode, all stations in the WLAN are IEEE 802.11g stations,

Table 2.2: Range and Rate Data for IEEE WLANs

Network	Rate	Indoor Range	Outdoor Range
Legacy	2Mbps	35 meters	70 meters
IEEE 802.11a/g	54Mbps	35 meters	150 meters
IEEE 802.11n	250Mbps	100 meters	200 meters

so that they can operate in a way that is more efficient but not compatible with IEEE 802.11b. In the IEEE 802.11b-compatible mode of the IEEE 802.11g standard, some stations in the WLAN are IEEE 802.11b stations, which must operate in a way that is compatible with IEEE 802.11b. The IEEE 802.11b-compatible mode supports IEEE 802.11b bitrates of 1, 2, 5.5 and 11 Mbps in addition to higher bitrates of up to 54Mbps. In the IEEE 802.11g-only mode, timing spaces are even smaller than those used in IEEE 802.11a (Table 2.1), leading to a slightly higher throughput than IEEE 802.11a. However, the use of IEEE 802.11g-only mode in practice is unlikely given the large installed base of IEEE 802.11b equipment already in use. After all, the main motivation for the use of IEEE 802.11g over IEEE 802.11a is that IEEE 802.11g is compatible with IEEE 802.11b while IEEE 802.11a is not. One would expect IEEE 802.11g stations to mostly operate in the IEEE 802.11b-compatible mode in the field.

WLANs have a much shorter range when compared to cellular networks, however they have data rates which are orders of magnitude greater. For example, IEEE 802.11 had a data rate of 2Mbps, an indoor range of ~35 meters and an outdoor range of ~70 meters. Both the IEEE 802.11a and IEEE 802.11g standards have data rates of 54Mbps and indoor range of ~35 meters and outdoor range of ~150 meters. The recent drafts of the new IEEE 802.11n standard define modifications to both PHY and MAC layers that would allow for more than 100Mbps of MAC layer throughput over an indoor range of ~100 meters and an outdoor range of ~200

meters. Raw data rates of up to 600Mbps are also supported. A key issue of the design of the IEEE 802.11n standard is its backward compatibility with existing IEEE 802.11 specifications where a 2.4GHz IEEE 802.11n device must support all mandatory IEEE 802.11g rates and a 5GHz IEEE 802.11n device must support all mandatory IEEE 802.11a rates. This is achieved through the use of a mixed-mode preamble which allows legacy IEEE 802.11a/g devices to detect packet transmissions using the legacy short PHY preamble while non IEEE 802.11a/g compatible receivers supporting IEEE 802.11n higher data rates can use long preambles. Additionally, the IEEE 802.11n standard introduces implicit and explicit beamforming to provide for extra performance gain. Other interesting features of IEEE 802.11n are enhanced MPDU aggregation (new maximum MPDU size of up to 65,535 octets compared to 2,304 octets for existing IEEE 802.11 counterparts), reduced interframe spacing by allowing packets in a burst to be separated by a $2\mu s$ compared to SIFS ($10\mu s$) and improved link adaptation where a receiver provides feedback to the sender based on received channel estimates and received signal-to-noise ratio. In Table 2.2 we summarize the data rates of the different IEEE 802.11 standards. In the remainder of this thesis, the terms WLAN, IEEE 802.11 and Wi-Fi are used interchangeably.

2.2.1 IEEE 802.11 Channel Access Mechanisms

In IEEE 802.11, the Medium Access Control (MAC) layer is responsible for arbitrating access of IEEE 802.11 compliant wireless devices to the shared wireless medium. The MAC layer for all IEEE 802.11a/b/g standards is the same. Two access modes are defined in the IEEE 802.11 standard, a centralized contention-free polling based Point Coordination Function (PCF) and a distributed contention based Distributed

Coordination Function (DCF). The DCF channel access mechanism is mandatory for all IEEE 802.11 compliant wireless cards while PCF is optional.

DCF DCF is the fundamental access method of IEEE 802.11 and is used in both ad hoc and infrastructure modes of operation. It enables fast installation with minimal management and maintenance costs and is a very robust protocol for best-effort service. The CSMA/CA mechanism is designed to reduce the packet collision probability when multiple stations try to access the same channel [std97]. In CSMA/CA, each station must sense the channel before it attempts transmission to ensure it is not already captured by another station. If the channel is idle for a DCF Interframe Spacing (DIFS) duration, the station immediately starts transmission. Otherwise, the station must perform a binary exponential random backoff in which it chooses a random backoff time interval that specifies the period a STA has to sense the channel idle (beyond a DIFS duration) before it can transmit its frame. The backoff time is defined as

$$Backoff\ Time = Backoff\ Slots \times aSlotTime, \tag{2.1}$$

where *Backoff Slots* is the number of backoff slots and is chosen randomly from [0,CW_{min}] and *aSlotTime* is the slot time duration whose value is a function of the underlying PHY layer specifications. CW_{min} is the initial contention window size, and it equals $2^n - 1$, where n is usually set to 5 at the first transmission attempt. A backoff counter keeps track of the backoff duration. During the backoff procedure, the counter is decremented while the channel is idle. However, the counter is frozen if the channel becomes busy again. Transmissions can start only after the counter reaches zero. As intuitively seen,

the CSMA/CA protocol reduces the probability of packet collisions by uniformly distributing the contending STAs over the timeline.

The IEEE 802.11 MAC requires the receiver to send a mandatory positive acknowledgement (ACK) back to the transmitter for every successfully received packet. ACKs are transmitted after a Short Interframe Spacing (SIFS) duration, which is shorter than DIFS to enable ACK transmissions before any new packet transmission is initiated. If no ACK is received within an SIFS duration, the packet is assumed corrupted and the sender attempts retransmission after the channel is sensed idle for a DIFS duration plus a new backoff duration that is randomly chosen from a contention window $[0,CW_{i+1}]$, where $CW_{i+1} = 2 \times CW_i$ and i is the number of unsuccessful packet retransmissions. In IEEE 802.11, there is no way to distinguish between packet collisions and transmission errors due to low channel quality, and colliding stations also double their contention window size. Once a data frame is successfully transmitted, the contention window size is reset back to CW_{min}. Fairness in DCF is achieved by preventing burst transmissions. In other words, a client with multiple packets is forced to undergo the backoff procedure for every frame transmission. While it increases packet delay, mandatory backoff reduces delay jitter and ensures that clients cannot dominate network usage.

The DCF also defines a more complex transmission mechanism, RTS/CTS which uses Virtual Carrier Sensing so that DCF can be considered as a form of reservation MAC protocol. A STA that wants to send a packet, first transmits a Request To Send (RTS) frame which indicates the total anticipated transmission time for the Head-Of-Line (HOL) data frame (including ACK). This

information is indicated by the duration field of an RTS frame. The receiving node then replies with a Clear To Send (CTS) frame and acknowledges its availability for the imminent packet transmission. All other clients that successfully decode the duration field in RTS and CTS packets store medium reservation information in their Network Allocation Vectors (NAV). For these clients, NAV is used in conjunction with physical carrier sensing to detect the medium availability. These clients will refrain from accessing the channel if NAV is non-zero or if carrier sensing indicates a busy medium. While RTS/CTS protection could not be justified for short packets because of the large overhead involved, the ability to reserve transmission time is important for long packet transmissions and packets with real-time multimedia content.

PCF In addition to DCF, the IEEE 802.11 standard also defines the Point Coordination Function (PCF), for infrastructure configurations. PCF provides a contention-free access period during which channel access is completely controlled by the Point Coordinator (PC) residing at the AP. At the beginning of the Contention-Free Period (CFP), NAVs are set to block all MSs from accessing the channel during a CFP. The IEEE 802.11 MAC allows for both DCF and PCF to coexist within the same BSS, where time is divided into repetitive periods referred to as superframes. The superframe duration should be an integer multiple of the beacon interval. CFP can start only at the beginning of a superframe when all stations (including the ones in power saving mode) wake up to listen for broadcast and multicast packets. As indicated earlier, CFPs are optional, thus in each superframe, a mandatory Contention Period (CP) could be preceded by an optional CFP.

During a CFP, the AP polls each MS in a prespecified order to exchange data frames. Only MSs which are CF-Pollable can be polled by the AP, while other non-CF-Pollable MSs remain silent during a CFP. The AP sends a special CF-End packet to indicate the end of a CFP. Below, a description of new Quality-of-Service features of IEEE 802.11e is provided.

2.3 Quality-of-Service Support in IEEE 802.11e

The widely used DCF access mechanism does not provide any Quality-of-Service (QoS) capabilities since all packets, regardless of higher layer application requirements, are transmitted using the same set of MAC contention parameters (i.e., DIFS, CW_{min} and CW_{max}). Under the DCF, STA might have to wait an arbitrarily long time to send a frame, so real-time applications such as voice and video may suffer [DC99]. To address this issue, the IEEE 802.11 QoS Working Group introduced the IEEE 802.11e amendment in November 2005. This standard defines a new enhanced coordination function: the Hybrid Coordination Function (HCF). HCF defines two new channel access mechanisms, namely, a centralized polling-based HCF Controlled Channel Access (HCCA) and a distributed contention-based Enhanced Distributed Channel Access (EDCA).

EDCA Unlike DCF, EDCA provides distributed and differentiated access to the wireless medium by defining eight user priorities which are further mapped into four access categories (ACs). ACs provide support for the delivery of traffic at QoS stations (QSTAs). Each AC_i ($i = 0, 1, 2, 3$), is in fact an enhanced variant of legacy DCF, which uses CSMA/CA with random backoff to access the

wireless medium. In EDCA, different ACs use different access parameters (e.g., minimum/maximum contention window size 'CWmin[i]/CWmax[i]', arbitrary interframe space 'AIFS[i]') to acquire prioritized access to the wireless medium. This is in contrast to the legacy DCF which only provides egalitarian access to the wireless medium.

A new concept, transmission opportunity (TXOP), is introduced in IEEE 802.11e. A TXOP is the maximum time period when a station has the right to initiate multiple frame-exchange sequences, separated by a short inter-frame space 'SIFS', to transmit traffic within the same AC. A TXOP is defined by a starting time and a maximum duration. A station cannot transmit a frame that extends beyond a TXOP. If a frame is too large to be transmitted in a TXOP, it should be split into smaller fragments.

HCCA The polling mechanism HCCA is similar to the legacy PCF. The HCCA uses a QoS-aware HC, which is typically located at the AP in infrastructure WLANs as a polling master to allocate TXOPs to itself and other QSTAs. Because of its polling-based mechanism, HCCA provides parameterized QoS guarantees for real-time traffic. The HC is allowed to start contention-free bursts at any time during the contention period after the medium remains idle for at least a Point Interframe Spacing (PIFS) interval, where PIFS < DIFS. In HCCA, the QoS AP (QAP) needs to compute the polling order and the amount of channel time (i.e., TXOPs) granted to a station for each poll (together called a "service schedule"), and poll each QSTA to initiate frame-exchange sequences. After acquiring the channel, the HC polls QoS-Pollable QSTAs. In order to be placed on the HC's polling list, a QSTA must issue a QoS reservation by means of

special QoS management action frames. A separate reservation must be made for each traffic stream which is described by a traffic specifications element (TSPEC). QSTAs may send TXOP requests during polled TXOPs or EDCA TXOPs. Upon receiving a QoS-Poll frame, the polled QSTA either responds with a QoS-Null frame, if it has no packet to send; otherwise, it responds with a QoS-Data+QoS-ACK packet. Similar to the EDCA case, the polled QSTA, during a polled TXOP, may initiate multiple frame exchange sequences. This gives the HCF the flexibility to support bursty QoS traffic. At the end of a TXOP, the HC gains control of the channel again, and it either sends a QoS-Poll to the next QoS station on its polling list, or releases the channel if there are no more stations to be polled.

EDCA vs. HCCA Although HCCA is recommended for parameterized QoS guarantees in IEEE 802.11 WLANs primarily because of its efficiency in providing deterministic QoS for multimedia applications, the polling-based HCCA does not allow for maximum power conservation at QSTAs and is less flexible due to several limitations, for example (i) energy dissipation due to unnecessary signaling which occurs when the scheduler at the QAP recomputes the service schedule every time a TS is added to or deleted from the iBSS, or when the physical transmission rate changes due to link rate adaptation, (ii) when two wireless LANs using HCCA operate on the same channel, it requires additional coordination between them [CSS05], and (iii) like DCF and PCF, EDCA functionality is mandatory for licensing IEEE 802.11e compliant WLAN cards while HCCA is optional [Eni]. Moreover, from a power consumption point of view, periodic polls in HCCA were found to be both bandwidth and energy inefficient

when used to serve uplink streams with variable packet inter-arrival times such as video and voice codecs with silence suppression [PCCM06]. However, the lack of a quantitative and efficient control of stations' medium occupancy (i.e, airtime usage) places a huge challenge for providing parameterized QoS using EDCA.

Based on the discussion above, we use the IEEE 802.11e contention-based EDCA for the energy efficient MAC and CAC framework, proposed in this thesis for QoS-enabled power saving APs.

2.3.1 Overview of the TSPEC Element

In order to initiate a traffic stream (TS) connection, a QSTA needs to send an ADDTS (add traffic stream) QoS request frame containing a traffic specification (TSPEC) element to the QAP. A TSPEC describes the QoS requirements of a TS. The major TSPEC parameters [std05a, CSS05], include: mean ρ and maximum P data rate (in bits per second), delay bound D (in milliseconds), maximum service interval $SImax$ (in microseconds), nominal MSDU size L (in octets), minimum PHY rate R (in bits per second), maximum burst size σ (in octets) and surplus bandwidth allowance (or surplus factor) SF. Among the different TSPEC fields, L and ρ are used to capture application requirements while medium requirements are captured by SF and R. SF accounts for packet error probability and packet drop rate and it refers to the minimum allocation of excess time by the scheduler such that packet drop rate is bounded. More details of the definitions of the different TSPEC fields can be found in the IEEE 802.11e standard [std05a]. The connection admission control unit at the QAP uses the set of TSPEC parameters in order to decide whether or not to accept

new TSs. Next, we present a literature review of various admission control schemes and resource allocation algorithms for both EDCA and HCCA modes.

2.3.2 QoS Provisioning in IEEE 802.11e

With the advent of QoS-support in IEEE 802.11e, wireless multimedia transmission across WLANs has been gaining increasing attention in recent years. The IEEE 802.11e standard sketched a simple reference scheduler which assigns transmission times with fixed length. The transmission time for each station, called transmission opportunity (TXOP), is estimated from various fields of the TSPEC. Unlike the inefficient reference scheduler which uses the minimum negotiated physical rate R to calculate TXOPs, the PRBAC scheme proposed in [GCZ05] considers physical rate variances due to station mobility and wireless channel characteristics. The basic idea is to use the long-term average physical rates for admission control and at the same time use the instantaneous physical rates to allocate TXOPs to individual stations.

Because of the fixed TXOP duration the reference scheduler is inefficient for services with variable bit rate (VBR) traffic. Also it can not account for physical rate adaptation to the radio channel variations. The subject of QoS support in IEEE 802.11e and the resource allocation issues arising from the unique real-time requirements of voice and video applications have attracted the attention of many researchers. Initial contributions were mainly concerned with the feasibility of the EDCA and HCCA mechanisms of the Hybrid Coordination Function (HCF) for multimedia transmission. Since it provides deterministic QoS guarantees over EDCA, HCCA was often used for channel access [vdSAH06]. Ansel *et al.* [ANT06] presented an approach for efficient scheduling on the QAP, based on measured queue sizes of

each traffic stream. The paper shows that the fixed transmission opportunity (TXOP) allocation, as in the simple scheduler of IEEE 802.11e standard, is not efficient for video transmission as it does not consider the bursty characteristics of video traffic; for this reason they propose improved scheduling schemes to alleviate this problem. In [FGTB04], a new variable (effective TXOP) is introduced and defined as the necessary TXOP which can statistically guarantee that the packet loss probability is less than a threshold. Given a desired packet loss probability, the effective TXOP duration of a newly arrived VBR flow is derived, then the same procedure as in the reference scheme of IEEE 802.11e is applied for admission control. In [GMN03], the mean TXOP is used as a guideline for allocating channel access time and a token bucket algorithm is used to allow nodes to vary their TXOPs over time according to their changing bit rate requirements. Stations' polling order is determined based on an earliest deadline first scheduling. Unlike these schemes, the work in [RPD07] uses a fixed TXOP value but offers multiple polls per stream per service interval. Polling decisions are made by the QAP and are based on queue length information received from the nodes and the timestamp of when the node was last given extra time in the HCCA. Additional polling is limited to the amount of time remaining in the HCCA period.

The contention-based EDCA of IEEE 802.11e has also attracted some attention due to the challenging additional contention constraint, compared to HCCA. Admission control schemes for EDCA can be mainly classified into measurement-based and analytical based. Measurement-based admission control algorithms are usually effective and simple to implement in practice. However, without a theoretical foundation, it is very difficult to achieve overall optimization. The work in [XL04] evaluated

the Distributed Admission Control (DAC) which was proposed by the IEEE 802.11e working group but was not supported in the final standard. In DAC, a QAP announces via beacons the transmission budget, which is the additional amount of time available for each AC during the next beacon interval. The transmission budget of an AC is calculated as the difference between the transmission limit of this AC and the measured amount of time occupied by the transmission of each AC during each beacon interval. The authors concluded that it is difficult to avoid performance vibration because a station always adjusts its transmission parameters at every beacon interval. Additionally, DAC can only protect existing flows when the traffic load is not very heavy; and finally, DAC lacks good mapping between TXOP parameters and the QoS requirements from applications.

A two-level protection and guarantee scheme was proposed in [XLC04] based on DAC. In the first level, tried-and-known, a new voice/video flow is first accepted tentatively, and then the throughput and delay performance are measured for some beacon intervals. If the average throughput and/or delay do not meet reasonable requirements, the flow will kill or reject itself. This level protects each existing voice or video flow from new and other existing QoS flows. The second level, early protection, protects the existing QoS flows from best effort traffic by dynamically adjusting the EDCA parameters to increase the initial contention window size and interframe space for low-priority best-effort traffic when the number of active stations is large. In this way, the number of collisions can be kept relatively small. Although more efficient than DAC, the scheme involves more adjustable parameters, and in addition, finding the optimized parameters is not a trivial task. Reference [GZ03] introduces a threshold-based admission control. In

this scheme, each station needs to measure the traffic condition on the wireless link. Depending on how the traffic condition is measured and computed, the admission control can be implemented in two ways, either based on the relative occupied bandwidth $= T_{busy}/T$, where T is a fixed sampling period, or using the *average collision ratio* $=$ *number of collisions/total number of transmissions.* For both parameters, the scheme sets lower and upper thresholds. The admission criterion is based on the selected thresholds. If the measured parameter is less than the lower threshold, then the scheme admits the inactive AC with the highest priority during the next period of T seconds. If the measured parameter is larger than the lower threshold and less than the upper threshold then it takes no action. Otherwise, the scheme stops the transmission of the lowest active AC during the next period of T seconds.

Analytical-based admission control schemes are based on analytical models, and it is possible to optimize the entire system. However, the analytical models are often based on unrealistic assumptions such as error-free physical channels. For example, Pong *et al.* [PM03] proposed a Markov chain model-based admission control, in which admission control is performed based on the predicted achievable throughput for each flow which is calculated as a function of the probability of a successful transmission, overall collision probability and overall idle probability. This model is based on the two-state Markov chain model proposed by Bianchi [Bia98]. Thus, the model assumes saturation conditions, where each station always has packets to transmit, an assumption that is not always true especially with multimedia traffic. The work in [BPCQ03] adjusts the contention window (CW) values for different stations so that the goals of admission control can be fulfilled. Every time a new station wants to join the network

with a throughput requirement of R_{n+1}, a new CW set will first be calculated for the existing n stations and the new station. This set of new CW values will be used to compute the throughput. If the resulting throughput meets the requirements, station $n+1$ is accepted and the new CW set is distributed to all the stations. Otherwise, station $n+1$ is rejected. Similar to the Markov chain based scheme above, this model considers only saturation conditions.

A different set of resource allocation and capacity scheduling schemes used the concept of effective (or guaranteed) bandwidth for VBR multimedia traffic. The effective bandwidth is defined as the channel resources required by a traffic stream in order to keep a given QoS performance metric below a certain threshold. Usually packet loss rate is used as the scheduling criterion. Verbiest *et al.* [VPV88] used statistical parameters, i.e., average bit rate, the bit rate variance and the peak bit rate, to calculate the aggregate bandwidth demands for a given number of multiplexed VBR multimedia streams given a packet loss performance. Their analysis is based on the well-known fact that if a sufficient number of sources is multiplexed, the resulting distribution will approach a normal distribution with an average equal to the sum of the individual averages and a variance equal to the sum of the individual variances. The authors in [vdSAH06] proposed methods to improve the number of admitted stations by creating multiple subflows from one global video flow, each with its own traffic specification. To simplify the design of the scheduler, their scheme shapes the traffic that is arriving at the MAC buffer through a twin leaky bucket scheme that removes the burstiness of the video traffic. Based on twin-leaky bucket analysis, the TSPEC information of a video flow is used to calculate the effective bandwidth of the flow. In order to achieve an optimal scheduling policy, a low complexity linear-programming

solution is proposed, which effectively allocates the optimal transmission opportunity to each generated subflow in order to maximize the utilization of the wireless medium under the contention-free period.

2.4 Power Management in IEEE 802.11

IEEE 802.11 standards included mechanisms that can be used by end stations to reduce their power consumption. The mechanisms operate at the media access control layer and permit the MS to turn off its transceiver for considerable periods of time. When associated with an AP in infrastructure mode, a MS can signal the AP that it is entering a low-power mode. The AP will then buffer packets destined to the MS until the station awakes again and becomes ready for packet-exchange. Different power management schemes of IEEE 802.11 standards/amendments use different mechanisms that a station, which awakens from sleep mode can use to initiate packet delivery from the AP. Below is a brief illustration of the various handset power saving mechanisms supported by IEEE 802.11.

2.4.1 Power Save Mode in IEEE 802.11b

The IEEE 802.11b specifications [std99] define a power saving mode (PSM) for handset power management. The AP buffers incoming frames destined for mobile stations in PSM and periodically announces its buffering status through the traffic indication map (TIM) contained in the beacon frames. The mobile station wakes up periodically to listen to the beacon frames. In the unicast case, once the bit corresponding to its association ID (AID) is set in the TIM, the mobile station initializes a PS-Poll frame

to the AP to retrieve data and the AP responds to each poll with one buffered frame. Multiple PS-Polls are allowed until all outstanding frames have been retrieved. In the broadcast/multicast (B/M) case, the existence of buffered B/M frames is indicated by setting the B/M traffic indication bit in the delivery TIM (DTIM), which is a special TIM sent out at a fixed number of beacon intervals. All B/M frames buffered at the AP are delivered immediately after the beacon frame containing DTIM. As opposed to the normal continuous active mode (CAM), a mobile station in PSM can often have opportunities to turn its network interface off to save energy, given that it has no data pending at the AP. For light to moderate traffic load, the legacy PSM can greatly reduce the energy consumption which extends the mobile stations' battery-lifetime.

2.4.2 Automatic Power Save Delivery in IEEE 802.11e

The automatic power-save delivery (APSD) is a power management defined as part of the IEEE 802.11e [std05a] to improve the power conservation of handsets by allowing for bursty frame transmissions, unlike the one packet per PS-Poll of legacy PSM. APSD defines two delivery modes, namely the UnScheduled APSD (US-APSD) and Scheduled APSD (S-APSD). A non-AP QSTA that is in sleep mode wakes up as soon as it has a packet to transmit. It can then begin an unscheduled service period (SP) by sending a trigger frame to the QAP, which is any uplink frame (e.g., QoS-Data or QoS-Null frames) associated with a trigger-enabled access category. Unscheduled SPs are contiguous periods of time following the acknowledgement of a trigger frame, during which a QSTA is expected to remain awake. Unscheduled SPs are acquired via contention. Access categories that a QSTA has designated for use as US-APSD

triggers and those enabled to be delivered using US-APSD are called trigger-enabled and delivery-enabled, respectively. A traffic category may be delivery enabled, trigger enabled or both. The QAP responds by transmitting at least one MAC protocol data unit associated with the delivery-enabled AC and destined to the same non-AP QSTA who initiated the US-APSD. A QSTA must signal the QAP prior to the use of US-APSD for one or more ACs. Signaling can be done either during (re)association or via admission control. Backward compatibility of legacy stations that do not support US-APSD is guaranteed if at least one of the ACs is non delivery-enabled [PCCM06].

A QSTA's scheduled SP starts at fixed intervals of time specified in the service interval field of the corresponding TSPEC element. A non-AP QSTA using scheduled SP wakes up at specified intervals to receive any DL unicast frames buffered at the QAP otherwise, in the absence of any buffered data, the QAP sends poll frames to the non-AP QSTA using S-APSD. The station subsequently wakes up at a fixed time interval to receive the buffered frames from the QAP.

2.4.3 Power Save Multi-Poll in IEEE 802.11n

The recent IEEE 802.11n [std06a] amendment defines a new power save protocol called Power Save Multi-Poll (PSMP), which has both a scheduled and unscheduled parts. PSMP operation is controlled using the PSMP Action Frame. The amounts of time used for downlink and uplink transmission are defined as part of the PSMP frame. PSMP-UTT (uplink transmission time) is the time scheduled for PSMP-capable stations in the BSS to transmit frames to the AP. PSMP-DTT (downlink transmission time) is the time scheduled for the AP to transmit frames to the stations within the BSS. A PSMP Sequence is a series of frames that starts with a PSMP

Action Frame and is followed by either PSMP uplink or downlink frames. While the UTT and DTT for a PSMP-capable station are negotiated as part of the admission control in Scheduled PSMP (S-PSMP), a station that uses UnScheduled PSMP (US-PSMP) must issue a trigger frame each time it requires the AP to schedule UTT and DTT for it. In PSMP, the time-line is viewed as an AP's own TXOP during which it sends a schedule to one or more stations in the BSS letting the PSMP-capable stations know when to be awake to receive downlink data frames and when they (individually) are allowed to begin transmitting. Subsequently, PSMP capable stations do not need to perform conventional carrier sensing prior to their individual transmission. By using a schedule, stations can doze for the maximum amount of time possible while missing no frames. US-PSMP functionality does not replace US-APSD, but rather extends it to add further functionality. PSMP capable stations can still use US-APSD and PSM trigger frames when appropriate.

2.4.4 Flexible Broadcast/Multicast Service in IEEE 802.11v

The IEEE 802.11v amendment [std06b] is currently under standardization efforts. It defines layer 2 management of IEEE 802.11 devices, such as monitoring, configuring and updating, in either a centralized or distributed manner. These new procedures allow wireless infrastructure to control parameters on wireless client adapters, e.g., identifying which network or access point to connect [Bin08].

The IEEE 802.11v draft proposes a flexible broadcast/multicast service (FBMS) to provide flexible management over the broadcast/multicast (B/M) traffic in power saving protocols. FBMS provides flexible delivery intervals for group addressed frames by allowing an AP to transmit group addressed frames as individually addressed

frames to a requesting STA. In FBMS, a B/M stream can choose to be served at a configurable delivery interval, which is a multiple of DTIM periods. Thus traffics of different B/M groups can be differentiated by selecting different delivery intervals. The mobile station does not have to expend extra energy to receive irrelevant B/M traffics. In FBMS, STA first advertises Directed Multicast Service (DMS) Capability. The non-AP STA sends a request to specify traffic classes for the directed multicast traffic delivery. If the AP accepted the traffic class specified by the non-AP STA, the AP shall transmit the requested group addressed (broadcast/multicast) traffic as the individual addressed (unicast) traffic to the requesting STA, and the AP may still transmit the group addressed traffic for other STAs in the BSS as normal. Other STAs will process group addressed traffic as normal while the requesting STA shall filter out group addressed frames that are transmitted as individually addressed frames.

The major disadvantage of the FBMS is that it significantly increases the delay of the B/M streams, and the saved energy is reduced when the number of B/M streams increases, as in this case some streams have to be delivered within the same DTIM periods. The FBMS does address the regular unicast traffic [HYM+07]. Following is a survey of the handset power saving problem in WLANs.

2.4.5 Mobile Station Power Saving in IEEE 802.11

In addition to handset power conservation support in IEEE 802.11, client station power saving has received a lot of attention in recent years as many researchers have explored power management strategies in WLANs [MBH01, THH03, CSE04, CSA99]. In [CV02] and [Cha03], the authors investigate application-specific protocols to reduce the power consumption of network interfaces for streaming media applications.

Krashinsky and Balakrishnan [KB05] address the issue of interaction between energy-saving protocols and TCP performance for web-like transfers, and present a bounded slowdown (BSD) protocol that guarantees bounded delay on transfer round-trip time while conserving energy. Their work is extended in [QS05], where a smart PSM is proposed to guarantee arbitrary user-desired delay performance. Reference [NAS04] uses different beacon periods for different HTTP clients to improve the network performance in terms of response time of HTTP applications. Anand *et al.* [ANF05] implemented a self-tuning power management (STPM) module in the Linux kernel that adaptively adjusts the power management schemes with application hints. The study in [HYM+07] analyzed the impact of background traffic on the power consumption of MSs. The authors designed a scheduled PSM protocol with scheduled, uninterrupted frame delivery procedures and differentiated multicasting. In their protocol, a beacon period is divided into multiple time slices, and all pending frames are scheduled to deliver at the appointed time slices. The TIM structure is redesigned to convey the slice assignment information. However, this work only considered DL traffic.

Though the above research contributions have done excellent work to improve power or network performance in WLANs, none of them addresses the vulnerability of IEEE 802.11 power management mechanisms against the behavior of a power-saving AP operating at low-power duty-cycles.

2.5 Wireless Mesh Networks

Multihop relaying in IEEE 802.11 WLANs has been emerging recently as a cost effective way of reducing the infrastructure cost of providing wired network connections

to WLAN hotspots [FT05]. Multihop relaying in cellular infrastructure networks had attracted a lot of attention from both industry and academia. For example, pre-installed ad hoc relay stations are proposed in [AMM01] to forward traffic between cells. The system in [AN00] allows direct communication between two nodes, using ad hoc networking, whenever they are within the same transmission range without using the cellular infrastructure. The authors in [CCW03] propose a similar scheme that maintains simplicity by restricting the number of hops between any pair of communicating nodes to two only. Other schemes and protocols are discussed in [Lin01, WQDO01, RBM02, MMV03].

An IEEE 802.11 (Wi-Fi) mesh network can be considered as a concatenation of multiple hotspots that forms a wireless backbone network which ultimately connects to the service provider's infrastructure. In fixed Wi-Fi WMNs, APs are typically mounted on street-lights or telephone poles. Each mesh AP (MAP) creates a small wireless coverage area called "a hop" and each hop can serve a number of mobile wireless clients or wired clients within a home network. Wi-Fi mesh deployments may also contain a number of mesh points (MPs), which only relay traffic in a WMN but do not provide wireless coverage for MSs. Each MP or MAP acts much like a router and automatically discovers neighboring mesh nodes and relays packets across several hops in the wireless backbone. In IEEE 802.11 based mesh networks, there are typically a number of APs connected to a wired backhaul of the service provider, called root APs (or portal APs). The rest of the MPs/MAPs use multihop communications to communicate with these root nodes. Packets are forwarded from the mobile stations and are relayed to their destination. ESS mesh networks form a rectilinear topology in the simplest case (generates least interference), or can be fully-connected mesh

(costly to deploy but more reliable).

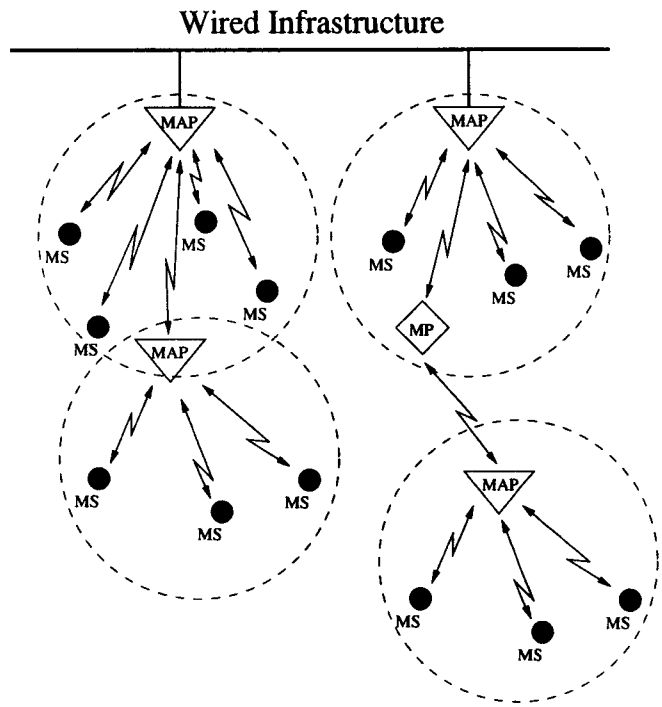


Figure 2.3: Solar Powered WLAN Mesh Network Example.

In conventional IEEE 802.11, APs are continuously powered using alternating current mains power or power-over-ethernet connections [FT05]. In future WMNs, wired power connections may not always be readily available. This is increasingly true as network coverage moves into more expansive outdoor areas. Example applications are emergency relief, mobile/military, temporary deployment and outdoor metro Wi-Fi. In such cases, fixed power connections can often be replaced by a solar powered design comprising a solar panel and a battery. Figure 2.3 shows an example of a WLAN mesh network. Four MAPs are shown in this network, two of which are portal mesh nodes and are connected to the wired distribution system, while the

other two (shown at the bottom of the ESS hierarchy) are solar MAPs (SMAPs). All MAPs provide both end user station Wi-Fi coverage and also perform backhaul traffic relaying. In addition to a solar MAP, the figure also shows a solar MP for performing traffic relaying services in the ESS.

In some WMN deployments routing can be done completely in an ad hoc manner where packet forwarding is performed by both clients and APs, e.g., the Ad Hoc On Demand Distance Vector (AODV) protocol. In classic Wi-Fi WMNs, MSs are battery-powered and need to operate in low-power sleep modes to conserve energy, however, routing among static APs is more efficient and stable than routing among client devices whose movements trigger dynamic paths across different hops. Other mesh routing protocols include (i) spanning tree routing, where the best path between a mesh node and a portal node could be determined using metrics like hop count and link transit delay; and (ii) Optimized Link State routing (OLSR) protocol which utilizes a proactive routing scheme and improves over the performance of its flooding-based parent by keeping the link state information distribution local to only a subset of nodes called multi-point relays. Other routing schemes for WMNs include Weighted Cumulative Expected Transmission Time (WCETT) [DPZ04] and its variants, e.g., WCETT with Load Balancing (WECTT-LB). The WCETT scheme combines two routing metrics by taking their weighted average and it aims at achieving a tradeoff between the individual stream selfishness and the delay and throughout experienced by the routed flows. The WECTT-LB enhances the basic WCETT by incorporating load balancing into the routing metric and provides a congestion aware routing and traffic splitting mechanism to achieve global load balancing in the network [MD07].

Due to the popularity of mesh solutions for Wi-Fi coverage extension, an IEEE

802.11 task group is developing the IEEE 802.11s standard for Extended Service Set wireless mesh networking [Dep04]. The first draft of the IEEE 802.11s standard was issued in March 2006 and it defines various mechanisms for improving the reach of WLANs with efficient wireless routing protocols and it allows APs to configure as a mesh network. In addition to a new path selection protocol "Hybrid Wireless Mesh" which was adopted by the draft, IEEE 802.11s also allows vendors to use inter-operable and proprietary path selection protocols while using the existing IEEE 802.11 addressing format to define a mesh network with autoconfiguration capabilities. Other capabilities of the IEEE 802.11s standard include topology discovery, path selection and forwarding, congestion control, channel allocation, traffic and network management and end-to-end multihop security mechanisms.

2.5.1 Infrastructure Power Saving in Solar Powered ESS Wireless Mesh Networks

The use of wireless links to connect mesh nodes to portal APs in a WMN allows for much flexibility in positioning the APs. On the contrary, providing electrical power for mesh nodes may place severe restrictions on their location. In classic infrastructure-based WLANs, there is very little need to conserve power in APs which are connected to a wired distribution system since they usually have access to external mains power or power-over-ethernet connections. In many wireless multi-hop mesh network applications, such as the SolarMESH Network at McMaster University [Uni04], emergency-relief networks, mission-critical/military and outdoor metro Wi-Fi installations, wired power connections may not be always readily available or could be very expensive to

provide. In such cases, fixed power connections can often be replaced by a solar powered design comprising a solar panel and a battery. Additionally, in practical WMNs deployments, average AP usage is usually far below AP maximum capacity. For this reason, a huge potential for infrastructure power saving is inherent in the operation of the network. In the solar powered case, since AP power saving is not permitted, solar panel/battery cost is often much higher than would be possible otherwise. For this reason, power saving on the AP would be highly desirable. Unfortunately, this is not possible since the existing IEEE 802.11 standard requires that APs remain active at all times.

2.5.2 IEEE 802.11 Power Saving APs

Two variations of IEEE 802.11 have recently been proposed [ZTZK04, LTZ05]. Reference [ZTZK04] first introduced IEEE 802.11-based power saving access points (PSAPs), intended for use in multihop battery and solar/battery powered WLAN applications. Three different frame design arrangements were considered for how a PSAP can dynamically modify its sleeping schedule to adaptively support current loading conditions while saving as much power as possible. In the SR (Sleep-Relay) and RS (Relay-Sleep) designs, a single dynamic movable boundary allows the PSAP to trade off upstream relaying and local contention subframes for power saving sleep periods. In the more complex SRS (Sleep-Relay-Sleep) arrangement two movable boundaries are used. The protocols achieve some degree of AP power saving in cases where legacy end stations are being accommodated. Unfortunately, there are many practical disadvantages to these procedures, mainly caused by IEEE 802.11's assumption that an AP is always active on its assigned channel. This restriction is a major impediment

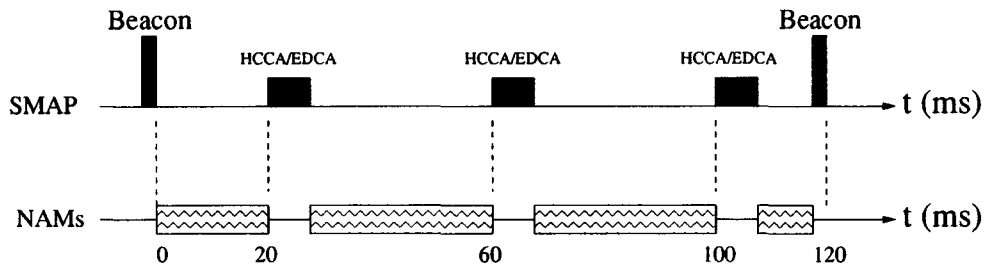


Figure 2.4: Power Saving MAP Example with HCCA/EDCA Service Periods.

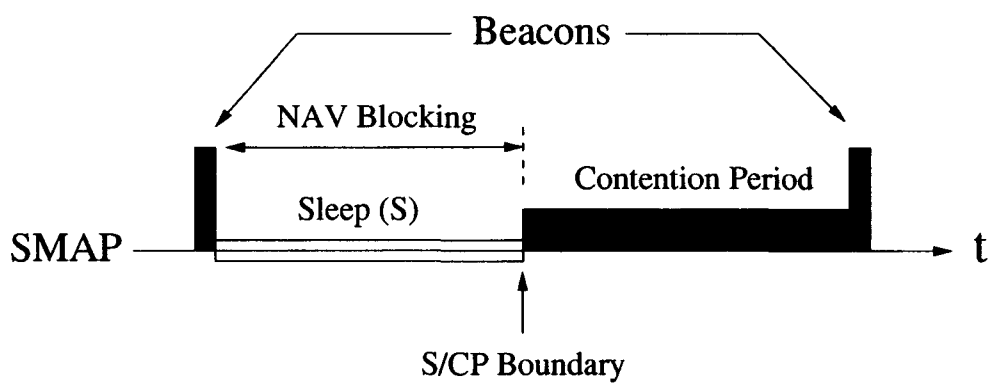


Figure 2.5: Best Effort MAP Power Saving with Movable Boundary.

to the development of practical power saving infrastructure.

The second approach [LTZ05] extends the IEEE 802.11 network allocation vector (NAV) mechanism and introduces a dynamic scheme for updating the sizes of the sleep and awake periods based on a network allocation map (NAM). The NAM is used to specify periods of time within the superframe when the PSAP is unavailable. During these periods the PSAP is assumed to be inactive and conserving power. Figure 2.4 shows an example of this type of activity for a single inter-beacon period. The channel activity is shown in the upper timeline, and the NAMs are shown on

the lower timeline. In this example three HCCA (or EDCA) periods have been scheduled and the AP advertises the NAMs as shown so that power saving can occur when the channel is not needed. An algorithm for dynamically updating the NAMs was proposed in [LTZ05] using least mean square optimization. However, the work in [LTZ05] only considers best effort traffic with no QoS guarantees.

A NAV blocking mechanism can be used by an AP to power save or switch to different channels. Single radio SolarMESH [Uni04] power saving APs take advantage of the fact that during conventional IEEE 802.11 contention free periods, associated MSs do not try to capture the channel without being polled. A PSAP, therefore, can turn off its transceiver to extract power saving or it can tune its antenna to a different channel to relay backhaul traffic to its parent node. An example configuration is shown in Figure 2.5. Here, a superframe length (t_{SF}) is equal to one beacon interval only. The PSAP goes to sleep right after the start of the CFP, and then wakes up before the CFP ends by a CF-End packet. MSs are allowed to compete for packet transmission during the contention period.

2.5.3 Challenges with Infrastructure Power Saving

Resource Assignment In solar powered WMNs, fixed power connections are replaced by a solar powered design comprising a solar panel and a battery. Therefore, node resource assignment consists of provisioning each node with a solar panel and battery combination that is sufficient to prevent node outage for the duration of the mesh network deployment. This assignment must use "geographic provisioning" to account for the solar insolation capability of the node location. In [BST08], the authors studied this resource assignment problem

with the objective of minimizing the total battery cost for a given energy source assignment. A methodology and algorithms for determining the resource assignment were given based on the target traffic flow profile which specifies the power consumption workload for which nodes are being configured. Shortest path and energy aware routing were also used to characterize the resource assignment.

Cost of Deployment Sizing of the battery and solar panel is important since they are often a significant fraction of the solar design's cost in sustainable energy solar-powered WMNs. Therefore, whenever possible, the decision of whether to provision all mesh nodes with a solar-design or to use fixed power connections should be considered. An alternative for reducing the cost of WMNs deployment is to use hybrid provisioning where some nodes are operated using sustainable solar energy while others use conventional fixed power connections. In [STS08], a cost model was introduced to optimize the hybrid provisioning of the nodes. Based on their study that included several North American locations with different meteorological situations, the authors in [STS08] concluded that in certain geographic locations a hybrid solar powered WLAN mesh node is the optimum cost configuration.

MS Handoff In [FT05], the authors studied the affect of infrastructure power saving on MS handoffs in IEEE 802.11 WLANs and showed that the extra delay caused by power saving exacerbates the handoff procedure. Since power saving APs may spend a considerable percentage of time in power saving mode, MS scanning for new APs may need a much longer time to find one, resulting in very long scanning delays. The same problem lengthens the handoff execution period. Authentication and bridging table update messages may suffer unnecessary long

delays while they are traversing multiple hops since receiver radios could be in PS mode. A number of AP activation methods are proposed in [FT05]. The protocols combine partial activation of the intermediate nodes as well as a smart time coordination between them to alert power saving mesh APs beforehand so that they can accommodate low latency scanning for imminent handoffs.

QoS Provisioning for Real-time Multimedia Traffic When considering the issues of network capacity, bandwidth allocation and QoS provisioning for real-time multimedia applications in contention-based distributed WLANs, infrastructure power saving is extremely challenging due to the time-varying bandwidth requirements for multimedia content, contention constraint, the packet-based network and most importantly, intense competition for the channel. Additionally, delay-aware QoS schedulers are needed to guarantee delay requirements for voice/video streams served by a power saving AP that operates at a less than 100% duty-cycle. Various effects of AP power saving on the handset power performance should be taken into consideration when scheduling awake and sleep periods of a PSAP. This is true since stations will have to awaken and fight the contention at the beginning of availability periods because they were unable to send packets generated during the AP's unavailability (sleep) periods. The focus of this thesis is to devise a framework for a power saving QoS-enabled AP (PSQAP) intended for use in low power infrastructure applications. The objective is to allow a PSQAP to power save while maintaining user QoS requirements and without significantly affecting station power saving.

The problem of low WLAN VoIP capacity and high handset power consumption is described below.

2.6 Voice over Internet Protocol over WLAN

WLANs were originally designed to carry data traffic for applications which are delay insensitive, such as emails, file transfers and web-browsing. As WLAN data rates increased significantly, applications which require strict guarantees on bandwidth and delay could be admitted by reserving the needed network resources. An example of such applications is VoIP, where voice calls are established over data networks with minimal cost since the infrastructure already exists, allowing for free voice communications. Many businesses have started using VoIP to place phone calls over their WLAN. In 2005 the VoIP over WLAN (VoWLAN) phone market reached nearly $72 million with Cisco Inc. leading the revenue market [WiF]. Although much less than cellular phone market sales, the VoWLAN phone market showed a 327% increase in sales compared to only a 13% increase in cellular mobile phone sales in year 2006 [WiF].

Although VoIP-capable networks are quickly growing as a percentage of the communications market, they still represent a small percentage of the overall voice market. Therefore, these networks must be able to communicate with other types of networks especially the PSTN. Gateways are boundary entities used to perform the translation required (e.g., signaling and voice coding formats) between the two networks. The enterprise network shown in Figure 2.6 has one LAN carrying voice and data traffic. Additionally, the figure shows a mobile VoIP phone with a WLAN interface that is accessing the LAN through an AP. The WLAN VoIP phone can communicate directly with other VoWLAN (or VoIP) phones, or can dial a plain old telephone that is residing on the PSTN. However, in the latter case, the call must be routed through the PSTN gateway to perform the necessary signaling and media conversion. It should be

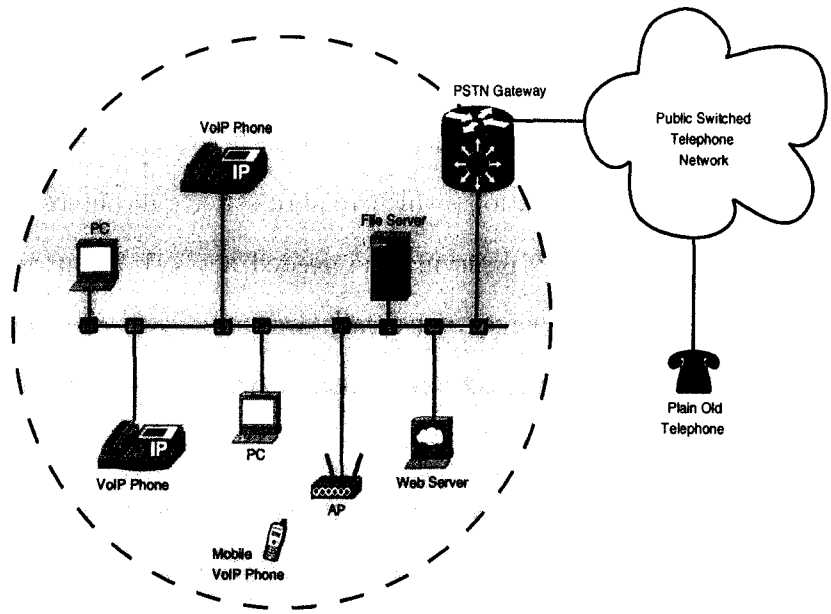


Figure 2.6: *A VoIP-Capable Network with a PSTN Gateway*

mentioned that Session Initiation Protocol (SIP) [RSC+02] is the de facto standard for establishing media and voice sessions in IP LANs including WLANs.

VoIP over WLAN (VoWLAN) combines two potent technologies: Wi-Fi and VoIP. Since it costs only less than 1/10th of what it costs using cell phones, VoWLAN proves to be a more cost effective technology than cellular VoIP; thanks to the widely installed Wi-Fi infrastructure [Bin08]. It also provides better voice quality due to higher data rates (e.g., 11Mbps, 54Mbps), provides converged infrastructure such as unified messaging (integrated email and voice mail) and it can extend to home office using a virtual private network (VPN). With dual-mode cellular/Wi-Fi phones becoming popular, wireless VoIP may become the more significant driver for VoIP than wireline VoIP.

VoIP capacity is typically very low in practical voice over WLAN networks. This is mainly due to the short audio payload used, combined with the large per packet overheads associated with various protocol layers. Many performance studies have shown that given the choice of small frame sizes in current IP phones (typical packetization values are $10 - 30ms$), the maximum number of simultaneous VoIP calls that can be placed in a single IEEE 802.11b [std99] AP cannot exceed about 15 calls regardless of the CBR vocoder selection [GK02, GK03, AEF+03]. Even when the polling-based PCF (or the IEEE 802.11e HCCA) channel access mechanism is used, the upper bound on the number of simultaneous calls is only about 17 [VCM01, KW00, KW01]. With header compression and downstream packet multiplexing, VoIP capacity can only be doubled, at best [SLLY02], [WLQL04]. These numbers are far below theoretical capacities [WLL05].

These capacity effects can be reduced by using larger voice packetization intervals [VCM01]. Unfortunately this is not always possible since it cuts into the end-to-end connection latency budget and the packet loss resilience of the call. In some cases however, this approach can result in significant capacity improvements without suffering any latency or error resiliency problems. In an enterprise setting for example, intra-enterprise VoIP connections will usually always experience very low latency and packet loss, and thus higher packetization intervals are reasonable. However, when one end of the call is outside the enterprise, this may not be the case. Sending/receiving larger packets at less frequent intervals is more energy and bandwidth efficient than when frequent smaller packets are used. However, tackling this problem in VoIP is more challenging than in other types of traffic (e.g., best-effort) since different QoS requirements (such as delay and jitter) must be taken care of when several

short codec frames are delayed in order to be packed into larger VoIP packets.

2.7 Summary

This chapter introduced wireless local area networks and provided an overview of various channel access mechanisms, power conservation schemes, and Quality-of-Service capabilities in IEEE 802.11 WLANs. A comprehensive literature review was provided for the issues of handset power saving and admission control in IEEE 802.11 WLANs. A detailed description of the infrastructure power saving problem in solar-powered ESS wireless mesh networks was given along with a survey of different challenges and the relevant literature. Finally, Voice over IP technology and the problems of low capacity and poor power performance of VoIP over WLAN were described.

Chapter 3

QoS-Enabled Power Saving Access Points

3.1 Overview

IEEE 802.11 is now being used in many situations where access point (AP) power saving would be highly desirable. Example applications are emergency relief, mobile/military, temporary deployment and outdoor metro-WiFi. In such cases, fixed power connections, which may not always be readily available, can often be replaced by a solar powered design comprising a solar panel and a battery. In the solar powered case, since AP power saving is not permitted in conventional IEEE 802.11, solar panel/battery cost is often much higher than would be possible otherwise. Additionally, in practical WMNs deployments, average AP usage is usually far below AP maximum capacity. For this reason, a huge potential for infrastructure power saving is inherent in the operation of the network.

This chapter presents and assesses the performance of a framework for a power

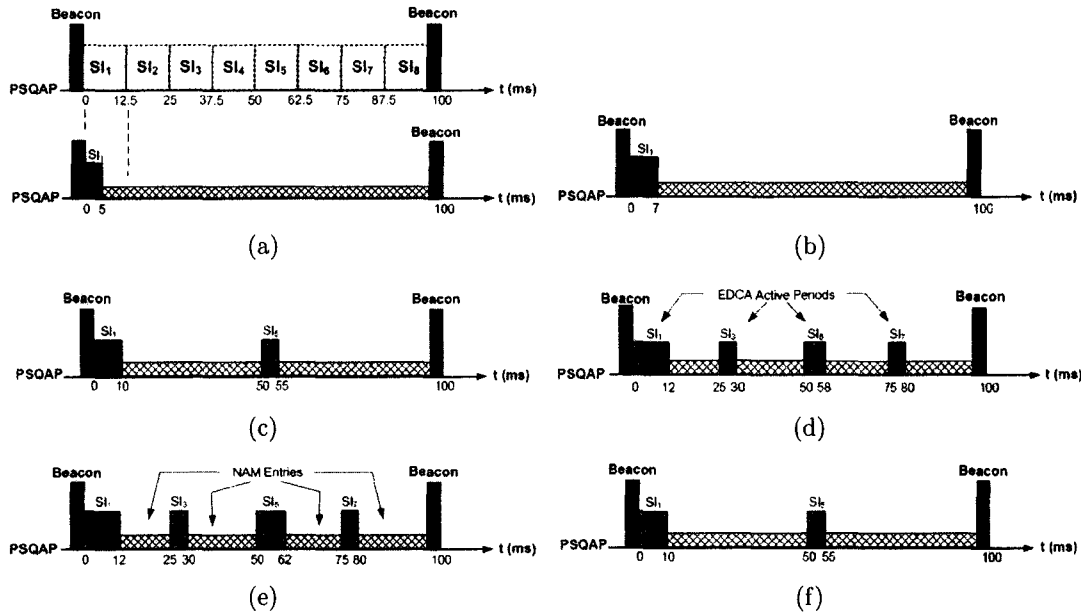


Figure 3.1: Dynamic Timeline Update using Dynamic Service Interval Activation

saving quality-of-service (QoS) enabled access point (PSQAP), intended for use in low power infrastructure applications. The proposed framework introduces infrastructure based power saving while preserving the QoS requirements for delay and loss intolerant real-time applications. An extensive simulation study is used to evaluate the integrated framework, which consists of a proposed energy efficient media access control (MAC) protocol and an adaptive connection admission control (CAC) scheme.

3.2 Power Saving QoS AP (PSQAP) Protocol

The energy-efficient MAC protocol of the proposed framework is shown in Figure 3.1. The proposed MAC protocol extends the concept of a Network Allocation Map (NAM), introduced in [LTZ05] for best effort traffic. A power saving AP includes a NAM field in its periodic beacon broadcasts, to coordinate traffic delivery and

power saving at both end stations and at the AP. The NAM provides a simple layout of the AP's current superframe, which consists of time intervals separated by boundaries that are marked as either fixed or moveable. A PSQAP operates on a smaller duty-cycle for power saving extraction. To achieve this, a PSQAP offers only the bandwidth required to serve the offered traffic load while it sleeps, i.e., conserves energy, otherwise. When real-time flows are present, a PSQAP schedules its awakening/sleeping pattern in a manner that satisfies the delay and packet loss requirements for the admitted flows. To achieve this, the proposed protocol employs a novel scheme for dynamic activation and de-activation of unscheduled service intervals, as described below.

3.2.1 Dynamic Service Interval Activation

The superframe is sub-divided into a pool of N equal service intervals (SIs). The value of N determines (among other things) the service interval length (t_{SI}) and the minimum delay bound for real-time streams that can be met by the PSQAP. For example, in the top part of Figure 3.1(a), a 100ms superframe with 8 SIs is shown. In this example, the minimum delay bound that can be met by the PSQAP is 12.5ms. Each SI consists of two sub-intervals, i.e., an EDCA activity sub-interval and an (AP power saving) sleep sub-interval. The boundary between these two subintervals is moveable and is dynamically updated as discussed below.

The initial default status of the N service intervals is "inactive" and the length of the activity sub-interval of an inactive SI is set to zero. Initially, the AP advertises a beacon NAM that defines a single active SI which is referred to as the "Initial Access Period" (IAP), as shown in Figure 3.1(a). The PSQAP may dynamically activate

additional SIs in order to accommodate newly accepted traffic streams (TSs) or to accommodate increased levels of best-effort traffic. New stations which join the basic service set (BSS) and sample the beacon frame will learn the current set of active SIs advertised in the NAM.

A QoS station (QSTA) may contend for channel access during the IAP to request the admission of a real-time TS. If the new TS is accepted, the PSQAP may activate additional SIs and adjust their activity sub-intervals in order to meet the requirements for the admitted TS. Once a TS is accepted, the PSQAP instructs the corresponding QSTA to contend for channel access using US-APSD in a specific set of active SIs that meet the requirements of the TS. Active stations are only permitted to access SIs that are either assigned to them by the PSQAP, or are randomly selected by best-effort stations after associating with the PSQAP. The PSQAP may also instruct active stations to remain awake during the next beacon frame transmission time to sample a newly updated NAM.

A simple example is now given for how the SI activation procedure works. In Figure 3.1(a), the PSQAP is initialized and SI_1 is active with an EDCA activity interval set to 5ms, leaving the remainder of the SI for AP power saving. In Figure 3.1(b), a new best-effort station $QSTA_1$ associates with the PSQAP and transmits during SI_1. The PSQAP extends the admissible part of SI_1 and advertises this in the NAM. In Figure 3.1(c), $QSTA_2$ has a new real-time TS (with a 50ms delay bound). As a result, the PSQAP adjusts the length of the access period of SI_1, activates SI_5, and advertises this in the NAM. The new set of active SIs in Figure 3.1(c) is a super-group of that in Figure 3.1(b). Notice that the schedule of $QSTA_1$ is not affected by the admission of $QSTA_2$.

Figure 3.1(d) shows $QSTA_3$ with a real-time TS which requires a 25ms delay bound. To meet this bound, the PSQAP activates both SI_3 and SI_7. The new set of active periods is a super-group of that in Figure 3.1(c). The service schedule for both $QSTA_1$ and $QSTA_2$ is not affected. Later on, a best-effort station, $QSTA_4$, joins the BSS in Figure 3.1(e). As part of the association signaling, the PSQAP sends a list that has SI_1 and SI_5 to $QSTA_4$ which randomly chooses SI_5. This means that $QSTA_4$ will confine its packet transmissions to SI_5. The right boundary of SI_5 is shifted to the right and the NAM is updated. After $QSTA_3$ terminates its TS (Figure 3.1(f)), and since no other stations are assigned to SI_3 or SI_7, the PSQAP deactivates both service intervals and updates the NAM. $QSTA_1$, $QSTA_2$, and $QSTA_4$ are not affected by this NAM update.

The selective (de)activation of SIs described above tries to ensure that when the timeline is updated, the new set of active SIs will always be either a super- or sub-group of the old set. This has the advantage of reducing the signaling required to update active QSTAs. However, when the channel utilization in a given active SI becomes very low it is advantageous for the AP to deactivate this SI in order to conserve energy. In this case, the PSQAP still needs to signal all affected QSTAs so that they are assigned to other active SIs in the superframe. We define a new QoS action frame that the PSQAP can use for this purpose. Finally, it is important to note that backward compatibility of the proposed MAC protocol is achieved through the use of NAMs. During unavailability (sleep) periods of a power saving AP, all legacy and QoS stations block their Network Allocation Vectors (NAVs) and refrain from accessing the channel, in a similar manner to virtual carrier sensing in conventional IEEE 802.11, until the NAV is reset at the beginning of the next EDCA activity

interval.

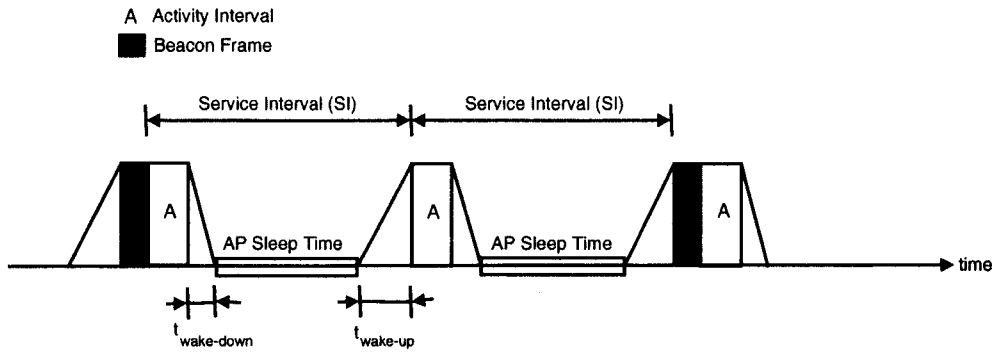


Figure 3.2: Awake-Doze State Transition Time Overhead.

3.2.2 Service Interval (SI) Constraints

There are practical limits on the values of the SI periods that can be used. This can be determined by considering the power consumed by the RF circuitry during the ramp-up (doze-to-awake) and ramp-down (awake-to-doze) periods. The potential sleep time for a PSQAP (per SI) can be calculated as

$$T_{sleep}^{perSI} = t_{SI} - t_A - t_{wake-up} - t_{wake-down} \quad \text{(s)}, \tag{3.1}$$

where t_A is the length of the EDCA activity subinterval (Figure 3.2). $t_{wake-up}$ and $t_{wake-down}$ are the time spent by the radio during the ramp-up and ramp-down, respectively. The power consumption for a PSQAP during an EDCA activity period is $\eta_{Tx}\ t_{DL}/t_A + \eta_{Rx}\ t_{UL}/t_A$, where η_{Tx} and η_{Rx} are the assumed power dissipation (W) in transmit and receive modes, respectively. t_{UL} and t_{DL} are the fractions of the activity subinterval spent in receiving uplink and transmitting downlink traffic from/to QSTAs, respectively.

Figure 3.2 above shows that awake-sleep and sleep-awake state transitions are accompanied with non-negligible overheads in terms of wasted channel time. Our analysis of PSQAP energy consumption results obtained by simulations shows that for a coarse-grained t_{SI} ($\simeq$ 20ms), the percentage of wasted channel time ranges from 5% to 15%, while it is in the range of 10% to 30% of t_{SI} when a fine-grained t_{SI} of length 10ms is used. For both cases, the PSQAP was found to suffer high power consumption due to turning the (internal radio) oscillators off during the sleep mode of the PSQAP. Modern chipsets may keep the oscillators running while the radio is off at the cost of dissipating extra power ($\simeq$ 6mW compared to 2mW) to save much of its wake-up transition time. This also suggests that a lightly loaded PSQAP (e.g., 15% duty-cycle) can achieve more energy saving by turning off the oscillators while in doze mode. On the other hand, it is a more energy-efficient choice for a loaded PSQAP with 50% duty cycle to keep the oscillators running while in doze mode. The reason for this is that the energy saving due to shorter wake-up times would be larger than the energy saving during the sleep interval if the oscillators are turned off.

This last observation suggests the existence of an adaptive control algorithm at the PSQAP that makes a decision on whether to turn the RF circuitry oscillators off before transiting to sleep mode based on real-time measurements of the achieved channel utilization. To conclude this section, it should be mentioned that the choice of having very few SIs in a superframe (e.g., $t_{SI} \approx$ 100ms), although efficient in terms of channel efficiency and power saving, would limit the delay bound for real-time streams that can be served by the PSQAP. This is true for real-time multimedia TSs especially in a multi-hop wireless mesh network where the one-hop delay limit is typically $\ll$ 100ms.

Table 3.1: The New QoS Action Frame.

Order	Information
1	Category = 1 (QoS)
2	Action = 4 (ASSIGN)
3	List of Assigned SIs

3.2.3 QoS Signaling for NAM Updates

The selective (de)activation of SIs described above tries to ensure that when the timeline is updated, the new set of active SIs will always be either a super- or sub-group of the old set. Thus reduces the signaling required to update active QSTAs. However, when the PSQAP deactivates an SI, for energy conservation purposes, the PSQAP still needs to signal all affected QSTAs so that they are assigned to other active SIs in the superframe. We define a new QoS Action frame that the PSQAP can use for this purpose. The proposed QoS Action frame is shown in Table 3.1, where the Category field has the value 1, corresponding to QoS. A new Action identifier ASSIGN with an enumerated value 4 is introduced, and is followed by the new list of SI(s) assigned by the PSQAP to QSTAs previously contending in the deactivated SI.

3.3 Connection Admission Control (CAC)

This section presents a CAC scheme for controlling the air-time usage of QSTAs served by a PSQAP. When a new TS_i is initiated, a test is first performed to check whether the upper delay bound (D_i) of the new stream can be met. If $D_i < t_{SI}$, then TS_i cannot be admitted. Since TXOPs are granted on a time basis, the CAC unit at the PSQAP then needs to convert the rate-based TSPEC parameters into time based

ones. All TXOPs are calculated over a fixed interval determined by the PSQAP over which the schedule is computed; in our case this interval is the SI length, i.e., t_{SI}. The CAC unit then searches for a combination of SIs that have sufficient resources to admit TS_i. For such an admissible SIs combination, the distance between the starting of any two successive SIs must be $\leq D_i$.

To check for sufficient channel-time resources, the CAC uses the mean bit-rate ρ_i and nominal MSDU size L_i, which are obtained from the $TSPEC_i$ elements, to determine the number of MSDUs generated by TS_i (per t_{SI}) as follows

$$N_i = \lceil \frac{\rho_i t_{SI}}{L_i} \rceil \qquad (MSDU/t_{SI}). \tag{3.2}$$

The modified bit-rate for this source including additional per packet overheads (i.e., MAC/IP/UDP/RTP headers and tails) is given by

$$\rho_i^{'} = \frac{N_i(L_i + O_i)}{t_{SI}} \qquad \text{(bit/s)}, \tag{3.3}$$

where O_i represents the overhead. The Transmission Budget (TB_i) required to serve all MSDUs arriving within the t_{SI} interval for TS_i is

$$TB_i = N_i^{'}(\frac{L_i}{R_i} + T_{overheads,i})SF^{max} \qquad \text{(s)}, \tag{3.4}$$

where $N_i^{'}$ is the average number of MSDUs calculated as in Equation 3.2 but with ρ_i replaced with $\rho_i^{'}$. R_i is the minimum physical transmission rate (bit/s) required to guarantee the QoS, $T_{overheads,i}$ denotes the required IEEE 802.11e MAC overheads (SIFS, AIFS, ACK, etc.), SF^{max} is the maximum value of the *SurplusFactor* for

TS_i [std05a], and $(\frac{L_i}{R_i} + T_{overheads,i})$ is the minimum required successful transmission time for a single frame for stream i including overhead. The *SurplusFactor* is used to compensate for the traffic loss caused by packet collisions or channel error.

Each admitted stream i has a counter, $T_{Counter,i}$, that keeps track of the number of time units that stream is expected to use the channel during the current SI. At the beginning of the jth SI, the counter is set to a value $TXOP(j)$, i.e., $T_{Counter,i}(j) = TXOP_i(j)$, where $TXOP_i(j)$ is the transmission opportunity granted to stream i in SI_j to serve all MSDUs belonging to that stream which arrived since last *TXOP*. $TXOP_i(j)$ is calculated in a similar way as in Equation 3.4 such that $TXOP_i(j) = N_i' \ (\frac{L_i}{R_i} + T_{overheads,i}) \ SF_i(j-1)$, where $SF_i(j-1)$ is the *SurplusFactor* for the ith stream in the (j-1)th service interval. For a given TS_i, the PSQAP updates $SF_i(j+1)$ using a running weighted average

$$SF_i(j+1) = fSF_i(j) + (1-f)\frac{T_{Access,i}(j)}{T_{Success,i}(j)}, \tag{3.5}$$

where $T_{Access,i}(j)$ and $T_{Success,i}(j)$ are the measured channel access times for TS_i regardless of success, and, all successful transmission times for TS_i as measured by the PSQAP, respectively. f is a smoothing factor and is used to control the effect that fluctuations in $T_{Access,i}$ and $T_{Success,i}$ can cause to the values of the *SurplusFactor*. By experimenting with various values of f, it was found that large values ($\simeq 1$) make the CAC more immune to instantaneous packet transmissions statistics observed on the wireless channel. In other words, the update of SF, hence the activity sub-interval length of a PSQAP, is done sluggishly. On the otherhand, the use of small f values results in a high packet-loss rate due to the rapid update of the activity sub-interval and the associated offered capacity.

For each stream i, we use a $T_{Limit,i}$ which is the maximum allowable channel-time occupancy for TS_i (during a service interval.) Whenever a packet belonging to that stream is sent, regardless of success, the same amount of access time is subtracted from the $T_{Limit,i}$. The QSTA is not allowed to transmit any frame belonging to TS_i once the corresponding $T_{Limit,i}$ is depleted. In other words, $T_{Limit,i}$ limits the length of the unscheduled TXOP that the PSQAP grants after receiving a trigger frame from QSTA$_i$. $T_{Remainder,i}(j)$ is calculated at the end of the jth SI as the difference between $T_{Limit,i}(j)$ and the actual number of time units used by QSTA$_i$ for packet transmissions during SI$_j$, i.e., it is the number of unused time units remaining for TS_i at the *end* of the jth SI. $T_{Limit,i}$ at the beginning of the (j+1)th SI is given by

$$T_{Limit,i}(j+1) = \min(T_{Remainder,i}(j) + T_{Counter,i}(j+1), N_i'(\frac{L_i}{R_i} + T_{overheads,i})SF^{max}) \quad \text{(s)}. \quad (3.6)$$

In an infrastructure-based WLAN, all packet transmissions go through the QAP, therefore it is valid to assume that a PSQAP has full knowledge of $T_{Limit,i}$ values for all admitted TSs. At the end of SI$_j$, the PSQAP calculates $TXOPBudget(j+1)$ which is the unused additional amount of access time available for QSTAs in the (j+1)th SI, such that

$$TXOPBudget(j+1) = \max(0, t_{CP} - \sum_{k=1}^{M(j)} T_{Limit,k}(j+1) - t_{BE}) \quad \text{(s)}, \quad (3.7)$$

where $M(j)$ is the total number of admitted real-time sessions in SI$_j$, and t_{CP} is the admissible portion of t_{SI} allocated for contention access. t_{BE} is the amount of access time (per SI) reserved for best effort (BE) traffic, and the term is the sum of the

maximum allowable transmission times for all QSTAs having ongoing real-time TSs in the (j+1)th SI. The reason a quota, t_{BE}, is reserved for BE traffic is to ensure that BE traffic is operational all the time and it does not suffer from starvation that could occur when channel bandwidth is offered only to real-time traffic flows. If $TXOPBudget(j)$ is depleted, a newly-arriving TS cannot be admitted in SI_j. To decide on the admission of a new stream, at least one admissible combination of SIs, Ψ, should exist, such that for each $SI_j \in \Psi$, $TB_i < TXOPBudget(j)$. Otherwise, TS_i is rejected.

Finally, from Equation 3.7, the sleep time for the PSQAP during SI_j is given by

$$T_{Sleep,j} = t_{SI} - \sum_{k=1}^{M(j)} T_{Limit,k}(j) - t_{BE} \qquad \text{(s)}. \tag{3.8}$$

Knowing the typical power consumption values for an operating IEEE 802.11 radio [ZTZK04] (see Table 3.2), the energy conservation achieved by a PSQAP during SI_j can be calculated as

$$\xi_{SI,j} = T_{Sleep,j}.(\eta_{Rx} - \eta_D) \qquad \text{(J)}, \tag{3.9}$$

where η_D is the assumed power dissipation (W) in doze mode.

3.3.1 Enforced Access Point Power Saving

An important feature of the proposed CAC is the ability to enforce power saving at the PSQAP by setting t_{CP} in Equation 3.7 to $t_{CP} = \alpha t_{SI}$, where we set the non-negative design parameter $\alpha < 1.0$ so that the PSQAP can use $(1-\alpha)\ t_{SI}$ as potential

sleep time. Generally, power control algorithms for wireless network applications statistically control time averaged energy consumption values in order to meet specific power consumption constraints. However, for a special class of WMN deployments, usually referred to as temporary deployments, the network is to be installed for a fixed time period, T_L [BST08]. In certain temporary solar-powered WMN deployments, long-term statistical power control might not be possible. Examples include situations where the deployment duration T_L is relatively short, or non-varying traffic profiles, or installations where issues of reliability and connectivity are of top priority compared to other performance metrics. In these cases, instantaneous and/or short-term infrastructure power control is crucial to the bandwidth offered at each node such that the network survives without node outage for the full deployment duration. Using the enforced power-saving feature of the proposed CAC scheme, an adaptive control algorithm could use as input the deployment duration (T_L), residual battery charge, solar panel size, historical solar insolation data and traffic design profile(s), to calculate the control parameter α that maps to the instantaneous maximum bandwidth that should be offered by the PSQAP to satisfy the no-outage constraint. The design and details of such a controller are outside the scope of this thesis.

3.4 PSQAP Framework Performance Analysis

In this section, the performance obtained using the proposed energy-efficient framework for PSQAPs is discussed. Performance metrics considered include power consumption for both the PSQAP and handsets, one-way packet delay and packet loss performance. Extensive simulation-based performance studies have been conducted

Table 3.2: Simulation Parameters

Parameter	Value
Number of APs	1 (PSQAP)
Voice codec / Payload size	G.711 ($64Kbps$) / 160 *bytes*
Smoothing factor (f) / SF^{max}	0.65 / 2.0
R	$11Mbps$
$t_{BE} : t_{SI}$	1 : 5
Maximum Retry Limit	3
Maximum MPDU size	2304 *bytes*
UPD / RTP header	20 / 12 *bytes*
MAC / IP header	34 / 8 *bytes*
Beacon interval (B)	$100ms$
SIFS / Slot time	$10/20\ \mu s$
T_{ACK} / T_{PHY}	$248\mu s$ / $192\mu s$
$\eta_{Tx}/\eta_{Rx}/\eta_D$	$750/500/8\ mW$

to evaluate the proposed framework. The simulation results for a PSQAP are obtained and compared with those of a conventional (i.e., no power-saving) AP.

3.4.1 Simulation Methodology

Simulations were performed using a discrete event simulator written in C for the contention-based EDCA to meet the specifications of both the IEEE 802.11e standard and the proposed scheme of selective SIs (de)activation. Also, the simple scheduler defined in IEEE 802.11e was modified to incorporate the proposed connection-admission control algorithm. The results presented are for a one-hop access network model with contention-based end station traffic. To achieve power management, the PSQAP maintains status information for each associated mobile station. The state transition timing of a network interface card (NIC) is controlled by the scheduling service of the MAC layer, which periodically queries the power state of the underlying PHY

layer and collects information to make doze-to-active and active-to-doze transition decisions.

Before each beacon transmission, the PSQAP calculates the length of EDCA-activity sub-interval for each service interval SI_j in a superframe, as explained in Section 3.3. Once the uplink traffic transmission time has been derived, the scheduler of the PSQAP calculates the start and end times of both the activity and sleep sub-intervals of all SIs. This information is then included in the NAM field of the beacon frame.

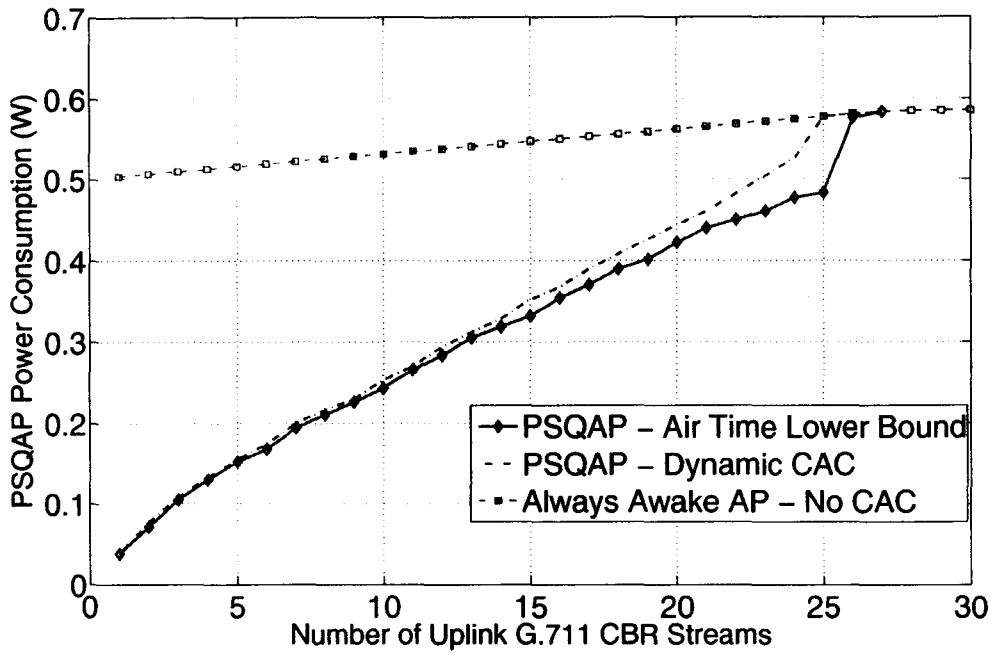
We first present results for CBR VoIP traffic, then for a traffic mix of VBR VoIP and BE (Email) traffic served by a PSQAP that employs our MAC and CAC framework. CBR calls arrive according to a Poisson process with an arrival rate, $\lambda = 0.075$ calls/s. All QSTAs are assumed to use the same voice codec (G.711 with 64Kbps vocoding rate) and the number of active voice terminals N is assumed constant over the simulation period of interest. For VBR VoIP, voice codecs are equipped with voice activity detection (VAD) and the talk-spurt on/off periods are assumed to be exponentially distributed with means of 1s and 1.35s, respectively (Brady's on-off model [Bra69]). As shown below, more VBR connections can be accommodated than CBR VoIP since the maximum capacity for VBR is that for CBR VoIP scaled by $(T_{on}+T_{off})/T_{on}$. Although the maximum permissible one-way connection latency for VoIP is about $150 \sim 200$ ms [Uni], the per hop delay in a multi-hop mesh network is far below this value. A typical value for the maximum delay bound of $30ms$ [WLL05] is assumed. Throughout the simulations, a packet is discarded if it exhausts the maximum retry limit (which was set to 3 [WLL05]), or if it exceeds the maximum delay bound (for VoIP traffic only.) A stringent upper bound on packet loss probability of

0.01 is assumed [WLL05], and t_{SI} = 20ms. The length of e-mail data messages follow a Cauchy(0.8, 1) distribution with exponential inter-arrival times. The average e-mail message length is 24 packets and the packet size is fixed at 160 bytes [PBLS99]. We assume a strict upper bound of 5s on the average e-mail transmission delay [KPP05]. Since our focus is on QoS satisfaction for the real-time streams, we only consider uplink traffic in our simulations, i.e., we assume uni-directional calls only.

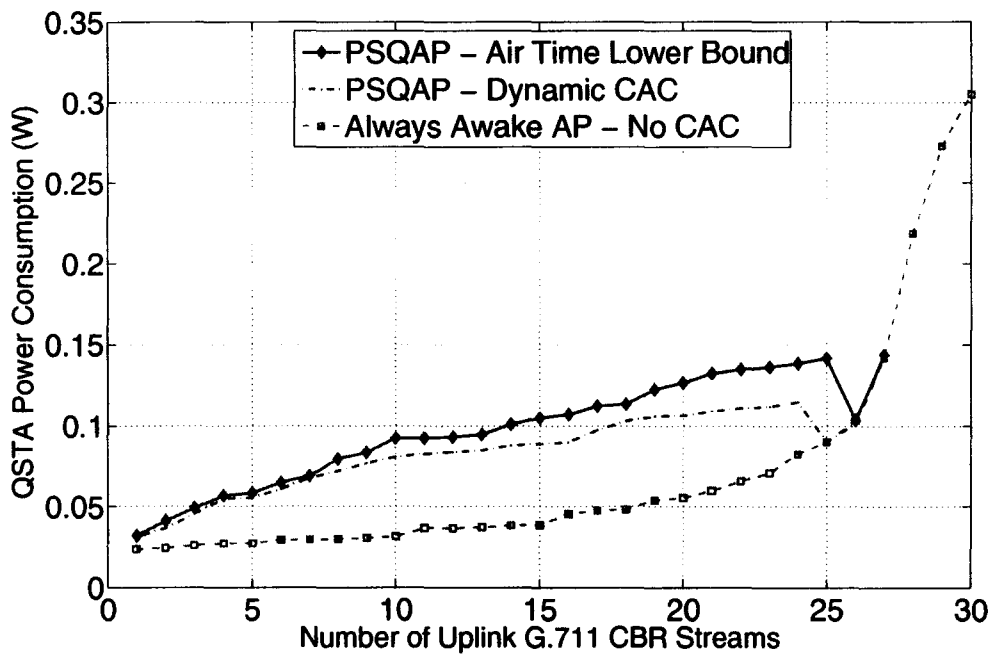
All results were obtained by averaging 10 simulation runs, and the simulation time for each was 500s, which was sufficient to eliminate any transient effects. Power consumption results were obtained with margins of ± 0.001 (95% confidence). Both the PSQAP and handsets are assumed to be within the same transmission range, therefore, RTS/CTS protection is eliminated throughout the simulations. Successful packet transmissions are acknowledged by destination nodes. The different simulation parameters used in the experiments are summarized in Table 3.2. The results show a significant improvement in AP power conservation with the use of our integrated adaptive CAC and MAC schemes for serving real-time traffic.

3.4.2 Simulation Results

Figure 3.3(a) shows the average power consumption for different APs with and without power saving. Clearly the energy efficient framework allows for a significant reduction of the AP power consumption for almost all ranges of VoIP loading. In numbers, a PSQAP can save about 55-90% of the total power dissipated by an always awake AP for low and medium loading, and up to 40% for high loading. For the sake of comparison, we also show an additional curve (labeled PSQAP - Air Time Lower Bound) for the mean AP power consumption lower bound which is obtained by



(a) CBR: Mean AP Power Consumption.



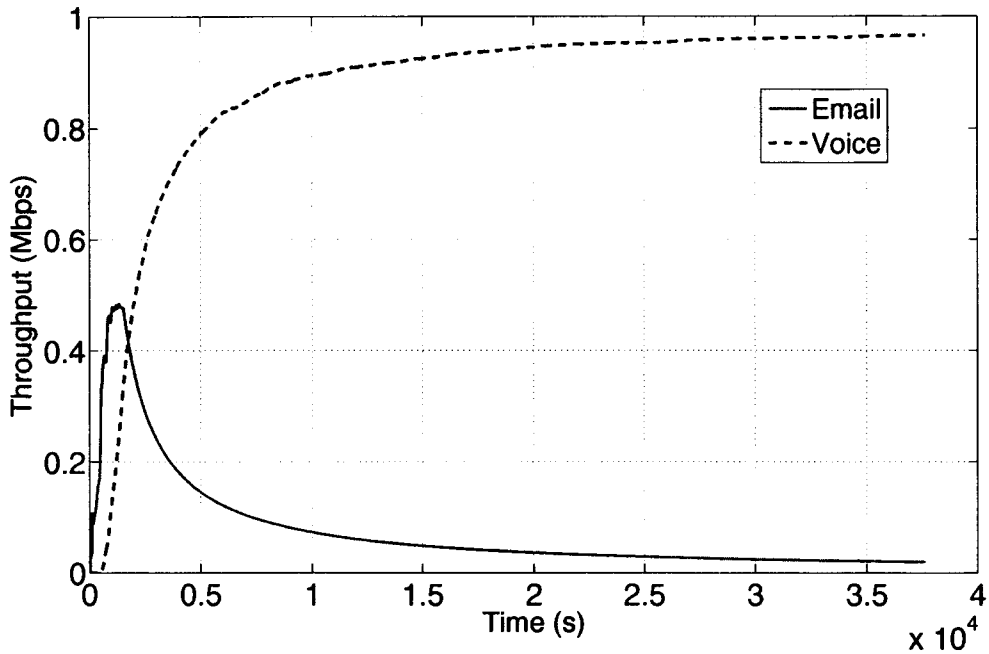
(b) CBR: Mean QSTA Power Consumption.

Figure 3.3: Framework Performance (CBR VoIP).

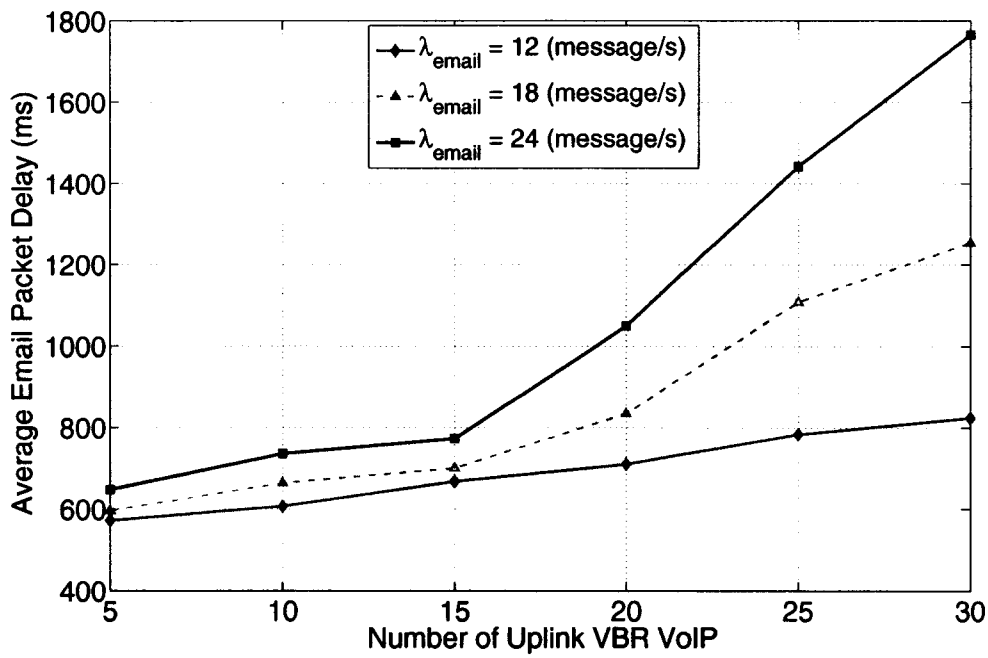
having the PSQAP continuously decrease the allocated air time for admitted streams until the packet loss probability reaches the 1% limit. In Figure 3.3(a) it can be seen that the PSQAP - Dynamic CAC curve closely approximates this bound for low and medium loading. Note that both curves perfectly match the curve for an AP which does not power save, when the number of admitted calls becomes very close to the maximum VoIP capacity which has been found to be 27.

From Figure 3.3(a), we can also see that the curves labeled "PSQAP-Air Time Lower Bound" and "PSQAP-Dynamic CAC" match the curve for an always awake AP after the $25th$ and $26th$ call are admitted, respectively. The reason for this is that, in the simulations we assumed that once the sum of available transmission budgets *TXOPBudget* in all SIs becomes insufficient to admit a new call, the CAC disables the power saving functionality and the PSQAP resumes operation as a conventional always awake AP. The difference of 1 call between the two "PSQAP-" labeled curves is due to the fact that "PSQAP-Air Time Lower Bound" curve is slightly more accurate in terms of the offered bandwidth to the "PSQAP-Dynamic CAC" curve which implements our proposed adaptive CAC algorithm.

Figure 3.3(b) plots the mean station power consumption for different numbers of admitted VoIP connections. From the figure, it can be seen that the greedy behavior of the PSQAP, by continuously trying to maximize its power saving, directly affects the mean power consumption at the stations. Again, we show a lower bound for station power consumption which is obtained by turning off PSQAP power saving. An analysis of the results in Figures 3.3(a) and 3.3(b) is useful for adjusting protocol specific parameters that would allow a PSQAP to save power while causing a minimal increase in the power dissipation from the stations' batteries.



(a) VBR VoIP and Email: Average Throughput.



(b) VBR VoIP and Email: Email Message Mean Delay.

Figure 3.4: Framework Performance (VBR VoIP and Email).

In the second experiment, we set the maximum duty cycle of the PSQAP to 90%, to force a minimum of 10% of t_{SI} for power saving at the PSQAP. For the traffic mix of VBR VoIP and e-mail, the traffic load is gradually increased. A total of 16 email users and 36 VoIP sources are progressively added at 3s intervals in order to show how a newly admitted flow impacts the performance of previously admitted flows. VoIP sources are scheduled only after all email users have joined the WLAN. Figure 3.4(a) shows the throughout for both VBR VoIP and e-mail traffic averaged over the total simulation time (1hr.) From the figure, we can see that at the beginning, the email traffic has high throughput; then as more real-time flows are admitted, it gradually drops as a result of the CAC and the high priority of voice traffic. Due to our choice of setting a quota for BE traffic, it can be observed that even when the traffic load becomes heavy, email traffic is not completely starved. From our simulation, the PSQAP awake time ratio was found to be (89%) which is very close to the maximum offered normalized bandwidth (i.e., 90%) while the channel busyness ratio was found to be 72%. The difference (i.e., 89% - 72%) is due to mandatory idle channel-time intervals in the IEEE 802.11e CSMA/CA MAC (i.e., SIFS, AIFS, and exponential random backoff.) Mean delay values for both traffic types, calculated from the simulation were found to be 4.26ms and 701.7ms. As expected these values are well below the specified delay bounds (i.e., 30ms for VoIP and 5s for email.)

Finally, we show (in Figure 3.4(b)) the email packet delay for a 75% duty-cycle PSQAP serving VBR and email traffic. Results are plotted for various numbers of uplink VBR flows. It can be seen that the data packet delay steadily increases as λ_{email} increases. However, the increase is much higher for larger λ_{email} values than for low or medium values of λ_{email}. The reason for this, is that in the presence of a

large number of VBR streams, the pre-allocated quota for email traffic is insufficient to accommodate newly arriving email data messages which are queued longer before they are transmitted.

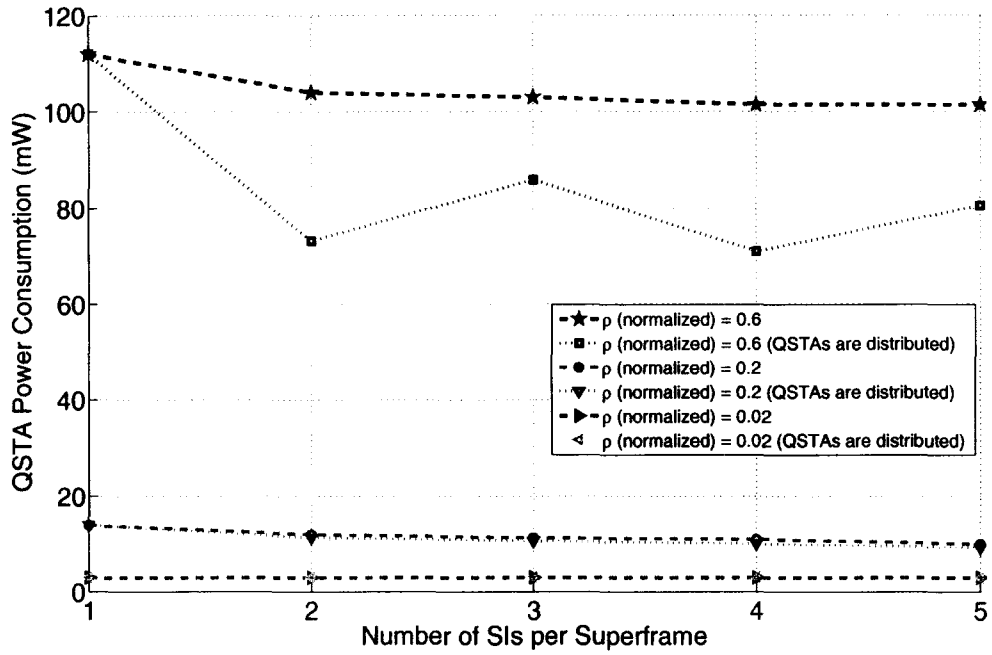


Figure 3.5: Mitigating Queuing Effect by Distributing QSTAs over Multiple SIs.

3.4.3 Improving on Handset Power Consumption

When a large number of IEEE 802.11 stations are contending for channel access, a significant portion of the station power consumption and UL packet delay is due to packet queuing. We present simulation results that show the value of distributing QSTAs over multiple SIs to mitigate these effects. Associated with the PSQAP are 20 data QSTAs receiving Poisson process packet arrivals; the data packet payload is 1000 bytes and the PSQAP operates at a 50% duty cycle. Figure 3.5 shows the mean QSTA power consumption results for two cases, (i) QSTAs are allowed to transmit

packets in any SI and, (ii) the channel activities of the 20 QSTAs are split over different SIs. For the latter case, a simple QSTA distribution policy was adopted. Based on the station ID, QSTAs with even station ID values contend for the channel in even SIs (i.e., SI_0, SI_2, and SI_4) while QSTAs with odd station ID values contend in odd SIs (i.e., SI_1 and SI_3). Results are plotted for various normalized load values and for various numbers of SIs per superframe.

In the case where stations are allowed to contend for channel access in any SI, it can be seen from the figure that for all normalized load values, increasing the number of SIs does not considerably lower the QSTA power consumption. Conversely, by distributing the QSTAs over 2 SIs we are able to reduce average station power consumption by a factor of 0.36 for a normalized load value of 0.6. This is a significant improvement in power consumption for the stations and should be used whenever QSTA power saving is important. The jumps in values for the second top curve in Figure 3.5 are because odd numbered QSTAs were allowed to share one even numbered SI with QSTAs having even station ID values. This was to ensure a fair and even channel access in cases where number of SIs is odd (but > 1). As a result, there are more QSTAs per SI for 3 SIs vs. 2 SIs and for 5 SIs vs. 4 SIs.

3.5 Conclusions

This chapter presented the use of an enhanced version of the IEEE 802.11e US-APSD access mechanism to build an energy-efficient framework for power saving QoS-enabled access points (PSQAPs). This type of framework is intended for use in solar/battery powered wireless mesh networks where power saving on the infrastructure side would be very useful. First, a media access control protocol was proposed, which

uses a novel scheme for selective activation/deactivation of service intervals. SIs are advertised in the beacons using a network allocation map. Using the proposed MAC protocol, a PSQAP can dynamically adjust its duty-cycle by activating/deactivating SIs without affecting the schedule of already admitted QSTAs. Then, a dynamic connection admission control scheme for controlling the air-time usage of the QSTAs was also included. One main advantage of the proposed framework is its backward compatibility since it requires no upgrade to IEEE 802.11 handsets. Additionally, a simple model control parameter was also extracted, which can be used to provide for instantaneous and short-term capacity provisioning and power control at the PSQAP in certain common situations.

Results obtained from simulations show that a WLAN power-saving AP which implements our energy-efficient framework can reduce its power consumption by roughly an order of magnitude, while guaranteeing the QoS requirements of real-time services. These improvements, however, are accompanied with doubling the power consumption values at the handsets. A simple scheme was introduced for distributing the handsets' channel access over different SIs. Simulation results obtained show that when fewer stations contend for the channel in an SI, handset power consumption decreases as a direct result of reduced queuing.

Chapter 4

Efficient Video Transmission over PSQAPs

4.1 Overview

With the advent of QoS-support in IEEE 802.11e [std05a], wireless multimedia transmission across WLANs has been gaining increasing attention in recent years. Among various applications, videoconferencing is one of the most attractive real-time multimedia applications for WLANs. However, due to the characteristics of wireless networks such as high error rate, time-varying channel conditions, limited bandwidth and limited battery power of wireless devices, wireless video applications face many challenges.

In this chapter, we investigate the technical challenges associated with the transmission of video applications over solar/battery powered wireless mesh networks. Using the energy-efficient framework presented in Chapter 3 as a start, we first propose and investigate a connection admission control (CAC) protocol for the video traffic.

The proposed CAC uses a discrete autoregressive model of order one (DAR(1)) to accurately predict the aggregate bandwidth required for multiplexed videoconferencing streams served by a PSQAP. The conflicting relationship between power performance at both the PSQAP and stations in the integrated MAC and CAC framework is studied thoroughly. New scheduling mechanisms are proposed and evaluated for efficient low-power delay-sensitive transmission of multiplexed videoconferencing applications. Analytical models and simulation experiments are used to assess the performance of the proposed CAC and video scheduling algorithms.

4.2 Videoconference Bandwidth Prediction

Although H.264 has been recently introduced as the new video coding standard, H.263 is still very widely used [NCCH06, NMNA06]. H.263 is a video standard that can be used for compressing the moving picture component of audio-visual services at low bit rates. It adopts the idea of *PB* frames, i.e., two pictures being coded as a unit. With this coding option, the picture rate can be increased considerably without increasing the bit rate much [Uni06].

The work in [Kou06, CKP08] used five different long sequences of H.263 encoded videos (from [FR01]) with low or moderate motion, in order to derive a statistical model which fits well with the real data. The length of the videos varied from 45 to 60 minutes and the data for each trace consisted of a sequence of the number of packets per video frame. The authors investigated the possibility of modeling the traces with a number of well-known distributions (gamma, lognormal, log-logistic, exponential, geometric, Weibull, Pearson V); the results in [Kou06] have shown that the use of the gamma distribution (which has been shown in the past to be the best fit for MPEG-1,

MPEG-2 and H.261 videos) is not a good choice for H.263 videoconference traces. The best fit among all studied distributions was achieved for all the traces with the use of the Pearson type V distribution. The superior quality of the Pearson V fit, in comparison to other distribution fits, has been shown with the use of powerful goodness-of-fit tests like Q-Q plots [LK91], Kolmogorov-Smirnov (K-S) tests [LK91] and Kullback-Leibler (KL) tests [BA02]. For each one of these movies, the VBR coding version (in Quarter Common Intermediate Format (QCIF) resolution) was used.

Although the Pearson V was the best fit among all distributions, the degree of goodness-of-fit for the Pearson V varies for all traces. The reason that the Pearson V distribution fit cannot be a perfect fit in any of the examined cases is that the high autocorrelation between successive video frames in a videoconference trace can never be perfectly "captured" by a distribution generating independent frame sizes according to a declared mean and standard deviation, and therefore none of the fitting attempts, as good as they might be, can achieve perfect accuracy. Still, very high accuracy can be achieved for *multiplexed* videoconference sources with the use of a Discrete Autoregressive Model of order one (DAR(1) model). A Discrete Autoregressive model of order p, denoted as DAR(p) [JL79], generates a stationary sequence of discrete random variables with an arbitrary probability distribution and with an autocorrelation structure similar to that of an Autoregressive model. DAR(1) is a special case of a DAR(p) process and it is defined as follows: let V_n and Y_n be two sequences of independent random variables. The random variable V_n can take two values, 0 and 1, with probabilities $1-\rho$ and ρ, respectively. The random variable Y_n has a discrete state space S and $P\{Y_n = i\} = \pi(i)$. The sequence of random variables

X_n which is formed according to the linear model: $X_n = V_n X_{n-1} + (1 - V_n)Y_n$, is a DAR(1) process.

A DAR(1) process is a Markov chain with discrete state space S and a transition matrix:

$$\mathbf{P} = \rho\mathbf{I} + (1 - \rho)\mathbf{Q}, \tag{4.1}$$

where ρ is the autocorrelation coefficient, $\mathbf{I}$ is the identity matrix and $\mathbf{Q}$ is a matrix with $\mathbf{Q}_{ij} = \pi(j)$ for $i, j \in S$.

Autocorrelations are usually plotted for a range W of lags. The autocorrelation can be calculated by the formula,

$$\rho(W) = E[(X_i - \mu)(X_{i+w} - \mu)]/\sigma^2. \tag{4.2}$$

Here, μ is the mean and σ^2 is the variance of the frame size for a specific video trace. A DAR(1) model was built in [Kou06] based on the Pearson V distribution, which was shown through a series of statistical tests to provide high accuracy in modeling traffic from multiplexed H.263 videoconference sources. This model was also used in [CKP08] in order to propose a new call admission control algorithm for videoconference traffic transmission over wireless cellular networks.

In this work, accurate predictions of the DAR(1) model are used to estimate the aggregate bandwidth requirements for multiplexed H.263 videoconference traffic served by a PSQAP that utilizes the proposed energy-efficient MAC and CAC framework. In our study, we use, without loss of generality, one of the five traces studied

in [Kou06] and [CKP08], a video stream extracted and analyzed from a camera showing the events happening within an office ("Office Cam"). The trace has a mean of 90.3 Kbps, a peak of 1 Mbps and a standard deviation of 32.7 Kbps.

4.3 Connection Admission Control and Resource Allocation

This section introduces a CAC scheme used by a PSQAP to efficiently control the airtime usage of videoconferencing QSTAs. The PSQAP employs the proposed energy-efficient framework to extract power saving while serving real-time video stations. Since TXOPs are granted on a time-basis, the CAC unit at the PSQAP needs to convert the rate-based TSPEC parameters into time-based ones. The DAR(1) model, explained in Section 4.2, uses the statistical data of multiplexed video traces to predict the channel resources required on a service interval (SI) basis. In other words, the capacity offered by the PSQAP to accommodate video traffic in the ith service interval, SI_i, is determined as $TB_i = f(M(i))$, where $M(i)$ is the total number of admitted H.263 videoconference stations in SI_i. $f(M(i))$ is calculated as

$$TB_i = [DAR(1)_{M(i)} + M(i) \ (\frac{O}{R} + T_{overheads})] \ SF(i-1) \quad (s), \quad (4.3)$$

where $DAR(1)_{M(i)}$ is the predicted value from the DAR(1) model, of channel-time (s) required to transmit uplink packets of $M(i)$ video stations in SI_i, O is the additional per video packet overheads, i.e., MAC/IP/UDP/RTP headers (in bits) and $T_{overheads}$ denotes the required per packet IEEE 802.11e MAC overheads including propagation

delay, PHY preamble, SIFS, binary exponential backoff, AIFS and ACK (s). R is the minimum physical transmission rate (bit/s) required to guarantee the QoS for a real-time flow and is obtained from TSPEC elements. $SF(i-1)$ is the average *SurplusFactor* for the multiplexed video streams in the (i-1)th service interval, and is used to compensate for the traffic loss caused by packet collisions or channel error. For the set of admitted video flows, the PSQAP updates $SF(i+1)$ using a running weighted average

$$SF(i+1) = \gamma SF(i) + (1-\gamma)\frac{T_{Access}(i)}{T_{Success}(i)}, \tag{4.4}$$

where $T_{Access}(i)$ and $T_{Success}(i)$ are the measured channel access times for the multiplexed video TSs regardless of success, and, all successful transmission times as measured by the PSQAP, respectively. γ is a smoothing factor and is used to control the effect that fluctuations in T_{Access} and $T_{Success}$ can cause to the values of the *SurplusFactor* and $0 \leq \gamma \leq 1$. By experimenting with various values of γ, it was found that large values ($\simeq 1$) make the CAC more immune to instantaneous packet transmissions statistics observed on the wireless channel. In other words, the update of SF, hence the activity sub-interval length of a PSQAP, is done sluggishly. On the otherhand, the use of small γ values results in a high packet-loss rate due to the rapid update of the activity sub-interval and the associated offered capacity.

TB_i is the maximum allowable channel-time occupancy for videoconferencing traffic in a given SI, SI_i. Whenever a video packet is sent, regardless of success or not, the same amount of access time is subtracted from TB_i. A QSTA is not allowed to transmit any video packet once the corresponding TB_i is depleted. To decide upon the admission of a new videoconference flow TS_j that arrived in SI_i, the centralized

Table 4.1: Simulation Parameters for a PSQAP Serving Superposed H.263 Traffic

Parameter	Value
Number of APs	1 (PSQAP)
Beacon interval (B)	$80ms$
R	$11Mbps$
Maximum Retry Limit	3
Maximum MPDU size	2304 *bytes*
UPD / RTP header	20 / 12 *bytes*
MAC / IP header	34 / 8 *bytes*
SIFS / Slot time	$10/20\ \mu s$
T_{ACK} / T_{PHY}	$248\mu s$ / $192\mu s$
$\eta_{Tx}/\eta_{Rx}/\eta_D$	$750/500/2\ mW$

CAC unit at the PSQAP first assumes as if TS_j was actually admitted, and then it checks whether the following inequality is true, $TB'_i < t_{CP_i}$. Otherwise, TS_j is rejected. TB'_i is calculated as in Equation 4.3 but with $M(i)$ replaced with $M(i)+1$, to predict the required bandwidth for serving $M(i)$ video flows in addition to the newly arrived TS_j. t_{CP_i} is the admissible portion of t_{SI_i} allocated for contention access and t_{SI_i} is the length of SI_i.

The enforced AP power saving feature already presented in Section 3.3.1, is also enabled in the CAC scheme proposed for multiplexed videoconferencing flows. In this scheme, $t_{CP} = \alpha t_{SI}$ and $0 < \alpha < 1.0$ so that the PSQAP can use $(1-\alpha).t_{SI}$ as potential sleep time. Below, we show how to use the prediction-based CAC algorithm for multiplexed H.263 sources to design efficient bandwidth allocation and capacity scheduling algorithms for low-power transmission of H.263 videoconferencing applications over power saving PSQAPs.

4.4 Performance Analysis

This section reports on results obtained from both analytical models and comprehensive simulation studies which were conducted to evaluate the power performance of the proposed energy-efficient framework, compared with well-known existing schemes and the theoretical optimal performance. Since the focus of this work is on QoS satisfaction of the real-time videoconferencing streams served by a PSQAP, only uplink traffic is considered in this study. This allows for a better analysis of the various effects on QoS perceived by video stations when the PSQAP alternates between sleep and awake periods. Performance metrics considered include average packet collision probability and power consumption for both the PSQAP and handsets. The performance study uses a video stream which was extracted and analyzed from a camera showing the events happening within an office ("Office Cam"). The trace has a mean of 90.3 Kbps, a peak of 1 Mbps and a standard deviation of 32.7 Kbps.

4.4.1 Simulation Methodology

Simulations were performed using a discrete event simulator "simlib" written in C for the contention based EDCA to meet the specifications of both the IEEE 802.11e standard and the proposed low-energy framework. Also, the simple scheduler defined in IEEE 802.11e was modified to incorporate the proposed accurate prediction-based admission control scheme. The results presented are for a one-hop access network model with contention-based end station traffic. To achieve power management, the PSQAP maintains status information for each associated mobile station.

Before each beacon transmission, the PSQAP calculates the length of EDCA-activity sub-interval, TB_i, using the DAR(1)-based CAC scheme as explained in

Section 4.3. Throughout the experiments, we do not account for excess time allocation due to packet errors and retransmissions. This is to better evaluate and compare the bandwidth allocation and PSQAP/handset power performance for different CAC schemes presented and considered in this study. For the prediction-assisted scheduling schemes, once the expected transmission time for UL video packets transmission has been derived, the scheduler located at the PSQAP calculates the start and end times of both the active and sleep sub-intervals of the SI and includes this information as part of the NAM field of the next beacon frame.

To accurately capture all the components of power consumption at both handsets and PSQAPs, we used an energy model derived from the work in [FT07], where $\eta_{Tx} = 750mW$, $\eta_{Rx} = 500mW$ and $\eta_D = 2mW$ are the assumed power dissipation in transmit, receive and doze modes, respectively. Modern chipsets may keep the oscillators running while the radio is off for the cost of dissipating extra power ($\simeq$ 6mW compared to 2mW) in order to save most of the wake-up time. In this case, it takes only about 0.5ms to wake-up RF circuitry rather than 2.5ms. However, in this work, we assume that oscillators are switched off while the radio is in sleep mode. Time for wake-down (i.e., active to sleep) transition is assumed to be 0.5ms. To better simulate the activities of the wireless interface card, the energy usage during the active-doze transition is modeled as 125μJ and the energy usage during the doze-active transition is assumed 250μJ, respectively [KTKS08]. The difference in values is due to the fact that when the radio is on, the current I is at its maximum value compared to $10 - 15\%$ only of its maximum value when the radio is off, while the voltage V remains constant during state transitions.

Some statistical data that indicate the burstiness of an H.263 trace include the

minimum, maximum and average video frame sizes which were found to be (from the video trace file) 254, 5191 and 903.8 bytes, respectively. The variance in frame size is 107157.06. A packet is discarded if it exhausts the maximum retry limit (which was set to 3 [WLL05]), or if it exceeds the maximum delay bound, which is assumed to be equal to the packet inter-arrival time, i.e., $80ms$. In other words, a video packet must be served before the arrival of the next packet, otherwise it is dropped. A stringent upper bound on packet loss probability of 0.0001 is assumed [DH99]. Throughout the study, the packet loss probability was ensured to be well below the specified upper bound. All results were obtained by averaging 10 simulation runs, and the simulation time for each was 45min., which was sufficient to eliminate any transient effects. Results were obtained with margins of ± 0.0012 (95% confidence interval). Both the PSQAP and handsets are assumed to be within the same transmission range, therefore, RTS/CTS protection is not used. Successful packet receptions are acknowledged by the destination nodes. The different simulation parameters used in the experiments are summarized in Table 4.1. The results presented in the following sections show a significant improvement in AP power conservation with the use of the proposed integrated infrastructure power-saving framework which consists of a prediction-based adaptive CAC and MAC schemes for serving real-time video traffic.

4.4.2 Selection of Video Capacity Provisioning Scheme

In the proposed energy-efficient framework, a PSQAP should provide adequate capacity in order to satisfy QoS requirements of real-time multiplexed H.263 video flows, before it can switch to sleep mode for power saving extraction. In this experiment, a superframe is assumed to have only one SI which consists of two sub-intervals, i.e., an

EDCA activity sub-interval and an (AP power saving) sleep sub-interval, as shown in Figure 4.1(a). The boundary between these two sub-intervals is moveable and is dynamically updated based on the offered traffic load. New stations which join the basic service set (BSS) and sample the beacon frame will learn the start/end boundaries of active/sleep sub-intervals advertised in the NAM.

First, we investigate and compare three design approaches that a PSQAP could use to decide on the time instant, within a SI, at which the PSQAP terminates the contention-based EDCA-activity sub-interval in order to start the power-saving (i.e., sleep) sub-interval. The objective is to determine which video capacity provisioning scheme is efficient for a PSQAP serving H.263 videoconference QSTAs. The study considers the power performance of both the PSQAP and MSs. The SI length, t_{SI}, is assumed $80ms$ which closely matches the inter-arrival time for H.263 video packets. For simplicity, a superframe interval is set to have one SI only, i.e., $T_{SF} = 80ms$. All three schemes considered in this experiment assume the contention-based US-APSD (EDCA) channel access mechanism during the activity sub-interval.

Prediction-based (DAR(1) Model) In the first approach, a PSQAP uses prediction from DAR(1) model, presented in Section 4.2, to make resource allocation decisions for multiplexed H.263 video streams. Based on the number of video streams being served, the scheduler at the PSQAP determines the EDCA interval length required to accommodate UL video packet transmissions such that the stringent packet loss bound of 10^{-4} is satisfied. Based on a specified maximum video-packet loss rate, the DAR model uses the bit-rate statistics of individual videoconference sources to predict the aggregate bandwidth requirements

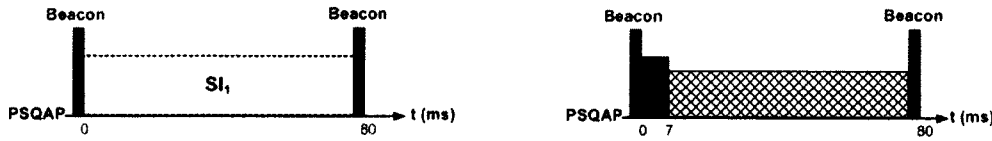
that will meet the target packet loss bound for the multiplexed H.263 videoconferencing flows. The aggregate bandwidth predictions calculated by the DAR model Because of the dynamic nature of the DAR(1)-model bandwidth prediction algorithm, the predicted aggregate capacity might not be identical in subsequent SIs, for a given number of admitted video streams. To better study the effectiveness of using DAR(1) prediction-based CAC, we have also studied the performance of two other well-known multimedia traffic modeling schemes. Therefore, in these experiments, instead of using DAR(1) prediction to calculate the length of EDCA activity sub-intervals, we use the analytical models provided by the authors in [VPV88] and [vdSAH06]. The model in [VPV88] uses statistical parameters, i.e., average bit rate, the bit rate variance and the peak bit rate, to calculate the aggregate bandwidth demands for a given number of multiplexed VBR multimedia streams given a packet loss performance. On the other hand, reference [vdSAH06] uses a twin-leaky bucket analysis to calculate the effective bandwidth of a flow based on its TSPEC information. Unlike the DAR(1) model, capacity estimates calculated by these schemes are static, i.e., calculations depend only on the statistical characteristics of traffic flows and are time-invariant (i.e., do not change for the same set of traffic streams).

Wait Until Complete In the second approach presented in this study, "Wait-Until-Complete", the PSQAP immediately sleeps following the completion of the last, i.e., N*th* video packet transmission, assuming there are N admitted video streams at the PSQAP. Since all real-time multimedia flows, including video streams, must establish a connection through the AP, the implementation of such a scheme is straightforward once the AP knows the total number of streams

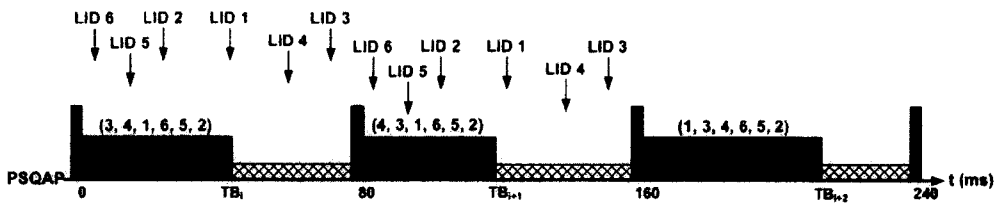
admitted.

Sleep on Empty Queues We also present and evaluate a third video capacity provisioning approach in which the PSQAP switches to sleep mode only when all QSTAs' queues are guaranteed to be empty. In "Sleep-on-Empty-Queues" scheme, the PSQAP does not necessarily sleep following the transmission of exactly N video packets. Instead, since the PSQAP already knows the exact packet arrival times at all individual stations (because video packets inter-arrival time is fixed at 80ms), the PSQAP sleeps only when it is sure that all queues are empty. The PSQAP is assumed to always monitor the wireless channel activity during the EDCA subinterval. After N packet transmissions, if the channel is sensed idle for a threshold, τ, then the PSQAP concludes there are no more packet transmissions and it can then terminate the activity period and start power saving. The value of τ is determined based on the EDCA parameters used for different priority access category traffic (i.e., $AIFS$, CW_{min} and CW_{max}).

The simple example in Figure 4.1(a) shows a PSQAP that is initialized with an EDCA activity interval set to 7ms, leaving the remainder of the SI for AP power saving. The PSQAP may dynamically update the size of EDCA activity sub-interval in order to either accommodate newly accepted video traffic streams (TSs) or to maximize its power saving extraction when the channel becomes lightly loaded. A QoS station (QSTA) may contend for channel access during the activity sub-interval to request the admission of a real-time video TS. Figure 4.1(b) shows the packet arrival times for 6 H.263 video QSTAs relative to the beacon transmission time. Although the number of handsets is assumed fixed in the figure, we can note that



(a) Structure of a Superframe with a Single EDCA SI.



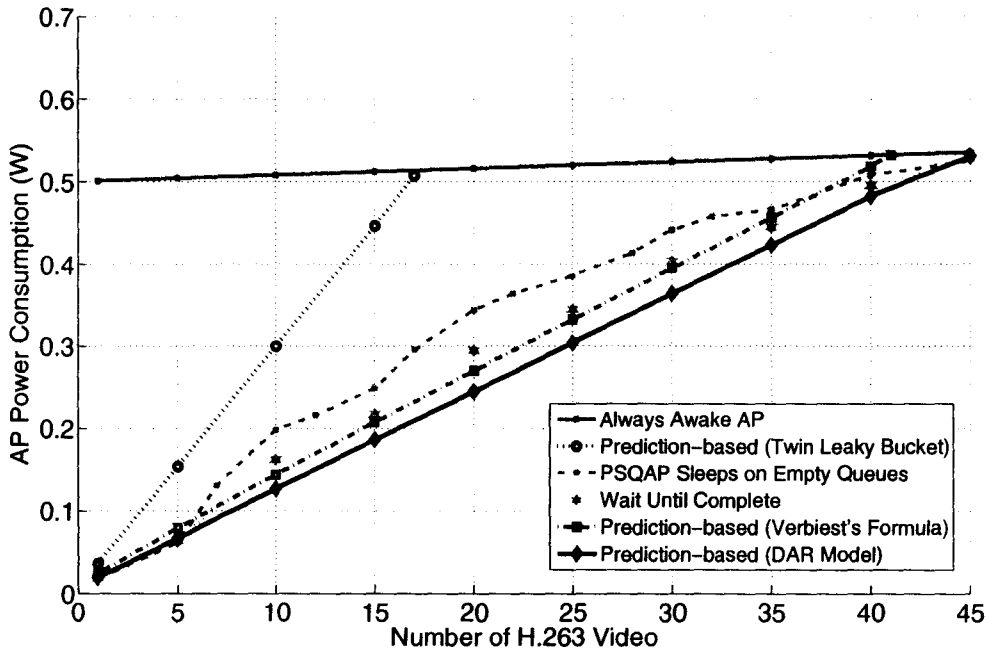
(b) Example of Relative Packets Arrival Times and Dynamic Capacity Provisioning for the Basic Scheme (One EDCA SI).

Figure 4.1: Superframe Structure and Timeline for a PSQAP Serving Videoconference QSTAs Using One US-APSD SI.

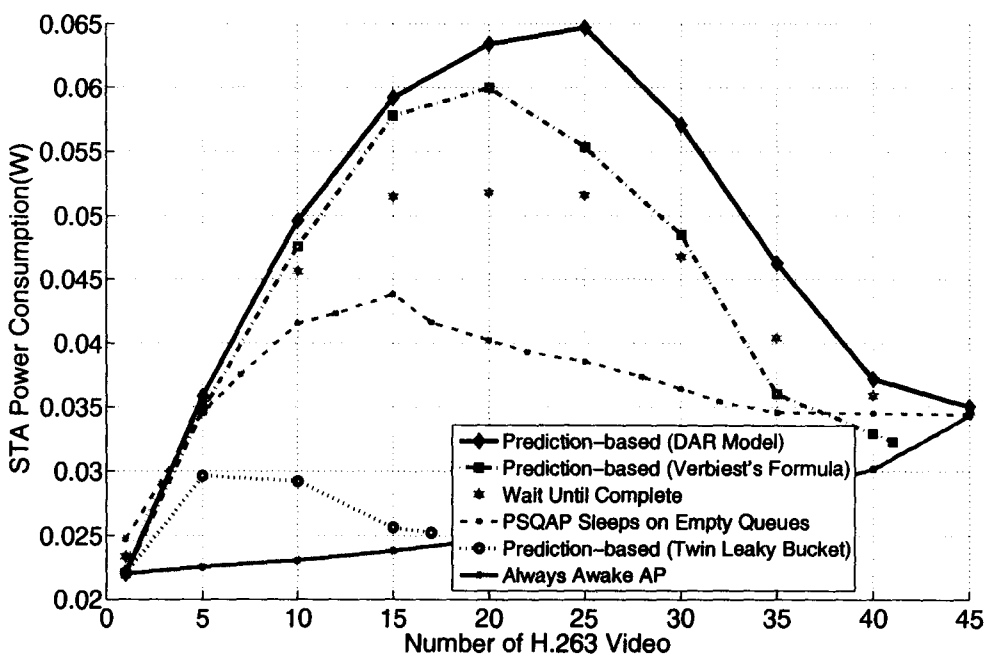
$TB_i \neq TB_{i+1} \neq TB_{i+2}$. This is due to the variable length of video packets which is also accurately captured by the DAR(1) model through a dynamic bandwidth prediction that updates every service interval.

Once a station successfully associates with the PSQAP, a station identifier (referred here as Logical Identifier "LID") is created and assigned by the PSQAP. From the figure, Stations LID 3, LID 4 and (LID 1 in superframe $i + 1$) generate their video packets while the AP is sleeping; therefore, these stations must awaken and contend for channel access at the beginning of every SI. As we will show later in this chapter, reducing contention level and the associated *queuing* is crucial for improving the handset power performance. Due to the random nature of the channel access in EDCA, it can be seen that the stations' channel access sequence (shown on top of each activity sub-interval in Figure 4.1(b)) varies across subsequent SIs.

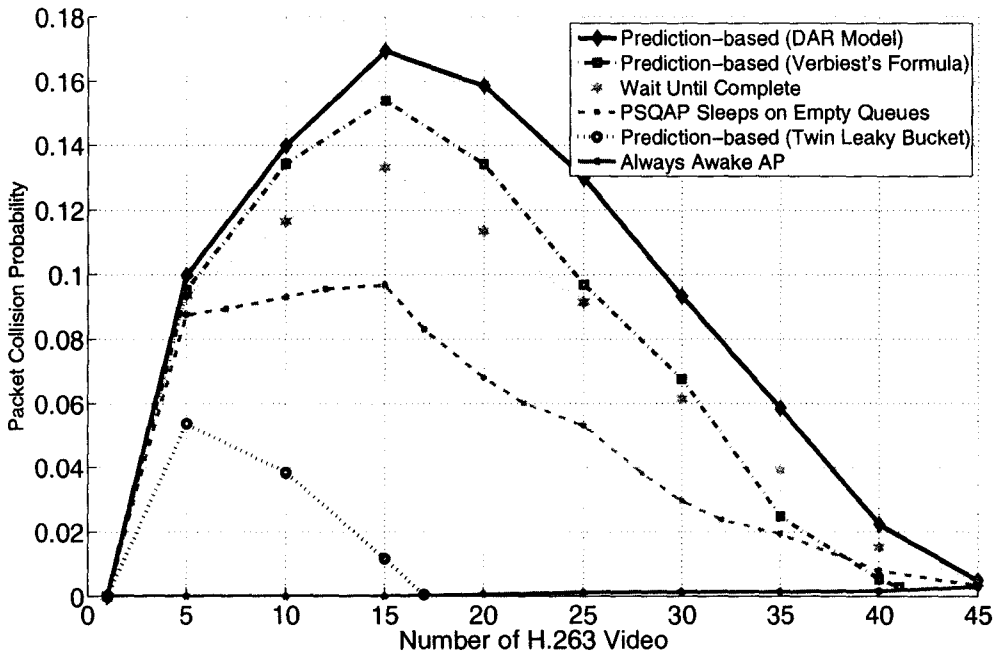
Performance results for the three video capacity provisioning schemes, namely, "Prediction-based (DAR Model)", "Wait-Until-Complete" and "PSQAP Sleeps on



(a) PSQAP Power Consumption.



(b) Video QSTA Power Consumption.



(c) Packet Collision Probability.

Figure 4.2: Power Performance Results For Basic Scheme (1 SI - EDCA).

Empty Queues" in addition to the other two bandwidth prediction algorithms "Prediction-based (Verbiest's formula)" and "Prediction-based (Twin Leaky Bucket based CAC)" are shown in Figure 4.2. In Figure 4.2(a), we can see that, except for the Twin-Leaky Bucket based algorithm, prediction-based video scheduling achieves better power saving at the AP side compared to "Wait-Until-Complete" and "PSQAP Sleeps on Empty Queues" schemes. Furthermore, it can also be seen that bandwidth prediction and capacity provisioning for video QSTAs using DAR(1) prediction are more energy-efficient for the power-saving AP than the other two capacity-estimation models in [VPV88] and [vdSAH06]. Because of accurate aggregate capacity calculations using the DAR(1) model, the proposed scheme is also capable of admitting more multiplexed video streams than the other two schemes (from the figure, the

"Prediction-based (DAR Model)" scheme can admit 4 more video flows (10%) than the "Prediction-based (Verbiest's formula)" scheme and about 27 more flows (150%) than the "Prediction-based (Twin Leaky Bucket)" scheme).

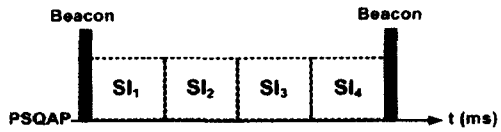
For the sake of comparison, we also show results for a conventional always awake AP (i.e., with no power saving). The effectiveness of the "Prediction-based (DAR Model)" scheme is even more evident compared to what the current IEEE 802.11 standard(s) can offer in terms of AP power saving. From Figure 4.2(a), a PSQAP that uses DAR-prediction for bandwidth allocation needs only $10-50\%$ of the total power consumed by an always awake AP to serve the same low/moderate video traffic load. In practical deployments of wireless mesh networks, mesh APs are rarely provisioned for high (i.e., saturation) load values. Therefore, power conservation possible through the proposed framework could significantly reduce the total cost of typical sustainable solar-powered WMN installations.

On the other hand, it can be seen from Figure 4.2(b) that the DAR(1) prediction-based capacity scheduling algorithm exhibits the highest handset power consumption values compared to other capacity prediction and bandwidth allocation schemes. This is true since for the proposed DAR(1) prediction-based scheduling algorithm, the difference between the estimated capacity (i.e., length of activity sub-interval) and the channel time actually occupied by video stations is relatively small. Hence, the PSQAP can significantly increase its power saving at the expense of queuing more video packets that are generated during the sleep sub-interval. These video packets will have to queue from the start of the next SI and fight contention. The more accurate the bandwidth-prediction is, the shorter is the awake sub-interval (the longer is the sleep period), the more packets are queued and more QSTAs will have to awaken

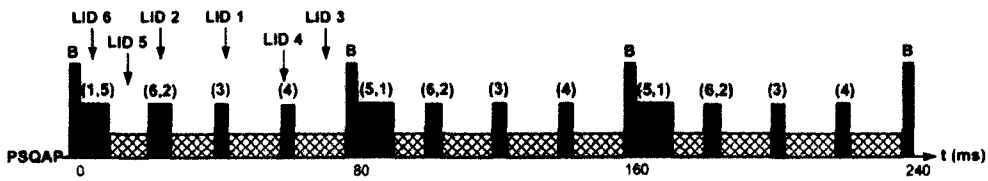
and contend at the beginning of SIs (Figure 4.2(c)). While the increased queuing results in an unnecessary increase in handsets' power consumption, the PSQAP consumes less power when packet transmissions of different QSTAs are packed along the time line and idle intervals are minimized. Because of its superior prediction quality, the average MS power consumption values for the DAR(1) model are higher than those achieved by the Verbiest's formula based CAC and are much higher than those of the Twin leaky bucket based CAC.

For all three curves of the prediction-based scheduling schemes shown in Figure 4.2(b), the handset power consumption scales well with large numbers of video flows admitted (relative to each scheme). From the figure, we can see that, for high load values, MS power consumption drops fast to reach that of a handset served by an always awake AP. The explanation for this effect is that as the number of admitted video stations increases, the offered bandwidth also increases to accommodate the increasing video load. The increase in AP's duty-cycle results in a corresponding and similar decrease in the queuing and contention observed by QSTAs at the beginning of SIs. When the number of admitted stations is *large* enough (depending on the algorithm used for capacity estimation), the contention level dramatically decreases and starts to approximate the contention experienced by stations served through a non-power saving AP (Figure 4.2(c)).

To conclude this section, we reemphasize that the use of prediction-based capacity provisioning lends itself as the most energy-efficient scheduling mechanism for a PSQAP serving variable bit-rate real-time multimedia traffic. Additionally, because of its proven estimation accuracy, the DAR(1) model is used to calculate the aggregate capacity requirements for multiplexed H.263 video streams. However, in order



(a) Structure of a Superframe Supporting Split Access with Multiple EDCA (US-APSD) SIs.



(b) Example of Relative Packets Arrival Times and Dynamic Capacity Provisioning for the Prediction-assisted with Split Access over 4 EDCA SIs Scheduling Scheme.

Figure 4.3: Superframe Structure and Timeline for a PSQAP Serving Videoconference QSTAs Using Split Access Scheme.

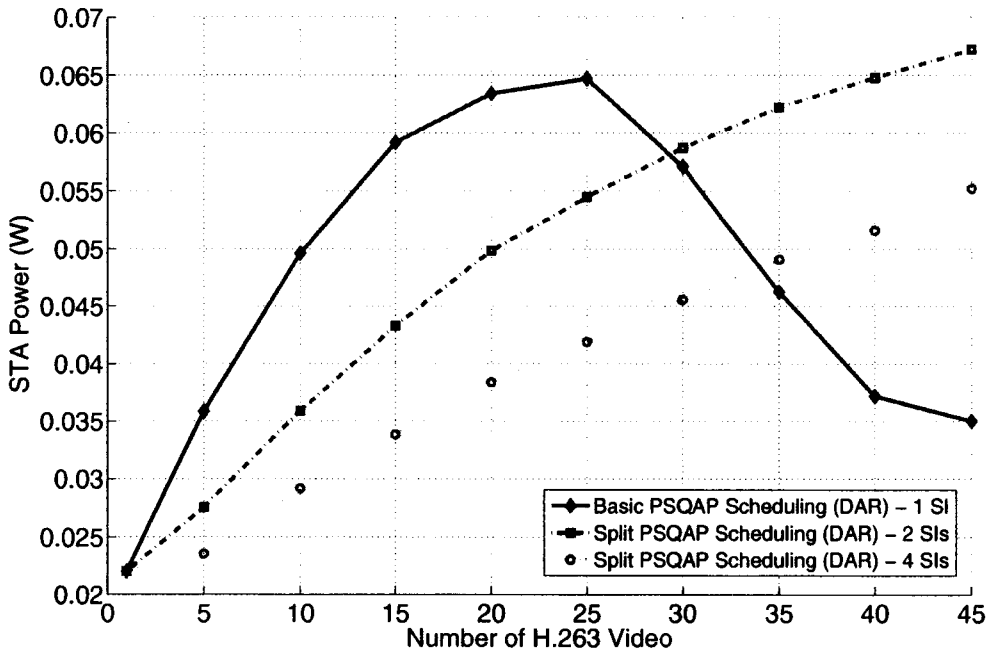
to optimize for power performance at the handset side, we need to modify the single SI design presented in this section such that the PSQAP schedules the offered capacity (i.e., EDCA activity sub-intervals) in a way that minimizes the contention and queuing experienced by QSTAs. Results presented in this section also show that it is impossible to simultaneously optimize the power performance at both the PSQAP and handsets. This is due to the conflicting relationship between power saving at both the AP and handsets in a contention-based channel access mechanism. However, in Sections 4.4.3 and 4.4.4 we show that a near-optimal low-energy design is possible and we also show how to successfully achieve it.

4.4.3 Performance of Split Access Over Multiple EDCA SIs

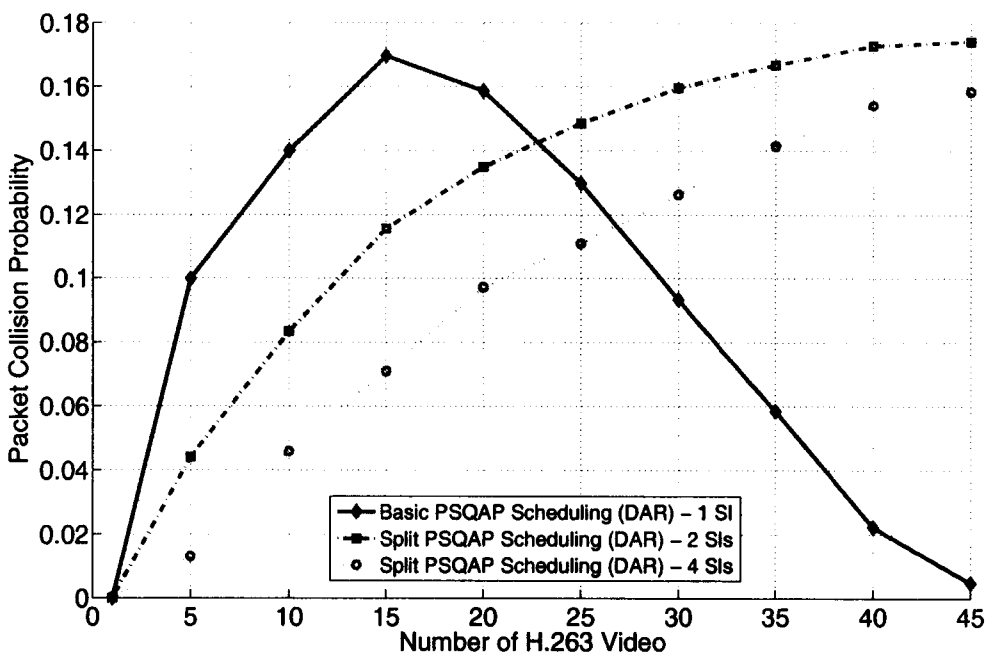
In this section, we study the performance of an improved version of the prediction-based with a single EDCA SI framework design for PSQAPs presented in Section 4.4.2.

In this improved scheme (Figure 4.3), the superframe interval is divided into a number of EDCA (US-APSD) SIs of shorter length, as shown in Figure 4.3(a). In order to overcome the unnecessary increase in handset power consumption due to contention and queuing effects, we propose a splitting scheme in which, instead of having all stations fight contention in a single SI, a load balancing is performed to ensure that video stations are uniformly distributed over different SIs (Figure 4.3(b)). Once a TS is accepted, the PSQAP instructs the corresponding QSTA to contend for channel access using US-APSD in a specific SI that meets the requirements of the video TS. Admitted QSTAs are only permitted to access SIs that are assigned to them by the PSQAP. For simplicity, all admitted video streams are assumed to have identical TSPEC parameters. Therefore, optimal load balancing could be easily achieved by distributing the video QSTAs evenly across different SIs in a superframe. A simple splitting criterion based on station's identifier is used, in which a station LID i, is assigned an SI_j where $SI_j = LID\ i\ mod\ N_{SI}$, and N_{SI} is the number of SIs in a superframe interval. For example, in Figure 4.3(b), Stations LID 1 and LID 5 are assigned SI_1, and Stations LID 2 and LID 6 are assigned SI_2 ($N_{SI} = 4$). Based on the number of H.263 videoconference QSTAs contending in each SI, the DAR(1) model-based connection admission control scheme presented in Section 4.3 is used to calculate the length of activity periods (as in Equation 4.3). According to the discussion in Section 4.4.2, this allows for minimizing the AP's power consumption.

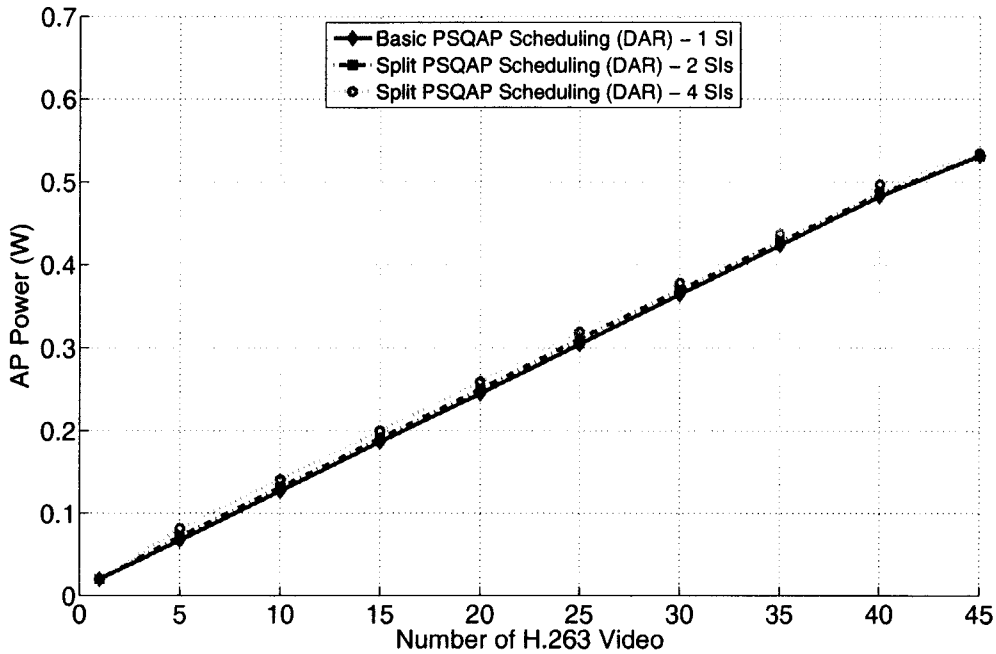
Results in Figure 4.4(a) show that prediction-based scheduling schemes with multiple SIs achieve a significant reduction in handset power consumption over the basic scheme with one SI presented earlier. In the same figure, it can be also seen that the more SIs used to accommodate stations' packet transmissions, the better is the



(a) Video QSTA Power Consumption.



(b) Packet Collision Probability.



(c) PSQAP Power Consumption.

Figure 4.4: Splitting Scheme.

power saving achieved at the handset side. In numbers, the use of 4 SIs, for example, saves about $20\% - 43\%$ while the use of 2 SIs saves about $3\% - 28\%$ of handset power consumption when compared to the similar prediction-based scheme which uses one SI only. However, we can see that this improvement is possible as long as the number of superposed H.263 TSs is < 30. Beyond this load value, the handset power consumption continues to increase linearly as the number of video streams increases and the multiple-SIs scheme (for both 2 and 4 SIs) loses the attractive scalability characteristic exhibited by the single SI scheme for high load values (Figure 4.2(b)). The reason is, unlike the case for a single EDCA SI, a video QSTA in the Split Access scheme is only allowed to access the channel in the SI specified by the PSQAP during connection setup time. Therefore, a video packet that is generated during an

awake period of the PSQAP cannot be transmitted if the arrival SI is different from the SI assigned to the handset. While Split Access with Multiple SIs significantly reduces queuing and collisions for low to moderate video loadings, the scheme fails to maintain this property as more stations queue from the start of the SIs.

From a PSQAP point of view, Figure 4.4(c) shows that both "Prediction-based with Single EDCA SI" and "Prediction-based with Multiple EDCA SIs" schemes exhibit almost the same AP power performance. This is expected since both schemes use the prediction-assisted CAC scheme in Section 4.3 to calculate the length of activity sub-intervals. The figure also shows that a power saving AP incurs extra power due to frequent on-off state transitions when using the multiple-SIs capacity provisioning approach. However, this extra power is negligible and can be ignored compared to the total PSQAP's power consumption.

4.4.4 Performance of Using IEEE 802.11n Unscheduled-PSMP

This section presents and evaluates the performance of a more sophisticated energy-efficient framework design that utilizes a modified version of the IEEE 802.11n US-PSMP power management mechanism to further improve the power consumption at the handset side. PSMP uses a feedback based channel access mechanism to serve packets. In the standardized US-PSMP, a PSMP-capable station awakens and sends (using conventional CSMA/CA) a trigger frame to inform the AP that it has one or more packets to send (or to indicate its availability for receiving packets at the DL). The queue length of the station is included in the trigger frame. Based on the queue information provided from all awake PSMP-capable stations, the scheduler unit at the AP allocates uplink and/or downlink transmission times for each station

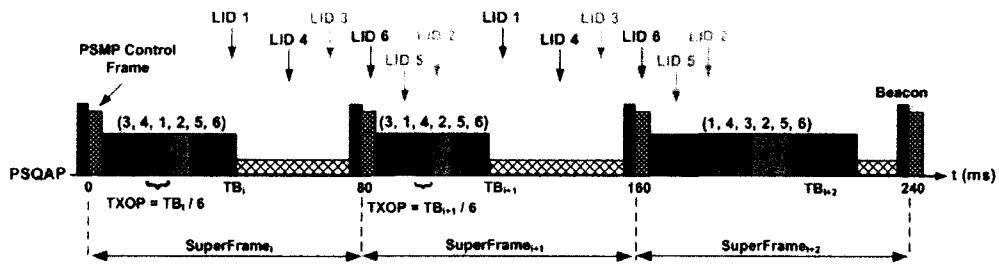
as part of a PSMP frame. A PSMP frame starts with a PSMP Action frame, which contains the time schedule to be followed by both the AP and stations during a PSMP frame. All downlink packet transmissions should complete first before uplink transmissions can take place. In the context of the original PSMP, the scheduled uplink and downlink slots are viewed as AP's owned intervals, therefore stations are not allowed to perform normal carrier sensing procedures. PSMP efficiently minimizes the contention by scheduling TXOPs that are large enough to empty stations' buffers and by eliminating the need to perform conventional CSMA/CA and the associated binary exponential backoff before transmitting a packet.

It's likely that if PSMP operation is implemented by vendors, it will be unscheduled mode that gets the call because it simply extends IEEE 802.11e US-APSD, which was adopted by the Wi-Fi Alliance. This, in addition to other reasons previously outlined in Section 2.3, are the motives behind choosing the unscheduled channel access mode of the PSMP for our proposed energy-efficient PSQAP framework. Recall that in standardized US-PSMP, trigger frames are sent using conventional CSMA/CA, which can severely affect the power performance of MSs that awaken and fight the contention at the beginning of SIs. Therefore, in this proposal for low-energy transmission of videoconferencing applications, we make use of the fact that video packets arrivals are almost fixed and hence a PSQAP can periodically assign PSMP uplink/downlink TXOPs to stations without the need to receive trigger frames. By using the AP's knowledge of packet arrival times at individual stations, we can safely eliminate the need for trigger frames while saving much of the channel capacity and station's power consumption. In order to determine the arrival-times for subsequent video packets, a PSQAP needs only to know the first packet arrival-time in an H.263

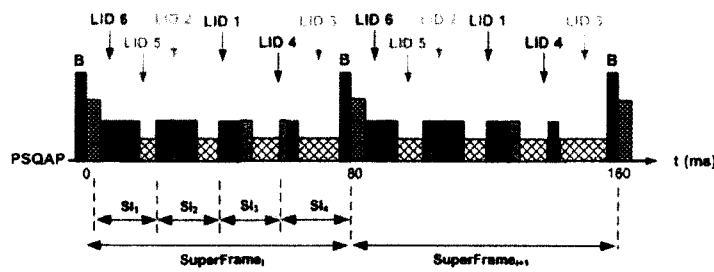
stream. The TSPEC element of IEEE 802.11e contains a filed "Service Start Time" that indicates for both the AP and the handset when QoS guarantees should be provided.

Below, we present and evaluate the performance of two different scheduling schemes that utilize our modified version of the original IEEE 802.11n US-PSMP. The proposed scheduling schemes differ in terms of (i) number of times a handset would need to awaken and attempt channel access before it can successfully transmit a video packet, (ii) number of SIs per superframe and (iii) the mechanism by which a PSQAP determines the length of UL TXOPs in a PSMP frame. For each scheme, the performance on both the PSQAP and handsets is studied, and we show that by combining accurate capacity-prediction using the DAR(1) model-based CAC algorithm (Section 4.3) and by minimizing the number of state transitions, a near-optimal power performance can be achieved for QSTAs and the power saving QAP.

Prediction Assisted US-PSMP with One Service Interval In the first PSMP-based capacity scheduling scheme, a PSQAP, which is assumed to have knowledge about exact video packet generation times, broadcasts a PSMP Action frame following the beacon transmission. During connection-setup time, the AP sets the value of the 'Service Start Time' field of the TSPEC element to indicate when it is ready to serve the QSTA's backlog. Each PSMP frame starts with a PSMP Action frame that contains the schedule for UL and DL TXOPs of individual QSTAs. The schedule information is formatted as a 2-D table where for each station's *LID*, the corresponding TXOP start time is recorded as an integer multiple of time units of length $32\mu s$ each, referenced to the PSMP



(a) Example of Relative Packets Arrival Times and Dynamic Capacity Provisioning of the US-PSMP with 1 SI Scheme.



(b) Example of Relative Packets Arrival Times and Dynamic Capacity Provisioning of a Scheme Uses Multiple US-PSMP SIs.

Figure 4.5: Timeline for a PSQAP Serving Videoconference QSTAs Using Unscheduled-PSMP SIs.

Action frame transmission time. In this scheme, the length of all TXOPs in a PSMP frame is identical and is determined as TB_i/N, where TB_i is the activity interval (i.e., PSMP frame) length allocated for N admitted video QSTAs and is estimated according to Equation 4.3.

While the IEEE 802.11n standard does not specify CSMA/CA within a PSMP frame, the proposed scheduling scheme does allow for conventional contention, although it significantly minimizes the resulting queuing and collision effects. This is true since in conventional US-PSMP, an AP would first receive a trigger frame that has the queue length information at the handset, and it then grants a TXOP that is only large enough to accommodate the station's packet transmission(s). However, in this work, trigger frames are eliminated to further reduce the power consumption overhead and instead we use prediction to estimate the TXOPs. Therefore, under the proposed scheme, queuing and contention between two or more QSTAs with neighboring TXOPs is inevitable if a TXOP is insufficient to accommodate a packet transmission. In this case, all neighbor stations, which have TXOPs assigned later in the PSMP frame, will have to wait for this station's transmission to complete first before they can start their TXOPs and CSMA/CA should then be utilized to avoid/reduce collisions.

The order in which TXOPs are allocated to QSTAs in a PSMP frame is calculated based on an Earliest-Deadline First (EDF) scheduling. In the first round of the algorithm, videoconferencing QSTAs are first sorted in an ascending order based on their packets' transmission deadlines. Since all videoconference streams are assumed to have identical delay bound and packet inter-arrival time of 80ms, the sorting is done based on packets' arrival times relative to

the target beacon transmission time (TBTT). TXOPs are then assigned in a reverse order, i.e., videoconferencing QSTAs with early deadlines are assigned TXOPs late in a PSMP frame, while QSTAs with late deadlines are assigned TXOPs early in a PSMP (Figure 4.5(a)). The idea here is that packets generated early in a superframe and thus have the earliest deadline(s) (e.g., LID 6, LID 5 and LID 2 in Figure 4.5(a)) are served within the same superframe, while other video packets with the latest deadline(s) (e.g., LID 3, LID 4 and LID 1) are guaranteed service early in the PSMP frame following the superframe when they arrived.

When the first round of the algorithm completes, the produced TXOP assignment is guaranteed to meet the specified packet delay bound for most of the videoconferencing QSTAs. However, for very few video stations, which usually occupy the middle of the sorted QSTAs list (and are thus assigned TXOPs in the middle of a PSMP frame), their assigned TXOPs could result in discarding video packets because of exceeding the packet transmission delay-bound. Therefore, a second round of the algorithm is performed in which the satisfaction of delay bound for each QSTA is tested against its assigned TXOP. For a station $QSTA_i$ (assumed to be located at position i in the sorted QSTAs list), if the assigned TXOP is found to violate the specified delay bound, then the algorithm searches for a potential candidate among all handsets that follow $QSTA_i$ in the list (i.e., QSTAs with later deadlines). Once such a candidate is found, both its TXOP and the $QSTA_i$'s TXOP are swapped, and the test restarts from the beginning of the list and only stops when the produced TXOP assignment satisfies the delay bound for all video stations.

It is important to notice that the proposed modification to the original US-PSMP, where queue status feedback is replaced with prediction to determine the TXOP length, poses challenges on how to determine the optimal TXOP length. As we will show later in this chapter through detailed simulation results, there is an interesting tradeoff between TXOP length, station power consumption, PSQAP power consumption and channel utilization. Through a comparative study, three different mechanisms are presented and evaluated for determining the TXOP length. The first mechanism is based on worst case scheduling, where the maximum video packet size is used to schedule TXOPs for video packets transmissions. In the second mechanism, prediction from DAR(1)-model is used to obtain a strong estimate of the total bandwidth that will be needed for all terminals, and the TXOP length of a handset is set to $1/N$ of that estimate. We show that this second mechanism is the most energy-efficient capacity scheduling approach and it does achieve near-optimal power performance for both the PSQAP and the handsets. The third mechanism is the limiting case, where TXOP $\rightarrow 0$, and in which videoconferencing QSTAs awaken at the beginning of an activity sub-interval to transmit their UL packets using conventional EDCA (US-APSD). Note that for the limiting case, the scheme works in an identical manner to the prediction-based with 1 EDCA SI scheme presented earlier in Section 4.4.2.

US-PSMP with Multiple Service Intervals (No-Prediction) We now present a scheme where a PSQAP schedules 4 SIs per superframe, each has a PSMP frame, to serve the video backlog at the handsets (Figure 4.5(b)). Similar to the US-PSMP variant presented above, the superframe length is assumed to

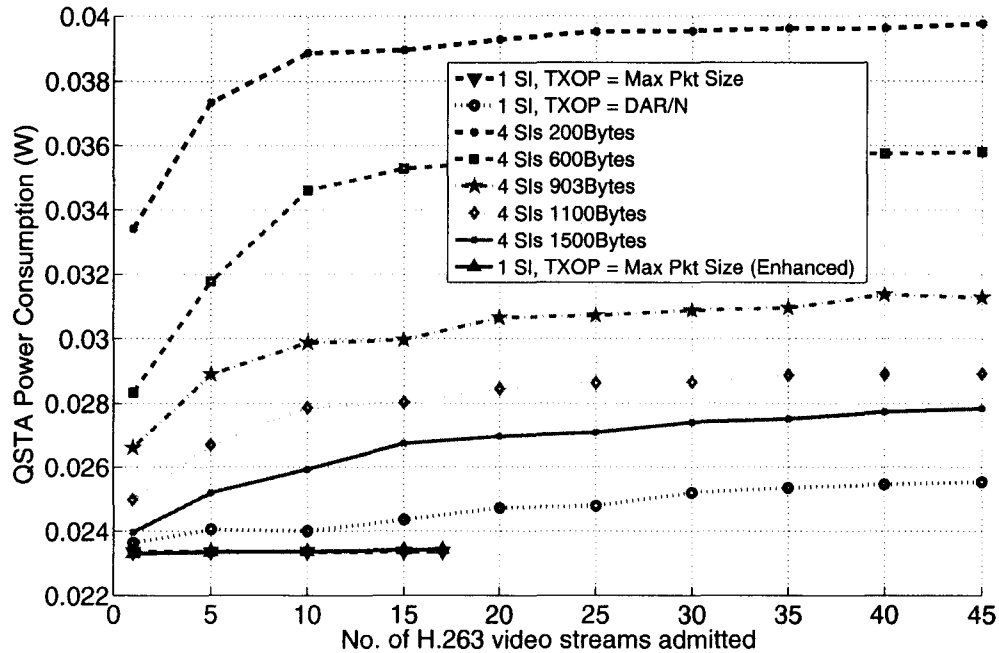
be 80ms but the service interval length, t_{SI}, is equal to 20ms. In this variant, the DAR(1) prediction-based CAC algorithm is not used for determining the length of TXOPs. However, we still rely on the PSQAP's knowledge of packet arrival times at individual stations. A video packet that arrives at the *ith* SI, SI_i, will be queued at the corresponding handset's buffer regardless of channel availability status. In the following SI, SI_{i+1}, the QSTA is assigned a fixed length (usually short) TXOP. A QSTA must confine its packet transmission within the length of the assigned TXOP. If the assigned TXOP is too short to accommodate the entire video packet, then the packet could be fragmented. Fragmentation should preserve the consistency and correctness of all higher layers information including sequence numbers and time-stamps, for a correct play back of the video content at the destination(s). The PSQAP in SI_{i+2}, based on the station's queue length information, will schedule another TXOP that has the exact length required to empty the buffer at the QSTA. Since in this scheme, it takes each video packet a maximum of 3 SIs (2 SIs in addition to the SI at which the packet arrives) to obtain service; the choice of having 4 SIs per superframe ensures that video packets are served well within the 80ms delay bound.

At the beginning of each SI, the PSQAP has a full knowledge of QSTAs with packet fragments that still need a second TXOP to be transmitted. Additionally, the PSQAP is also aware of video packets that were generated in the immediate previous SI and therefore are awaiting for their first fixed-length TXOPs. At the start of a given SI, SI_i, if the scheduler decides that no video packets are in need of PSMP-TXOPs in SI_i, the PSQAP remains in the *doze*

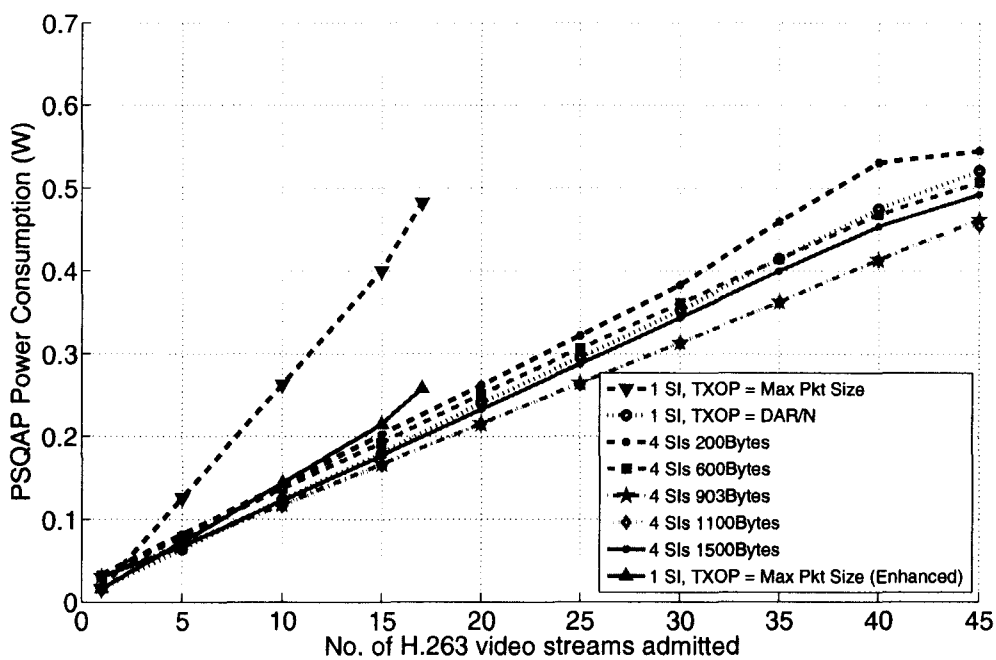
state to extract more power saving. Figure 4.3(b) shows the packet arrival times for 6 videoconference QSTAs, the PSQAP's timeline and TXOPs allocated to serve video packets in the 4 SIs/PSMP frames. In the figure, Stations LID 6 and LID 5 generate packets at SI_1. The PSQAP assigns fixed length TXOPs (indicted in red and green colors) to both stations in the following SI, i.e., SI_2. However, since Stations LID 3 and LID 4, whose packets arrived in SI_4 of the previous superframe, still have packet fragments in their queues, the PSQAP schedules their TXOPs earlier than the TXOPs for LID 6 and LID 5. This is done to ensure delay bound satisfaction of video packets.

4.4.5 Experiments and Results

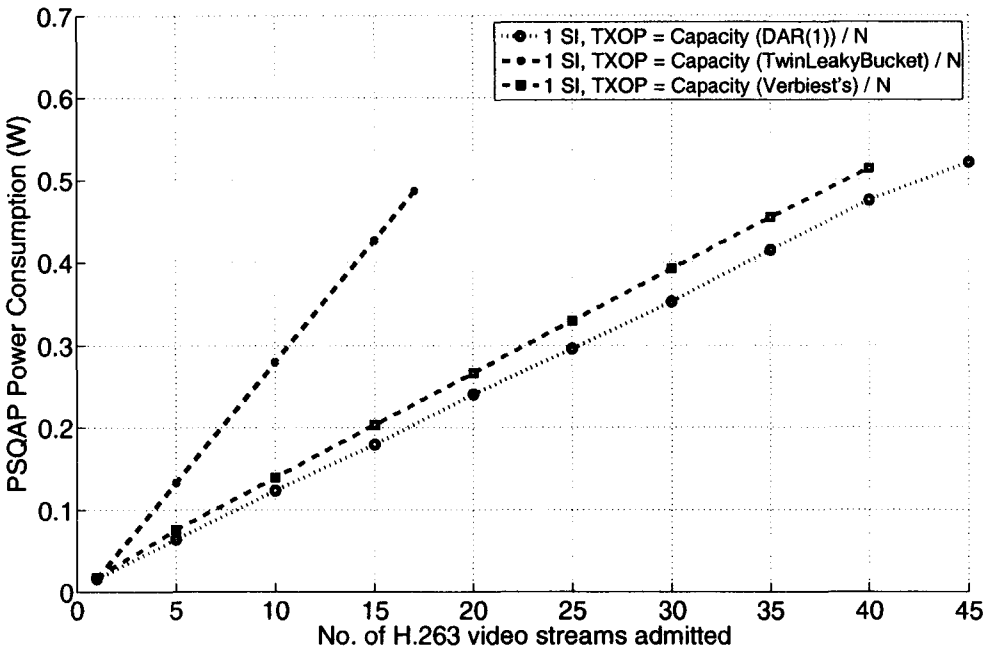
Figure 4.6 shows the results obtained from a comparative simulation study of the two unscheduled PSMP-based energy-efficient designs and their variants introduced in Section 4.4.4. In Figure 4.6, the curve labeled "1SI, TXOP = Max Pkt Size" corresponds to the use of the maximum video packet size to schedule TXOP intervals. As can be seen from Figure 4.6(a), this mechanism closely approximates the optimal power performance for handsets and thus it represents the lower bound. This is true since a video packet transmission will always complete within the assigned TXOP and before the next station's TXOP begins, i.e., no queuing overhead. On the other hand, it can be seen that the same scheme is extremely energy inefficient from a PSQAP's perspective (Figure 4.6(b)). This is also expected since for most of the video packets, the transmission completes well before the TXOP ends, leaving the PSQAP in idle waiting and the channel unutilized until the beginning of the next TXOP in the PSMP frame. We also show results for an enhanced variant of this scheme, indicated by the



(a) Video QSTA Power Consumption.



(b) PSQAP Power Consumption.



(c) PSQAP Power Consumption Results for Prediction-Assisted with 1 SI Scheme Using DAR(1) Mode, Verbiest's Formula, and Twin-Leaky Bucket Analysis.

Figure 4.6: Power Performance Results for PSMP-based Video Capacity Scheduling Schemes.

curves labeled "1SI, TXOP = Max Pkt Size (Enhanced)". For this variant, a PSQAP also uses worst-case scheduling (i.e., TXOPs are scheduled based on the maximum video packet size); however, once a station completes its packet transmission, the PSQAP makes a decision whether to sleep or to remain awake for the remainder of the TXOP. The decision is made based on the viability of power saving extraction achieved by sleeping compared to the energy-overheads of state transitions required for sleeping the internal radio circuitry. As shown in Figure 4.6(a), the handset power performance is identical for both worst-case scheduling schemes. However, a significant PSQAP power saving is achieved by the enhanced version. Worst-case scheduling schemes do not exploit the statistical multiplexing characteristics of variable-rate multimedia

traffic such as H.263 videoconferencing streams. Therefore, a serious problem with these mechanisms is their very inefficient usage of channel resources which is clearly reflected in the maximum number of H.263 video flows admitted (18) compared to 45 for other curves in Figure 4.6.

Figure 4.6(a) also shows the power performance results for the single SI unscheduled PSMP-based Scheduling Scheme that uses the DAR(1)-based CAC algorithm. The results show the value of combining accurate capacity-prediction based CAC with the energy-efficient MAC protocol presented in Chapter 3. From Figure 4.6(a), it can be seen that handsets' power performance using this scheme closely approximates the lower bound on handset power consumption. In numbers, the handset power drain under the prediction-assisted US-PSMP scheme (with 1 SI) is only $2mW$ more than the lower bound, when the PSQAP is operating at the right-end of the video traffic load range.

On the PSQAP side, capacity scheduling using prediction-assisted PSMP achieves very good power performance. Under the proposed scheme, the PSQAP consumes at most 15% more power than the lower value achieved by the curve labeled "4 SIs, 903 Bytes" which corresponds to the use of US-PSMP with the 4 SIs Scheduling Scheme that assigns TXOPs based on the statistical mean packet size. However, when comparing both schemes, the proposed algorithm saves about 22% of the handset power (Figure 4.6(a)). In general and as can be seen from Figure 4.6(a), the use of the prediction-assisted scheme with 1 SI achieves better handset power performance compared to all other PSMP schemes that schedule 4 SIs/PSMP frames in a superframe. This interesting observation relates to the fact that handsets incur a significant power consumption when transitioning between awake and sleep states.

Therefore, it is more energy-efficient to have each QSTA wake up exactly once per superframe to send its video backlog. For the same reason we can also explain the improved handset power performance when using a larger TXOP size under the Multiple SIs scheme in Figure 4.6(a). The larger the TXOP allocated for a MS, the more likely its packet transmissions will fit into this first TXOP and the MS can save the extra power that it would have consumed in case it needs to awaken for a packet fragment transmission. The handset power consumption reaches its lower bound when the TXOPs are scheduled based on the maximum video packet size. On the other hand, we can also see that AP power consumption also decreases for larger TXOP sizes as shown in Figure 4.6(b). This is true since in this case the AP won't need to schedule additional TXOPs for MSs. However, it can be seen that this trend stops when TXOP $>$ 1100Bytes, and the reason is that for bigger TXOP values the AP spends a significant portion of its awake time in an idle waiting state following the completion of relatively short packet transmissions.

Finally, Figure 4.6(c) also shows the results when using models in [VPV88] and [vdSAH06] to estimate the required capacity and determine the TXOP length. The superior prediction accuracy of the proposed framework using the DAR(1)-based CAC algorithm achieves a lower power consumption on the PSQAP side without compromising the QoS requirements of H.263 multimedia streams.

4.4.6 Analysis of the Handset Power Performance

This section provides analytical models for the power performance of a QSTA served by a PSQAP. The objective is to better analyze the various effects of a PSQAP alternating between awake and sleep periods on handset power saving. By capturing the

major energy expenditure components of a handset, it is possible to assess and verify the effectiveness of using the proposed energy-efficient PSQAP framework, which combines accurate bandwidth predictions and a modified variant of US-PSMP, for low-energy transmission and efficient capacity provisioning for real-time videoconferencing applications. Our derivation of the average handset power consumption is based on power calculations for a randomly chosen MS. In the remainder of this section, we refer to such a station as the "tagged" station.

MS Power Consumption Under the Prediction-assisted (Single EDCA SI) Scheme

First, an analytical model is presented for the power consumption of an H.263 videoconferencing station served by a PSQAP. The PSQAP is assumed to use the scheme presented in Section 4.4.2. In this scheme, the PSQAP uses the prediction-based CAC (Equation 4.3) to estimate the aggregate bandwidth requirements for multiplexed H.263 videoconference streams. QSTAs use the contention-based EDCA channel access mechanism to send video packets. Video packets that arrive during the activity sub-interval are processed instantaneously (provided that the channel is idle and the time remaining till the end of the activity sub-interval is sufficient to send a video packet, or a fragment of it). On the other hand, video packets generated during the sleep period are queued and stations awaken at the beginning of the following SI. Throughout the analysis, we use Ψ_s to indicate the group of video packets generated during the sleep period of a superframe and Ψ_a for video packets that are generated during the activity interval (denoted by TB_i, TB_{i+1} and TB_{i+2} in Figure 4.1(b)). The total number of admitted H.263 videoconferencing streams is denoted by N of which N_s denotes the number of QSTAs $\in \Psi_s$ and similarly, N_a denotes the number

of QSTAs $\in \Psi_a$.

We assume that arrival times of the first packets in video streams are Poisson distributed while the inter-arrival time of subsequent packets in each flow is fixed at 80ms intervals. Additionally, packet arrivals are assumed uniformly distributed over the time line, thus they are uniform over a superframe interval, T_{SF} (s). The handset power consumption η_{MS} can be written as

$$\eta_{MS} = No.\ of\ superframes\ per\ second \times \zeta_{superframe} \quad \text{(W)}, \quad (4.5)$$

where $\zeta_{superframe}$ is the average energy expenditure of a mobile station in a T_{SF} interval and is equal to

$$\zeta_{superframe} = \Pr\{MS \in \Psi_s\}\ \zeta_s + \Pr\{MS \in \Psi_a\}\ \zeta_a \quad \text{(J)}. \quad (4.6)$$

$\Pr\{MS \in \Psi_s\}$ and $\Pr\{MS \in \Psi_a\}$ are the probabilities that the tagged QSTA belongs to Ψ_s and Ψ_a, respectively, where $\Pr\{MS \in \Psi_s\} = N_s/N$ and $\Pr\{MS \in \Psi_a\} = N_a/N$; and $N_a + N_s = N$. ζ_s and ζ_a are the average energy (J) expended by a QSTA $\in \Psi_s$ and by a QSTA $\in \Psi_a$, respectively. The PSQAP is assumed to handle at most N_{max} concurrent H.263 videoconferencing sessions at its assumed maximum data rate. To calculate N_a, we assume that the capacity of the PSQAP is discretized and normalized such that a new H.263 video stream consumes, on average, $1/N_{max}$ of this capacity. When N simultaneous video flows are admitted, the PSQAP needs to roughly set $t_A = T_{SF}N/N_{max}$, where t_A is the activity sub-interval length which can be also written as $t_A = T_{SF} - t_S$. From the simulation results presented in Figure 4.2(a), N_{max} was found to be

45. Thus, N_a can be calculated as $\frac{N \times t_A}{T_{SF}} = \frac{N^2}{N_{max}}$.

For a handset that generates a packet during the sleep period, ζ_s can be expressed as

$$\zeta_s = \zeta_{Pkt,s} + P_c\, \zeta_{Pkt,c} + T_{Q,s}\, \eta_{Rx} + \zeta_{Doze} + \zeta_{sleep-wake} + \zeta_{wake-sleep} \quad \text{(J)}. \quad (4.7)$$

$\zeta_{Pkt,s}$ and $\zeta_{Pkt,c}$ are the energy expenditure (J) for a MS $\in \Psi_s$ while successfully transmitting a packet and during a packet collision, respectively. P_c is the packet collision probability for a station $\in \Psi_s$. η_{Rx} is the power (W) consumed during receive/listen. ζ_{Doze} is the energy expenditure (J) while the handset is in the power saving mode (normalized to a superframe interval) and is calculated as the product of the time in which the MS is sleeping and the power consumption during the *doze* mode (η_{Doze}). $\zeta_{sleep-wake}$ and $\zeta_{awake-sleep}$ are the energy expended (J) during the doze-to-awake and awake-to-doze state transitions, respectively. To calculate P_c, recall that at the beginning of each activity period, all N_s stations that have packets in their queues awaken to contend for the channel. Also, an average of $N_s \frac{N_a}{N}$ new video packets ($\in \Psi_a$) are generated whilst N_s video packets are being served. This makes the average number of QSTAs contending for the channel at the beginning of an SI equal to $N_s(1+\frac{N_a}{N})$. For a tagged station that is randomly chosen, the packet collision probability can be written as

$$P_c = \frac{N_s(1+\frac{N_a}{N}) - 1}{CW_{min}[VIDEO] - 1}, \quad (4.8)$$

where $CW_{min}[VIDEO]$ is the initial contention window size for the VIDEO access category as defined for the IEEE 802.11e EDCA. Note that only first-order packet collisions are considered in this analysis while higher-order collisions are ignored. The average queuing time $T_{Q,s}$ experienced by a tagged station $MS_i \in \Psi_s$ can be calculated as

$$T_{Q,s} = \frac{1}{N_s(1+\frac{N_a}{N})} \sum_{k=1}^{N_s(1+\frac{N_a}{N})-1} k\ (T_{Pkt} + T_{DIFS}) = \frac{(T_{Pkt} + T_{DIFS})(N_s(1+\frac{N_a}{N}) - 1)}{2} \quad \text{(s)}, \quad (4.9)$$

where T_{Pkt} is the packet transmission time (s) and it equals $\frac{Payload+O}{R} + (T_{ACK} + T_{SIFS} + T_{DIFS})$. $\zeta_{Pkt,s}$ and $\zeta_{Pkt,c}$ are calculated as

$$\zeta_{Pkt,s} = \frac{Payload + O}{R} \times \eta_{Tx} + (T_{ACK} + T_{SIFS} + T_{DIFS}) \times \eta_{Rx} + \eta_{Rx}\ (1 - P_c)\ aSlotTime\ \frac{CW_{min}[VIDEO] - 1}{2} + \eta_{Rx}\ P_c\ aSlotTime\ \frac{2CW_{min}[VIDEO] - 1}{2} \quad \text{(J)}, \quad (4.10)$$

and

$$\zeta_{Pkt,c} = \frac{Payload + O}{R} \times \eta_{Tx} + (T_{SIFS} + T_{DIFS}) \times \eta_{Rx} + \eta_{Rx}\ aSlotTime\ \frac{CW_{min}[VIDEO] - 1}{2} \quad \text{(J)}. \quad (4.11)$$

$aSlotTime$ is the IEEE 802.11 MAC layer slot duration and η_{Tx} is the power (W) consumed during transmit mode. *Payload* is the payload length of a video packet (in bits) and O, is the per packet overheads, i.e., UDP/RTP/IP/MAC

headers and tails (in bits). T_{ACK}, T_{DIFS} and T_{SIFS} are the channel time (s) for an acknowledgment frame, DIFS and SIFS, respectively.

Similarly, we can calculate ζ_a for a handset that generates video packets during the awake period as

$$\zeta_a = \zeta_{Pkt,a} + T_{Q,a}\,\eta_{Rx} + P_{c,a}\,\zeta_{Pkt,c} + \zeta_{Doze} + 2\,(\zeta_{sleep-wake} + \zeta_{wake-sleep}) \quad \text{(J)}. \tag{4.12}$$

$\zeta_{Pkt,a}$ is the energy expenditure (J) for a $MS_j \in \Psi_a$ while successfully transmitting a packet and is written as

$$\zeta_{Pkt,a} = \frac{Payload + O}{R} \times \eta_{Tx} + (T_{ACK} + T_{SIFS} + T_{DIFS}) \times \eta_{Rx} + \eta_{Rx}\; aSlotTime\; \frac{CW_{min}[VIDEO] - 1}{2} \quad \text{(J)}. \tag{4.13}$$

$T_{Q,a}$ is the average queuing time experienced by a tagged station $MS_j \in \Psi_a$ and it can be written as

$$T_{Q,a} = \frac{N_s N_a}{N}\frac{T_{Q,s}}{N_a} + (1 - \frac{N_s}{N})(\frac{N_a - 1}{2})\;(T_{Pkt} + T_{DIFS}) \quad \text{(s)}. \tag{4.14}$$

As explained earlier, the term $\frac{N_s N_a}{N}.\frac{1}{N_a}$ represents the probability that a tagged station $MS_j \in \Psi_a$ generates a video packet while all or some of the N_s QSTAs are still contending for the channel at the beginning of an activity period, and $T_{Q,s}$ is the average queuing time for such a station (s). The term $1 - \frac{N_s}{N}$ is the complementary probability, i.e., the probability that MS_j generates a video packet after all N_s packets have been transmitted. $\frac{N_a - 1}{2}$ is the number of other

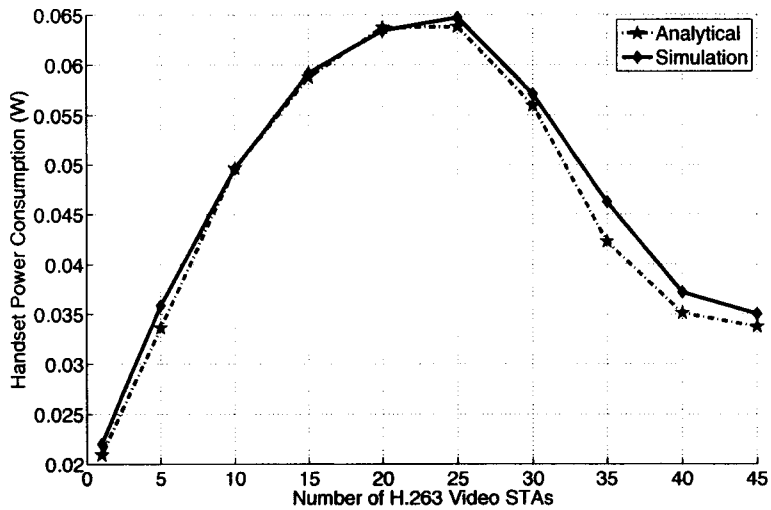


Figure 4.7: Analytical vs. Simulation Results for Handset Power Consumption Under the Prediction-assisted 1 SI (EDCA) Scheduling Scheme

packet transmissions that station MS_j will have to wait, on average, before it successfully sends its packet.

$P_{c,a}$ is the packet collision probability for stations $\in \Psi_a$ and is calculated only for those stations that generate packets while N_s packets are contending for channel access. Thus, $P_{c,a} = P_c \frac{N_s}{N}$.

By substituting Equations 4.6 to 4.14 into Equation 4.5, the average handset power consumption can be calculated. To verify the analysis presented above, the model is used to calculate an approximation to the average MS power consumption assuming the prediction-assisted EDCA (1 SI) scheduling scheme in Figure 4.1. The average packet size (from H.263 trace file) is assumed for video payload (i.e., $Payload$ = 903Byte), $CW_{min}[VIDEO] = 64$, $aSlotTime = 20\mu s$ and O = 2704bit. $\zeta_{sleep-wake}$ and $\zeta_{wake-sleep}$ are modeled as 125μJ and 250μJ, respectively. The values for other parameters are listed in Table 4.1.

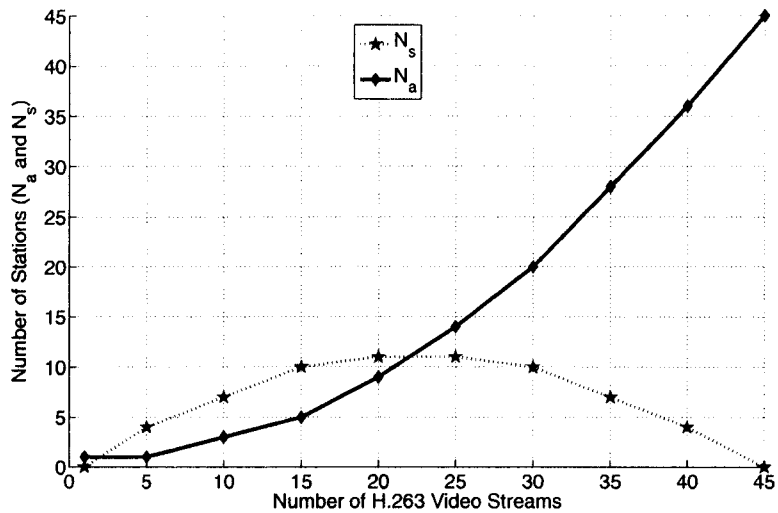


Figure 4.8: Number of QSTAs in Ψ_a and Ψ_s vs. Total Number of H.263 Video Stations Admitted

Figure 4.7 shows the results obtained from the analytical model compared to simulation results previously shown in Figure 4.2(b) for the curve labeled "Prediction-based (DAR Model)". From the figure, it can be seen that analytical results closely match the results from simulations. The effect of queuing and contention on handset power performance is obvious in the range of 5-25 video streams. In this range and as shown in Figure 4.8, most of the video packets are generated while the PSQAP is sleeping. This is true since for a small to moderate number of video flows, the ratio of the awake period to the total superframe interval is ≤ 0.5. Hence the quantity $\Pr\{MS \in \Psi_s\}\ \zeta_s$ dominates the power consumption in Equation 4.6 (and Equation 4.5). However, as more video handsets are admitted, N_s diminishes as a result of the PSQAP increasing the offered capacity in order to accommodate the increase in video traffic load. Consequently, the $\Pr\{\text{MS} \in \Psi_a\}$ also increases and the term $\Pr\{MS \in \Psi_a\}\ \zeta_a$ starts to dominate the handset power consumption. From Figure 4.8, we can

also notice that N_s reaches zero when N_{max} video stations simultaneously exist at the PSQAP.

MS Power Consumption Under PSMP-based with Multiple SIs Scheme

We also present a model for the power consumption of an H.263 videoconferencing station served by a PSQAP which is assumed to utilize the unscheduled PSMP-based with Multiple SI scheduling scheme presented in Section 4.4.4. The power consumption derivation is also based on Equation 4.5. For this scheme, the average energy expended by the handset during a superframe interval, $\zeta_{superframe}$, can be defined as

$$\zeta_{superframe} = \zeta_{PSMP} + \zeta_{Pkt(Payload<TXOP)} + 2(\zeta_{awake-sleep} + \zeta_{sleep-wake}) + \zeta_{Doze} \quad \text{(J)}. \quad (4.15)$$

ζ_{PSMP} is the MS energy expenditure (J) while sampling a PSMP Action frame. $\zeta_{Pkt(Payload<TXOP)}$ is the energy expenditure (J) during the transmission of a H.263 video packet of length *Payload* (in bits) and $TXOP$ is the size (bits) of the fixed length PSMP TXOP that is first allocated to a newly arrived video packet in the PSMP-based with Multiple SI scheduling scheme (Figure 4.5(b)). We can write

$$\zeta_{Pkt(Payload<TXOP)} = \frac{\min(Payload, TXOP) + O}{R} \times \eta_{Tx} + (T_{ACK} + T_{SIFS}) \times \eta_{Rx} \quad \text{(J)}. \quad (4.16)$$

Equation 4.15 only considers the case where video packet payload is less than

Table 4.2: Analytical vs. Simulation Results for Handset Power Consumption Under the PSMP-based with 4 SIs Scheduling Scheme

TXOP	Simulations	Analytical
200 Byte	0.0396W	0.0408W
5191 Byte (Max.)	0.0233W	0.0241W

the first assigned PSMP-TXOP. However, as explained in Section 4.4.4, if the assigned TXOP is too short to accommodate the entire video packet, the packet could be fragmented. In this case, the PSQAP uses the station's queue length information to schedule another PSMP-TXOP (in the following SI) that has the exact length required to send the remaining packet fragment. The station will then need to awaken and sample the PSMP in the next SI to know its UTT. The station can then sleep to conserve energy and then awaken at its scheduled PSMP-TXOP for packet transmission. To account for this latter case, Equation 4.15 can be rewritten as

$$\zeta_{superframe} = 2\,\zeta_{PSMP} + \zeta_{Pkt(Payload<TXOP)} + \zeta_{Pkt(Payload>TXOP)} + 4(\zeta_{awake-sleep} + \zeta_{sleep-wake}) \quad \text{(J)}. \quad (4.17)$$

$\zeta_{Pkt(Payload>TXOP)}$ is the energy expended by the handset to transmit a packet fragment and is expressed as

$$\zeta_{Pkt(Payload>TXOP)} = \frac{(Payload - TXOP) + O}{R} \times \eta_{Tx} + (T_{ACK} + T_{SIFS}) \times \eta_{Rx} \quad \text{(J)}. \quad (4.18)$$

The model was verified by calculating the handset power consumption for the two limiting cases shown in Figure 4.6(a) for the unscheduled PSMP-based scheme. Table 4.2 shows the results obtained from both the analytical model and simulations for an H.263 videoconferencing QSTA under the unscheduled PSMP-based with Multiple SIs Scheduling Scheme. Simulation results in Table 4.2 are the average of the power consumption values shown in Figure 4.6(a) for the curves labeled "4 SIs 200 Bytes" and "1 SI, TXOP = Max Pkt Size", respectively. Results from the table show that the analytical model closely approximates the results from simulations. The excess in handset power consumption when a small TXOP size is used can be explained in light of Equation 4.17. Since a TXOP of size 200bytes is likely insufficient to accommodate an H.263 video packet transmission (the minimum packet size = 234 Byte), the MS still needs to awaken for a second PSMP-TXOP. As a result, the MS samples a second PSMP Action frame in order to know its UTT and then awakens again for packet transmission. The corresponding extra power components, ζ_{PSMP} and $2\ (\zeta_{awake-sleep} + \zeta_{sleep-wake})$, are shown in Equation 4.17.

4.5 Conclusions

This chapter investigated the technical challenges associated with the transmission of video applications in solar/battery powered wireless mesh networks. An enhanced version of the IEEE 802.11n US-PSMP power management mechanism was used to provide near-optimal power performance for a power saving QoS-enabled AP (PSQAP) serving videoconferencing flows. A connection admission control scheme was first

proposed and integrated with the energy-efficient MAC protocol presented in chapter 3. The proposed CAC uses a discrete autoregressive model of order one DAR(1) to accurately predict the aggregate bandwidth requirements for multiplexed H.263 videoconferencing streams. A comparative study was conducted to evaluate several algorithms for performing video capacity prediction and resource allocation under the contention-based US-APSD channel access mechanism. Findings from this study were then used to devise new scheduling mechanisms that further improve the handset power performance by distributing packet transmissions over multiple US-APSD SIs. Analytical models and simulation experiments were used to assess the performance of two video scheduling schemes that were based on a modified version of the IEEE 802.11n unscheduled PSMP power management.

Results obtained from simulations show that bandwidth allocation using the proposed DAR(1)-based CAC algorithm achieves the best power-conservation performance at the PSQAP. The results also confirm the conflicting relationship between power performance at both the PSQAP and stations in a contention-based MAC protocol such as the one used in the proposed energy-efficient framework. Findings obtained from the comparative study also show that queuing and contention, which are the result of stations queuing from the start of SIs, are the main reasons for high handset power consumption values. The use of prediction-assisted scheduling combined with multiple SIs can roughly halve the handset power consumption. However, the achieved improvements do not extend to high loading values and power consumption levels are at least double the lower-bound. A near optimal power performance for both the PSQAP and handsets was achieved by using a modified version of the US-PSMP with one SI, where the length of the activity sub-interval is determined

using DAR(1) prediction.

Chapter 5

WLAN VoIP Capacity Allocation Using AVP

5.1 Overview

WLAN VoIP capacity is known to be very low due to the short audio payload used, combined with the effects of overheads at various protocol layers. An IEEE 802.11b access point (AP) for example, can support only about 10 G.711 voice connections using a 20ms packetization interval without advanced header compression [GK02]. With header compression and downstream packet multiplexing, VoIP capacity can only be doubled, at best [SLLY02, WLQL04]. These numbers are far below theoretical capacities [WLL05]. The low VoIP over WLAN capacity was also found to adversely affect the power performance of both the AP and handsets. This poor performance can be substantially improved if the connection latency margin is known in advance and is used in the selection of RTP packetization parameters [VCM01]. Unfortunately this is not always possible since it cuts into the end-to-end connection latency budget

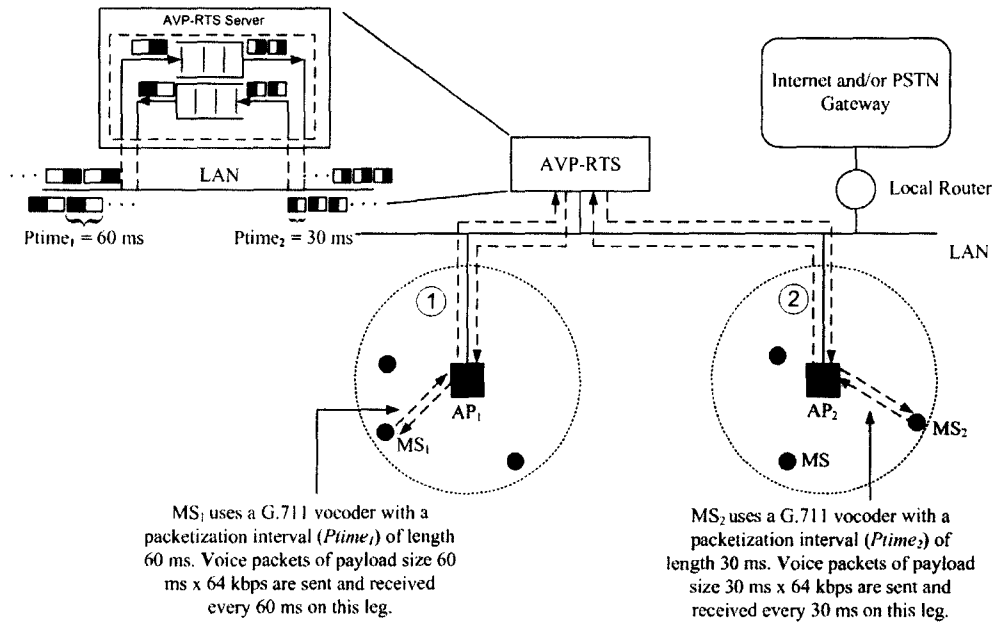


Figure 5.1: WLAN Enterprise Setup With Adaptive Voice Packetization Real-Time Server.

and the packet loss resilience of the call.

This chapter investigates and evaluates the use of an adaptive voice packetization server (AVP-RTS) which splits the RTP VoIP connection into two separate call legs. In this way each call leg can be assigned different packetization parameters, thus assigning the capacity and power saving gains asymmetrically across the connection. Algorithms are proposed and compared for performing this assignment. Results obtained from both simulations and analytical models are also reported at the end of the chapter. The performance study of the AVP scheme considers both VoIP capacity and power saving improvements, AVP delay overhead and call blocking probability.

5.2 VoWLAN Adaptive Voice Packetization

In Figure 5.1 an example setup is shown for an enterprise wireless LAN supporting real-time packetized VoIP. A set of APs provides wireless coverage for a population of mobile stations (MSs). VoIP calls made can be either intra- or inter-enterprise. The proposed AVP-RTS server is placed on the switched subnetwork that services the APs as shown in Figure 5.1. When VoIP connections are established they are routed through the AVP-RTS server, which negotiates and concatenates two separate RTP [SCFJ03] sessions. As a result, once the connections are established, uplink voice packets arriving at an AP from the MSs are routed to the AVP-RTS server before they are forwarded to the destination AP. This is also true for all intra-enterprise voice traffic. In the example shown, MS_1 initially establishes a VoIP call to MS_2. This call is routed to the AVP-RTS server which establishes the call to the end destination. The media path is shown, consisting of two legs, numbered 1 and 2.

In order to set the packetization parameters, the server needs to have some information regarding the latency of the call path. This information is inferred from whether or not the placed call is intra- or inter-enterprise. In more sophisticated schemes the latency along the call path could be estimated first. Using this information the AVP-RTS server can assign the packetization intervals for each leg of the call. In the figure these intervals are shown as $Ptime_1$ and $Ptime_2$, respectively. The idea of having the AVP-RTS server break each end-to-end voice connection into two separate legs allows for negotiating asymmetric packetization intervals at both legs of the same voice call, under the control of the AVP-RTS server. An asymmetric assignment is useful when one of the end APs of a VoIP connection is more heavily loaded than the other end AP of the same connection. In this case, assigning a larger

packetization interval length for the MS associated with the heavily loaded AP can be used to assign additional "capacity and power saving gain" to that AP. In order to do this exactly the server would need to know the available latency margin, LM, for each connection and then negotiate the packetization intervals of both handsets accordingly. In this work, the latency margin is defined as the difference between the maximum permissible round-trip connection latency, UB_{Delay}, which is usually about $300 \sim 400$ms [Uni] and the measured round-trip delay budget including propagation delay, maximum de-jitter/playout buffering, etc.

The internal configuration of the AVP-RTS server is depicted at the top left corner of Figure 5.1. The device maintains a separate pair of FIFO queues for each VoIP session. Within each pair, a single queue is used for voice packets arriving from each MS. The server re-packetizes incoming RTP payload at each leg in a manner that is transparent to the handset at the other leg of the same VoIP call. The use of larger packetization intervals may result in an increase in the playout delay of VoIP packets. Therefore, the required maximum playout/de-jitter delay is accounted for by incorporating it as a component of the subtracted delay budget when calculating LM. It should be noted that playout delay normally increases with the variation in packet delay, which is expected to be small for intra-enterprise VoIP connections.

5.3 Problem Formulation

In this section, an IEEE 802.11 WLAN enterprise deployment such as that shown in Figure 5.1 is assumed. Now assume that we are given, (i) the number of active voice calls taking place between APs, (ii) the assumed connection round-trip latency characteristics and vocodec parameters of each voice call, and, (iii) the utilization

(i.e., relative loading) of each AP. Then the objective is to assign the packetization intervals for each VoIP session optimally. A min-max objective function is adopted such that the utilization of the most heavily loaded AP in the network is minimized. Throughout our analysis, the VoIP capacity is defined as the maximum number of VoIP connections that can be accommodated before the first new call is blocked because of insufficient resources. Assuming there are M APs in the WLAN enterprise, the objective is to obtain,

$$\min \max\{U_i\}, \quad i = 1, \ldots, M$$

subject to :

$$MinPtime_{j,1} \leq Ptime_{j,1} \leq MaxPtime_{j,1}, \quad j = 1, \ldots, n_i$$

$$MinPtime_{j,2} \leq Ptime_{j,2} \leq MaxPtime_{j,2}$$

$$Ptime_{j,1} + Ptime_{j,2} + AD(Ptime_{j,1}, Ptime_{j,2}) + D_{network} \leq UB_{Delay} \quad (5.1)$$

where $AD(Ptime_{j,1}, Ptime_{j,2}) =$

$$\left(\left\lceil \frac{\max(Ptime_{j,1}, Ptime_{j,2})}{\min(Ptime_{j,1}, Ptime_{j,2})} \right\rceil - 1 \right) \times \min(Ptime_{j,1}, Ptime_{j,2}) \quad (5.2)$$

and U_i is the utilization of AP_i and is given by

$$U_i = \sum_{j=1}^{n_i} (T_{DL}^{j,1} + T_{UL}^{j,1}) \times \gamma_{j,1}, \quad (5.3)$$

n_i is the number of active VoIP sessions admitted by AP_i, $\gamma_{j,1}$ is the number of voice packets sent by, (i) the MS involved in the jth VoIP session, and, (ii) AP_i (per second), and is dependent on the packetization interval used at the leg between the MS and

AP_i for that call ($Ptime_{j,1}$). This is such that $\gamma_{j,1} = 1/Ptime_{j,1}$. $T_{DL}^{j,1}$ and $T_{UL}^{j,1}$ are the transmission times for downlink (DL) and uplink (UL) packets at the call leg between the MS and AP for the jth call, respectively. $MinPtime_{j,1}$ and $MaxPtime_{j,1}$ are the lower- and upper-bounds of the packetization interval values supported by the handset associated with the ith AP and used in the jth call. LM_j is the available round-trip latency margin for the jth VoIP call.

$AD()$ is the VoIP packet adaptation delay incurred at the AVP-RTS server for the jth VoIP session as a result of using asymmetric $Ptime_{j,1}$ and $Ptime_{j,2}$ values. In conventional VoIP, access delay within the WLAN is the major local delay incurred. Access delay of a VoIP packet is the time between its arrival to the AP (or when it is generated at the MS) until it is either successfully transmitted over the WLAN or dropped at the head of the queue because it has exhausted the retry limit for retransmission. With the AVP-scheme, the local delay also includes the packet adaptation delay incurred at the AVP-RTS server. The packet adaptation delay is the time from the arrival of a VoIP packet from MS to the AVP-RTS server until the time at which the payload of this packet is forwarded to the MS at the other end of the call. As a result of negotiating asymmetric $Ptime$ values by the AVP-RTS server, two MSs having an active call j may use different packetization intervals (i.e., $Ptime_{j,1} \neq Ptime_{j,2}$). Equation 5.2 calculates the packet adaptation delay as the time that a voice packet of length $\min(Ptime_{j,1}, Ptime_{j,2})$ arriving from one MS will have to wait in its corresponding queue at the server for packets to arrive so that they all can be packed into a larger packet of length $\max(Ptime_{j,1}, Ptime_{j,2})$, which will then be forwarded. With feasible packetization interval values of $10, 20, \ldots, 100$ms, the average packet adaptation delay varies between 0 and 90ms.

$D_{network}$ is the network delay which includes all delay components, experienced by VoIP packets, other than those mentioned in the latency margin constraint 5.1 (e.g., propagation delay, channel access delay, etc). To simplify the analysis of channel access delay component of $D_{network}$, MSs are assumed to utilize the scheduled automatic power save delivery (S-APSD) of IEEE 802.11e since the use of scheduled packets delivery mechanisms for VoIP traffic can guarantee the minimum access delay for voice packets because MSs access the shared channel at pre-scheduled time intervals without need for contention [BIMR01, CWX+03, HIIM02]. The propagation delay in a typical WLAN enterprise setup is also negligible (compared to the packetization intervals) since a wired distribution system (e.g., Ethernet) is used to transport traffic, including VoIP streams, between different APs in the enterprise (Figure 5.1).

The optimization function in Equation 5.1 minimizes utilization (i.e., capacity fraction needed to support all currently admitted CBR VoIP streams at an AP) of the AP having the maximum utilization among all APs in the enterprise WLAN. In other words, the min-max optimization function above maximizes the number of VoIP sessions that the most heavily loaded AP in the enterprise can simultaneously accommodate before the first new call is blocked. In Equation 5.3, the utilization of the ith AP, U_i, is defined as the sum of time needed by the AP to transmit and receive voice packets to/from all n_i stations in an interval of 1s where each station carries a single VoIP call. It is obvious that the number of packets sent/received for a VoIP call depends mainly on the packetization interval value(s) used by the two call parties. Therefore, the min-max AP utilization in a WLAN is found by searching for the optimal pair of packetization interval values for all ongoing VoIP sessions in the enterprise. Note that each active call (e.g., call j) needs a pair of packetization

interval values ($Ptime_{j,1}$, $Ptime_{j,2}$) which corresponds to voice packets generated at the two end MSs of that call (Figure 5.1).

The min-max optimization problem defined above is subject to several constraints: (i) the pair of packetization interval values ($Ptime_{j,1}$, $Ptime_{j,2}$) for any given call must be within the lower and upper bounds of the packetization intervals used by both end MSs of the call (first and second constraints in Equation 5.1), (ii) for any given active call, the total delay budget of that call (due to voice packetization delay at AP and MS, packets adaptation delay incurred at the AVP-RTS server because of asymmetric $Ptime_{j,1}$ and $Ptime_{j,2}$ values, in addition to other network delay components) must be less than or equal to the maximum permissible round-trip connection latency, UB_{Delay} (i.e., the third constraint in Equation 5.1).

The optimization in Equation 5.1 is global, i.e., is applied over all active calls in the WLAN enterprise at the same time. The non-linear ceiling $\lceil\ \rceil$ operator in Equation 5.2 adds significant complexity to this problem. Although the fractional objective function can be easily transformed to an equivalent convex problem [BV04], unfortunately, it is difficult to solve because of the non-linearity in the latency margin constraint in Equations 5.1 and 5.2, and given the fact that both $Ptime_1$ and $Ptime_2$ are typically integer multiples of 10ms [sip], [sno].

An algorithm Min-Max AP Utilization ($MMAU$) is proposed and is shown in Figure 5.2. The algorithm employs a local optimization technique that searches for a suboptimal pair of $Ptime_1$ and $Ptime_2$ values of every newly requested VoIP call. The solver works in a way similar to a typical call admission control scheme. For each newly requested call, the algorithm starts by determining the APs that the two voice MSs are associated with (i.e., AP_1 and AP_2). Then it initializes the packetization

```
% Initialize Ptimes for both MSs with the maximum possible
% values based on LM and vocoder constraints.
1:  AP1 ← ID of AP that MS1 is associated with.
2:  AP2 ← ID of AP that MS2 is associated with.
3:  Initial_Ptime1 ← min(LM/2, Max_Ptime1);
4:  Initial_Ptime2 ← min(LM/2, Max_Ptime2);
5:  Optimal_Ptime1 ← Initial_Ptime1; Optimal_Ptime2 ← Initial_Ptime2;
6:  minAudioFrame ← 10ms;
7:  Best solution found so far = ∞;
8:  Begin
9:    if (Initial_Ptime1 = = Initial_Ptime2){
10:     Update Utilizations of end APs;
11:     Exit; % Near optimal Ptimes are found.
12:   } else if ( Utilizations(AP1) ≥ Utilizations(AP2) ) {
13:     for (Ptime1 = Initial_Ptime1; Ptime1 > Initial_Ptime2; Ptime1− = minAudioFrame)
14:     {
15:       Ptime2 = Initial_Ptime2;
16:       while (Ptime1 & Ptime2 do not satisfy LM constraint)
17:         Ptime2 = Ptime2 - minAudioFrame;
18:       if ( Max(Utilizations(AP1,AP2)) < Max(Best solution found so far) )
19:         Update best solution of Ptimes and best AP utilizations found so far;
20:     };
21:   } else if ( Utilizations(AP2) > Utilizations(AP1) ) {
22:     for (Ptime2 = Initial_Ptime2; Ptime2 > Initial_Ptime1; Ptime2− = minAudioFrame)
23:     {
24:       Ptime1 = Initial_Ptime1;
25:       while (Ptime2 & Ptime1 do not satisfy LM constraint)
26:         Ptime1 = Ptime1 - minAudioFrame;
27:       if ( Max(Utilizations(AP1,AP2)) < Max(Best solution found so far) )
28:         Update best solution of Ptimes and best AP utilizations found so far;
29:     };
30:   };
31: End
```

Figure 5.2: Min-Max AP Utilization Voice Packetization Algorithm (MMAU)

interval length of each MS involved in the new call (i.e., $Ptime_1$ and $Ptime_2$) with a feasible solution which is selected as the minimum of the maximum permissible packetization interval supported by that MS and half of the available latency margin of the new connection. These initial values are the initial optimal (suboptimal in the context of heuristic algorithms) values for $Ptime_1$ and $Ptime_2$, and are updated during the progress of the algorithm as better feasible solutions are found.

After the initialization phase, the algorithm proceeds (lines 9, 10 and 11) by checking if initialization values for $Ptime_1$ and $Ptime_2$ are the same. If so, a near-optimal solution is found and the algorithm terminates since the use of the symmetric initial values of $Ptime_1$ and $Ptime_2$ satisfies the latency constraint in Equation 5.1. If initial *Ptime* values are different, the algorithm (in line 12) checks if AP_1 is more heavily loaded than AP_2. If this is the case, the algorithm searches for the largest possible value of $Ptime_1$ that would guarantee the admission of the new call while achieving the minimal loading of AP_1 (lines 13 through 20). This is done by starting with $Ptime_1 = Initial\_Ptime_1$ and iteratively decrementing $Ptime_1$ until it reaches $Initial\_Ptime_2$. In each iteration of the for loop (line 13), the algorithm tries to find the largest possible value of $Ptime_2$ that (given the current value of $Ptime_1$ in the loop iteration) satisfies the latency constraint and would produce a better solution (i.e., lower utilization of AP_1 and AP_2). If none is found, $Ptime_1$ is decremented by the minimum frame size (assumed to be 10ms) and the same steps are repeated until either a better solution is found or $Ptime_1$ reaches the value of $Initial\_Ptime_2$. On the other hand, if AP_2 is more heavily loaded than AP_1, the algorithm searches for the largest possible value of $Ptime_2$ that would guarantee the admission of the new call while achieving the minimal utilization of AP_2 (lines 21 through 30).

```
% Initialize Ptimes for both MSs with the maximum possible
% values based on LM and vocoder constraints.
1:  AP1 ← ID of AP that MS1 is associated with.
2:  AP2 ← ID of AP that MS2 is associated with.
3:  Initial_Ptime1 ← min(LM/2, Max_Ptime1);
4:  Initial_Ptime2 ← min(LM/2, Max_Ptime2);
5:  Optimal_Ptime1 ← Initial_Ptime1; Optimal_Ptime2 ← Initial_Ptime2;
6:  minAudioFrame ← 10ms;
7:  Best solution found so far = ∞;
8:  Begin
9:    if (Initial_Ptime1 = = Initial_Ptime2){
10:     Update Utilizations of end APs;
11:     Exit; % Near optimal Ptimes are found.
12:   } else if ( Initial_Ptime1 < Initial_Ptime2 ) {
13:       Ptime1 = Initial_Ptime1;
14:       for (Ptime2 = Initial_Ptime2; Ptime2 > Initial_Ptime1; Ptime2− = minAudioFrame)
15:       {
16:         if (Ptime1 & Ptime2 satisfy LM constraint)
17:           if ( Sum_of_Utilizations(AP1,AP2) < Optimum solution found so far )
18:             Update optimal solution of Ptimes and optimal (i.e.,
19:             minimum) AP utilizations found so far;
20:       }
21:   } else if ( Initial_Ptime1 > Initial_Ptime2 ) {
22:       Ptime2 = Initial_Ptime2;
23:       for (Ptime1 = Initial_Ptime1; Ptime1 > Initial_Ptime2; Ptime1− = minAudioFrame)
24:       {
25:         if (Ptime1 & Ptime2 satisfy LM constraint)
26:           if ( Sum_of_Utilizations(AP1,AP2) < Optimum solution found so far )
27:             Update optimal solution of Ptimes and optimal (i.e.,
28:             minimum) AP utilizations found so far;
29:       }
30:   };
31: End
```

Figure 5.3: MaxNetworkCap Voice Packetization Algorithm

In contrast to the *MMAU* algorithm, which favors the AP having the highest load, we consider the *MaxNetworkCap* algorithm which is a variant of *MMAU*. *MaxNetworkCap* ensures fairness among different APs when assigning packetization interval values for different active calls. For a given number of active VoIP calls, the algorithm attempts to maximize network VoIP capacity by minimizing the total loading of all APs (i.e., load balancing). To achieve this, it uses a local optimization that searches for a suboptimal pair of $Ptime_1$ and $Ptime_2$ of a newly requested VoIP call to minimize the sum of the mean loading of all APs. The algorithm initialization is identical to *MMAU*. After the initialization phase, the algorithm proceeds (lines 9, 10 and 11 in Figure 5.3) by checking if initialization values for $Ptime_1$ and $Ptime_2$ are the same. If so, a near-optimal solution is found and the algorithm terminates since the use of the symmetric initial values of $Ptime_1$ and $Ptime_2$ satisfies the latency constraint in Equation 5.1. If initial *Ptime* values are different, the algorithm (line 12) checks if $Initial\_Ptime_1$ is less than $Initial\_Ptime_2$. If this is the case, the algorithm searches for the largest possible values of $Ptime_1$ and $Ptime_2$ that would guarantee the admission of the new call while achieving the minimal sum of AP_1 and AP_2 utilizations (lines 13 through 20). This is done by fixing $Ptime_1 = Initial\_Ptime_1$ and setting $Ptime_2 = Initial\_Ptime_2$. Then, the algorithm iteratively decrements $Ptime_2$ by the minimum frame size (assumed 10ms) until it reaches $Initial\_Ptime_1$. In each iteration of the for loop (line 14), the algorithm searches for the largest possible value of $Ptime_2$ that (given the value of $Ptime_1$) satisfies the latency constraint and would produce a better solution (i.e., lower sum of AP_1 and AP_2 utilizations).

On the other hand, if $Initial\_Ptime_2$ is less than $Initial\_Ptime_1$, the algorithm fixes $Ptime_2$ to $Initial\_Ptime_2$ and sets $Ptime_1 = Initial\_Ptime_1$ and iteratively searches for

1: **Begin**
2: $Ptime_1 \leftarrow \min(\min(Max\_Ptime_1, Max\_Ptime_2), LM/2)$;
3: $Ptime_2 \leftarrow Ptime_1$;
4: **End**

Figure 5.4: *Split-LM* Voice Packetization Algorithm

the largest possible value of $Ptime_1$ that would guarantee the admission of the new call while achieving the minimal sum of AP_1 and AP_2 utilizations (lines 21 through 30).

We also include the case where the handsets on both ends of a VoIP connection simply split the available latency margin on both sides of the call (i.e., $Ptime_1 = Ptime_2 = \min(\min(MaxPtime_1, MaxPtime_2), LM/2)$ (Figure 5.4). Note that *Split-LM* can yield the *optimum* solution if and only if both $MaxPtime_1$ and $MaxPtime_2$ are greater than or equal to $LM/2$.

5.3.1 Worked Example of Packetization Interval Assignment

In this section, a worked example is given of the proposed assignment schemes. A setting similar to the one shown in Figure 5.1 is also assumed but with three APs connected to the same LAN. A newly arriving VoIP call is to be established between MS_1 (already associated with AP_1) and MS_2 which is associated with AP_3. We assume that initially the utilizations of the three APs $\{AP_1, AP_2, AP_3\}$ are $\{0.36, 0.52, 0.74\}$, respectively. We also assume that the available round-trip latency margin between AP_1 and AP_3 $(LM_{1,3}) = 180$ms. This is typical of an enterprise scenario.

Initial AP Utilization = { **0.36**, 0.52, **0.74** } and $LM_{1,3}$ = 180ms.

The values assigned for $Ptime_1$ (AP_1) and $Ptime_2$ (AP_3) using the proposed algorithms are shown below.

1. *Basic* VoIP:

$Ptime_1 = Ptime_2$ = 20ms.
AP Utilizations = { **0.422**, 0.52, **0.82** }

2. Search steps for *Min-Max AP Utilization (MMAU)*:

Step	$Ptime_1$ (ms)	$Ptime_2$ (ms)	$Ptime_1$+$Ptime_2$ + $AD(Ptime_1$+$Ptime_2)$
1	50	90	190ms $\geq LM_{1,3}$
2	40	90	210ms $\geq LM_{1,3}$
3	30	90	180ms $\leq LM_{1,3}$

After Step 3, AP Utilization = { **0.405**, 0.52, **0.761** }

3. Search steps for *MaxNetworkCap*:

Step	$Ptime_1$ (ms)	$Ptime_2$ (ms)	$Ptime_1$+$Ptime_2$ + $AD(Ptime_1$+$Ptime_2)$
1	50	90	190ms $\geq LM_{1,3}$
2	50	80	180ms $\leq LM_{1,3}$

After Step 2, AP Utilization = { **0.391**, 0.52, **0.764** }

4. *Split-LM*:
$Ptime_1 = Ptime_2$ = min(min(100ms, 50ms), 90ms) = 50ms.
AP Utilization = { **0.391**, 0.52, **0.771** }

Figure 5.5: Worked Example. MS_1 and MS_2 use a G.711 vocoder at 64Kbps. $Min\_Ptime_1 = Min\_Ptime_2$ = 20ms, Max_Ptime_1 = 50ms and Max_Ptime_2 = 100ms.

First, the *Basic* VoIP packetization scheme uses the same conventional packetization interval value (20ms) for both $Ptime_1$ and $Ptime_2$. As a consequence, AP_3 utilization under the *Basic* scheme is the largest among all other schemes (Figure 5.5). Next, the *MMAU* algorithm favors AP_3 over AP_1 because of its higher load. Therefore, the algorithm tries to allocate the largest possible value for $Ptime_2$ in order to minimize the resulting utilization for AP_3 due to the admittance of the arriving call. For the *MMAU* algorithm, steps 1 and 2 fail to find a pair of $Ptime_1$ and $Ptime_2$ values that satisfy the latency margin constraint (i.e., infeasible). As expected, the *MMAU* algorithm gives the minimum resulting utilization for AP_3 (**0.761**). The *MaxNetworkCap* algorithm in Figure 5.5 aims at minimizing the sum (U_{AP1} + U_{AP3}) and therefore it tries to allocate the largest possible values for both $Ptime_1$ and $Ptime_2$ without favoring one over the other. It can be noted that the value of (U_{AP1} + U_{AP3}) under *MaxNetworkCap* is the minimum compared to the others. Finally, *Split-LM* scheme uses a straightforward method for apportioning the available latency margin of a VoIP connection. It selects the minimum of three values, namely $MaxPtime_1$ (50ms in the example), $MaxPtime_2$ (100ms in the example) and half of the round-trip available latency margin of the connection between the two end MSs ($0.5 \times LM_{1,3} = 90$ms). This gives the same value (50ms) for both $Ptime_1$ and $Ptime_2$.

5.3.2 MS Power Consumption

In this section, we briefly discuss the energy expenditure for a MS during a VoIP call. The energy calculation used in this chapter is based on the Scheduled-APSD mechanism from IEEE 802.11e. In S-APSD, at the connection setup time, MS negotiates a polling schedule for a QoS traffic stream (TS) with the AP. Based on the negotiated

schedule, an IEEE 802.11e enabled MS awakens at predetermined polling instants. The time interval between the start of successive polling instants is called the polling cycle (or service interval). The timeline is divided into superframes (which is assumed to be 100ms). A superframe is further subdivided into an integer number of equal service intervals.

In a typical VoIP application, the AP polls MS once every packetization interval duration (i.e., service interval = *Ptime*). To further conserve energy, the S-APSD allows MS to switch to doze (sleep) mode between polling instants.

The MS wakes-up just prior to each polling instant, receives a QoS-Poll frame from the AP, sends/receives UL/DL voice packets, acknowledges successful reception and then returns to sleep mode until the next polling instant. The energy expended during the awake period depends on number of packets transmitted/received at UL/DL streams (a packet at each stream), length of packets (function of voice codec rate and *Ptime* value) and PHY layer transmission rate.

In our energy calculations we also assume that MS awakes at the beginning of each superframe interval to sample the periodic beacon frames. As a result, the total power consumption of MS during a VoIP session is the sum of the total energy expenditure for sending/receiving voice packets at both the UL and DL streams of the call, sampling periodic beacon frames and the energy dissipated while the MS is sleeping between consecutive polling intervals (energy calculations are normalized to an interval of 1s). This is computed using typical $\eta_{Tx}/\eta_{Rx}/\eta_D$ power consumption values for IEEE 802.11 chipsets, where η_{Tx}, η_{Rx} and η_D are the power consumed by the MS in transmit, receive/listen and *doze* (sleep) modes, respectively. Power consumption is obtained by multiplying the different power consumption factors by

the times spent in that particular power consumption state.

5.3.3 Station Mobility When Using AVP

In the proposed AVP-scheme, both legs of a VoIP connection are anchored at the AVP-RTS server which sets the VoIP packetization parameters based on information regarding the latency budget of the call path. In Section 5.2, we assume that this information is inferred from whether or not the placed call is intra- or inter-enterprise. In a typical WLAN enterprise setup, intra-enterprise (IN_IN) calls have much smaller latency budget than inter-enterprise (IN_OUT) calls.

When MS performs a horizontal handoff by crossing an IEEE 802.11 BSS boundary, the newly established VoIP leg between the MS and the new AP will still be anchored at the AVP-RTS server. When this happens the call type (i.e., IN_IN or IN_OUT) does not change. The packetization intervals that were originally negotiated could remain, or alternately, new *Ptime* values could be negotiated using standard SIP/SDP signaling procedures [RSC+02, HJ98]. This latter case would be much more complex.

5.3.4 AVP Support in PSQAPs

In Section 5.3, it was assumed that the proposed AVP-RTS server is placed on the switched subnetwork that services the APs as shown in Figure 5.1. In a typical WLAN enterprise setup, a wired distribution system (e.g., Ethernet) is used to transport traffic between different APs in the enterprise.

For a proper incorporation of the AVP scheme into the energy-efficient framework proposed for power-saving QoS-enabled access points in Chapters 3 and 4 of

this thesis, we need to carefully consider the various distinguishable characteristics of wireless mesh networks. For example, because of the lack of a wired distribution system in WMNs, one or more mesh points (MPs) and/or mesh APs (MAPs) will have to carry out the functionalities of an AVP-RTS server. In a WMN deployment, routing is performed in a multi-hop traffic relaying fashion, hence mesh nodes that are likely to be common for all/most VoIP connections are good candidates to perform the adaptive voice packetization functions. Additionally, propagation and queuing delays are no longer negligible, unlike the basic case discussed in Section 5.1. The more hops in a WMN that VoIP packets traverse are, the more this will cut into the end-to-end connection latency and limit the capacity and power saving improvements otherwise possible through AVP. In this case, rendezvous reservation protocols could be used to permit low power multihop infrastructure nodes along the call path to schedule overlapping awakening periods in order to meet applications' QoS requirements [TBJ00, VCTS07].

As regards to problem formulation, the min-max optimization presented in Section 5.3 is still valid for the WMN deployment case, however, the local-optimization solver deployed by both the *MMAU* and *MaxNetworkCap* algorithms need to be revisited. Recall from Section 5.3 that both algorithms search for a sub-optimal AP's loading, U_i, by considering possible feasible *Ptime* assignments for the end APs only. While this is sufficient in a WLAN enterprise similar to the one shown in Figure 5.1, it does not reflect the fact that *Ptime* assignments also affect the loading (and power performance) of all intermediate nodes along the VoIP connection in a WMN setup. Therefore, both heuristics should consider the loading (utilization) of intermediate and end nodes while determining the optimal values for $Ptime_1$ and $Ptime_2$.

Table 5.1: AVP Simulation Parameters

Parameter	**Value**
Number of APs	5
Voice codec	G.711 (64Kbps)
Call arrival rate (λ)	0.1call/s
Average call duration (Exponential)	180s
Calls per run	5000calls
PHY Layer Transmission Rate (R)	11Mbps
Latency margin (LM)	180ms (intra-enterprise calls)
	60ms (inter-enterprise calls)
Maximum MPDU size	2304Bytes
UPD / RTP header	20 / 12 Bytes
MAC / IP header	34 / 8 Bytes
Beacon interval (B)	100ms
SIFS / Slot time	10/20 μs
Time for ACK (T_{ACK})	248μs
Time for PHY Layer Preamble (T_{PHY})	192μs
Guard Time (T_G)	0.5ms
$\eta_{Tx}/\eta_{Rx}/\eta_D$	750/500/8 mW

5.4 Experiments and Results

This section discusses the performance obtained using the AVP-RTS server. First, results from extensive simulation study that show the VoIP capacity and power saving improvements are shown. Following this we provide analytical models and results which characterize the scalability and call blocking probability of the proposed scheme.

5.4.1 Simulation Methodology

A Matlab code was developed for the contention-free HCCA to meet the IEEE 802.11e standard and the proposed AVP scheme specifications. The results were obtained by averaging 10 simulation runs, the simulation time for each was 14 hours. Results were obtained to within ± 0.007 with 95% confidence. The different simulation parameters used in the experiments are summarized in Table 5.1.

We first analyze the influence of the available round-trip latency margin for voice connections on the achieved capacity. The effects on MS power saving when using the AVP-RTS-based schemes are also included. Then simulation results are shown which better examine the effect of our choice of voice codec and underlying IEEE 802.11 PHY on the performance of the proposed voice packetization algorithms. Following this, we present results that compare the performance of the $MMAU$ scheme to an optimal *Ptime* assignment and show that $MMAU$ can give reasonable sub-optimal assignments in typical WLAN settings. Finally, analytical models and results are provided which characterize the scalability of the proposed schemes.

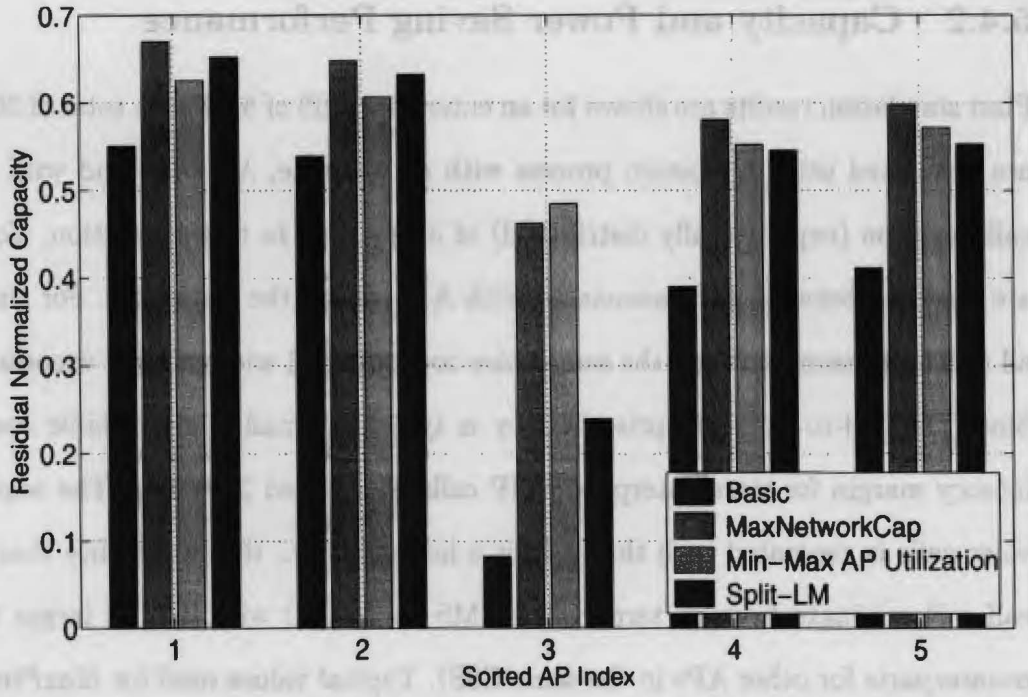


Figure 5.6: Average Residual Normalized Capacity per AP for Different VoIP Packetization Interval Assignment Schemes Assuming Intra-ESS VoIP Calls Only Scenario.

Table 5.2: Average Power Consumption per MS Assuming Intra-Enterprise Calls Only

Ptime Assignment Scheme	Avg. Power (W)
Basic Scheme	0.06
Min-Max AP Utilization (MMAU)	0.032
MaxNetworkCap	0.03
Split LM	0.038

Table 5.3: Residual Network Capacity (Normalized) Assuming Intra-Enterprise Calls Only

Ptime Assignment Scheme	Aggregated APs Residual Capacity
Basic Scheme	0.15
Min-Max AP Utilization (MMAU)	0.56
MaxNetworkCap	0.59
Split LM	0.50

5.4.2 Capacity and Power Saving Performance

First simulation results are shown for an enterprise ESS of 5 APs. A total of 5000 calls are generated using a Poisson process with arrival rate, λ, of 0.1 and with average call duration (exponentially distributed) of 3 minutes. In this simulation, VoIP calls are initiated between MSs associated with APs within the same ESS. For simplicity, all MSs are assumed to use the same voice codec (G.711 with 64Kbps vocoding rate). Since the end-to-end enterprise latency is typically small, the available round-trip latency margin for intra-enterprise VoIP calls is assumed $\geq$ 180ms. The sequence of voice calls is generated such that AP_3 is a hot-spot (i.e., the probability that a voice call will originate from or terminate at MS associated with AP_3 is larger than its counterparts for other APs in the same ESS). Typical values used for *MaxPtime* (i.e., maximum values for packetization interval) are uniformly distributed between 20ms and 100ms.

Figure 5.6 shows the residual capacity per AP for the different assignment schemes. Clearly the *MMAU* algorithm gives the highest residual capacity (i.e., minimum loading) for the hot spot node (AP_3). In numbers, the *MMAU* algorithm allows for a residual capacity of about 50% of AP_3's total VoIP capacity. This residual is a bit more than 600% of that obtained by Basic VoIP and as high as 250% of that obtained by the *Split-LM* scheme.

Table 5.2 shows the average power dissipation during an active call for MSs under each packetization interval assignment scheme. It is clear that *MMAU* saves about 50% of the energy expended at the MS side compared to the conventional (basic) voice packetization, which means doubling the battery life-time for voice MSs. While the *MaxNetworkCap* algorithm is as energy efficient as the *MMAU*, *Split-LM* consumes

Table 5.4: Average Power Consumption per MS Under a Call Mix of Intra- and Inter-Enterprise Calls

Ptime Assignment Scheme	**Avg. Power (W)**
Basic Scheme	0.06
Min-Max AP Utilization (MMAU)	0.042
MaxNetworkCap	0.042
Split LM	0.045

Table 5.5: Residual Network Capacity (Normalized) Under a Call Mix of Intra- and Inter-Enterprise Calls

Ptime Assignment Scheme	**Aggregated APs Residual Capacity**
Basic Scheme	0.15
Min-Max AP Utilization (MMAU)	0.36
MaxNetworkCap	0.37
Split LM	0.36

more power than both former algorithms. Table 5.3 also shows that the *MMAU* scheme can increase the VoIP capacity of the WLAN by more than threefold compared with the basic scheme.

In the next experiment we characterize the performance of the AVP-RTS server in cases where not all the active connections enjoy large round-trip latency margins. Five access points are split over two ESSs such that APs 1, 2 and 3 are co-located in ESS 1 while AP_4 and AP_5 are co-located in ESS 2. A total of 5000 calls are generated randomly according to a Poisson arrival process with a mean rate, λ, of 0.1 and with an average call duration of 3 minutes (call duration is exponentially distributed). The call arrivals consist of a mix of intra-ESS (25%) and inter-ESS (75%) calls. VoIP calls are initiated only between MSs associated with any pair of different APs in both ESSs (i.e., no intra-AP calls are considered). The available round-trip latency margin

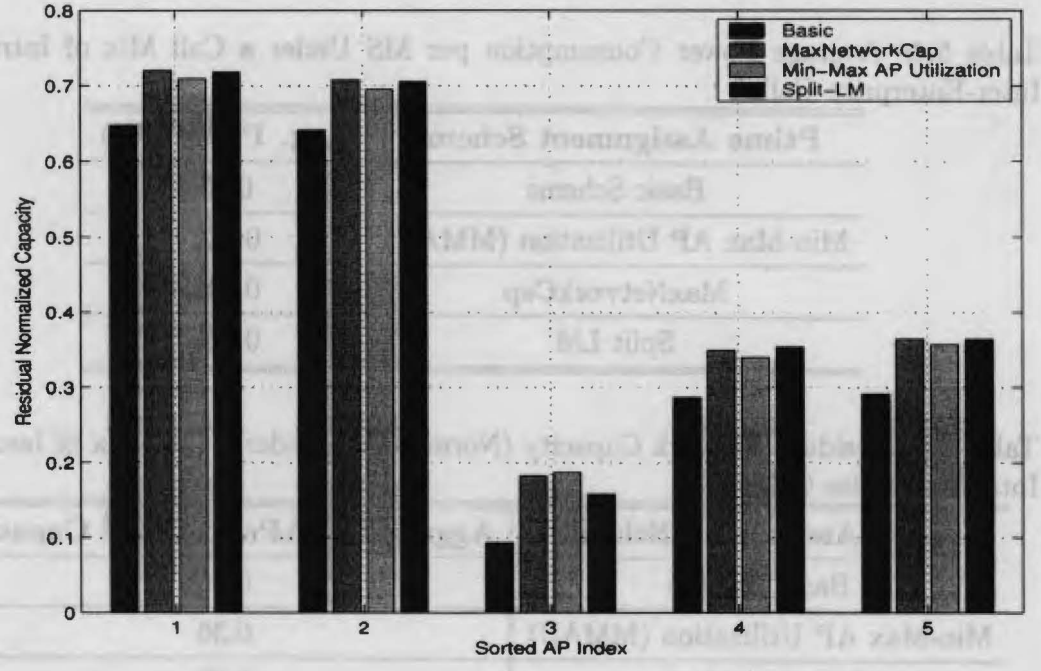


Figure 5.7: Average Residual Normalized Capacity per AP for Different VoIP Packetization Interval Assignment Schemes Assuming a Call-Mix of Intra- and Inter-Enterprise Calls.

for intra-ESS calls is assumed to be $\geq$ 180ms while it is $\leq$ 60ms for inter-ESS calls. Again, voice calls are generated in a way that makes AP_3 the most heavily loaded. *MaxPtime* values used are uniformly distributed between 20ms and 100ms.

In Figure 5.7 the *MMAU* algorithm achieves a residual capacity for AP_3 that is only 200% and 115% of that obtained by the *Basic* and *Split-LM* schemes, respectively. Although the achieved capacity improvements for AP_3 are less than those obtained in Figure 5.6 above, the *MMAU* algorithm is able to obtain the highest residual capacity for hot spot AP_3 compared to the other schemes.

The limited capacity improvements in the presence of inter-ESS calls are expected since the available latency margin for inter-ESS VoIP calls is much smaller than that for intra-ESS calls. For intra-ESS calls, the small latency budget allows the *MMAU*

scheme to apportion the margin in a way (i.e., by using large values for $Ptime_1$ and $Ptime_2$ of individual VoIP connections) that improves the capacity of hotspot APs. However, when inter-ESS VoIP calls are introduced to the total call mix, the AVP-RTS server does not have much to do because of the limited latency margin which directly lowers the ceiling of the calculated values for $Ptime_1$ and $Ptime_2$. Similarly, we can explain the limited performance improvements of the AVP-RTS server in terms of power conservation at the MS side (Table 5.4) and the aggregate residual capacity of the entire network (Table 5.5) when the call-mix contains a significant portion of inter-ESS calls.

IEEE 802.11a/g allows for a higher system capacity than IEEE 802.11b, due to its shorter PHY preamble and contention time slot (see Table 2.1). Although it uses the same MAC mechanism(s) as 802.11b, 802.11a/g maximum data rate is about five times larger than that of 802.11b. Using the MAC parameters listed in Table 2.1, we have performed simulations to validate the VoIP capacity improvements achieved when using the AVP-RTS scheme. Here, we assume a WLAN enterprise identical to the one used in Section 5.4.2. In Tables 5.6, 5.7 and 5.8, VoIP capacities for the most heavily loaded AP (AP_3) are shown for the Basic scheme that uses 20ms packetization intervals as well as for the other three AVP heuristic algorithms assuming different CBR codecs.

As expected, due to its higher data rates(s), IEEE 802.11a [std05b] achieves much higher capacities than IEEE 802.11b for all voice codecs. However it is interesting to note that for any given choice of vocodec, the VoIP capacity achieved by any packetization scheme for 802.11a (at 54Mbps) is larger than that of 802.11b (11Mbps) for the same packetization scheme by more than 5 times. This is due to the smaller

slot time, PHY header and ACK frames of 802.11a. In the 802.11g mode, timing spaces even smaller than those used in 802.11a (Table 2.1), result in a slightly higher capacity than 802.11a.

Two other important observations can be noted. First, given a transmission mode, the capacity for ordinary VoIP does not decrease much as the data rate decreases. For example, for the 802.11a/g transmission mode and G.711 voice codec, even when the data rate drops from 54 to 18Mbps, the capacity for Basic VoIP only drops by about 40%. This is because the change of data rate only affects the transmission time of the payload, which corresponds to a small proportion of the total transmission time. The major part, such as PHY, IFS and ACK, is data rate independent. Similarly, we can explain why the lower rates of G.726 and G.729 voice codecs fail to bring about corresponding higher capacities if the *Basic* packetization scheme is used.

Conversely, the *MMAU*, *MaxNetworkCap* and *Split-LM* algorithms make better use of higher PHY data rates to achieve much larger VoIP capacities than that of the *Basic* algorithm. Since the proposed *MMAU* algorithm continuously tries to maximize the payload size, the transmission time of the payload contributes significantly to the total packet transmission time. Therefore, VoIP capacities for all AVP-heuristic algorithms, especially the *MMAU* scheme, change dramatically when changing the underlying PHY rate. This observation is valid regardless of the CBR codec used. Again, for the same reason we can relate the dramatic increase in VoIP capacity for *MMAU*, *MaxNetworkCap* and *Split-LM* schemes when a lower rate voice codec is used given the same transmission mode and data rate.

For the various transmission modes, PHY data rates and voice codecs, the three

Table 5.6: VoIP Capacity of 802.11b and 802.11a Derived from Simulations for *Basic*, *MMAU*, *MaxNetworkCap* and *Split-LM* Packetization Schemes Assuming G.711 Codec (64Kbps)

MAC	Basic	MMAU	MaxNetowrkCap	Split-LM
802.11b (11Mbps)	16.1 calls	0.6277 (43 calls)	0.6094 (41 calls)	0.4270 (27.9 calls)
802.11b (5.5Mbps)	12.7 calls	0.5656 (27.6 calls)	0.5502 (26.7 calls)	0.3956 (19.9 calls)
802.11a (54Mbps)	105.6 calls	0.58 (250.6 calls)	0.56 (239.8 calls)	0.39 (170.7 calls)
802.11a (36Mbps)	89.3 calls	0.53 (188.6 calls)	0.51 (182.1 calls)	0.35 (136.7 calls)
802.11a (18Mbps)	60.9 calls	0.45 (108.3 calls)	0.43 (105.7 calls)	0.30 (89.1 calls)

adaptive VoIP packetization schemes can improve VoIP capacity over that of the *Basic* scheme. It can be also seen that the achieved improvements are higher for lower rate codecs. For example, the *MaxNetworkCap* algorithm allows for roughly 250% capacity improvements for G.729 codec while the same packetization algorithm yields only 125% improvements when G.711 codec is used. The *MMAU* heuristic scheme achieves the highest VoIP capacity improvements for AP_3 which is the most heavily loaded AP. The results show that capacity improvement of about 300% is possible when a G.729 codec is used.

Next, the *MMAU* algorithm is also applied to two WLAN setups identical to the one shown in Figure 5.1 and its performance is compared, in terms of utilization of hot-spot node, to that of an algorithm that assigns *Ptimes* optimally. Brute force enumeration was used to obtain the optimal assignment of packetization interval values for VoIP calls generated in both networks. Because of the exponential complexity

Table 5.7: VoIP Capacity of 802.11b and 802.11a Derived from Simulations for *Basic*, *MMAU*, *MaxNetowrkCap* and *Split-LM* Packetization Schemes Assuming G.726 Codec (32Kbps)

MAC	Basic	MMAU	MaxNetowrkCap	Split-LM
802.11b (11Mbps)	17.8 calls	0.70 (56.9 calls)	0.68 (53.9 calls)	0.48 (32.6 calls)
802.11b (5.5Mbps)	14.8 calls	0.66 (40.6 calls)	0.64 (38.8 calls)	0.45 (25.6 calls)
802.11a (54Mbps)	120.7 calls	0.66 (356.4 calls)	0.64 (335.2 calls)	0.4 (215.7 calls)
802.11a (36Mbps)	106.1 calls	0.63 (283.8 calls)	0.61 (269.1 calls)	0.42 (181.3 calls)
802.11a (18Mbps)	77.9 calls	0.56 (176.1 calls)	0.54 (169.2 calls)	0.38 (123.6 calls)

of brute force, the number of generated calls is limited to 3 so that an optimum solution for the *MMAU* optimization problem can be obtained. In each WLAN setting, three calls were generated with random parameters (i.e., with random sources, destinations, round-trip latency margins, $MaxPtime_1$ and $MaxPtime_2$). For simplicity, the G.711 codec is used. S_i is defined as the probability that a generated VoIP call will originate from AP_i and D_i as the probability that a call will be destined to AP_i (where $i = 1, 2, 3$). In the first WLAN setting, we assume that $S_1 = S_2 = S_3$ and $D_1 = D_2 = D_3$ (i.e., all APs in the enterprise have almost the same VoIP traffic load). Note that $S_1 + S_2 + S_3 = 1$ and $D_1 + D_2 + D_3 = 1$. On the other hand, AP_1 is assumed to be the hot spot in the second setup. This is achieved by setting $S_1 > S_2 = S_3$ and $D_1 > D_2 = D_3$.

Tables 5.9 and 5.10 compare the average utilization for AP_1 achieved by using the *MMAU* algorithm against the optimal solution obtained by brute force for the two cases. Table 5.9 corresponds to the WLAN configuration where all APs are

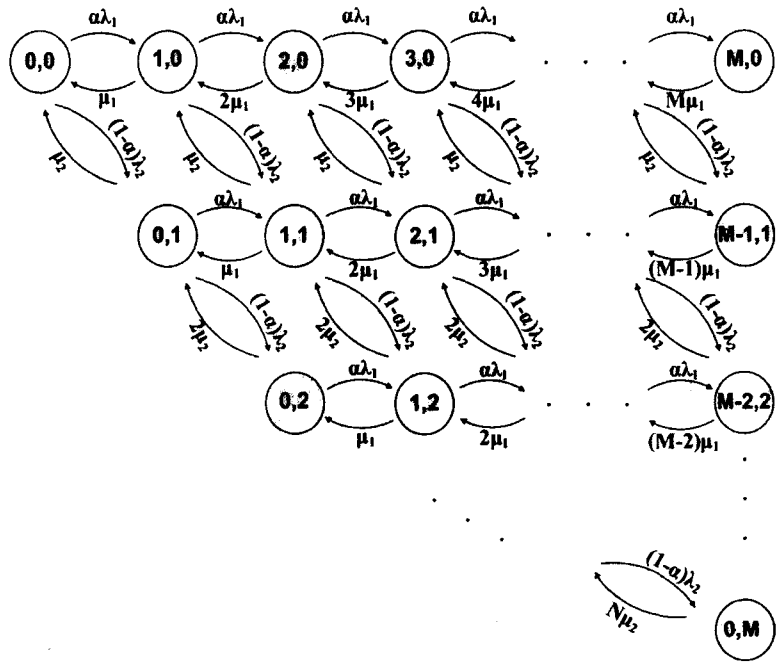


Figure 5.8: Markov Process for the Basic (Ordinary) Packetization Scheme.

equally loaded, whereas Table 5.10 shows results for the second scenario where AP_1 is assumed to be the most heavily loaded node. Results from Table 5.9 indicate the optimal solution obtained for the *MMAU* optimization problem. As for the second case, Table 5.10 indicates that the proposed algorithm was unable to yield the optimal *Ptime* assignment (because AP_1 has a higher utilization under the proposed *MMAU* scheme than it has under the brute force enumeration method). However, it can be seen that the suboptimal utilization obtained is very close to the optimal one.

Table 5.8: VoIP Capacity of 802.11b and 802.11a Derived from Simulations for *Basic*, *MMAU*, *MaxNetowrkCap* and *Split-LM* Packetization Schemes Assuming G.729 Codec (8Kbps)

MAC	Basic	MMAU	MaxNetowrkCap	Split-LM
802.11b (11Mbps)	19.3 calls	0.75 (75.4 calls)	0.73 (70.1 calls)	0.5 (38.4 calls)
802.11b (5.5Mbps)	17 calls	0.73 (62.7 calls)	0.71 (58.6 calls)	0.48 (33.6 calls)
802.11a (54Mbps)	135.3 calls	0.74 (520.7 calls)	0.72 (477.9 calls)	0.49 (266 calls)
802.11a (36Mbps)	123.6 calls	0.73 (456 calls)	0.71 (419.4 calls)	0.49 (238.8 calls)
802.11a (18Mbps)	98.3 calls	0.7 (331.8 calls)	0.68 (308.2 calls)	0.45 (184 calls)

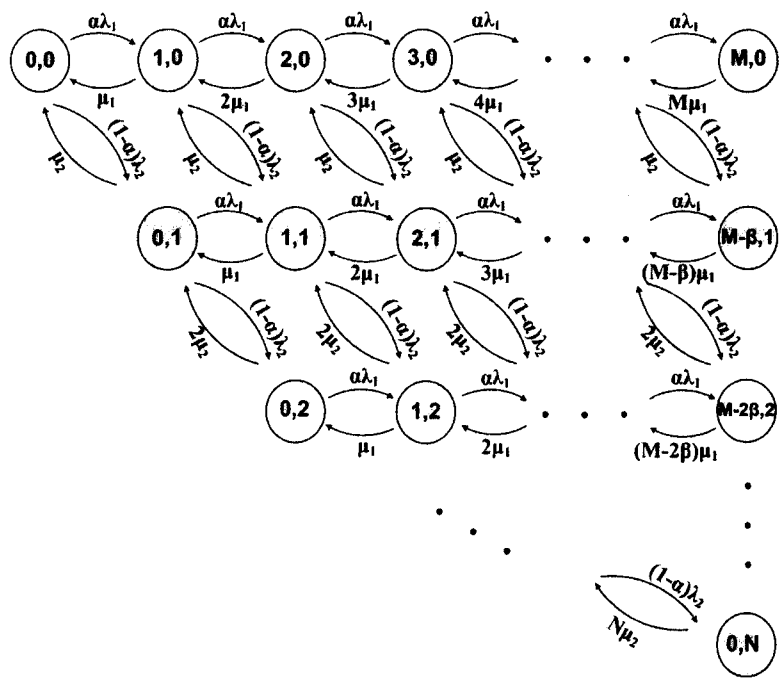


Figure 5.9: Markov Process for the *Split-LM* Packetization Scheme.

Table 5.9: Utilization of AP_1 Under Both Heuristic (MMAU) and Optimal Min-Max AP Utilization Algorithms (All APs Have the Same Traffic Load)

Ptime Assignment Scheme	**AP_1 Utilization**
MMAU (Heuristic)	0.09
Min-Max AP Utilization (Optimal)	0.09

Table 5.10: Utilization of AP_1 Under Both Heuristic (MMAU) and Optimal Min-Max AP Utilization Algorithms (AP_1 Has the Highest Traffic Load)

Ptime Assignment Scheme	**AP_1 Utilization**
MMAU (Heuristic)	0.17
Min-Max AP Utilization (Optimal)	0.15

5.4.3 Call Blocking Performance

In this section the performance of the adaptive packetization schemes is examined based on blocking probability. An enterprise setting consisting of two ESSs is considered for this study. We allow for two different types of arrivals, namely, intra-enterprise and inter-enterprise calls. Intra-enterprise calls, referred to as IN_IN, are established between two users who reside on the same enterprise LAN (i.e., ESS1 or ESS2). An IN_IN call enjoys a round-trip available latency margin (LM) of 180ms. Inter-enterprise calls, referred to as IN_OUT, are established between two users residing in two different LANs. IN_OUT calls are assumed to have a limited round-trip latency margin of 60ms. If a new call arrives and does not find sufficient resources at either end AP, then the call is blocked.

Analytical models are developed for new call blocking probability to assess the *Basic* and the *Split-LM* VoIP packetization schemes. The calls mix is composed of a fraction, α, of IN_IN calls whereas the rest are IN_OUT. The above case can be

modeled as a Markov process, provided that new call arrivals for both call classes are Poisson (with rates equal to λ_1 and λ_2, for IN_IN and IN_OUT calls, respectively). Call durations (i.e., service times) are assumed to be exponentially distributed with means of $1/\mu_1$ (s) and $1/\mu_2$ (s) for IN_IN and IN_OUT calls, respectively. We assume that an access point under the *Basic* packetization scheme is capable of supporting a maximum of M VoIP calls concurrently. Given the conventional fixed packetization interval values of 10ms or 20ms, the value of M can be easily calculated.

Figure 5.8 shows the Markov process for the *Basic* packetization scheme. In this figure, the state $s(i,j)$ indicates i active intra-enterprise calls and j inter-enterprise calls where $i,j = 0 \ldots M$. From the figure the new call blocking probability is given by,

$$B_{new} = \sum_{i=0}^{M} Pr\{s(M-i,i)\}. \tag{5.4}$$

It should be clear that if $\lambda_1 = \lambda_2$ and $\mu_1 = \mu_2$, then Equation 5.4 above reduces to the Erlang B loss formula.

We also calculate the blocking probability of the *Split-LM* packetization algorithm. To do this, all voice handsets are assumed to have common values for $MaxPtime_1$ and $MaxPtime_2$ for both types of call arrivals. This assumption is necessary to calculate the bandwidth requirements for a VoIP call (whether it is IN_IN or IN_OUT) under the *Split-LM* algorithm. Figure 5.9 provides a Markov process describing the *Split-LM* packetization mechanism. In this figure, the ratio of bandwidth requirements of an IN-OUT call to that of an IN_IN call is assumed to equal $\beta \geq 1$. Accordingly, an AP can admit a maximum of M IN_IN calls or N IN_OUT calls simultaneously where $M/N = \beta$. In Figure 5.9, the state $s(i,j)$ indicates i active intra-enterprise calls and

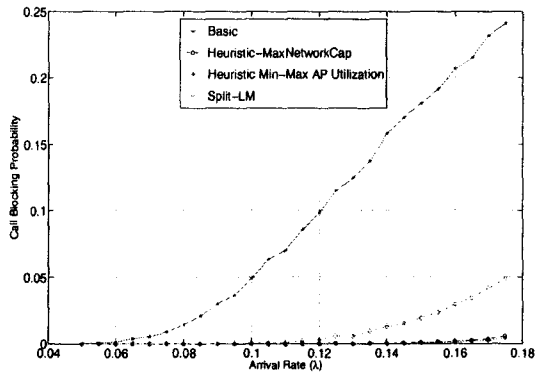
j inter-enterprise calls such that $i = 0 \ldots M, j = 0 \ldots N$, $M/N = \beta$ and $\beta \geq 1$. From the figure, it can be seen that the new call blocking probability is given by,

$$B_{new} = \sum_{i=0}^{N} Pr\{s(M - i\beta, i)\}. \tag{5.5}$$

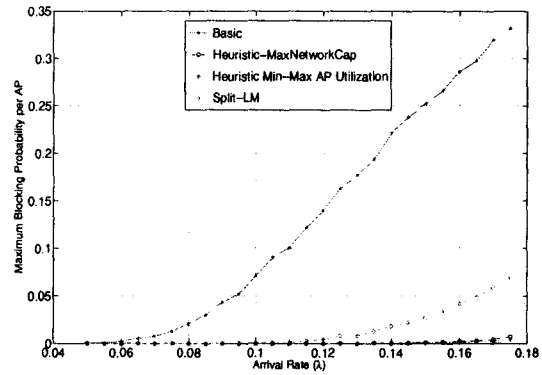
Figures 5.10, 5.11 and 5.12 compare the new call blocking probability for different call compositions and under different call arrival rates. An arriving call originates from AP_i in one of the two ESSs according to a predefined probability δ_i where $\overline{\delta} = \{0.1, 0.1, 0.6, 0.1, 0.1\}$. Note that the δ_is are configured in a way that makes AP_3 a hot spot.

In each figure, both the total blocking probability and the maximum AP blocking probability (i.e., blocking probability for AP_3) are shown for the *Basic*, *MaxNetwork-Cap*, *MMAU* and *Split-LM* schemes. From the figures, the total and maximum new call blocking probabilities for all algorithms increase with the values of λ. In the same figures, both total and maximum new call blocking probabilities, when the *Basic* scheme is used, are independent of the IN_IN percentage of total calls, α. This is due to the nature of the *Basic* algorithm that uses the same voice packetization interval for all VoIP streams whether they are IN_IN and/or IN_OUT calls. The *MaxNetworkCap* and *Split-LM* algorithms also achieve a better blocking probability than that of the *Basic* scheme, while *MMAU* adaptive VoIP packetization algorithm yields the least new call blocking probabilities for all values of λ and α.

Since the length of the packetization intervals calculated by the three AVP algorithms (*MMAU*, *MaxNetworkCap* and *Split-LM*) are functions of the available latency margin of the calls, the blocking probabilities of the three algorithms increase as the percentage of IN_OUT calls of total call-mix increases. This is mainly because

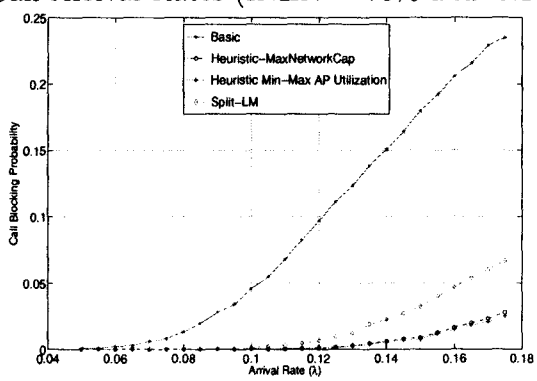


(a) Blocking Probability for Entire WLAN Enterprise Setting.

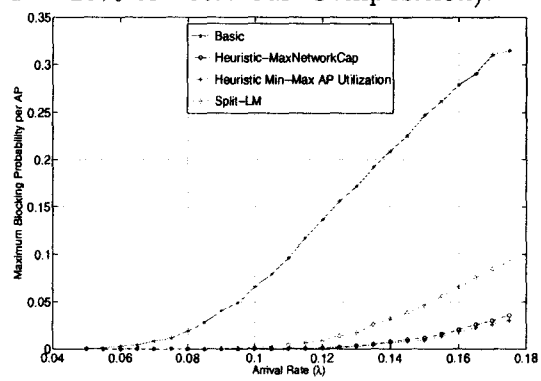


(b) Maximum AP Blocking Probability.

Figure 5.10: Total System and Maximum AP Blocking Probability Under Different Call Arrival Rates (IN_IN = 75% and IN_OUT = 25% of Total Call Composition).

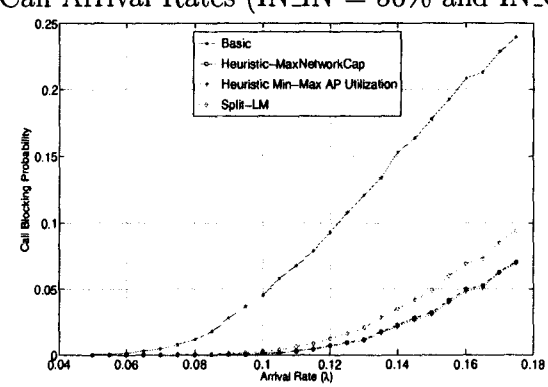


(a) Blocking Probability for Entire WLAN Enterprise Setting.

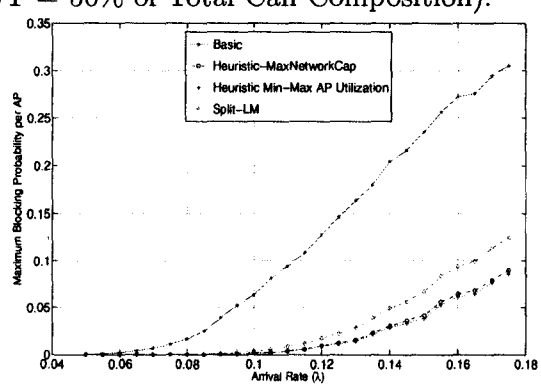


(b) Maximum AP Blocking Probability.

Figure 5.11: Total System and Maximum AP Blocking Probability Under Different Call Arrival Rates (IN_IN = 50% and IN_OUT = 50% of Total Call Composition).



(a) Blocking Probability for Entire WLAN Enterprise Setting.



(b) Maximum AP Blocking Probability.

Figure 5.12: Total System and Maximum AP Blocking Probability Under Different Call Arrival Rates (IN_IN = 25% and IN_OUT = 75% of Total Call Composition).

IN_OUT calls have a limited latency margin which restricts the maximum packetization intervals that are negotiated by the AVP-RTS server. From Figures 5.10(b), 5.11(b) and 5.12(b), the superiority of the *MMAU* algorithm over all other schemes in minimizing the call blocking probability of AP_3 is clear. The value of having an AVP server to adaptively packetize voice is well justified over splitting the latency margin, a technique that smart voice handsets may deploy to improve bandwidth efficiency. The *MMAU* algorithm can achieve a maximum AP blocking probability that is only 15% of that obtained by *Split-LM* for $\alpha = 75\%$; blocking probability increases as α decreases.

5.5 Conclusions

This chapter proposed and investigated the use of an enterprise-based adaptive voice packetization server (AVP-RTS) which splits the RTP VoIP connection into two call legs. In this way, voice coding parameters at each end of the call are negotiated separately and the server can allocate the available latency margin (and the ensuing capacity gain) asymmetrically across the call.

When considering both enterprise WLAN endpoints, the problem was formulated as a min-max optimization so that the maximum capacity gain can be allocated to those WLANs which have the minimum spare capacity. Several heuristic algorithms (*MMAU*, *MaxNetCap* and *Split-LM*) were proposed for solving this optimization problem and their performance was assessed and compared. The simulation results presented show that by using the AVP-RTS server we can significantly improve the multi-AP VoIP capacity for certain typical IEEE 802.11 situations.

The results also show that the actions of the AVP server can significantly reduce

average handset power consumption compared with the typical default VoIP parameter case. The call blocking probability of the adaptive voice packetization scheme was analyzed using a Markov model. Results obtained from this model also showed the efficiency of using the AVP-RTS server in minimizing both the total- and maximum-AP new call blocking probabilities for a typical VoWLAN installation.

Chapter 6

Conclusions and Future Work

6.1 Conclusions

In this thesis we considered Quality-of-Service provisioning in low power wireless mesh networks. For this type of deployment, classical fixed power connections are replaced by a battery operated or solar powered design. In this architecture, a power saving mesh node can operate at less than 100% utilization to conserve energy and reduce cost. The scheduling of unavailability (i.e., sleep) periods of a power-saving AP poses several challenges for QoS guarantees due to the extra delay and loss experienced by end-user traffic during the channel unavailability periods.

We presented a framework for a power saving quality-of-service (QoS) enabled access point (PSQAP), intended for use in low power infrastructure applications. The proposed framework introduces infrastructure based power saving while preserving the QoS requirements for delay and loss intolerant real-time applications. The framework, which consists of an energy efficient media access control (MAC) protocol and an adaptive connection admission control (CAC) scheme, uses an enhanced version

of the mandatory IEEE 802.11e US-APSD access mechanism. This facilitates the application of the framework to existing IEEE 802.11-compliant devices. Using a simulation-based study of a PSQAP serving real-time multimedia traffic streams, we quantified the timing constraints for scheduling both the awake and sleep periods of a power saving AP. It was found that careful scheduling of the awake/sleep periods of an AP is essential for both minimizing the AP power consumption and successfully meeting the QoS requirements of real-time applications. The findings highlight the fundamental differences between infrastructure based power-saving for best-effort traffic and real-time audio/video flows.

A media access control protocol, which uses a novel scheme for selective activation/deactivation of service intervals, was proposed. We showed that our scheme of selective SI (de)activation minimizes the signaling required to update the active QSTAs on AP's duty-cycle changes. We also proposed a dynamic connection admission control scheme for controlling the air-time usage of QSTAs. The proposed CAC scheme is located at the PSQAP and requires no upgrade to the IEEE 802.11 handsets. We summarized our recommendations for choosing the optimal service interval values. The recommendations allow for serving real-time flows having a relatively small delay bound without sacrificing much of the channel-time and power consumption for frequent AP state transitions. Through simulations we showed that a power-saving AP which implements the proposed energy-efficient framework can reduce its power consumption by roughly an order of magnitude, while guaranteeing the QoS requirements for real-time services. These improvements, however, are accompanied with doubling the power consumption values at the handsets. A simple scheduling scheme was introduced for distributing the handsets' channel accesses over

different SIs. Simulation results obtained showed that when fewer stations contend for the channel in an SI, the handset power consumption decreases as a direct result of the reduced queuing.

An architecture was proposed to allow for efficient and low-energy transmissions of multiplexed videoconferencing flows over power saving QAPs. Under the contention-based US-APSD mechanism, we showed that the use of accurate capacity predictions is more energy efficient, compared to "Wait-until-Complete" type of scheduling algorithms, for a PSQAP serving bursty video applications. A discrete autoregressive model DAR(1) was used to estimate the aggregate bandwidth requirements for multiplexed H.263 videoconferencing streams. We presented the results of a simulations-based comparative study that evaluated several algorithms for performing video capacity prediction and resource allocation. We also showed that the US-APSD based architecture falls short of achieving a satisfactory power performance at the handsets. The results also confirmed the conflicting relationship between power saving requirements at both the PSQAP and stations in a contention-based MAC protocol. We proposed a more handset power-efficient mechanism whereby a modified version of the US-PSMP with one SI was used with a DAR(1)-based CAC scheme to achieve a near optimal power performance at both the PSQAP and handsets. An analytic characterization of the handset power performance of both architectures was also provided.

To overcome the problem of low capacity of VoIP over WLAN networks, we proposed and investigated the use of an enterprise-based adaptive voice packetization server (AVP-RTS) which splits the RTP VoIP connection into two call legs. In this way, voice coding parameters at each end of the call are negotiated separately and

the server can allocate the available latency margin (and the ensuing capacity gain) asymmetrically across the call. When considering both enterprise WLAN endpoints, the problem was formulated as a min-max optimization so that the maximum capacity gain can be allocated to those WLANs which have the minimum spare capacity. Several heuristic algorithms (*MMAU*, *MaxNetCap* and *Split-LM*) were proposed for solving this optimization problem and their performance was assessed and compared. The simulation results presented show that by using the AVP-RTS server we can significantly improve (at least double) the multi-AP VoIP capacity for certain typical IEEE 802.11 situations. The results also show that the actions of the AVP server can significantly reduce average handset power consumption compared with the typical default VoIP parameter case. The call blocking probability of the adaptive voice packetization scheme was analyzed using a Markov model. Results obtained from this model also showed the improved efficiency of using the AVP-RTS server in minimizing both the total- and maximum-AP new call blocking probabilities for a typical VoWLAN installation.

6.2 Future Work

Finally, this work has spurred several research directions including the incorporation of handset handoffs into the proposed framework. In this case, the bandwidth and timing requirements of an imminent horizontal handoff, resulting from a handset crossing an IEEE 802.11 boundary, can be defined as a TSPEC element that a PSQAP processes according to the procedures defined in the proposed framework. This work can also be expanded by considering downlink traffic in addition to the uplink traffic only scenarios which have been studied in this work. Some challenging

examples might include asynchronous uplink and downlink traffic belonging to the same multimedia sessions. Relaying traffic in multi-hop low power mesh networks can be also a challenging problem in terms of scheduling the service intervals and activity periods for power-saving mesh nodes along the multimedia path.

Bibliography

[AEF+03] F. Anjum, M. Elaoud, D. Famolari, A. Ghoshand, R. Vaidyanathan, A. Dutta, P. Agrawal, T. Kodama, and Y. Katsube. Voice Performance in WLAN Networks An Experimental Study. In *Proc. of IEEE Global Telecommunications Conference (GLOBECOM)*, volume 6, pages 3504–3508, December 2003.

[AMM01] R. Ananthapadmanabha, B.S. Manoj, and C.S.R. Murthy. Multi-hop Cellular Networks: The Architecture and Routing Protocols. In *Proc. of the 12th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, volume 2, pages G–78–G–82, September/October 2001.

[AN00] T. Adachi and M. Nakagawa. Capacity Analysis for a Hybrid Indoor Mobile Communication System Using Cellular and Ad Hoc Modes. In *Proc. of the 11th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, volume 2, pages 767–771, 2000.

[ANF05] M. Anand, E.B. Nightingale, and J. Flinn. Self-tuning Wireless Network Power Management. *Wireless Networks*, 11(4):451–469, 2005.

[ANT06] P. Ansel, Q. Ni, and T. Turletti. FHCF: A Simple and Efficient Scheduling Scheme for IEEE 802.11e Wireless LAN. *Mobile Networks and Applications*, 11(3):391–403, 2006.

[BA02] K.P. Burnham and D.R. Anderson. *Model Selection and Multi-Model Inference, 2nd Ed.* Springer-Verlag, New York., 2002.

[Bia98] G. Bianchi. IEEE 802.11-Saturation Throughput Analysis. *IEEE Communications Letters*, 2(12):318–320, 1998.

[BIMR01] R.O. Baldwin, N.J. Davis IV, S.F. Midkiff, and R.A. Raines. Packetized Voice Transmission Using RT-MAC, a Wireless Real-time Medium Access

Control Protocol. *SIGMOBILE Mobile Computing and Communications Review*, 5(3):11–25, 2001.

[Bin08] Advances in Wireless Local Area Networks. http://www.comsoc.org/tutorialsnow/Bing/index.html, April 2008.

[BPCQ03] A. Banchs, X. Perez-Costa, and D. Qiao. Providing Throughput Guarantees in IEEE 802.11e Wireless LANs. In *Proc. of the 18th Interntaional Teletraffic Congress*, September 2003.

[Bra69] P. Brady. A Model for Generating On-Off Speech Patterns in Two-Way Conversation. *Bell Systems Technical Journal*, 48(7):2245–2272, 1969.

[BST08] G.H. Badawy, A.A. Sayegh, and T.D. Todd. Solar Powered WLAN Mesh Network Provisioning for Temporary Deployments. In *Proc. of IEEE Wireless Communications and Networking Conference (WCNC)*, pages 2271–2276, Las Vegas, NV, USA, 2008.

[BV04] S. Boyd and L. Vandenberghe. *Convex Optimization.* Cambridge University Press, 2004.

[CCW03] R.-S. Chang, W.-Y. Chen, and Y.-F. Wen. Hybrid Wireless Network Protocols. *IEEE Transactions on Vehicular Technology*, 52(4):1099–1109, July 2003.

[Cha03] S. Chandra. Wireless Network Interface Energy Consumption: Implications for Popular Streaming Formats. *Multimedia Systems*, 9(2):185–201, 2003.

[CKP08] S. Chatziperis, P. Koutsakis, and M. Paterakis. A New Call Admission Control Mechanism for Multimedia Traffic over Next-Generation Wireless Cellular Networks. *IEEE Transactions on Mobile Computing*, 7(1):95–112, January 2008.

[CSA99] J.-C. Chen, K.M. Sivalingam, and P. Agrawal. Performance Comparison of Battery Power Consumption in Wireless Multiple Access Protocols. *Wireless Networks*, 5(6):445–460, 1999.

[CSE04] Y. Chen, N. Smavatkul, and S. Emeott. Power Management for VoIP over IEEE 802.11 WLAN. In *Proc. of IEEE Wireless Communications and Networking Conference (WCNC)*, volume 3, pages 1648–1653, 21-25 March 2004.

[CSS05] C.-T. Chou, S.N. Shankar, and K.G. Shin. Achieving Per-Stream QoS with Distributed Airtime Allocation and Admission Control in IEEE 802.11e Wireless LANs. In *Proc. of IEEE INFOCOM*, volume 3, pages 1584 – 1595, March 2005.

[CV02] S. Chandra and A. Vahdat. Application-specific Network Management for Energy-Aware Streaming of Popular Multimedia Formats. In *Proc. of the General Track of the Annual Conference on USENIX Annual Technical Conference (ATEC)*, pages 329–342, Berkeley, CA, USA, 2002.

[CWX+03] X. Chen, C. Wang, D. Xuan, Z. Li, Y. Min, and W. Zhao. Survey on QoS management of VoIP. In *Proc. of International Conference on Computer Networks and Mobile Computing (ICCNMC)*, pages 69–77, October 2003.

[DC99] D.-J. Deng and R.-S. Chang. A Priority Scheme for IEEE 802.11 DCF Access Method. *IEICE Transactions on Communications*, E82-B(1):96–102, 1999.

[Dep04] I. S. Department. IEEE 802.11s ESS Mesh Networking Working Group, 2004.

[DH99] D.A. Dyson and Z.J. Haas. A Dynamic Packet Reservation Multiple Access Scheme for Wireless ATM. *Mobile Networks and Applications (MONET) Journal*, 4(2):87–99, 1999.

[DPZ04] R. Draves, J. Padhye, and B. Zill. Routing in Multi-Radio, Multi-Hop Wireless Mesh Networks. In *Proc. of the 10th Annual International Conference on Mobile Computing and Networking (MobiCom '04)*, pages 114–128, New York, NY, USA, 2004. ACM.

[Eni] Eningeering a WLAN Network. A White Paper, Nortel Networks. http://www.nortel.com/promotions/whitepaper/wlan/collateral/NN108980-082304.pdf.

[FGTB04] W.F. Fan, D. Gao, D.H.K. Tsang, and B. Bensaou. Admission Control for Variable Bit Rate Traffic in IEEE 802.11e WLANs. In *Proc. of the 13th IEEE Workshop on Local and Metropolitan Area Networks (LANMAN)*, pages 61–66, 25-28 April 2004.

[FR01] F.H.P. Fitzek and M. Reisslein. MPEG-4 and H.263 Video Traces for Network Performance Evaluation. *IEEE Network*, 15(6):40–54, November/December 2001.

[FT05] S. Fashandi and T.D. Todd. Real-Time Handoff in Solar/Battery Powered ESS Mesh Networks. In *Proc. of the IEEE 16th International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, volume 3, pages 1489–1494, 11-14 Sept. 2005.

[FT07] A. Farbod and T.D. Todd. Resource Allocation and Outage Control for Solar-Powered WLAN Mesh Networks. *IEEE Transactions on Mobile Computing*, 6:960–970, August 2007.

[Gas05] M. Gast. *802.11 Wireless Networks: The Definitive Guide.* O'Reilly Media, Inc., 2nd edition, 2005.

[GCZ05] D. Gao, J. Cai, and L. Zhang. Physical Rate Based Admission Control for HCCA in IEEE 802.11e WLANs. In *Proc. of the 19th IEEE International Conference on Advanced Information Networking and Applications (AINA)*, pages 479–483, Washington, DC, USA, 2005.

[GK02] S. Garg and M. Kappes. On the Throughput of 802.11b Networks for VoIP. Technical Report: http://www.research.avayalabs.com/techreportY.html/ ALR-2002-012, Avaya Laboratories, Basking Ridge, NJ, USA., 2002.

[GK03] S. Garg and M. Kappes. Can I Add a VoIP Call? In *Proc. of IEEE International Conference on Communications (ICC)*, volume 2, pages 779–783, May 2003.

[GMN03] A. Grilo, M. Macedo, and M. Nunes. A Scheduling Algorithm for QoS Support in IEEE 802.11 Networks. *IEEE Wireless Communications*, 10(3):36–43, June 2003.

[GZ03] D. Gu and J. Zhang. A New Measurement-based Admission Control Method for IEEE802.11 Wireless Local Area Networks. In *Proc. of the 14th IEEE Personal, Indoor and Mobile Radio Communications (PIMRC)*, volume 3, pages 2009–2013, 7-10 Sept. 2003.

[HIIM02] T. Hiraguri, T. Ichikawa, M. Iizuka, and M. Morikura. Novel Multiple Access Protocol for Voice over IP in Wireless LAN. In *Proc. of the Seventh International Symposium on Computers and Communications (ISCC)*, pages 517–523, 2002.

[HJ98] M. Handley and V. Jacobson. SDP: Session Description Protocol (RFC - 2327), 1998.

[HYM+07] Y. He, R. Yuan, X. Ma, J. Li, and C. Wang. Scheduled PSM for Minimizing Energy in Wireless LANs. In *Proc. of IEEE International Conference on Network Protocols (ICNP)*, pages 154–163, 16-19 October 2007.

[JL79] P. A. Jacobs and P. A. W. Lewis. Time Series Generated by Mixtures: The DARMA(p,q) Process. *Advances in Applied Probability*, 11(2):266–267, June 1979.

[KB05] R. Krashinsky and H. Balakrishnan. Minimizing Energy for Wireless Web Access with Bounded Slowdown. *Wireless Networks*, 11(1-2):135–148, 2005.

[Kou06] P. Koutsakis. A New Model for Multiplexed VBR H.263 Videoconference Traffic. In *Proc. of IEEE Global Telecommunications Conference (GLOBECOM)*, pages 1–6, November 2006.

[KPP05] P. Koutsakis, S. Psychis, and M. Paterakis. Integrated Wireless Access for Videoconference from MPEG-4 and H.263 Video Coders with Voice, E-mail, and Web Traffic. *IEEE Transactions on Vehicular Technology*, 54(5):1863–1874, September 2005.

[KTKS08] A.M. Kholaif, T.D. Todd, P. Koutsakis, and M.N. Smadi. QoS-Enabled Power Saving Access Points for IEEE 802.11e Networks. In *Proc. of IEEE Wireless Communications and Networking Conference (WCNC)*, Las Vegas, NV, USA, 2008.

[KW00] A. Kopsel and A. Wolisz. A Performance Comparison of Point and Distributed Coordination Function of an IEEE 802.11 WLAN. In *Proc. of Seventh International Workshop on Mobile Multimedia Communications (MoMuC)*, October 2000.

[KW01] A. Kopsel and A. Wolisz. Voice Transmission in an IEEE 802.11 WLAN Based Access Network. In *Proc. of IEEE WoWMoM*, pages 24–33, Rome, Italy, July 2001.

[Lin01] C.R. Lin. On-demand QoS Routing in Multihop Mobile Networks. In *Proc. of the Twentieth Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM)*, volume 3, pages 1735–1744, 2001.

[LK91] A.M. Law and W.D. Kelton. *Simulation Modeling & Analysis, 2nd Ed.* McGraw Hill Inc., 1991.

[LTZ05] Y. Li, T.D. Todd, and D. Zhao. Access Point Power Saving in Solar/Battery Powered IEEE 802.11 ESS Mesh Networks. In *Proc. of the Second International Conference on Quality of Service in Heterogeneous Wired/Wireless Networks (QSHINE)*, 22-24 August 2005.

[MBH01] J.P. Monks, V. Bharghavan, and W.-M.W. Hwu. A Power Controlled Multiple Access Protocol for Wireless Packet Networks. In *Proc. of the Twentieth Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM)*, volume 1, pages 219–228, 2001.

[MD07] L. Ma and M.K. Denko. A Routing Metric for Load-Balancing in Wireless Mesh Networks. In *Proc. of the 21st International Conference on Advanced Information Networking and Applications Workshops (AINAW '07)*, pages 409–414, Washington, DC, USA, 2007. IEEE Computer Society.

[MMV03] W. Matthew, J. Miller, and N. Vaidya. A Hybrid Network Implementation to Extend Infrastructure Reach. *Technical report, University of Illinois*, January 2003.

[NAS04] S. Nath, Z. Anderson, and S. Seshan. Choosing Beacon Periods to Improve Response Times for Wireless HTTP Clients. In *Proc. of the Second International Workshop on Mobility Management & Wireless Access Protocols (MobiWac)*, pages 43–50, New York, NY, USA, 2004.

[NCCH06] S.X. Ng, J.Y. Chung, P. Cherriman, and L. Hanzo. Burst-by-Burst Adaptive Decision Feedback Equalized TCM, TTCM and BICM for H.263-assisted Wireless Video Telephony. *IEEE Transactions on Circuits and Systems for Video Technology*, 16(3):363–374, March 2006.

[NMNA06] E.A.V. Navarro, J.R. Mas, J.F. Navajas, and C.P. Alcega. Performance of a 3G-Based Mobile Telemedicine System. In *Proc. of the 3rd IEEE Consumer Communications and Networking Conference (CCNC)*, volume 2, pages 1023–1027, 8-10 Jan. 2006.

[PBLS99] Q. Pang, A. Bigloo, V.C.M. Leung, and C. Scholefield. Service Scheduling for General Packet Radio Service Classes. In *Proc. of IEEE Wireless Communications and Networking Conference (WCNC)*, volume 3, pages 1229–1233, 1999.

[PCCM06] X. Perez-Costa and D. Camps-Mur. AU-APSD: Adaptive IEEE 802.11e Unscheduled Automatic Power Save Delivery. In *Proc. of IEEE International Conference on Communications (ICC)*, volume 5, pages 2020 – 2027, June 2006.

[PM03] D. Pong and T. Moors. Call Admission Control for IEEE 802.11 Contention Access Mechanism. In *Proc. of IEEE Global Telecommunications Conference (GLOBECOM)*, volume 1, pages 174–178, December 2003.

[QS05] D. Qiao and K.G. Shin. Smart Power-saving Mode for IEEE 802.11 Wireless LANs. In *Proc. of the 24th Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM)*, volume 3, pages 1573–1583, 13-17 March 2005.

[RBM02] T. Rouse, I. Band, and S. McLaughlin. Capacity and Power Investigation of Opportunity Driven Multiple Access (ODMA) Networks in TDD-CDMA based Systems. In *Proc. of IEEE International Conference on Communications (ICC)*, volume 5, pages 3202–3206, 2002.

[RPD07] N. Ramos, D. Panigrahi, and S. Dey. Dynamic Adaptation Policies to Improve Quality of Service of Real-time Multimedia Applications in IEEE 802.11e WLAN Networks. *Wireless Networks*, 13(4):511–535, 2007.

[RSC+02] J. Rosenberg, H. Schulzrinne, G. Camarillo, A. Johnston, J. Peterson, R. Sparks, M. Handley, and E. Schooler. SIP: Session Initiation Protocol (RFC - 3261), June 2002.

[SCFJ03] H. Schulzrinne, S. Casner, R. Frederick, and V. Jacobson. RTP: A Transport Protocol for Real-Time Applications, RFC 3550, July 2003.

[sip] Sipura SPA-841 IP Telephone Guide. http://www.sipura.com/Documents/SPA841AdminGuide0105-DRAFT.pdf.

[SLLY02] H.P. Sze, S.C. Liew, J.Y.B. Lee, and D.C.S. Yip. A Multiplexing Scheme for H.323 Voice-over-IP Applications. *IEEE Journal on Selected Areas in Communications (JSAC)*, 20(7):1360–1368, September 2002.

[sno] Snom 360 IP Telephone. http://www.snom.com/snom360_voip_phone.html.

[std97] Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications. IEEE Standard 802.11-1997, IEEE, November 1997.

[std99] Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications. IEEE Standard 802.11b-1999, IEEE, 1999.

[std05a] Media Access Control (MAC) Enhancements for Quality of Service (QoS). IEEE Standard 802.11e-2005, IEEE, November 2005.

[std05b] Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications. IEEE Standard 802.11a-2005, IEEE, 2005.

[std06a] Amendment to IEEE Std. 802.11: Enhancements for Higher Throughput. IEEE Draft 1.06 802.11n-2006, IEEE, 2006.

[std06b] Amendment to IEEE Std. 802.11: Wireless Network Management. IEEE Draft 0.05 802.11v-2006, IEEE, 2006.

[STS08] A.A. Sayegh, T.D. Todd, and M.N. Smadi. Resource Allocation and Cost in Hybrid Solar/Wind Powered WLAN Mesh Nodes. *Wireless Mesh Networks: Architectures and Protocols*, pages 167–189, 2008.

[TBJ00] T.D. Todd, F. Bennett, and A. Jones. Low Power Rendezvous in Embedded Wireless Networks. In *Proc. of the 1st ACM International Symposium on Mobile Ad Hoc Networking & Computing (MobiHoc)*, pages 107–118, Piscataway, NJ, USA, 2000.

[THH03] Y.-C. Tseng, C.-S. Hsu, and T.-Y. Hsieh. Power-Saving Protocols for IEEE 802.11-based Multi-hop Ad Hoc Networks. 43(3):317–337, 2003.

[Uni] International Telecommunication Union. Recommendation G.114, 1996. One-Way Transmission Time.

[Uni04] McMaster University. The SolarMESH Network. http://owl.eng.mcmaster.ca/ todd/SolarMESH/, 2004.

[Uni06] International Telecommunication Union. Recommendation H.263, 1996., March 2006.

[VCM01] M. Veeraraghavan, N. Cocker, and T. Moors. Support of Voice Services in IEEE 802.11 Wireless LANs. In *Proc. of the Twentieth Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM)*, volume 1, pages 488–497, Anchorage, AK, USA, 2001.

[VCTS07] S. Venugopal, W. Chen, T.D. Todd, and K. Sivalingam. A Rendezvous Reservation Protocol for Energy Constrained Wireless Infrastructure Networks. *Wireless Networks*, 13(1):93–105, 2007.

[vdSAH06] M. van der Schaar, Y. Andreopoulos, and Z. Hu. Optimized Scalable Video Streaming over IEEE 802.11 a/e HCCA Wireless Networks Under Delay Constraints. *IEEE Transactions on Mobile Computing*, 5(6):755–768, June 2006.

[VPV88] W. Verbiest, L. Pinnoo, and B. Voeten. The Impact of the ATM Concept on Video Coding. *IEEE Journal on Selected Areas in Communications (JSAC)*, 6(9):1623–1632, December 1988.

[WiF] Consumer Adoption of Wi-Fi Telephony Driving Up Wi-Fi Handsets Sales. http://www.tmcnet.com/planetpdamag/articles/.

[WLL05] W. Wang, S.C. Liew, and V.O.K. Li. Solutions to Performance Problems in VoIP Over a 802.11 Wireless LAN. *IEEE Transactions on Vehicular Technology*, 54(1):366–384, Jan. 2005.

[WLQL04] W. Wang, S.C. Liew, Pang Q., and V.O.K. Li. A Multiplex-Multicast Scheme That Improves System Capacity of Voice-over-IP on Wireless LAN By 100%. In *Proc. of the Ninth International Symposium on Computers and Communications (ISCC)*, volume 1, pages 472–477, 28 June-1 July 2004.

[WQDO01] H. Wu, C. Qiao, S. De, and Tonguz O. Integrated Cellular and Ad Hoc Relaying Systems: iCAR. *IEEE Journal on Selected Areas in Communications*, 19(10):2105–2115, October 2001.

[XL04] Y. Xiao and H. Li. Evaluation of Distributed Admission Control for the IEEE 802.11e EDCA. *IEEE Communications Magazine*, 42(9):S20–S24, September 2004.

[XLC04] Y. Xiao, H. Li, and S. Choi. Protection and Guarantee for Voice and Video Traffic in IEEE 802.11e Wireless LANs. In *Proc. of the Twenty-third AnnualJoint Conference of the IEEE Computer and Communications Societies (INFOCOM)*, volume 3, pages 2152–2162, 7-11 March 2004.

[ZTZK04] F. Zhang, T.D. Todd, D. Zhao, and V. Kezys. Power Saving Access Points for IEEE 802.11 Wireless Network Infrastructure. In *Proc. of IEEE Wireless Communications and Networking Conference (WCNC)*, volume 1, pages 195–200, 21-25 March 2004.