Extending MAC-layer QoS from Wired to Wireless Segments, and from Single Cell to Multi-cell Overlapping Environment

Li Zheng

A dissertation submitted in fulfilment of the requirements for the University of Technology Sydney for the Degree of

Doctor of Philosophy

2013

Supervisor: Professor Doan B. Hoang

Faculty of Engineering and Information Technology,
University of Technology Sydney,
NSW 2007, Australia

~ Dedicated to my mother
and to the memory of my father ~

ABSTRACT

Internet services today, especially real-time ones, require at least the same level of Quality of Experience (QoE) over radio links, as that provided by the wired parts for which most of the IP-based multimedia applications were originally designed. This leads to a problem of service quality consistency across the radio and wired segments of the network. In order to provide satisfactory QoE, provision of underlying layer Quality of Service, QoS, for multimedia applications over networks has been a long-standing and critical topic, especially within the radio access segment of the next generation wireless/mobile environment. With the assistance of QoE and multimedia adaption at the application layers, QoS components, such as QoS framework and service differentiation mechanisms, are better defined at MAC and IP layers; It is where radio limitation is commonly defined by time slot/division, frequency, code division, sector, and direction gain etc.

The problems of the QoS frameworks today vary from insufficient level of control, to scalability and complexity, as well as IP-mobility related problems. In the layer of medium access control (MAC), QoS fairness, especially in a highly loaded and/or overloaded system, is an issue in Contention Schemes for Service Differentiation. The shortcoming of efficiency (overhead and request collision) and complexity of polling schemes for Service Differentiation have also been observed. These two schemes would perform even worst in a multi-cell overlapping environment, rather than a single cell environment, without any central resources management.

This thesis investigates QoS supports for wireless network. The primary goal is to design effective QoS mechanisms/framework, in the context of medium access control in WLAN 802.11, to ensure wireless connectivity for multimedia traffic. Firstly, we propose a hybrid architecture following the principles of Differentiated Service (DiffServ) model over the core part of the network, and the principles of Integrated Services (IntServ explicit control) model locally over the wireless access segment. We then present in detail an example solution consistent with the hybrid QoS architecture principles, with an admission control core of Fair Intelligent Congestion Control (FICC). Within the framework, we analysed contention and polling schemes as the candidates of service differentiation in the MAC layer, and based on these results, we found our proposed the Multi-cycle Polling mechanism can actually meet the QoS requirements. We finally draw our attentions on Service Differentiation Scheme in a Multi-cell

Overlapping WLAN Environment. Due to widely use of WLAN almost in every smart-handset and notebook today and its shortcoming in spectrum efficiency and interference, the Co-existence/overlapping among WLAN themselves and other systems would become even more challenging. In fact, there is more and more serious interference incidents reported particularly in Metro transport systems in 2013. In this research, the graph colouring technique for grouping assignment is applied and the novel overlapping coordination scheme has been proved to effectively support QoS Service Differentiation in the interference environment. With the admission control core of FICC and the proposed Service Differentiation Schemes under the Hybrid framework, the wireless QoS issues have been well addressed for both single-cell and multi-cell environments.

LIST OF PUBLICATIONS

The following is a list of publications by the author of this thesis together with his academic and industrial supervisors during the tenure of her research study.

(1) Li Zheng, Doan B. Hoang; "Further Analysis and Tuning of Registered Multi-cycle Polling in Wireless Medium Access Management", 15th ACM International Conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems, Cyprus Island, October 2012.

(2) Li Zheng, Doan B. Hoang; "Wireless Hybrid QoS Architecture with an enhance of fair intelligence congestion control mechanism", journal of Wireless Engineering and Technology, Scientific Research Publishing, Volume 03, Number 03 (June 2012)

(3) Li Zheng, Doan B. Hoang; "A QoS Mechanism of Registered Multi-cycle Polling in Wireless Medium Access Control"; IEEE-RIVF International Conference on Computing and Telecommunication Technologies, Hanoi, November, 2010.

(4) Li Zheng, Doan B. Hoang and Ming Li. "Applying Fair Intelligent Congestion Control in a Hybrid QoS Architecture for Wireless Environment"; Seventh International Conference on Information, Communications and Signal Processing (ICICS 2009), Macau.

(5) Li Zheng, Doan B. Hoang. "Overlapping Impacts and Resource Coordination for High-density Wireless Communication"; IEEE-RIVF International Conference on Computing and Telecommunication Technologies, Danang, July 2009

(6) Li Zheng, Doan B. Hoang." Applying Graph Colouring in Resource Coordination for a High-density Wireless Environment"; IEEE 8th International Conference on Computer and Information Technology, Sydney, July 2008.

(7) Li Zheng, Doan B. Hoang. "Performance Analysis for Resource Coordination in a High-density Wireless Environment"; 13th IEEE Symposium on Computers and Communications, Morocco, July 2008

(8) Li Zheng, Arek Dadej, Steven Gordon. "Hybrid Quality of Service Architecture for Wireless/Mobile Environment"; IFIP TC6: Interworking 2002 on Converged Networking Data & Real-Time Over IP; Oct 2002, Perth.

(9) Li Zheng, Arek Dadej, Steven Gordon. "Fairness of Distributed Coordination Function for Multimedia Applications"; the 7th International Symposium on Digital Signal Processing for Communication Systems, DSPCS'2003, December 2003, Gold Coast, Australia.

ACKNOWLEDGEMENTS

The past few years have been a ‘period of intensive discovery’ for me. Firstly, I would like to express my gratitude to my supervisors Professor Doan B. Hoang at iNEXT – UTS Centre for Innovation in IT Services and Applications, Faculty of Information Technology, University of Technology, Sydney, Australia in the last few years of part-time research study.

My heartfelt thanks to my academic advisor Doctor Ming Li, at the School of Engineering and Information Technology, Deakin University, Geelong Victoria 3217, Australia, who has given me enormous guidance in participating in various studies at these research area. Without these guidance and support, I wouldn’t have been able to fulfil my research goal.

I am indebted to my colleagues in our office of Radio Access Engineering, SingTel Optus, Australia, for their valuable time and effort in contributing ideas from our practical deployment experiences. I am greatly influenced by their ‘clear thinking’. The ongoing motivations kept me never stop learning.

I also want to thank the academic fellows at the University of South Australia, Adelaide, Australia, where lots of research ideas and results were formed. This research would never have been possible without the scholarship funded by the University and Nortel Networks generously.

Finally, I want to thank my family for their supports and patience during my studies in the past years. Particularly, I want to thank my fiancée, Doctor Min Ling, for her endless love and encouragement throughout the tenure of my study.

CERTIFICATE OF ORIGINARITY

I hereby certify that the work in this thesis has not previously been submitted for a degree nor has it been submitted as part of requirements for a degree except as fully acknowledged within the text.

I also certify that the thesis has been written by me. Any help that I have received in my research work and the preparation of the thesis itself has been acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

…………………………………………….

Li Zheng

TABLE OF CONTENTS

ABSTRACT i
LIST OF PUBLICATIONS iii
ACKNOWLEDGEMENTS iv
CERTIFICATE OF ORIGINARITY v
LIST OF FIGURES ix
LIST OF TABLES xi
LIST OF EQUATIONS xi
Chapter 1 Introduction 1
1.1 Demands on Services Quality 1
1.2 The Main Contributions 2
1.3 Structure of Dissertation 5
Chapter 2 Background and Related Works 8
2.1 User Experience and Quality of Experience 8
2.2 Quality of Experience Model 10
2.3 Radio Communication Characteristics and its Quality 15
2.4 IP-Based Common Network Architecture and its traffic Characteristics 17
2.5 The Hierarchical QoS Structure 22
2.5.1 QoS Definitions 23
2.5.2 Application Layer QoS and Multimedia Adaption 25
2.5.3 The Framework of QoS Network Layer 27
2.5.4 Control Plane Mechanisms 32
2.5.5 Data Plane Mechanisms 35
2.5.6 Summary 38
Chapter 3 Hybrid Quality of Service Architecture and FICC Admission Control 40
3.1 Hybrid Quality of Service Framework 40
3.1.1 Review on Existing Methods 40
3.1.2 The Framework Requirements 42
3.1.3 The Proposed Architecture 44
3.2 Fair Intelligent Congestion Control as the Core of Admission Control 50
3.2.1 IEEE 802.11 WLAN and its MAC QoS Supports 51
3.2.2 Admission Control Mechanism for 802.11 Wireless LANs 53
3.2.3 Fair Intelligence Congestion Control 62
3.2.4 Simulation Analysis 65
3.3 Summary 69
Chapter 4 Service Differentiation in Contention Medium Access 71
4.1 The Basic of Contention Medium Access Schemes 71
4.2 Analysis in Service Differentiation 73
4.2.1 The Existing Approaches 74
4.2.2 The EDCF of IEEE 802.11e 76
4.2.3 Summary 79
4.3 Fairness Nature and Service Differentiation 80
4.3.1 Back-off Behaviour and Transmission Probability 80
4.3.2 Fairness in Transmission Opportunities 82
4.4 Performance Analysis on Fairness 84

4.4.1 Network Topology and Traffic Configuration 84
4.4.2 Simulation Observation of Fairness Issues with Comparison Between Overloaded and underloaded Scenarios 88
4.4.3 Discussion on Further Effects and Inheritance 92
4.5 Summary.. 95
Chapter 5 Service Differentiation in Polling Medium Access 96
5.1 The Basic Natures of Polling Medium Access 96
5.2 Service Differentiation for Session-based Applications Using Polling Schemes 97
5.3 Analysis on Existing Polling Enhancements 99
5.3.1 Previous Enhancements 99
5.3.2 802.11e Polling Enhancement 101
5.3.3 Summary.. 103
5.4. Registered Multi-cycle Polling 104
5.4.1 Analysis on Information Update and its mechanism 104
5.4.2 Registered Polling 109
5.4.3 Performance Analysis of Registered Polling 114
5.4.4 Further Performance Observations for Registered Polling.. 121
5.4.5 Multi-cycle Enhancement 129
5.4.6 Performance Analysis of Registered Multi-cycle Polling 132
5.4.7 Performance Analysis of Registered Multi-cycle Polling in a Scenario of Large Data Applications 134
5.5 Summary.. 137
Chapter 6 Service Differentiation and Resources Coordination in Multi-cell Overlapping BSS Environments 139
6.1 Introduction of OBSS 139
6.2 Literature Review and Observations on OBSS Interference 140
6.3 OBSS Coordination Essential 145
6.4 Concept of a Novel OBSS Coordination 148
6.5 OBSS Coordinated Resource Allocation Scheme and Protocol.. 154
6.5.1 OBSS Coordinated Resource Allocation Scheme 155
6.5.2 OBSS Resource Allocation Protocol 159
6.6 OBSS Grouping Assignment Using Graph Colouring.. 164
6.6.1 Graph Technique Background 164
6.6.2 Graphical Model 165
6.6.3 Mapping to Graph Theoretic Problem.. 166
6.6.4 Network Model and Properties 169
6.6.5 The Problem Solution and its Completeness.. 171
6.6.6 The Colouring Algorithm.. 172
6.6.7 Summary.. 177
6.7 Performance Analysis 178
6.7.1 Analysis on Messages Transfer and Its Overhead 178
6.7.2 Analysis on Resources Coordination.. 181
6.7.3 Performance Analysis on Uplink Period 182
6.7.4 Performance Analysis on Downlink Period.. 190
6.7.5 Performance Analysis on Second Group Assignment.. 196
6.8 Conclusion 200
Chapter 7 Conclusion and Future Works.. 201

7.1 Summary of Results 201
7.2 Further Development and Motivation 202
References and Bibliography 204
Appendix B Control & Management Mechanisms for OBSS Coordination 224
B.1 Overlapping Messages Collection 224
B.1.1 Specifications of Overlapping Information 224
B.1.2 Overlapping Information Collection Protocols 226
B.1.3 Entities Processes for Record Collections 230
B.2 Allocation Message Transfer 233
B.3 Slot Duration Re-adjustment 235
B.4 Mutual Synchronization across OBSS 238
Appendix C an Implementation of Messages Transfer for OBSS Coordination in the context of IEEE 802.11 240

LIST OF FIGURES

Figure 1 QoE and QoS Interaction 13
Figure 2 WLAN Network Components 18
Figure 3 General Network Architecture of Mobile/Wireless Environments 19
Figure 4 Traffic Requirements in Metric of Bandwidth and Delay 22
Figure 5 Entity-Relationship Model of QoS System 25
Figure 6 Hybrid QoS Architecture 48
Figure 7 QoS Agents 49
Figure 8 A Signalling Diagram for Flow Admission Procedure 54
Figure 9 State Transition Diagram of TS Lifecycle 58
Figure 10 Time Sequence Diagram of TS Setup 59
Figure 11 Time Sequence Diagram of Failed TS Setup detected within client MAC... 61
Figure 12 Time Sequence Diagram of TS Deletion (client-initiated) 61
Figure 13 Goodput comparisons between FICC and DiffServ 66
Figure 14 Queue Performance Comparisons between FICC and DiffServ 67
Figure 15 Resources Allocation of Bandwidth in FICC 67
Figure 16 Resources Allocation of Queue Length in FICC 68
Figure 17 Delay Distributions in FICC and DiffServ 69
Figure 18 DCF Access Procedure 72
Figure 19 EDCF Access Procedure 78
Figure 20 Simulated Network Topology 85
Figure 21 Total Throughputs, Load and Data Dropped 88
Figure 22 Rate of Packets Sent by Clients 89
Figure 23 Fairness between Voice 2 Client and AP 89
Figure 24 Multiple Streams Performance in AP 91
Figure 25 Videoconference Performance 93
Figure 26 Controlled Contention Interval [IEEE 802.11e D3 2003] 102
Figure 27 Simulated Network Topology for Registered Polling 116
Figure 28 Streams and Total Throughput of Registered Polling 118
Figure 29 Stream Delay Performance of Registered Polling 119
Figure 30 Total Network Throughputs with Long Superframes 124
Figure 31 Average Total Throughputs of Three Long Superframes 125
Figure 32 Voice Delay with 100ms Superframe 125
Figure 33 VoIP Delay Cumulative Distribution Function With 400ms Superframe .. 126
Figure 34 VoIP Delay with 10ms Superframe 127
Figure 35 Total Throughputs with 10ms Superframe 127
Figure 36 Average Total Throughputs with Different Superframes 128
Figure 37 VoIP Delay with 60ms Superframe 129
Figure 38 Multi-cycle Enhancement 131
Figure 39 Throughput Comparison for Registered Multi-cycle Polling 133
Figure 40 VoIP Delay of Registered Multi-cycle Polling 134
Figure 41 Video Conferencing Delay Distributions 135
Figure 42 VC and Bulk Data Throughput in Registered Multi-cycle Polling (no VoIP) 135
Figure 43 VC and Bulk Data Application in PCF (no VoIP) 136
Figure 44 Non-AP-Overlapped: $client_{2,1}$ is not reachable from AP_1. 141

Figure 45 AP-overlapped Scenario: $client_{2,1}$ is reachable from AP_1. 142
Figure 46 Frame Collisions of Two Delayed CFPs 143
Figure 47 Neighbourhood Capture 145
Figure 48 Supporting Distribution Systems 151
Figure 49 an Example Superframe 155
Figure 50 User priority to Access Category mappings 157
Figure 51 Resources Allocation Protocol 161
Figure 52 an Example of the OBSS First Grouping 162
Figure 53 an Example of the OBSS Second Grouping 163
Figure 54 Graph Varieties 166
Figure 55 Graphs Used for Modelling 170
Figure 56 Example of Maximum Degree First on a Planer Graph 177
Figure 57 Simulation Layouts 183
Figure 58 VoIP Delay Performances in Even-loaded Uplink Scenario 185
Figure 59 Total Throughput Performances in Over-loaded Uplink Scenario 186
Figure 60 Throughput Performances in Over-loaded Uplink Scenario 187
Figure 61 VoIP Delay Performances in Over-loaded Uplink Scenario 187
Figure 62 VoIP Delay Performances in Over-loaded Uplink Scenario with SD 189
Figure 63 Throughput Performances in Over-loaded Uplink Scenario with SD 189
Figure 64 Delay Performances for VoIP and Audio in Downlink Scenario 193
Figure 65 Throughput Performances in Downlink Scenario 194
Figure 66 VoIP Delay in Downlink Scenarios with Different Superframe Length 195
Figure 67 VC Throughputs in Downlink Scenarios with Different Superframe Length 196
Figure 68 An Example of Second Group Assignment 197
Figure 69 VoIP Delay Performance in Neighbouring BSS A 198
Figure 70 Audio Stream Delay Performance n BSS B 198
Figure 71 Video Conferencing Downlink Throughputs in BSS B 199
Figure 72 Bulk Data Throughput in Downlink Scenario with Second Assignment 199
Figure 73 Overlapping Message Collection Hierarchies 226
Figure 74 Time Sequence Diagram for Message Collection 227
Figure 75 Processes Relationship 230
Figure 76 State Transition Diagram of a Client Process 230
Figure 77 State Transition Diagram of an AP 231
Figure 78 State Transition Diagram of a RM 232
Figure 79 Allocation Message Transfer 234
Figure 80 Traffic Report Collection Process in AP 237
Figure 81 Slot Re-adjustment Protocol 238

LIST OF TABLES

Table 1 Possible PSNR to MOS Conversion ... 12
Table 2 TSPEC and Status field contents in the MLME-ADDTS.response ... 59
Table 3 Parameters of Multimedia Traffic Sources ... 86
Table 4 Simulation Attributes ... 87
Table 5 Simulation attributes used in Registered Polling experiments ... 115
Table 6 General Overheads of Service Differentiation Schemes ... 120
Table 7 EDCF QoS Parameter Set element format ... 157
Table 8 Traffic Configurations for Uplink OBSS Scenarios ... 185
Table 9 Traffic Configurations for Downlink Scenarios ... 193
Table 10 Overlapping Records Structures ... 225
Table 11 Overlapping Report Messages Structures ... 226
Table 12 Allocation Message Information Elements in A Beacon ... 233
Table 13 Traffic Information Structure ... 236
Table 14 Management Frame Format of Message Transfer in IEEE 802.11 ... 240
Table 15 BSS Traffic Report request (Information elements) in IEEE 802.11 ... 241
Table 16 OBSS Allocation Frame Format in IEEE 802.11 ... 243

LIST OF EQUATIONS

Equation 1 ITU E Model ... 10
Equation 2 Extended E Model ... 11
Equation 3 Linear Congestion Function ... 62
Equation 4 Probability of Transmission ... 81
Equation 5 Probability of Successful Transmission ... 81
Equation 6 Graph Colouring Technique ... 168

~ Being beaten is often a temporary condition, giving up is what makes it permanent ~

(Marilyn Vos Savant)

Chapter 1 Introduction

Since its inception and throughout its evolution, the worldwide Internet has extended its reach to new communication technologies and systems not long after each became available. Therefore unsurprisingly, the radio links have long been a part of the Internet, or that emerging digital mobile systems have generated a lively interest with respect to their seamless integration into the Internet.

Wireless/mobile computing is a rapidly emerging technology providing users with network connectivity without any necessity of wires. For example, Wireless local area networks (WLANs), like their wired counterparts, are being developed to provide high bandwidth in a limited geographical area. WLANs are presented as an alternative to the high costs wired LANs, primarily due to installation and maintenance. Physical and environmental necessity is another important factor to take into account in favour of WLANs. The usage of these WLAN networks are now largely covered many areas such as university campuses, corporate office space, downtown stores, etc. and supplements to the large mobile network worldwide.

1.1 Demands on Services Quality

Due to its popularity, network users find an ever growing set of multimedia applications and services/operators from which to choose from. The customer finds itself in s strong position, being able to choose among different service providers. Their subsequent choices are then largely likely to be influenced by experienced service quality, which could be described by Quality of Experience, QoE, combining user perception, experience and expectation. In order to provide satisfactory QoE for the users, providers need to react quickly on quality problem and require more intelligent under layers service quality network-centric mechanism, commonly referred as Quality of Service (QoS). For example, voice/video transmission exploits the reduced communication cost due to statistical multiplexing but must also deal with unpredictable throughput and delay, which implies QoE, using varied QoS mechanisms.

Supporting enhanced services over radio links is even more challenging than improving the already problematic performance of conventional applications. With wide deployment of radio system, the number of users will increase and almost certainly there will be demands, as for the wired networks, to be able to run real-time applications such as streaming video and audio. If one tries to extend these applications with a satisfactory QoE, to be able to run in the radio networks, there is certainly a need for the networks to be able to support QoS. Ideally, users of radio networks would want the same QoE that they have with wired networks. However, to meet these objectives, the radio networks must overcome some difficulties that are never encounters in the wired networks.

There are several factors in wireless/mobile networks, which make it very difficult to provide QoS guarantees. Firstly, resources in radio networks are scarcer than in wired networks. Radio links, in general, provide much lower bandwidths than the wired counterpart. This disparity is expected to hold in future even though rapid progress is being made for high-speed radio transmission, mainly due to the physical limitation of the radio media. Secondly, radio channels are inherently unreliable and prone to location-dependent, time-varying, and bursty errors due to noise, multi-path fading, shadowing and interferences. Even with channel coding, diversity combining and power control techniques, their unreliability is much higher than that of wired links.

It is very popular that researchers look at QoE and multimedia adaptation technique recently for the network quality problem. Using QoE E Model, the network resource is better defined and mapped into underlying layers. Wireless network bottleneck always commonly refers to MAC/PHY underlying layer component, such as time slot/division, frequency, code division, sector, and direction gain etc. Lack of bandwidth and poor services at the last are then transferred and defined by these components. It is where our research focuses to really solve such problem fundamentally. However, QoE and multimedia adaptation would help as assistant technique, via rate adaption, negotiation, re-transmission and codec, mainly purely over application and IP/TCP layers.

1.2 The Main Contributions

To meet the demands of QoS support for multimedia applications, particularly in the emerging radio systems, this dissertation investigates the QoS problems in wireless/mobile networks. The primary goal is to design effective MAC QoS mechanisms and framework to ensure radio connectivity for multimedia traffic. We present the QoS mechanisms/framework in the context of medium access control in WLAN 802.11 MAC and extend it from a single cell environment; to a multi-cell overlapping environment. This dissertation has its main contributions as follows:

Firstly, the research starts with the discussion on user QoE, which is the driver of QoS. It then presents hybrid QoS architecture in details with an example solution of 802.11 WLAN exploiting the 802.11e QoS extensions, on the base of analysis in existing QoS frameworks for wired network and radio network. The drawbacks of the proposed wireless QoS frameworks are pointed out, varying from insufficient level of control to scalability and complexity. Our proposed hybrid architecture follows the principles of Differentiated Service (DiffServ) model over the core part of the network and the principles of Integrated Services (IntServ explicit control) model locally over the radio access segment. The explicit resource management part in WLAN involves two steps: stream admission control/resource reservation and service differentiation.

With the Hybrid framework, a Fair Intelligent Congestion Control (FICC) is applied and built as the core of admission control, to intelligently predict per-queue fair share for all traffic aggregates, using feedback control to keep the Resources Manager operating at a desirable operating point at all times. It also allows overselling bandwidth, when the network is not congested, to make efficient use of the network resources. At the interface between wired and wireless networks, FICC enables effective admission controls. While the tasks of handling stream activation, keeping track of load and resource status, and performing AC decision policy have been well studied (such as those in RSVP/IntServ), we also investigate the admission control (AC) with focus on AC signalling and flow classification in the context of 802.11.The AC signalling is mainly accomplished by means of MAC (Medium Access Control) -level QoS profile negotiation mechanism defined in 802.11e. Once resources are reserved for a stream, application data frames must be classified in order for the service differentiation mechanism to be applied in the Access Point (AP).

As the second contribution of the dissertation, the Service Differentiation plays an important role in ensuring that the high priority (session streams) frames have transmission opportunities (TXOP) satisfying their QoS requirements as promised at the time of stream admission/registration.

The contention scheme, as the basic option of Service Differentiation, is firstly studied. Its fundamental rationale is Distributed Coordination Function (DCF) as defined in the IEEE 802.11. Many papers have studied the performance of DCF or modified DCF and proved that service differentiation can be supported with features of fairness and efficiency. However, none of the papers have analysed QoS fairness with respect to multimedia streams especially in a highly loaded and/or overloaded system. In overload conditions, traffic streams with light and frequent frames tend to dominate access and AP transmission (downlink) suffers more than other non-AP station (uplink). This problem is likely to affect performance of an infrastructure network as well as create asymmetry of throughput by bi-directional applications. More sophisticated mechanisms are needed to control access of wireless stations to radio resources. Such mechanisms have to be capable of controlling admission of traffic streams to the resource-limited network, and providing the greater transmission opportunity requirements of some stations (e.g. AP).

Polling scheme is able to maintain the balance of uplink and downlink and supposed to support real-time application. While most researches have paid attentions to the importance & implementation of polling-list scheduling/queuing technique, we emphasize the necessity of traffic information update in AP (for current frames arrival) for an intelligent polling, which can be obtained by either request from stations or stream registration. While the polling scheme with Request update (such as other proposed Controlled Contention/Resource Request) provides explicit information, it has the shortcoming of efficiency (overhead and request collision) and complexity. We then propose a Registered Multi-cycle Polling Scheme based on information update by stream registration. The registered polling performs well, particularly in a highly loaded scenario, with its implicit traffic information, which is available in the process of stream registrations/admission control, as part of our Hybrid QoS framework above. Through the performance analysis, we enhance the Registered scheme with Multi-cycle Polling mechanism. The enhancement provides satisfied delay performance for real-time applications, while it is still able to improve medium utilisation by adopting an optimal superframe length justified by simulation results.

The final contribution is to extend the study from the single-cell environment to multi-cell environment, where there are the situations of overlapping WLAN and none of contention, polling or combined scheme can provide Service Differentiation due to the co-channel interference. While attentions are usually paid on channel selection, load balancing and power control to avoid the problem, this study emphasizes that the overlapping avoidance is insufficient in a high-density scenario and it is necessary to have overlapping coordination to support Service Differentiation. Comparing to other coexistence/coordination schemes addressing the interference issue between WLAN and Bluetooth, our scheme focuses on solving interference problem among a high density WLAN domain where central resources management is already in place in a Hybrid framework. It also provides scalability and channel utilisation with re-use strategies, comparatively.

The scheme proposes to separate uplink from downlink transmission by assigning two individual periods in a superframe. During uplink period, the entire network operates in contention scheme offering its features of collision avoidance and efficiency. During downlink period, each cell is at least assigned into a group, which is given a time span for transmission and only transmits in that span to avoid collision. The design also provides options of polling schemes for each cell during the downlink period.

The timeslot assignment (to assign uplink/downlink and time span duration) is dynamically adjusted based on the update information on the overlapping topology and traffic condition, which are collected via a set of pre-defined protocols. Therefore, the scheme achieves the efficiency and balance (uplink/downlink) by meeting the topology and traffic changes. Through the simulation and analysis study, performance is further enhanced by assigning second groups to busy cell and adopting optimal values of the performance tuning parameters, such as superframe length and information collection frequency. Further study in applying graph colouring technique for grouping assignment suggests that when an overlapping network is deployed in such way where it can be represented by a planar graph, the assignment is not only NP (Nondeterministic Polynomial) -completeness, but also 4-colourable, which means insurance of our assignment solution and its enhancement for the network efficiency.

There are nine international publications arising from this works, listed in page ii.

1.3 Structure of Dissertation

The contents of this dissertation are organized as follows. Chapter 2 provides background information of service quality of radio/mobile systems and explains the reasons for their degraded performance, which directly lead to poor QoE from the aspect of end customers. We identify their radio limitations especially with current emergence of multimedia applications, and then present existing QoS approaches to these problems in a hierarchical structure.

In Chapter 3, we investigate the architecture approach for the problem, and derive a set of general requirements for any universal performance enhancing approach and verify that they can best be met by a suitably designed hybrid scheme. The Hybrid architecture is then specified in the context of IEEE 802.11 WLAN technology. In particular within the Hybrid framework, the Fair Intelligent Congestion Control FICC is adopted as an effective rate-based congestion control scheme that addresses both fair bandwidth sharing among traffic classes and congestion problems encountered in the existing QoS architecture.

Chapter 4 describes the first option of service differentiation mechanisms, which are key components in the Hybrid architecture. Several proposed contention-based schemes are studied. As the main concern of all these schemes, the fairness feature is well discussed with extensive performance measurements under a variety of multimedia application scenarios. We compare and contrast our findings with previous research and recommend the most appropriate usage of this option.

Chapter 5 describes the second option of service differentiation mechanisms, i.e. the polling-based schemes. We introduce the difference in nature and usage of the polling-based schemes from the contention-based schemes. And compared to other existing schemes, a novel Registered Multi-cycle polling scheme is presented with the extensive performance measurements showing it is the best approaches to support service differentiation in the hybrid architecture for multimedia applications while maintaining efficiency, which is important in the wireless environment.

In Chapter 6, we further discuss on the service differentiation mechanisms in the overlapping Basic Service Set (OBSS) environment, where multiple BSSs are deployed and interfered. We observe that neither polling nor contention scheme can provide service differentiation in such environment and point out the coordination essential between the BSSs. We then propose a novel solution for coordinated resources allocation schemes, where the usage of polling and contention schemes is coordinated among all BSSs, in order to support QoS without interference. The novel solution is supported by the well-defined message collection mechanisms and group assignment, using graph colouring technique. The performance measurements are also presented in details for uplink and downlink scenarios. These results closely approximate those obtained in the single BSS scenarios as discussed in Chapter 4 & 5.

Finally, Chapter 7 summarizes our results and identifies the original research contributions of this dissertation. We show that our architecture is capable of jointly providing wireless QoS support either in a single BSS or in the overlapping BSS environments. We conclude with directions for further research on coexistence between 802.11 WLAN and HiperLAN/2 WLAN with possible prediction and distributed method to support their Quality of Service.

~ You do what you can for as long as you can, and when you finally can't, you do the next best thing. You back up but you do not give up ~

(Charles Yeager)

Chapter 2 Background and Related Works

In this chapter, background material for the dissertation as well as a critical review on the related research work and development are presented. The discussion starts with end user experiences, QoE, Quality of Experience, and its relation with QoS. The digital radio link characteristics are described in Section 2.3, which is root cause of poor QoE today. Section 2.4 outlines the common network architecture and its modern multimedia traffic, which deteriorates the service quality problem. Existing approaches that are used to provide QoS within Internet are introduced in Section 2.5. In view of the shortcomings of the existing approaches, the requirements for a universal approach are also presented.

2.1 User Experience and Quality of Experience

Today, communication bandwidth is not still always infinite. Powerful devices, compelling applications and falling subscription prices have led to a growth of demand for broadband services at an unprecedented rate. For example, Coda Research and Cisco expect global Mobile Broadband traffic volume to roughly double every year, with approx. 418 million users generating 1.8 megabytes by 2017 [Code Research 2009], [Cisco 2010]. Users find an ever growing set of IP-based applications and access networks from which to choose. With the availability of residential broadband, IP-based digital TV is also offered via Ethernet and asymmetric digital subscriber line (ADSL).

Such a multitude of offers makes prices decrease, and competition between service and/or network providers increase. The customer finds itself in a strong position, being able to choose between different competing providers. Given similar pricing schemes, which are a primary decision aid for many users, their subsequent choices are then likely to be influenced by expected and experienced quality (i.e., through personal ratings of the perception and price-worthiness of a service). Since the introduction of packet switching network, quality has been a long-standing problem. Particularly, this problem becomes eminent in the context of interactive web applications and file downloads, where high latency and long waiting times caused by low quality network access directly translate into user annoyance and churn. Consequently, the providers' interest in how users perceive usability, reliability, quality, and price-worthiness has increased. A provider needs to be able to observe and react quickly on quality problems, at best before the customer perceives them and considers churn. Particularly, even in today 3G/4G mobile network, it seems there is no guarantee at all a good quality service is provided, even either for GSM voice or more advanced video streaming.

This development represents a huge challenge for network operators and service providers to provide satisfactory user experience. This rises the question of how network-level quality measurements and control relate to the user perception or satisfactory of a service. The network-level quality is generally defined in Quality of Service (QoS), using simple parameters such as bandwidth, loss, delay or other network related parameters to evaluate quality of service. However, they only imply user experience. Particularly, nowadays real-time multimedia applications are being deployed on IP networks and these technical parameters can no longer assess accurately the service quality/user experience, as it is perceived by human. Users expect to have good perceptual quality that can be derived from many factors including not only technical parameters but also users experience. Network operators need to control their resource while maintaining user satisfaction, which will result in user fidelity and benefit for the company. Therefore, they need to take into account not only (QoS) but also a recent new concept, Quality of Experience (QoE). The most widespread definition of QoE originates from the International Telecommunication Union, ITU-T SG 12 [ITU 2006], which describes QoE as ”overall acceptability of an application or service, as perceived subjectively by the end user”, which ”may be influenced by user expectations and context.”

The term QoE is relatively new to multimedia understanding. It is the overall acceptability of an application or service, as perceived subjectively by end-users. It is basically a subjective measurement of end-to-end performance at the service level, from the point of view of the users. It is commonly used in Operators defining systems, web and network services and also referring to as a metric for evaluating the Internet Services as a whole. In [Susan et al. 2000] [Ickin et al 2012] QoE is referred to as; what a customer experiences and values to complete his tasks quickly and with confidence. It also proposes a Quality of experience benchmark for e-business. [Moorsel 2001] [Heddaya 2002] considered QoE the perception metric elements of the network and performance relative to expectations of the users/subscribers. [Agboma and Liotta 2012] and [Ickin et al 2012] further define the most current user experience from the aspect of smart handset users on the 3G/4G networks.

2.2 Quality of Experience Model

QoE cannot be simply specified from a single parameter as low, fair, good and excellent but commonly be specified in several parameters such as resolution, colour, waiting time, etc. It is the subjective assessment to assess perceived quality by humans users. It is a kind of mathematical models that relates certain network and/or content properties to values, which express user satisfaction. In the system model, multiple coexisting users have various sessions, each consisting of different types of flows: audio, video and file transfer. As a result, three different quality models are considered to reflect the impact of resource constraints on user perceivable quality. In order to simplify the design, user utility functions are considered defined under a single metric to optimize user perceivable quality. A typical user-related measure is the mean opinion score (MOS) [ITU 2008], as a unifying metric. In [ITU 2008], such measure can be determined from subjective ratings by real users or predicted from objective measurements of properties of the delivered goods such as audio, video, or files.

MOS was originally proposed for estimating voice quality; it can be used as a general-purpose quality score, ranging from 1 (poor) to 5 (excellent).

ITU-T's E-Model [ITU 2003] is used to determine audio distortion, which provides a parametric estimation based on loss and delays of audio packets, and defines an R-factor that accounts for the various impairments on voice quality. The E Model has the following components:-

$$\text{User R Factor} = Ro - Is - Id - Ie + A$$

Equation 1 ITU E Model

Which results in an R factor of between 0 and 100.

The components of R are:-

Ro - representing the effects of noise and loudness ratio

Is - representing the effects of impairments occurring simultaneously with the speech signal

Id - representing the effects of impairments that are delayed with respect to the speech signal

Ie - representing the effects of “equipment” such as DCME or Voice over IP networks

A - the advantage factor, used to compensate for the allowance users make for poor quality when given some additional convenience

[Clark 2001] further extends E model to more closely approximate the user’s perspective by taking take into account both recency and delay, as shown in Equation 2.

User R Factor = 94 - Ie (end of call) – Id;

Where Ie (end of call) = Ie(av) + (k(I1 - Ie(av))) e -y/t3

Equation 2 Extended E Model

It is assumed that the I1 represents the exit value from the last significant burst of packet loss, y represents the time delay since the last burst, t3 is a time constant of typically 30-60 seconds and k is a constant (set to a nominal value of 0.7). Default value, Ro – Is of 94 for many of the E Model parameters can be assumed (per ITU G.107). Delays of less than 175mS have a small effect on conversational difficulty; whereas delays over 175mS have a larger effect, such Id = 4 + (delay – 175) / 9. Such passive monitoring model that incorporates the effects of burst packet loss and recency estimates the transmission quality of a Voice over IP (VoIP) network, and produces results that correlate well with user ranking as shown in its statistical experiments.

Recent studies in [Bettermann and Rong 2011] argue that among 21 E-model parameters [ITU 2003] the most significant threat to quality of VoIP at the packet level is packet loss. As packet loss occurs in bursts, the distribution of that packet loss (packet burstiness) affects perceived quality even more. It then point out that the probability of packet loss by a system with all paths is the probability of simultaneous packet loss on all paths, while burst ratio is the dependent relationship to the probability of transitioning to the loss state from the receive state. Using packet loss probability, the burst ratio, and knowing the codec used, the E-model is computed for a MOS estimate. The result interestingly shows that a user experience of 'very satisfied' may be achieved with a dispersity routing system of six paths despite each path experiencing loss probabilities up to 0.45.

[Tao et l. 2008] proposed a model to determine user satisfaction based on video quality. This model evaluates the impact of packet losses on video quality while accounting for parameters such as video codec, bit rate, packetization, and content characteristics. Quality is given in terms of the average Peak Signal-to-Noise Ratio (PSNR) of the video signal; and PSNR is logarithmic ratio between the maximum value of a signal and the background noise. It applies a non-linear curve mapping PSNR to MOS that de-emphasizes the impact of small losses when video quality is at extreme values. According to this study, there exist heuristic mappings of PSNR to MOS as shown in Table 1.

PSNR [dB]		MOS
> 37		5 (Excellent)
31-37		4 (Good)
25-31		3 (Fair)
20-25		2 (Poor)
< 20		1 (Bad)

Table 1 Possible PSNR to MOS Conversion

In a practise exercise of network design, to keep user-perceived service quality above an acceptance threshold, a network provider needs to know how network-level QoS parameters translate into user-level QoE perception and vice versa. In fact, the QoS Application Layer is only concerned with parameters such as resolution, frame rate, video codec type, audio codec type, layering strategy, sampling rate and number of channels, etc. The network Layer is concerned with parameters such as jitter, delay, burstiness, latency and packet loss, etc. To choose the right approaches of such measure and translation has been a very active research topic on its own [Wong et al. 2012][Piamrat 2009][Ketyko et al. 2010][Chen et al. 2009]. The specification, translation and mapping of these parameters are meaningful on the underlying layers of QoS, which is defined by Seven Layer OSI model, as discussed in the later sections.

The relationship between QoS and QoE is well discussed in [Siller 2010]. It considers that a better QoE can be achieved by considering the network and application layer QoS when arbitration is used, as shown in Figure 1. Ultimately Service quality is governed by the arbitrators, which control the existing and standardised mechanisms. The Interaction between layer QoS and QoE is managed and controlled by an entity referred to as the "Layer Arbitrator". It is provided by the feedback passed between each layer. The decisions taken by the arbitrators are based on look up tables or by implied intelligence derived from feedback. [Qiao 2011][Dohoon 2010] also examine such Management and arbitration at the application layer.

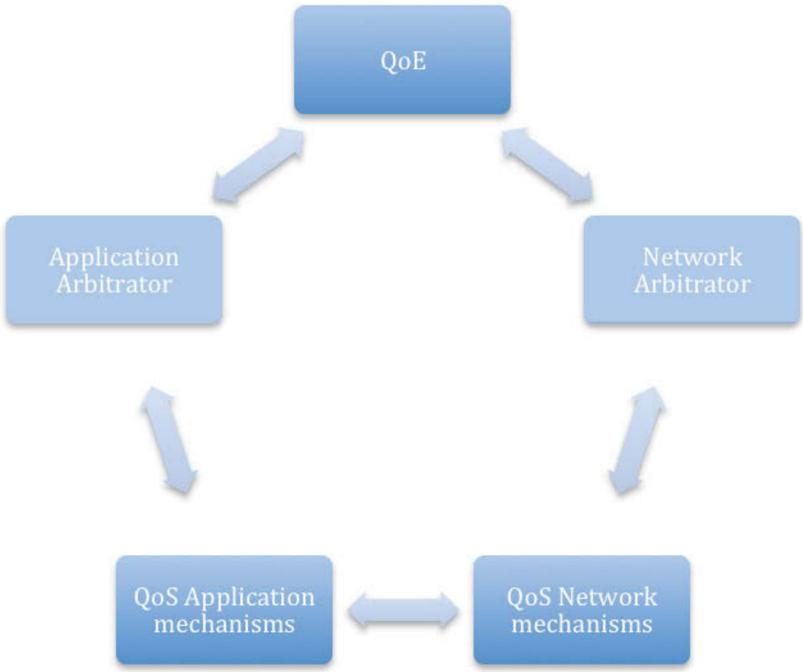


Figure 1 QoE and QoS Interaction

More importantly, in another term, user-centric QoE should gain supports from underlying layers application/network-centric QoS. QoS is designed to overall manage the limited network resources, which is root cause and origin of poor QoE by users. QoS is further classified by different mechanisms and control frameworks, which are core engine of providing the service quality. And QoS components are mostly based on MAC, PHY and IP layers. An overall approach, combining traditional QoS, Application QoS (such as Multimedia Adaptation) and QoE, should be considered in order to fundamentally solve the problem of service quality in the packet switching Internet today.

Even more obviously, such poor QoE by users are more common in the radio environments, either defined by QoS or QoE even today. GSM/2G/3G/4G mobile network optimization has been large and complex engineering up to today, since the introduction of the first mobile handset 30-40 years ago. Very often, Operators manage/optimize QoS resources in MAC/PHY via better division & management in term of time, code, frequency, sector, direction and etc. Many tools, QoS mechanisms and algorithm are built and used among these factors.

Particularly, WLAN has been become so widely equipped and used almost in every smart-handsets and notebooks today. And It is a legacy technology, and its shortcoming in transmission liability, spectrum efficiency and interference remain for long time as long as we have all these notebooks and smart-handset in use on the markets. In fact, there is more and more serious interference incidents reported particularly inside the Metro transport systems in 2013. Co-existence among WLAN itself and other systems will become even more challenging because there will be more smart-handset and notebooks in use and higher density deployments.

The root cause of such poor quality over radio environment in Mobile and WLAN systems are due to its Radio Communication Characteristics, which we discuss in the next section. And such limited radio resources are commonly defined by radio Characteristics factors of timeslot, code, frequency, sector, direction gain etc. in PHY/MAC layers, which QoS components mainly focus on.

On the other hand, it is very true that, without the QoE introduction, engineering hardly really knows how well the end-user benefits from their works indeed, because they commonly deal with objective QoS parameters. However, if the network offers poor service, the users firstly notice by the subjective QoE. QoE could provide a clear contract and direction for operators to work on the problem with the existing QoS tools, mechanisms and algorithms at the underlying layers.

In fact, the introduction of QoE comes with the background that the current Internet, particularly over its wireless segment, often offers varied network conditions, which are caused by limitations of underlying components, such as time slot division, frequency division, code division, sector, and direction gain etc. While we emphasize the importance of QoE technique along with other higher layer technique such as multimedia adaptation, our research in this thesis mainly continue pay attention on the main cause of the wireless quality problem at MAC level and with the objective QoS parameters, look at QoS framework, MAC mechanisms, and service differentiation to manage the limited resources defined at MAC/PHY layers. However, there is a close relationship between QoS parameters and QoE parameters. The degree of importance and priority of each QoS parameter are influenced by the QoE and hence providing guidance to where and what to focus in adapting QoS framework. However, this will be the focus of further research

2.3 Radio Communication Characteristics and its Quality

Radio communication mainly consists of Wireless LAN (WLAN) and Mobile (Cellular) systems whereby both possess a common feature that electromagnetic waves (rather than some form of wire, also known as radio) are used to carry the signal over part of the entire communication path. Mobile system is in its Four Generation (4G) stage currently under deployment while WLAN has higher bandwidth. There are some considerable differences between these systems. WLANs offer relatively high bandwidths that are shared among users within a limited coverage area, usually inside a building. Mobile systems offer low bandwidth links over a larger area with both indoor and outdoor coverage. They are built to enable mobility and promote bandwidth sharing between adjacent coverage areas, a characteristic that has only appeared in WLANs. Mobile systems are originally designed for real-time voice telephony and stepping into new generation for data communications, while WLANs are meant for data communications from the beginning. Nevertheless, many of these distinctions are getting blurred due to the fact that traditional voice networks are merging with the Internet while Mobile systems are evolving to higher capacities and variable bit rate allocations.

Both WLAN and Mobile system must share the Spectrum with, occasionally malicious, external RF sources, as well as with neighbouring systems of the same type. Their error rates are thus higher and their available bandwidth is lower than those of the shielded and isolated wired links. Due to their relatively small coverage areas, signal propagation delays between communicating devices are small compared to those of geostationary satellites, hence transmission time usually dominates total delivery delay. Terrestrial obstructions such as buildings, furniture and people, cause both indoors and outdoors links to suffer from multi-path fading. In addition, mobility, a key characteristic of mobile systems, constantly changes the fading and interference characteristics of a link. Therefore, the error behaviour of wireless parts of these systems varies in a faster and unpredictable manner. This becomes a serious weak point for a good quality communication from end to end. Quality guarantee over the radio parts is crucial in both WLAN and mobile systems. Due to these similarities, to assess the capabilities of these radio links, the dissertation has main focus on WLAN as discussed below.

The wireless link between the wireless station (STA) and the Access Point (AP) usually has very different characteristics compared to a conventional wired link. The available bandwidth on the wireless link is usually much lower than on a wired link. For example, in the century of Gbits per second Internet today, most WLANs only offer bandwidth in the units of Mpbs. The wireless link is often asymmetric, giving more capacity downlink than uplink. Even given with enough bandwidth, wireless link also suffers with propagation delay, which is a large portion of transmission delay. From the application's perspective, transmission delay has to be taken into account when calculating round-trip time (RTT) of this application. RTT is the time needed for the network to transmit a packet to the network with given bandwidth. Propagation delay is defined as the time spent for a single bit to traverse from one host to the other. Therefore, RTT is the sum of propagation delays and transmission delays for both uplink and downlink. The propagation delay in the wireless network is also much longer than in wireline networks. This delay is usually quite large in wireless environment, which decreases the user's ability to work interactively as the requests and replies are not as available as rapidly with a wireline connection.

Wireless communication is usually prone to transmission errors because of highly variable conditions in the natural environment and some unpredictable conditions such as handoff and interference. The transmission errors are bursts of distorted bits causing the transmitted data frame to be useless. There are several lower layers approaches to deal with the erroneous link. Forward Error Correction (FEC) is used in the radio link layer to correct the bit errors by using an encoding algorithm, which detects and corrects the bit errors at the receiving host. Automatic Repeat Request (ARR) is a link level re-transmission mechanism to ensure that the frames are eventually delivered without errors and it gives more accurate data transmission. Although bit error rate can be improved from as high as 10^{-3} to 10^{-7}, this link-level error detection mechanisms cause excess delays. Those remaining transmission errors can then be perceived as packet losses by upper protocol layer, which would normally initiate re-transmission of the lost packets. This is likely to occur during handover. Furthermore, if microcell or picocell is used to increase the capacity, the handoff rate would increase. Handoff would play a big impact on QoS as higher handoff rate can cause high error rate over radio interface. This is also caused by the unpredictable conditions such as the mobility of users, the terrain condition (fading and shadowing), interference and multi-path affects. These challenging issues have already been discussed in the previous sections.

It can be seen that natural properties of wireless link lead to the significant concerns on communication performance, such as less bandwidth, more delay (and jitter) and higher error rates as compared to wired communication. The new generation wireless environment no longer only transmits circuit-switched voice signal; An increasing number of modern Internet multimedia applications are starting to migrate to wireless communication, requiring as the same quality as they would in a wired environment. These applications have much higher performance demands on the network, such as higher bandwidth with limited packet loss, time bounded delay and jitter. Those excess delay and packet loss caused by wireless link are crucial to multimedia communication quality.

In addition, for delay sensitive or bandwidth intensive applications, resource constraints over the air interface may become a significant technical challenge in the design of new generation wireless multimedia networks. Provided that the total resources available over the air interface are, on average, sufficient to meet the total resource requirements of the user application sessions admitted to the system, the level of QoS desired/expected by users can be provided on an end-to-end basis by means of service differentiation (i.e. sacrificing the system performance for services tolerant of longer delay and higher rates of data loss in order to meet the quality of service specified for the other, less tolerant services). In order to keep constant quality of wireless multimedia application as that in the Internet, enhancements for quality controls are crucial research topic for new generation wireless communication.

2.4 IP-Based Common Network Architecture and its traffic Characteristics

Firstly, WLAN network architecture is studied in this section. Both HIPERLAN and IEEE 802.11 have a similar network infrastructure as shown in Figure 2. In the infrastructure configuration, there are some fixed points called Access Point (AP). A group of stations using the same radio frequency is called Basic Service Set (BSS). Within one BSS, there are several stations (STA). The STA communicates through the AP via air interface. AP connects to the Internet via the LAN as Assess Network, which is an IP-based network.

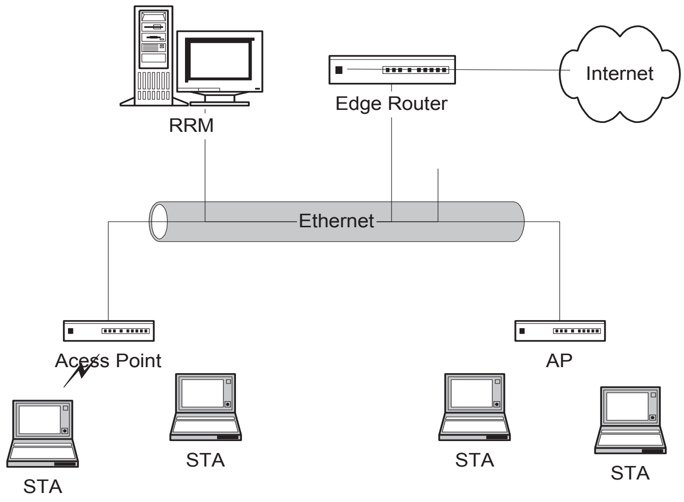


Figure 2 WLAN Network Components

A typical mobile/WLAN system is made up of three parts: a core network, an access network and the last radio hop as it can be seen in Figure 3. The core network maintains the connection to the Internet cloud. The access network provides the high-speed transportation among the Base Stations (or Access Points) and the core network. The last radio hop is established between the Base Stations and Mobile/wireless Terminals. The "glue" that holds all wired sub-networks together in the core network is its network layer protocol, Internet Protocol (IP), which provides global host addressing and basic data delivery. IP was designed with the explicit goal of supporting the integration of heterogeneous networks into a global entity; hence its implementation requirements and service offerings are minimal. IP provides end-to-end delivery of variable sized messages, called datagrams, which may be reordered, duplicated or lost without any warning. Thus, it is simple and stateless. There is no doubt that IP is also the correct proof of choice as a universal network layer.

In access network and the last radio hop, IP is also desirable to an extent in supporting multimedia application and it allows smooth interconnection with core network with building solutions, which are already available from traditional fixed network. Firstly, IP is ubiquitous today; applying IP into mobile/WLAN environments as prevalent fixed network has economic perspective. Secondly, the simplicity of IP can keep the network simple and stateless while pushing the complexity to the limit. Therefore, it provides more controls and freedom for new services in individual mobile/WLAN sub-system. Finally, as most end user applications are IP-based nowadays, the usage of IP stack in mobile terminal can have many advantages. In addition to providing the simplification, it also ensures applications behave inherently and consistently to mobile/WLAN environment and enriches the entire mobile application market, which is a key element for successful mobile/WLAN evolution. However, Since IP does not enhance the service provided by individual links in any way; end-to-end performance is limited by the worst link on the path, which are the wireless parts.

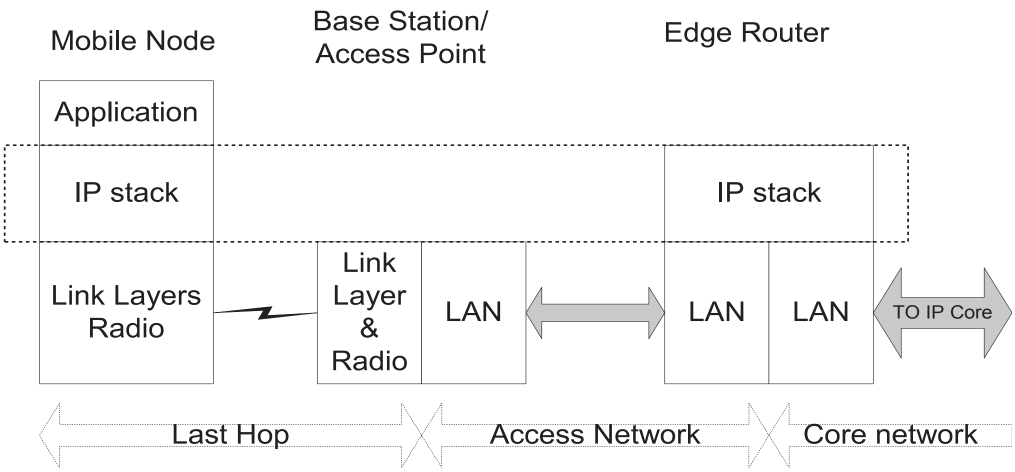


Figure 3 General Network Architecture of Mobile/Wireless Environments

On the other hand, new generation wireless communication is expected to include evolutionary capabilities and features. It should support enhanced multimedia (voice, data, video, and remote control) with usability on all popular modes (cellular telephone, e-mail, paging, fax, video conferencing, and web browsing) providing board bandwidth at high speed. Nowadays, the high speed of Internet has enabled people not only to work faster in the same applications but also to start building completely new applications and services. In particular, increased bandwidth has enabled real time applications such as sending voice and video information over IP networks. Non-real time applications have also reduced the time required for completing theirs tasks, making them usable in the new context. Each application has their own characteristics that may be defined by the use of a set of attributes. Value ranges between a minimum and maximum values often define the attributes. Such wide range of applications ranges from streaming video, voice over IP and videoconference to file transfer. To address the importance of QoS to these applications, discussions on individual application are presented as follows:

- Voice over IP

Voice over IP is also called packet voice. Voice streams are transmitted over IP packets instead of via a pre-scheduled channel in the traditional circuit-switched domain. This new means for voice transmission has significantly changed the telecommunication market today due to its low transmission cost. Packet payload size depends on the considered speech codec and the packet rate. The adaptive multi-rate speech codec (AMR) [Fings 2001] is a multi-mode codec with bit rates between 4.75 and 12.2 Kb/s. Processing is done in 20 ms frames taking 160 samples per frame. It offers high speech quality under a wide range of transmission conditions. With a packet rate of 12.2 Kb/s the packet size has a payload of 32 bytes. Furthermore the header comprising of RTP (Real Time Protocol), UDP (User Datagram Protocol) and IP has 60 bytes using IPv6. The mean call duration can be up to 120 seconds.

- Video conference

Videoconference is a Variable Bit Rate (VBR, one concept of ATM) real-time application, integrating a diverse range of traffic sources such as video, voice, and data over a single communication channel; such traffic sources differ a lot in their traffic patterns as well as in their performance requirements. The ITU recommendation H.323 defines the components, procedures and protocols necessary to provide audio-visual communication. H.323 enforces the use of Real Time Protocol (RTP) of the Internet Engineering Task Force (IETF) for flow transmission (flow set-up signalling). The bandwidth for video image depends on parameters such as frame rate, image size and image quality required. The image size and quality depend on the type of terminal and the concrete application. A typical image (176*144 pixels) with a transmission rate of 12 frames per second requires 600Kb/s producing a mean size of packets of 6 Kbytes/frame [Brain D1.1 2001]. The speech is slightly simpler and arrives normally at 16Kb/s with a frame size of 0.625 mini-second. The timing restrictions for such application are 200ms for end-to-end delay and 20 second for set-up time.

- Video & Audio Streaming

Video and music distribution companies have expressed interest in marketing video and music by providing downloadable channels utilising wireless services. The quality of the music transmission needs to be good and the combination of higher bit rate and compression techniques should ensure high quality reception. This type of service is not tolerant to delays. Video frames to be transmitted at the common rate of 30 frames/second are firstly sampled and quantitated, pixel-by-pixel, providing a constant data rate. For different transmission standards, capacity of up to 180 Mb/s is needed to transmit a signal, which is too large for the Internet. Compression is used to reduce the transmission rate of such signal. Moving Pictures Experts Group (MPEG) developed different coding standards (MPEG1, MPEG2 and MPEG4) for compression of audio and video signals. The compressed transfer rate depends on the variance of the image and audio quality and the compression rate, ranging from 4Mb/s to 30 Mb/s for video and 128Kb/s to 968Kb/s for audio [IEC13818-3 2002].

- Data Traffic

This is a general category of best-effort service that is originally designed in the Internet. This application is delay tolerant and will probably be priced based depending on the Mbytes used in the transfer. The most common data applications used today are web browsing, file transfer and e-mail. According to traffic statistics quoted in [Mah 1997], WWW is the most popular application in terms of both IP datagrams and total amount of data transferred and its popularity is increasing on a daily basis. In WWW browsing, the user employs a client application to look at pages stored at a WWW server. Pages consist of text, embedded objects (images or multimedia stream) and links to other pages. The sizes of pages vary significantly. For example, a page request can be a short data file while an image file is the largest file in one simple piece. The inter-arrival time between two pages can vary significantly depending on the user's demands. Files can be transferred through the Internet by using File Transfer Protocol (FTP). FTP in turn uses TCP to copy files among hosts. The file transfer itself is unidirectional instead of interactive, and it is not time bound (or have loose time restriction as hourly). Throughput is the main performance indicator for file transfer besides lost. Email is an instant message transmitted as a file between hosts. The minimum size of a traditional email can be 500 bytes [Brand and Aghvami 1998], including header information and the message context itself. Due to the increasing number of attachments in business traffic nowadays, mean email size can reach up to several Mbytes.

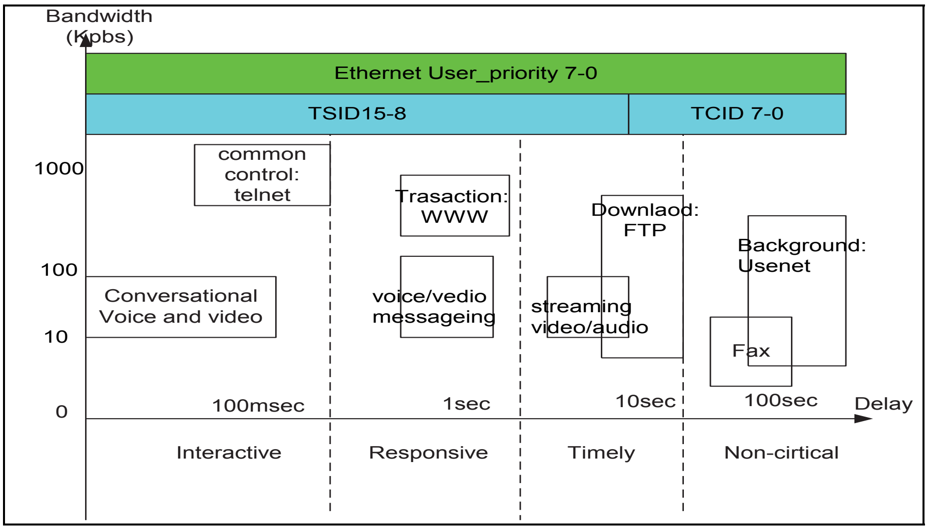


Figure 4 Traffic Requirements in Metric of Bandwidth and Delay

As discussed earlier, each application has its own characteristics and requirements for a good quality communication as illustrated in Figure 4 with metric of bandwidth and delay. In order to address the QoS needs of each application, we classify applications based on whether the application needs a session set-up dialog to be carried out before the data is actually transmitted on the network. A session is the small period of time that the user stays on the same destination for the application. The application can use various types of session layer protocols, such as H.323 (ITU), Session Initiation Protocol (SIP) and Real Time Streaming Protocol (RTSP) to set up a request dialog before transmission. Therefore, Voice over IP, Videoconference, Audio & Video streaming are classified as session-based applications while data traffic, including WWW, FTP and Email are classified as non-session-based applications. QoS of session-based applications need to be guaranteed for the duration of the session and are either negotiated between the application and network entities at the session set-up time, or implicit in the type of application. Non-session-based traffic does not need hard QoS guarantees, thus explicit resource reservation is not necessary. The traditional "best effort" service is sufficient for this class of traffic.

2.5 The Hierarchical QoS Structure

In this section, we introduce the existing QoS architectures and present the critical literature reviews on QoS enhancements over wireless environment. Provision of QoS has certainly been one of the major challenges for new generation wireless network, although it has been the most active research areas in networking for the last decade. A strong interest in research on QoS really started with the development of ATM networks at the end of the 1980s. The first version of Internet QoS, RSVP/IntServ (Resource Reservation Protocol / Integrated Services) architecture, was finalised in 1997, by the IETF [Braden et al. 1997, Wroclawski 1997a, Shenker et al. 1997]. This model was then criticised for its complexity and potentially bad scalability, which resulted in a new IETF proposal to deal with the network QoS problem: DiffServ [Black et al. 1998]. Since then, there have been many new proposals with the aim to arrive at a technically extremely lightweight solution by relying on economic mechanisms to manage network resources. Furthermore, other basic research with regard to QoS mechanisms, e.g., packet classification or packet scheduling can be regarded as related work as it is necessary to understand these basic functions when trying to inter-work with heterogeneous QoS systems in different sub-network systems such as Ethernet, WLAN and ATM.

2.5.1 QoS Definitions

There are many definitions of QoS in the literature [Fiedler et al. 2010] [Stiller 1996] and [Vogel et al. 1995]. But the definition used there is fairly self-contained. The purpose of QoS is to guarantee delivery of service/application, which users expect, especially when resources are limited, or demands are overwhelming. In other words, QoS provides preferential treatment to some arbitrary amount of different network traffic. It describes "the set of those quantitative and qualitative characteristics of a distributed system, which are necessary in order to achieve the required functionality of an application" [Vogel et al. 1995].

The Internet was originally designed to support a large number of users as a simple, best effort network, which ensures that all users receive a share of available bandwidth for a wide range of data applications. However, to successfully carry a significant quantity of real time services such as audio or video-conferencing, the traditional Internet is not sufficient. An enhanced network that provides quality of service can transmit some packets preferentially to achieve, for example, low transfer delay. Parameters that can be controlled include packet delay, packet loss, packet errors, available bandwidth and inter-packet delay (jitter). These parameters may be controlled absolutely ("guaranteed") or they may be controlled within some probabilistic framework to satisfy real time services needs. This is why QoS is needed, especially in wireless environment, which has new demands of real time services while suffering from its natural physical properties.

QoS is also a concept that needs to be considered on all layers and system components, which participate in the process ranging from data generation to its presentation [Khorsandroo et al. 2012][Fiedler et al. 2010] [Aurrecoechea et al. 1998] [Xiao and Ni, 1999]. We use an entity-relationship model, as shown in Figure 5 below, to introduce the conceptual model of a QoS system and its basic building blocks. The QoS system can be considered in two levels. The application level QoS determines the degree of QoS supports the application needs from the network QoS for a good quality communication and how a network provider exploits the technical features offered by the chosen network QoS architecture. The network QoS is at the heart of a QoS system as it represents the main constraint for the properties of a QoS system; while the actual selection of the system properties within this constraint is done by the application layer QoS. Such selection can be negotiable.

The Internet is networks of networks that have evolved into a global information infrastructure nowadays since it has allowed for a smooth evolution in time and space and continues to do so. Therefore, the first design principle of network QoS can be considered on a network wide basic and expressed firstly as QoS framework. It should represent the decisions, such as where to place the intelligence - in the end systems or the intermediate nodes and what levels of this intelligence is required. The network QoS can be further divided into QoS declarations and procedures. The QoS declarations describe the static part of the architecture, which contains properties like service classes, performance parameters (such as throughput, delay and loss) and their specification units (packet number for fixed-size packet network, bits or bytes for variable-size packet networks). They form the passive, static part of QoS architecture. On the one hand, they capture the terms by which network and users communicate using control path components and, on the other hand, they represent target values for data path components to achieve. QoS procedures constitute the dynamic part of the QoS architecture. This involves mechanisms on the data path as well as on the control path within networked nodes. The control path mechanisms are used to invoke a QoS treatment of a certain unit of networked-wide service. The data path mechanisms enforce whatever QoS has been agreed upon by sender, receiver and network provider.

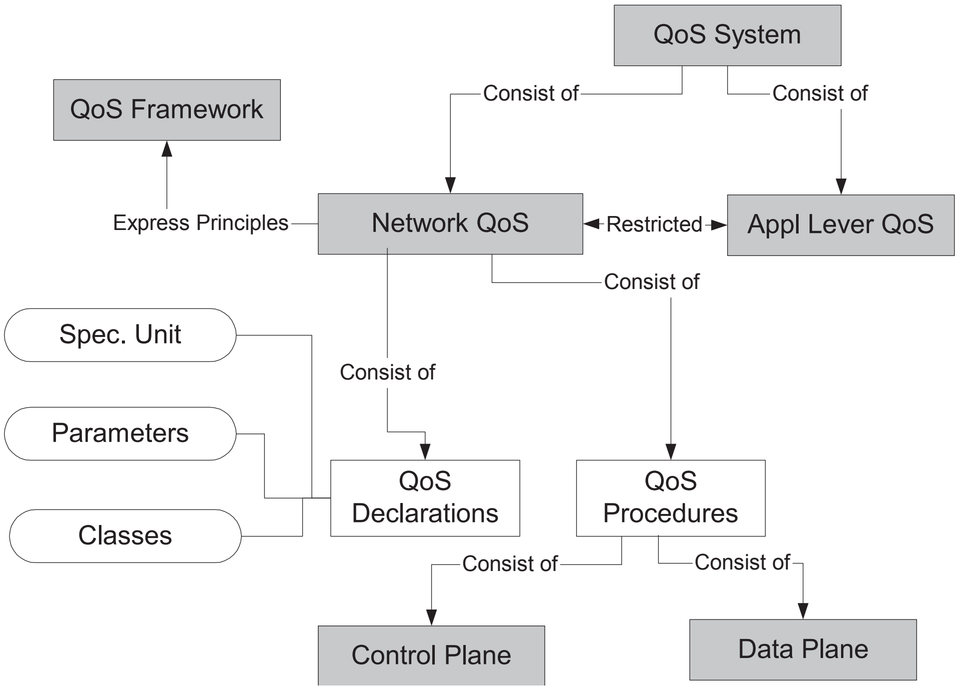


Figure 5 Entity-Relationship Model of QoS System

Details of the shaded entities as shown in Figure 5 are presented in the following discussions. They are the factual components of a QoS system whereas all the other entities can be considered as context. However, QoS system is also heterogeneous according to the nature of the Internet. Further discussion on Interworking issues among these entities will be covered in the Section 2.5.6 “Summary”.

2.5.2 Application Layer QoS and Multimedia Adaption

As discussed earlier in Section 2.3, while legacy applications remain large portions of wireless traffic; new applications that are highly intelligent are actually the driving forces for new generation wireless network. Both of them have different attributes and different QoS requirements as mentioned previously. A set of interfaces is needed to differentiate between application and underlying protocol stack and each of them should address a specific type of application [Benslimane et al 2008] [Bloch 2008]. In [BRAIN D1.2 2001], legacy applications are defined as type A; its interface is the simplest, because these applications access IP services by directly interacting with the classical transport layer. Most of the real time applications can be classified as type B; they require the usage of various session layer protocols (e.g. H.323, SIP), which may be even partly embedded into the applications. Therefore, type B applications directly have to manage QoS and mobility related issues by themselves. Type C and D are more futureorientated, they use high-level API (Application Programmer Interface), which is based on session layer protocol. In order to support all the different types of applications, application QoS should provide information, such as what QoS supports the application requires and what resources/ components the network QoS can supply, and a negotiation cycle is carried out.

At the beginning of a negotiation cycle (locally or among peers), the user specifies wishes, describing a user's satisfaction with the application performance, with respect to factors such as cost, synchronization and quality of presentation. Such user-perceived QoS parameters are then mapped down through the different system components (such as capabilities of multimedia devices and codec availability). As an example, a user may choose an advanced streaming video; such wishes may be described and mapped onto actual values in terms of visual quality, frame rate and resolution. Visual quality can be translated into peak signal to noise ratio, ranging from 30-25db; Frame rate decides the degree of motion, such as TV-rate, still image or single image; Resolution represents different clarification standard. Based on these parameters, demands such as bandwidth, delay or jitter can be calculated.

Negotiation between application and network QoS can then be undertaken based on demands. The application can start executing only if an agreement (contract) can be achieved. Otherwise, the user has to be informed. Thus it would be better to offer users only the range of parameters that the system is able to guarantee. In the mean time, the resource user agrees that he is not going to send more traffic over the network than was specified within the QoS contract. If the user does not operate according to the given QoS, the network may give no guarantee. Re-negotiation might extend to re-define a contract with QoS parameters if some parts of networks cannot cope with the original specification. So it is possible to offer different types of agreements or to assign probabilities to the agreements (e.g. downgraded service types or portion of bandwidth). It is very important that the contract is not static, and that the contract can be re-negotiated and monitored dynamically in order to adapt to changes. For example, this re-negotiation happens if the network’s state changes due to the congestion or handover of a wireless station to a new access point, which cannot guarantee the old contract.

Hence, the application QoS acts as an interface or a set of interfaces between user and network QoS not only statically but also dynamically. Due to the continuous new development of smart applications for wireless environment today, application QoS may even contain more intelligence (such as API, or codec) so as to meet the new demand and challenges.

There are also another significant trend of multimedia adaptation technique, which enable multimedia applications seamlessly adapt to changing network conditions, and provide high quality particularly for streaming applications. Adaptive bitrate streaming technique is commonly used in streaming multimedia over computer networks, by detecting a user's bandwidth and CPU capacity in real time and adjusting the quality of a video stream accordingly [Akhshabi et al 2011]. The single source video can be encoded at multiple bit rates using an encoder. The player client switches between streaming the different encodings depending on available resources over the network. It helps user maximize use of available bandwidth presented over IP layers, at a cost of higher operational complexity and possible higher traffic volume used.

Compared to similar but more vendor-centric implementation solutions such as HLS (HTTP Living Streaming) by Apple, Smooth Streaming by Microsoft, or HDS by Adobe, Dynamic Adaptive Streaming over HTTP (DASH), also known as MPEG-DASH, is an international latest MPEG standard [Lederer et al 2013][Liu et al 2013][Müller et al 2011]; it enables high quality streaming of media content over the Internet delivered from conventional HTTP web servers. By breaking the content into a sequence of small HTTP-based file segments, the content is made available at a variety of different bit rates. An MPEG-DASH client, the client automatically selects from the alternatives the next segment to download and play back based on current network conditions, so it provide high quality play back without stalls or re-buffering events. A number of researches including [Lederer et al 2013] [Müller et al 2011] put efforts on details of how a multimedia file is partitioned into one or more segments and delivered to a client. Such segment usually carries information of timing, URL, media characteristics such as video resolution and bit rates, which are the key factors for the performance of the implementation. On another hand, if the network could be presented with more available resources from the underlying layers and/or mechanisms, DASH could always have greater chance to capture high bit rate for the best communication.

2.5.3 The Framework of QoS Network Layer

Many strategies have been used to address the specific aspects of the QoS problem, such as SIP, which is heavily utilised in application QoS as stated in Section 2.5.2.

However, the Internet interconnects multiple administrative domains and is the concatenation of domain-to-domain data forwarding. QoS is only useful, and therefore most likely to be paid for, if it exists on an end-to-end delivery basis. Most QoS solutions/components alone do not solve the real-time service problems since they cannot guarantee that packets in any one flow are delivered across the Internet in a timely fashion. The Internet requires an end-to-end mechanism to ensure timely packet delivery through the network, even over the wireless links. This can only be achieved if there is control framework for timely packet delivery located at most router within the Internet. As discussed earlier in Section 2.4, IP, which is the most pervasive protocol in the network layer within the Internet, is an end-to-end transportation service in most cases; it is the best candidate network wide when playing the major role in providing the end-to-end QoS delivery. Therefore, the framework is designed in the network layer.

The planned framework focuses on the distribution of intelligence and control between topological components in the network and levels of such intelligence. The main concern here is the issue of host-centric versus network-centric. On the one hand, the locus of intelligence could be in the end systems in order to manage resources fairly between competing users of the network. Or it could be in the network as it may be the case, where dumb terminals are simply selecting the services offered by the network without amending them in any way. In fact, solutions, which compromise between those two extremes, are likely to be better choices. Although there is a range of possibilities on how to balance or distribute intelligence between the network and end systems, such as ATM QoS, Tenetsuite, the HeiTS/HeiRAT architecture [Dohoon 2010] and the QoS-A architecture [ITU-T 1996, Banerjea et al. 1996, Vogt et al. 1998, Campbell et al. 1994], the two major candidates, integrated services (IntServ) and Differentiated Services (DiffServ), have emerged as the principal frameworks for providing end-to-end QoS.

- Integrated Services

In the early 1990s, the IETF realised that the Internet's egalitarian best-effort model is not really suited for the emerging real-time applications, mainly for multimedia applications, if the network becomes significantly loaded. The study of this problem results in the development of the IntServ (Integrated Services) framework [Braden et al. 1994]. The philosophy of the IntServ model is to make an inescapable requirement for routers to reserve resources in order to provide special QoS for specific user packet streams or flows. This, in turn, requires flow-specific state in the routers [Bernet et al.

2000]. A flow is defined as "a distinguishable stream of related datagrams that result from a single user activity and require the same QoS" [Braden et al. 1994]. It uses a network-centric approach, yet maintains a high flexibility for users to choose from different services in a continuous way. It effectively locates some intelligence within the network in order to keep up reservation state and perform admission control. It also often couples with RSVP (Resource reSerVation Protocol) as the signalling protocol.

The IntServ framework is targeted at largely diverse traffic. In particular, the main traffic distinction is between real-time and elastic traffic, with the latter representing the traditional best-effort traffic. Best-Effort traffic is assumed to maintain an important role as the default service and applications using Best-Effort do not need any modification within IntServ. Real-time traffic is further categorized by the fact whether it is tolerant or intolerant to loss, where a late packet is considered lost. An even further distinction is made with respect to the adaptability of real-time traffic to delay rate variations respectively. The IETF has considered various QoS classes. And only two main services i.e. Guaranteed Service [Shenker et al. 1997] and Controlled Load Service [Wroclawski 1997a], are used for real-time traffic. Controlled Load Service provides approximately the same QoS under heavy load as a best-effort network would under light load; while Guaranteed Service provides an assured level of bandwidth, a firm delay bound and no queuing loss for conforming packets of a data flow.

Integrated Services is implemented by four components: the signalling protocol (e.g. RSVP), the admission control routine, the classifier and the packet scheduler. Applications requiring Guaranteed Service or Controlled-Load Service must set up the paths and reserve resources among routers before transmitting their data via signalling protocol. After joining the associated multicast IntServ group, routers are able to receive signalling request message containing a traffic specification (TSpec) from the traffic sender. Such information is used to determine traffic QoS requirements. The admission control routines would then decide whether a request for resources can be granted based on the traffic QoS requirements and local knowledge (computing resources available, cost constraints). It is then responsible for initiating its own reservation by generating a response message, which travels towards the sender along the reverse path of the request message. Details of reservation path set-up procedure will be covered later. When a router receives a packet, the classifier will perform classification and put the packet in a specific queue based on the classification result. The packet scheduler will then schedule the packet accordingly to meet its QoS requirements. The classifier and the packet scheduler are considered as data path mechanisms that will be discussed later.

The IntServ model then leads a fundamental change to the Internet architecture. However, researchers soon discover several problems with the architecture especially within some particular environments. It does not scale well in the Internet core, although it was originally designed with end-to-end principle on per application flow basis. It places a huge storage and processing overhead on the routers due to the giant amount of state information increasing proportionally. It also requires several complex additional components in the routers, such as RSVP, admission control, classification and packet scheduling. In the view of a domain (associated multicast IntServ group), ubiquitous deployment is required. RSVP functionality at the bottleneck nodes of a domain and tunnelling of RSVP messages over other part of the domain are the minimum requirements [Braden et al. 1997].

- Differentiated Services

Differentiated Services (DiffServ) can be seen as a response to the resulting complexity and scalability of the IntServ framework. It outlines a framework with the scope of inter-domain, which allows for bilateral contracts by Service Level Agreements (SLA) at each border [Black et al. 1998][Evans and Filsfils 2007]. The essence of DiffServ is to keep the core of the networks simple by only regarding traffic aggregates for establishing QoS context within core routers. The traffic aggregates are formed by adequately deciding at the edge of a DiffServ domain. A given traffic aggregate is forwarded in a router of the DiffServ domain by specified PHBs (Per-Hop Behaviours). By concatenating PHBs, it is possible to build sensible services, thereby allowing for an edge-to-edge scope to eventually achieve an end-to-end QoS.

PHBs are at the heart of the DiffServ architecture. They specify an externally observable forwarding behaviour of a router for a given traffic aggregate [Nichols et al. 1998]. These Behaviour Aggregates (BA) are identified by a DSCP (DiffServ Codepoint); DSCP is located in the upper 6 bits of the old IPv4 ToS (Type of Service) byte, which DiffServ now renames as the DS (DiffServ) field (in IPv6, DSCP is contained in the Traffic Class octet). DSCP is used to indicate the forwarding PHB treatment, which a packet should receive in that router. PHBs are normally standardized; Expedited Forwarding (EF) and Assured Forwarding (AF) are most commonly used. EF minimises delay and jitter, and provides the highest level of aggregate Quality of Service [Jacobson et al. 1999]. This means that a service based on EF requires strict policing or shaping at ingress edge routers to a DS domain. Excess AF traffic is not delivered with as high probability as the traffic 'within profile', which means it may be demoted but not necessarily dropped [Heinanen et al. 1999]. The AF PHB group consists of 12 PHBs that have a certain relationship with each other. There are 4 AF classes with three drop-precedence each. PHBs may be implemented by different scheduling, policing, shaping and buffer management mechanisms, which will be further discussed later.

The service provided is defined by means of a Service Level Agreement (SLA). The SLA is a contract between the service provider and the customer, which is established either statically or dynamically. It specifies the overall performance and features, which can be expected by a customer, as long as SLA adheres to a traffic conditioning agreement (TCA). The TCA specifies what constitutes the customer's traffic, i.e., which packets classify for this SLA, the traffic profile (e.g., a rate and a burst size), shaping or policing rules for the traffic, how to mark packets (i.e., which PHB to use) etc. It can also be established between two domains in order to provide a consistent DiffServ model. An End-to-End QoS can then be obtained by the concatenation of per-domains services (Per-Domain Behaviours, PDB) and SLA between adjoining domains along the path that traffic crosses in going from source to destinations. In such a case, Bandwidth Broker (BB) is introduced for dynamic SLAs to make a logically centralised admission control decision for an entire DiffServ domain in an automated manner [Nichols et al. 1999]. The design of a BB that gives strong guarantees and simultaneously achieves good resource utilisation is still an open research issue today. Such DiffServ/BB framework normally couples tightly with Policy-Based Management (PBM), which actually plays the role of inter-BB signalling (detailed in a later chapter). In such DiffServ/BB framework, by concatenating several SLAs, it is possible to build an end-to-end service.

The main differences between DiffServ and IntServ are simplicity, scalability and dynamic control. DiffServ does not attempt to give explicit end-to-end guarantees. Instead, in each congested network elements, traffic with a higher class of priority (codepoint/DS) has a higher probability of getting through, or in case of delay priority, is scheduled for transmission before traffic that is less delay-sensitive. The model does not require signalling protocols to control the mechanisms that are used to select the proper PHB in each node. Consequently, the amount of state information, which is required to be maintained per node, is proportional to the number of service classes and not proportional to the number of application flows. This reduces the computation and makes it simple and scalable. However, with regard to end-to-end principle, DiffServ/BB framework is introduced, which requires the usage of PDB, dynamic SLAs and PBM. Such solution is still very much under development according to capabilities, efficiency and complexity.

- Multi-Protocol Label Switching

Multi-Protocol Label Switching (MPLS) is similar to DiffServ in some respects as it also marks traffic at ingress boundaries in a network, and un-marks at egress points [Callon et al. 1999]. Unlike DiffServ, which uses the marking to determine priority within a router, MPLS markings (20-bit labels) are primarily designed to determine the next router hop. MPLS simplifies the routing process (decreases overhead to increase performance) while it also increases flexibility with a layer of indirection. It represents the convergence of connection-oriented forwarding techniques and the Internet's datagram routing protocols.

MPLS is more of a 'traffic engineering' protocol than a QoS protocol and can be considered as a data path technology, which introduces label switching into IP networks similar to ATM, but is explicitly tuned for IP transfer. MPLS routing is used to establish 'fixed bandwidth pipes' analogous to ATM or Frame Relay virtual circuits. The difference is arguable since the end result is service improvement and increased service diversity with more flexible, policy-based network management control whilst other QoS protocols also provide these advantages.

2.5.4 Control Plane Mechanisms

As discussed in Section 2.5.1, implementation of network QoS framework relies on consistent inter-working between the QoS procedures and declarations. QoS procedures encompass both the necessary interaction to invoke QoS treatment as well as the mechanisms that are employed to ensure the QoS assurance assigned to a unit of service, e.g., an application flow. The QoS declarations, on the other hand, contain the QoS specifications desired by a user invoking QoS treatment for a unit of service. The components of QoS procedures can be further classified depending upon whether they operate on the data or control path. Control path encompasses all actions that employ control on the way data is transferred over the network. The controlled components that actually affect the data transfer are located on the data path. The functionalities of control path mechanisms nowadays extend not only to signalling, but also to architecture of resources control via signalling. They are concerned with the configuration of network nodes with respect to which packets get special treatment and what kind of rules are to be applied to the use of resources. Different signalling protocols exhibit a large variety of characteristics, e.g., whether they are sender or receiver-oriented (initiating the process of QoS request), and whether they require multicast support.

- Resource Reservation Protocol (RSVP)

RSVP is a signalling protocol that provides reservation set-up and control to enable the IntServ [Braden et al. 1997]. It provides the highest level of QoS in terms of service guarantees, granularity of resource allocation and detail of feedback to QoS-enabled applications and users. By using RSVP, hosts are able to request a specific QoS from the network. It propagates the QoS request to all the routers along the path and additionally maintains state information within routers and hosts to provide the requested service. It can therefore be regarded as a state establishment and maintenance protocol [Zhang et al. 1993]. The protocol is placed on top of IP, thus not requiring any routing mechanisms in RSVP itself, but uses unicast and multicast routing mechanisms provided by the network layer. RSVP does not transfer application data but operates as a control protocol only.

RSVP starts by sending a PATH message from the sender that contains this traffic specification (TSpec) information to the destination address. Each RSVP-enabled router along the downstream route establishes a ‘path-state’. To make a resource reservation, receivers send a RESV (reservation request) message “upstream”. In addition to the TSpec, the RESV message includes a request specification (Rspec) that indicates the service types required and a filter specification that characterizes the packets for identification. When router along the upstream path receives the RESV message, it uses the admission control process to authenticate the request and allocate the necessary resources.

However, there are several shortcomings with RSVP, apart from having the signalling overhead as the main concern, the standard RSVP seems to provide limited adaptation capability to network dynamics. Although in combination with the periodic exchange of the refreshed messages to refresh reservations periodically by the receiver(s) according to the dynamic network changes, it once again induces more signalling overhead as a trade-off. Also, non-RSVP routers create a 'weak-link' in the QoS chain where the service fall back to 'best effort', even though the non-RSVP routers may not be the bottlenecks along the data path. The more routers are RSVP-enabled, the complexity of such operations are more significant.

- Policy-Based Management (PBM)

PBM focuses on QoS provision and configuration issues. Policy specifies the regulation of access to network resources and services by users, applications, or hosts, either conditionally or unconditionally based on administrative criteria [Rajan et al. 1999]. Policies control which users, applications, or hosts should have access to which resources and services and under what conditions. Instead of configuring individual network devices, the service provider regulates the network through PBM infrastructure, which provides supports for allowing administrative intentions to be translated into differential packet treatment of traffic flows.

As discussed previously, PBM couples tightly with DiffServ/BB/PDB across domains. PBM normally consists of four components. They are policy database, policy server, policy enforcement points (PEP) and management console. Each domain may contain one or more policy servers, which make policy and configuration decisions for network elements in the domain. It has access to a policy database as well as authorisation and accounting databases. In the policy database, policy is specified following a rule of "if certain condition happens, then take certain action". A management console serves as the network operator interfaces to the policy server through a set of Policy API in order to update and monitor policy changes in the policy database. The policy server consists of a set of policy decision points (PDP). PDP determines which actions are applicable to which packets. Policy enforcement points (PEP) perform the enforcement and execution of policy actions. PEP is also called policy client, which normally locates in router/switch. PDP communicates with PEP via Common Open Policy Service (COPS) [Durham et al. 2000].

The QoS policy and PBM are introduced in order to provide a uniform QoS control over the policy domains. By exchanging the SLA between the policy domains, the control mechanism of QoS can be extended to the entire network. Furthermore, SLA can be extensively used between policy client and policy server in a centralized approach. A central policy server then provides a user interface, which can exchange the dynamic SLA negotiation with policy client in a secured communication channel.

- Subnet Bandwidth Manager

The QoS 'chain' is end-to-end between sender and receiver, which means that every router along the route must have support for the QoS technology in use, as described previously. However, some Layer 2 technologies, such as Asynchronous Transfer Mode (ATM), have always been QoS-enabled. Whereas other more common LAN technologies, such as Ethernet, are not originally designed to be QoS-capable. This has resulted in the development of the 'Subnet Bandwidth Manager' (SBM) to define the mapping between upper-layer QoS protocols and services with those Layer 2 technologies like Ethernet.

SBM is a signalling protocol that allows communication and coordination between network nodes and switches in the SBM Framework and enables mapping to higher-layer QoS protocols [Seaman et al. 1999]. A fundamental requirement in the SBM framework is that all traffic must pass through at least one SBM-enabled switch. A 3-bit value (a component of an 802.1Q header) is used in IEEE 802.1p; it can represent an 8-level priority value [Yavatkar et al. 1999].

2.5.5 Data Plane Mechanisms

Following the discussion on different control path mechanisms, the second component of QoS procedures, the data path, is presented in this section. Data path mechanisms are the basic building blocks on which QoS is built. They enforce the QoS goals for units of service as projected on the control path. In general, by transforming the arrival processes of packets on all the interfaces of a packet switch into a controlled departure process, the data path satisfies the goals established on the control path. In short, they implement the actions that routers need to take on individual packets in order to enforce different levels of service.

Within the basic data path, operations in routers normally include components of fundamental packet forwarding operations, queue management and scheduling. The fundamental packet forwarding operation is made of classification, marking, metering, policing and shaping. When a packet is received, a packet classifier initially determines which flow or class it belongs to based on the content of some portion of the packet header according to certain specified rules. The packet is then marked by the marker in a certain field within the packet to label the packet for type differential treatment later. Traditionally, the classification can be complex-based on structured and/or multidimensional classifiers, which are required for application flows characterized by IP addresses and port numbers. However, modern packet stamping mechanisms, which allow packets to be marked by an unstructured, flat identifier, such as a DS field, would always be preferred to keep the classification task simple. A meter is used to measure temporal properties of the traffic stream against a traffic profile. It decides that the packet is in profile or out of profile [Zhao et al. 2000]; before passing the state information to other traffic management elements. Out of profile packets may be dropped, remarked for a different service, or held in a shaper temporarily. In profile packets are kept in different service queues for further processing. A shaper can delay some or all of packets in a packet stream in order to bring the stream into compliance with its traffic profile by using a buffer. A dropper can be implemented as a special case of a shaper by setting shaper buffer size to zero packets in order to drop/police out-of-profile packet. Thereafter, the packets are all conditioned before queuing and scheduling.

Queue management [Hoang and Phan 2007] [Phan and Hoang 2006] controls the length of packet queues by dropping or marking packets when necessary or appropriate, while scheduling determines which packet to send next and is used primarily to manage the allocation of bandwidth among flows. Therefore, queue management can control packet loss that is caused by network congestion. The size of buffers must strike a balance between allowing for bursty traffic arrivals and reasonable queuing delay. For the buffer management of output queues, two major decisions have to be made: when to discard a packet and which packet to discard. It couples tightly with scheduling scheme, for example, in the First-In-First-Out scheme; the scheme dictates that a buffer only drops packets, which are at the bottom of queue, when the queue is full. A further design decision for the buffer management is with regard to the isolation of buffers, with one buffer for all traffic and another for each application flow at the extremes. Due to the nature of wireless/mobile network in the MAC operation (such as CSMA/CD), it is unlikely to use queue management to control and avoid network congestion, which will be discussed in the next section. Instead, at network end-points, TCP protocol plays an important role, which uses adaptive algorithms such as slows tart, adaptive increase and multiplicative decrease.

Scheduling plays an important goal in packet delay control, particularly in queuing delay. Packet delay has three parts: propagation, transmission and queuing delay. It is difficult to improve propagation delay and per-hop transmission delay due to their natures, such as the distance, the medium, the speed of light and the packet size. The queuing delay is the waiting time that a packet spends in a queue before it is transmitted. This delay is determined mainly by the scheduling policy. Packet scheduling is concerned with the decision, which packet to send next on a given link if there are a number of buffered packets waiting for service. Such scheduling schemes are quite well developed, such as First Come First Serve (FCFS), Priority scheduling [Golestani 1990], Weighted Fair Queuing (WFQ) [Davin and Heybey 1990] [Shah et al 2011] and Weighted Round Robin [Saha et al. 1996; Shreedhar and Varghese 1996]. Besides delay control, link sharing is another important goal of scheduling. The aggregate bandwidths of a link are normally shared among multiple entities, such as different organizations, multiple protocols (TCP, UDP), or multiple services (FTP, telnet, real-time streams). An overloaded link should be shared in a controlled way, while an idle link can be used in any proportion. Schedulers that integrate hierarchical bandwidth sharing into their schemes have been previously described in [Floyd and Jacobson 1995] and [Stoica et al. 1997]. A typical development in the field of scheduling is described in [Stoica and Zhang 1999], where scheduling of a whole network, instead of just a link, is introduced. Alternative studies also includes [Nyandoro et al] and [Hassan et al]. The studies in [Nyandoro et al] has also shown that power control, as another alternative, could achieve service differentiation in 802.11 wireless networks by utilizing the concept of capture, where the high-power packet is still received correctly even when it collides with a low-power packet at the receiver. in recent [Hassan et al], game theory is used to achieve QoE particularly for wireless VoIP application, in a non-cooperative environment.

The mechanisms discussed above mostly operate normally at the IP layer because IP is the best layer to provide end-to-end QoS in a wired network. However, in most situations of wireless communication, the wireless last hop communication operates at the link layer instead of at the IP layer. Different technologies define different link layers specification, such as IEEE 802.11, as one of link layer models for WLAN. Therefore, data plane mechanisms in wireless last hop may not be standardized as above in IP layer. Indeed, data plane mechanisms couple tightly with individual link layer specifications to provide service differentiation over the wireless last hop. They are designed mainly with supports of link layer facilities. In Chapter 3, we will study a novel QoS architecture in the case of IEEE 802.11 WLAN and give details of service differentiations (data plane mechanisms), which are built upon the IEEE 802.11 link/MAC layer.

2.5.6 Summary

It can be found that there are several network QoS frameworks in use and the number of possible QoS components, which can build up these frameworks, is very large. These varied protocols, mechanisms and services are originally not designed to work together. Therefore, it is much more critical to mix and match their capabilities in a variety of possible frameworks to achieve the goal of end-to-end and top-to-bottom QoS-enabled communications. Interworking challenges in such heterogeneous QoS systems actually exist in many ways.

Heterogeneity may come from the fact that different QoS systems may be based on different QoS frameworks and thus use different QoS declarations. If such QoS frameworks are at a peer level, for example, one of the peering QoS systems uses ATM and the other one uses RSVP/IntServ, then the very basic communication system functionalities (data forwarding, routing, and addressing), need to be translated between the two frameworks. Among those functionalities, data forwarding behaviours (QoS declarations) translation is more critical for maintaining a consistent QoS from end to end. Although QoS declarations are often not that different but rather similar on a conceptual level, there is a considerable amount of work, especially within the IETF. They are mostly for the most prominent combinations of QoS frameworks, such as RSVP/IntServ over ATM, DiffServ over ATM and RSVP/IntServ over DiffServ.

The inter-working problems among different QoS systems can be considered not only based on a peer model, but also based on an overlay model. There is some individual research on the realm of MPLS based on RSVP as a trigger for setting up label-switched paths as stated in [Fourmaux and Fdida 1999]. In these heterogeneous QoS systems, the overlaid QoS architecture is not normally required to deal with how the underlying QoS architecture ensures the contracted QoS guarantees on the data path, but only that these guarantees are ensured. Therefore, an edge device is used to mediate between different QoS contexts on the data paths by re-labelling data packets. The re-labelling from one QoS context to another, of course, depends on the labels being used to establish QoS context inside heterogeneous QoS frameworks. In RSVP/IntServ, labels are represented by the 5-tuple (source address, destination address, source port, destination port, protocol), which is contained in the header of an IP datagram. For DiffServ, the DSCP is the label for interior nodes whereas for boundary nodes it depends on the respective SLAs, and in particular the respective TCAs, which constitute a label.

The re-labelling is not just the process of simply exchanging labels among each other; the process must provide a high flexibility to support techniques like aggregation and foresting at edge devices between the different QoS frameworks. Furthermore, re-labelling module of an edge device should also provide for the “hooks” to establish the QoS context that is demanded by control path procedures.

In a wireless communication environment, the inter-working problem becomes more challenging according to the wireless communication natures. It is noted that the wireless communication environment may be decomposed into wired part and wireless components. The design of a proper QoS framework, which can cover both wired part and wireless part, and provide end-to-end QoS, is still under discussion. The best solution seems to be a hybrid model, consisting of different QoS frameworks, that requires a smooth translation. Furthermore, as with the wireless hop, communication is not necessary in order to operate in IP layer, but additional data plane mechanisms are required in such environment, which will be discussed in the next chapter.

Although much discussion has been focused on QoS framework in this chapter, service differentiation (data plane mechanisms) over wireless last hop is also another critical issue to provide end-to-end QoS for wireless communication. It couples tightly with link layer specifications, which are unique in different wireless technologies. Specification designers pay significant attentions to define efficient and effective facilities to provide link layer service differentiation while developers [Kabara and Calle 2012; Maniatis et al. 2000; Dixit et al. 2001; Koodli and Puuskari 2001] still have large space to make choices to build up data plane mechanisms. While this area itself is independent research topic, it also has strong connections with control plane mechanisms and QoS framework. This dissertation firstly focuses on the QoS framework and control plane mechanisms for next generation wireless communication; then study service differentiation (data plane mechanisms), both by investigating the technical details of WLAN as a study case.

~ Success is never final and failure is never fatal. It's the courage that counts ~

(George Tilton)

Chapter 3 Hybrid Quality of Service Architecture and FICC Admission Control

In this Chapter, a hybrid QoS architecture framework suitable for new generation wireless IP networks is proposed in order to achieve the goals of QoS consistency across the network, as well as the scalability and simplicity of QoS control. We will then examine our hybrid QoS control model and study the relevant issues further for a radio access network based on the context of IEEE 802.11 WLAN with 802.11e. In particular, within the Hybrid framework, the Fair Intelligent Congestion Control FICC [Hoang and Phan 2007] [Phan and Hoang 2006] [Hoang 2006] is adopted as an effective rate-based congestion control scheme that addresses both fair bandwidth sharing among traffic classes and congestion problems encountered in current QoS architecture. Its Performance is studied with comparison with DiffServ model.

3.1 Hybrid Quality of Service Framework

3.1.1 Review on Existing Methods

Wireless multimedia services, especially real-time ones, require at least the same level of Quality of Service as that provided by the wired Internet for which most of the IP-based multimedia applications were originally designed. This leads to a problem of Quality of Service (QoS) consistency across the wireless and wired segments of the network [Maniatis et al. 2000][Bettermann and Rong 2011][Dixit et al. 2001]. Provision of Quality of Service (QoS) guarantees for multimedia applications over IP networks is rapidly becoming a critical research and design issue, especially within the radio access segment of the next generation wireless/mobile environment.

In Particular, for delay sensitive or bandwidth intensive applications, resource constraints over the air interface may become a significant technical challenge in the design of new generation wireless multimedia networks. Provided that the total resources available over the air interface are, on average, insufficient to meet the total resource requirements of the user application sessions admitted to the system, the level of QoS desired/expected by users can be provided on an end-to-end basis by means of service differentiation (i.e. sacrificing the system performance for services tolerant of longer delay and higher rates of data loss in order to meet the quality of service specified for the other, less tolerant services).

Therefore, radio access network is usually seen as the bottleneck in the end-to-end data path. QoS mechanisms over the wireless segment of the network have to be designed with consideration given to the entire (end-to-end) network QoS. Various end-to-end QoS architectures for networks involving wireline core/transport and wireless access segments have been proposed and discussed in the literature. Some researchers [Antonio et al. 2001] [Mahadevan and Sivalingam 2000a][Evans and Filsfils 2007] argue for DiffServ principles uniformly applied end-to-end throughout the entire network, while others [Banchs et al 2001] are of the view that more subtle and explicit QoS control mechanisms are required at the radio access level.

The BRAIN project group, which consists of Nokia, Ericsson, Siemens, Sony, BT, etc., is one of the dominant research groups in this area and also conduct research into area of IP Mobility Management in general. In its reports [BRAIN D1.2 2001] and [BRAIN D2.2 2001], the ISSLL (Integrated Services over Specific Link Layers) as QoS framework is proposed. This idea originated from the IETF ISSLL working group. The group recommends that the DiffServ architecture from the core network is expanded to the last hop using RSVP within the last wireless hop. However, the reports cover more on general concepts, which is most applicable to standard core network; whereas most individual access networks and last hops have their own QoS needs according to their technical specifications, especially at the date plane and control plane. Thus, BRAIN has suggested some good ideas to follow the new trend of Internet QoS community to apply DiffServ in core network and suggest per-flow traffic management.

A number of solutions involving the use of RSVP/IntServ [Braden et al. 1994] end-to-end QoS control over DiffServ based networks, have also been published for hybrid QoS architectures [Robles et al. 2001; Bernet et al. 2000; Conforto et al. 2002; Mahadevan and Sivalingam 2000a][Evans and Filsfils 2007]. In these studies, Resource Reservation Protocol (RSVP) is used with the Integrated Services message set as the signalling protocol for explicit resource reservations, enabling it to be used to provide end-to-end service for the user. However, the use of RSVP not only requires additional QoS mapping mechanism between RSVP and DiffServ, but also raises several issues including signalling overhead and set-up delays on roaming events. Also, in a wireless environment, there is no point in offering a tight bound performance at the current stage, due to the radio characteristic, for all the traffic. Therefore, it seems inappropriate to use RSVP for a worst-case delay link.

A complete DiffServ model has been used from end-to-end in the entire wireless network [Antonio et al. 2001; Mahadevan and Sivalingam 2000b; Yoon et al. 2000][Teitelbaum and Shalunov 2011]. It is known that even though the DiffServ model does not attempt to give an explicit end-to-end guarantee, it provides QoS in a simple, scalable and efficient way that is ideal in a busy network. Therefore, it is wise to apply DiffServ in core networks and transport networks as discussed above. As there is a high probability that the last hop will be congested, the DiffServ model can fit better in term of efficiency. However, handover operation is based on reservation or call admission control and there are also needs for special mechanism to manage delay sensitive traffic, high priority traffic (such as handover) or some other critical traffic. While the simple and scalable DiffServ QoS control model is suitable for the core part of the network, more explicit, admission and reservation based QoS mechanisms are required in the wireless access segment of the network where the resources available and the levels of traffic aggregation render the DiffServ principles less effective. Thus, the DiffServ model is inadequate for the entire picture.

3.1.2 The Framework Requirements

The limitations of the proposed QoS architectures above vary from insufficient level of control implemented within the DiffServ model only, particularly over the resource-limited air interface, to scalability and complexity, as well as IP-mobility related problems with solutions involving end-to-end use of the RSVP/IntServ model. The new development of next generation ireless/mobile network architecture is also given consideration on interaction between private wireless network and public mobile network or among different wireless/Mobile technologies (IEEE 802.11/HiperLAN, CDMA2000/UMTS, etc.), as one of design principles of QoS consistency among domains and systems. Owing to the shortcomings of existing approaches in providing QoS over next generation wireless communication, we recommend a few general criteria that should be satisfied by any universal solution in solving these performance problems. Such requirements are met in designing of a novel wireless QoS architecture in the next section based on the above rationales.

End-to-end Principle: The end-to-end principle is one of the architectural principles, not only for the Internet, but also for the next generation wireless/mobile network. On its way from one client to another client, the traffic has to pass different services within the networks. Notably, the end-to-end Service is conveyed over several sub-networks such as the last hop, access network and core network. In UMTS 1999 [ETSI UMTS 2000], the concept of bearer service is defined characteristically and functionally to achieve this goal. A bearer service includes all aspects that enable the provision of a contracted QoS. The end-to-end principle on the application level uses the bearer services of the underlying network(s), which are the Radio Bearer Service, Iu Radio Bearer Service and Backbone Radio Bearer Service.

One of the reasons for breaking the end-to-end principle into several underlying services is that underlying networks have different performance limitations according to QoS. For example, as discussed above, 3G/4G radio interface is one of the main bottlenecks for QoS along the whole path while the QoS techniques in the access and the core networks are much more mature today. In order to support this, the last hop network should offer some kind of additional services for QoS. However, provision of specific functions within the network often makes it difficult for the network to evolve towards support for new services. Therefore, support also needs to be minimal and should adhere to a simple design principle. All underlying networks should be independent of specific transport layer/ applications and the type of IP packets being transported should provide only an IP delivery service. The design should also minimise the usage of special functions to maintain transparency so that the difference of the framework/ architecture among the underlying networks is minimal, hence allowing for smooth inter-working. A balance between simplicity principle and the end-to-end QoS principle must be properly handled within the last hop, access network and core network.

- Simplicity and Resource Efficiency: This has been discussed in the previous paragraph as one of the trade-offs of some complicated End-to-End QoS designs. As discussed above, the bandwidth over the wireless links is extremely limited compared to the wired data path; a good solution should make efficient use of the wireless link and host resources. It should minimise its state and processing requirements, and avoid wasted overhead and retransmissions. The mechanisms used should be as close as possible to optimal for each particular wireless link.

Being simple is a very important criteria for the acceptance and deployment of such an approach today.

- Extensibility: Solution should limit its functionality to be independent of applications. Unforeseen requirements may occur in new applications. To serve these unanticipated needs, any solution should be easily adaptable in order to extend to cater to new needs. Thus, applications should be shielded from the complexity of underlying QoS specifications and QoS management.
- Compatibility: In order to achieve a seamless transition from legacy, a best effort end-to-end system that provides full support for QoS, a general and comprehensive QoS architecture should be able to support different kinds of end-systems as well as underlying network technologies. The interfaces and components of the architecture should be as generic as possible. This principle becomes critical in solution deployment in industry today.

3.1.3 The Proposed Architecture

In the following studies, we present the novel QoS framework in general according to these requirements. The designs in its two major parts, the core network and the last hop, are firstly discussed and the Hybrid framework is then presented based on the analysis given.

Differentiated Services at the Core Network

QoS architecture at the core network should be able to deliver quantitative differentiated services pertaining to suitable network control granularity; scalable and efficient network state management. In order to gain architectural scalability, detailed control information (e.g., per-flow states) and supporting control mechanisms (e.g., per-flow queuing) are not practical in the design of core networks. Consequently, the resulting level of service differentiation between service classes is often qualitative in nature. As such, network practitioners have to use quantitative provisioning rules to automatically engineer a network that experiences persistent congestion or device failure while attempting to maintain service differentiation.

A more dynamic form of provisioning is needed to compensate for the coarser-grained state information and the lack of network controllability if QoS is to be effectively realized. However, unlike traditional telecommunication networks where traffic characteristics are well understood and well controlled and long-term capacity planning can be effectively applied, internet traffic is more diverse and bursty, often exhibiting long range dependence [Willinger et al. 1994]. As a result, there is a need to design measurement-based dynamic control algorithms that can perform well under diverse traffic conditions. Another important challenge facing bandwidth management is the complexity associated with the rate control of traffic aggregates in core networks, which may comprise of flows exiting at different network egress points. This problem occurs when ingress rate control can only be exerted on a per traffic aggregate basis, (i.e., at the root of a traffic aggregate's point-to-multipoint distribution tree). Under such conditions, any rate reduction of an aggregate would penalize traffic flowing along branches of the tree that are not congested.

Based on the rationale discussed above, we considered applying differentiated services model in the core network. The model aggregates individual flows into several classes either on their entrance to the network, or when they cross-administrative domains. Flows may be rate limited, shaped or marked to conform to specific traffic profiles. These profiles are either negotiated between users and network providers (for aggregation on entrance into the network) or between neighbouring domains (for aggregation between domains). Inside a domain, each router only needs to select a Per Hop Behaviour (PHB) for each packet based on its class. State aggregation into a few classes means that this approach scales well, but the guarantees that may be provided are not as fine grained as with integrated services, which is neither economical nor practical for the core network.

The architecture intentionally leaves the definition of PHBs and their implementations open to allow different schemes in different domains along the entire data path. The services provided by this architecture are meant to offer various generic QoS levels as opposed to application specific guarantees; hence the decision to map traffic classes instead of flows to PHBs. Only entry points to a network must be aware of both application requirements and PHB semantics to perform flow aggregation into classes. However, when resources are limited in some of these domains, traffic policing, mean rate limiting, shaping and marking would be performed at these points based on the traffic profiles. Therefore, depending on the PHBs available, end-to-end services may not be fully offered in a pure DiffServ environment where resources are limited in some of its domains.

Explicit Control over the Last Hop

Most traffic profiles are normally static at the entry point of the domains. Both stable allocations for real-time applications such as streaming video and best effort allocation for bursty data applications such as web transactions are likely if a differentiated services model is used over the wireless last lop. However, due to the increasing diversity of applications, device programmability emerging, resources limitation on the wireless links and the nature of wireless channels, a stable allocation service can be easily "overrun" by non real-time sensitive data applications. Under such conditions, lower priority packets take advantage of service differentiation by transiting their packets using the higher priority service class. This practice leads to the "tragedy of the commons" phenomenon.

To avoid a total withdrawal of resources from the standard traffic classes with lower QoS requirements, e.g., other than streaming, there is a share reserved for interactive traffic from the pool of radio resources in the cell. In times of high load traffic flows with more demanding QoS requirements are allowed to displace flows belonging to applications with lower QoS requirements, but only up to a certain limit. The limits are specified by the maximum allowed number of active sessions for the regarded traffic class. When this limit is reached, the requested QoS is not accepted, but degraded to the next lower prioritized class.

While the limitation on standard traffic classes is in place, the bandwidth reservation and control mechanisms are still desired for those interactive traffic, even though they involve a difficult trade-off between guaranteeing the full length of bandwidth reservation and inhibiting excessive bandwidth hogging. Hard reservation guarantees bear the complexity of admission control when multi-tiered service quality is required. This requires applications to declare the session length in advance, which none of the widely deployed applications can easily provide. The absence of mobile device participation in the control algorithm makes it hard for bandwidth reservation. However, the natures of the applications may define certain QoS requirements. For example, a VoIP requires a stable bandwidth reservation while a file download may prefer a possible maximum allocation over a short term. The reservation based on these implications may slightly relieve users from declaring session lifetime, and gives early warning of any pending allocation degradation while keeping potential arbitrage.

The resource reservation for a flow is initiated by the receiver of the data flow and resources are requested for simplex flow only. Two-way reservations are emulated by making two simplex flows in opposite directions. The explicit path establishment and resource reservations rely on a signalling mechanism between the receiver and the Admission Controller (AC). However, this signalling is not necessary a RSVP. Even though RSVP is not a routing protocol, it relies on existing and future routing protocols to determine where packets get forwarded. The proper signalling mechanism is only concerned with the QoS of those packets that are forwarded; and should have options for its operation layer (IP-based or MAC-based), The selection should depend on the nature of networks, which may provide simplicity and effectiveness for signalling.

The signalling mechanism transports and maintains traffic control and policy control parameters that are opaque to the signalling. Its job is to transport these parameters from node to node. The actual processing of the traffic control and policy control parameters is performed by the relevant traffic control and policy control modules present in a node. These modules should at least include Admission Control (AC), QoS profile negotiation and radio resource status monitoring. In particular, AC is responsible for handling activation and deactivation of flow requests, keeping track of traffic load and radio resource utilisation status, and performing QoS renegotiation.

Overall Hybrid Architecture

While the DiffServ model is useful in providing efficient and scalable QoS control within the network segments characterised by high volume of available resources and high aggregation of traffic (i.e. core/transport network), it fails to provide subtle enough tools for controlling QoS where the resources are extremely limited and the levels of traffic aggregation are low, e.g. in the wireless access network. The last hop (wireless access) radio resource management is unable to rely solely on mechanisms providing differentiated treatment of packets that belong to different application sessions. In order to avoid degradation of QoS as the traffic generated by the users within the same access network increases, a mechanism is needed at the access network level to control the total resource requirements of the sessions admitted to the system (explicit admission control), and to reserve the amount of resources required by each session. Such a mechanism, which operates on a session-by-session basis, is characteristic of the IntServ QoS model. The above argument serves as a brief justification for our choice of a hybrid model where DiffServ principles is applied over the core/transport network domain, and IntServ principles is applied locally to the QoS control over the wireless access segment.

The proposed QoS control architecture (Figure 6) is comprised of the Differentiated Services part in the core/transport network segments, and explicit resource management (admission control and reservation) part in the radio access network. Our architecture does not presume any specific QoS control model in the remote network where correspondent node (the other party in the application session) is located; it assumes that it is the other network's responsibility to guarantee, at its end, a QoS level consistent with that in the remaining parts of the path.

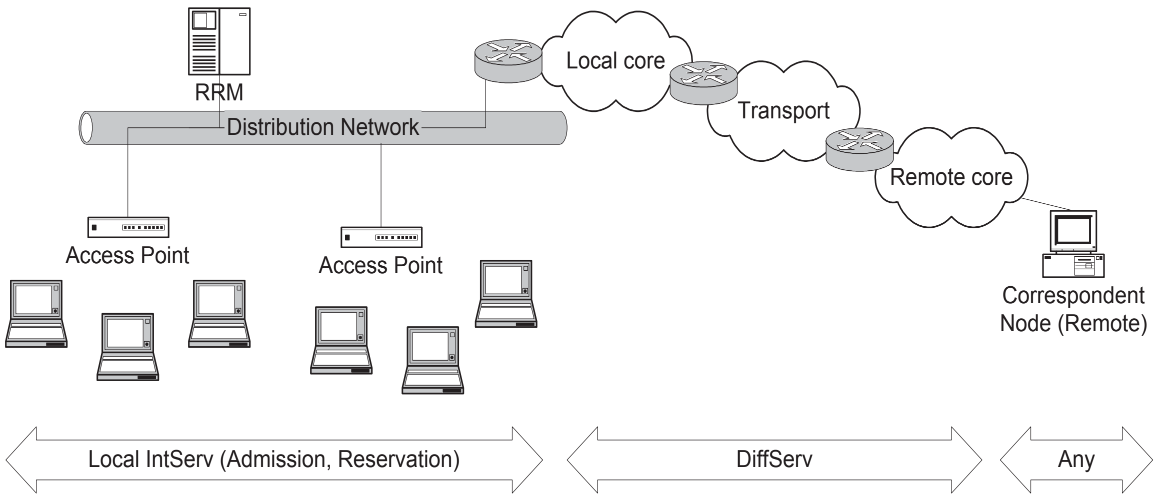


Figure 6 **Hybrid QoS Architecture**

The explicit resource management is localised to a single radio access network domain, where stately and fine-granularity control mechanisms operating at the level of individual flows and application sessions can be applied without causing scalability and complexity concerns. The resource management at the access network level is based on functional blocks typical of the IntServ model, i.e. admission controller and packet (frame) classifier/scheduler, with multiple queues and service disciplines used to enforce QoS guarantees given to the flows (sessions) upon admission. However, unlike the IETF IntServ architecture, it does not use explicit end-to-end path establishment and resource reservations such as those available with RSVP.

Limitation in its scope of the last hop means that the signalling required by the reservation based mechanism does not have to be implemented by means of explicit application level signalling protocol. It can be easily implemented at the medium access control level as part of MAC requests. In the case of the 3G access networks, the MAC-level admission control and reservation signalling become parts of the radio resource management necessary to handle admissions of individual terminals and changes of their link states. In the case of 802.11 WLANs, the explicit resource management cannot be easily and reliably achieved with the standard MAC data and control frames and random access based Distributed Coordination Function commonly implemented in the the 802.11 products. Signalling is required involving exchange of control packets between the wireless station, Access Point (AP) and WLAN-wide radio resource management entities in a suitable mechanisms, which will be discussed later in this Chapter.

Exchange of QoS signalling at the local radio access network level requires that all major entities in the network (wireless stations, access points, edge router) are equipped with QoS agents (Figure 7). The role of QoS agents (management plane processes) is to capture QoS requirements known to the application agents and/or application session control processes, and to facilitate the transfer of the QoS related information to QoS processors (classifiers, schedulers), as well as exchange of QoS signalling with other QoS (Resource Management) entities.

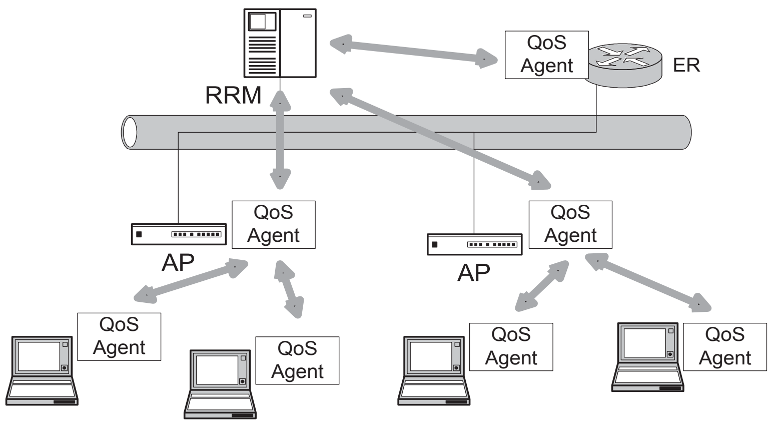


Figure 7 QoS Agents

Applications, particularly session-based, normally begin with session set-up procedure. The QoS parameters need to be guaranteed for the duration of the session and are either negotiated between the application and network entities at the session set-up time, or implicit in the type of application. Examples of session-based traffic include Voice over IP and video streaming sessions. Non-session-based traffic does not need hard QoS guarantees, thus explicit resource reservation is not necessary. The traditional “best effort” service is sufficient for this class of traffic. The amount of on-air time for transmissions of a specific class except non-session-based traffic is limited by the admission control procedures. These procedures not only permit explicit and signalling based control for session-based traffic, but also permit use of non-session-based traffic without explicit admission control up to some specified limit.

Central to the admission control, the Fair Intelligent Congestion Control (FICC), an aggregate intelligent congestion control scheme is adopted particularly at the edge devices to ensure that the domain is not congested to the point that it cannot maintain the agreed level of QoS. FICC plays a key role in admission control by suggesting an optimal amount of traffic that should be admitted to maintain an agreeable QoS level. The main purpose of FICC is to achieve the fair bandwidth allocation, minimum buffer queue length variation and simple implementation. It provides the required per-flow QoS response. The flow-level admission control focuses on maintaining fairness between individual flows locally so that the available resources are shared fairly with the pre-assigned allocation, while also trying to ensure delay performance when possible. As the DiffServ part for the core network has been well studied in the literature, we have chosen to further examine the resource management/explicit admission control in the radio access network as the key part of the hybrid QoS architecture for the mobile/wireless environment. The analyses below are given in the context of 802.11 WLAN, and the admission control over WLAN is largely built on the 802.11 MAC QoS extensions.

3.2 Fair Intelligent Congestion Control as the Core of Admission Control

The Admission Control functionality is located at the Radio Resource Manager and is responsible for admission control of session-based application streams. The admission decisions are made on the basis of stream QoS requirements and the current RRM’s knowledge of the resource usage (reservation) status in the WLAN. The decision calculations are based on the algorithms of FICC, which would ensure the effectiveness and the efficiency of traffic management. The admitted streams are then registered with the edge router for the purpose of mapping between the 802.11e stream QoS descriptors (TSPEC) and stream identifiers (TSID), the user priority levels on the Ethernet distribution network and the DiffServ DSCPs visible at the edge of core/transport network. The QoS signalling between the wireless client Radio Resource Manager, the edge router and the Access Point is accomplished by Admission Control mechanisms, built on 802.11MAC level standards, such as TSPEC negotiation facilities defined in 802.11e QoS supports, as discussed below.

3.2.1 IEEE 802.11 WLAN and its MAC QoS Supports

The IEEE 802.11 Wireless LAN (WLAN) is becoming a significant wireless part of the Internet constituency as a popular alternative to high installation and maintenance cost incurred over wired LAN infrastructures in recent years. However, the IEEE 802.11 WLAN is originally designed as "wireless Ethernet" by virtue of its best effort service provisioning based on Ethernet-like medium access control (MAC) protocol. The inherent property of radio network even leads to further QoS problem. Limitation of available bandwidth, propagation delay and higher error rates, caused by interference, are incurred thereby requiring a higher standard for QoS supports within the WLAN.

Standard MAC Functionalities

The 802.11 MAC layer specifies two modes of medium access operation: Distributed Coordination Function (DCF) and Point Coordination Function (PCF) [IEEE P802.11 1999]. In the DCF mode, the only mode of operation implemented in most of the commercially available 802.11b products, stations contend for transmission opportunities by following the principles of Carrier Sense Multiple Access with Collision Avoidance (CSMA/CA). Prior to transmission, a station must listen to the channel. If the medium (channel) is sensed free for a time interval greater than the DCF Inter Frame Space (DIFS), then the station can transmit a DATA frame. Otherwise, the transmission is deferred until the channel is free, after which an additional backoff time is applied. The backoff interval is randomly selected from a range of 0 to Contention Window (CW), with the initial CW of 32, doubled for every new attempt to transmit the same frame (up to the max value of 1024). To further reduce the probability of collisions, especially when hidden terminals may be present in the network, Request To Send (RTS) and Clear To Send (CTS) frames may be exchanged between the source and destination stations prior to the transmission of a DATA frame, causing all stations that listen for either the RTS or the

CTS to defer their transmissions. The DCF is simple and performs well under light traffic loads, but may not be sufficient to allow for any service differentiation.

In the PCF mode of operation, the AP alternates between Contention Periods (CP), where the DCF access rules are used, and Contention Free Periods (CFP), where the AP explicitly allocates transmission opportunities to wireless clients by polling. The CFP is started by a broadcast of a special beacon frame from the AP; this forces all wireless stations to enter a mode where they can transmit only in response to a poll. The AP sends poll frames, possibly piggybacked on the downlink frames, to stations that have been placed on the polling list; these in turn respond with acknowledgement and uplink frames. At the end of the CFP (signalled by a CF-End frame), stations revert to their normal operation under the DCF rules (i.e. enter the Contention Period), until the next CFP. PCF offers a certain level of service differentiation (the AP can schedule transmissions to and from selected stations), but fails to provide the level of service differentiation control necessary to deliver QoS guaranteed service.

802.11e QoS extensions

Several limitations of the basic DCF and PCF in WLAN applications, where QoS control is of primary concern, have led to the research in IEEE (through its 802.11e working group) on QoS extensions to the 802.11 MAC layer. The goal of this extension is to enhance the access mechanisms of IEEE 802.11 and provide a distributed access mechanism that can provide service differentiation. With the resulting documents, we will present a summary of the new features outlined in the 802.11e [IEEE P802.11e 2002].

The first enhancement offered by the 802.11e is the Enhanced DCF (EDCF). In the original 802.11 DCF, all clients enjoy equal access to the radio resource. The 802.11e EDCF offers up to 8 prioritized traffic categories (TCs) with TC7 having the highest priority. Frames to be transmitted can be marked with one of the TCs and forwarded to a queue specific to that TC. Frames at the head of each of the eight queues contend for a transmission opportunity using the DCF rules, except that the lower the TC priority, the longer the station has to sense the medium idle before gaining its transmission opportunity (TXOP). TXOP is a new concept/feature in 802.11e MAC defined as an interval of time when a client has the right to initiate transmissions, and is measured by a starting time and a maximum duration. Therefore, the 8 queues can be seen as 8 virtual stations contending for access to the radio medium. Allocation of different waiting times applied to the different queues ensures that the higher priority frames are given greater transmission opportunities more often than the lower priority frames.

The second major enhancement in 802.11e is the new contention-free access mechanism called Hybrid Coordination Function (HCF). Under the HCF, the AP can in addition to polling clients during the CFP, poll clients (by allocating them a TXOP) during the CP. The AP can also allocate itself TXOPs. Upon granting the TXOP, the start time and the maximum duration of transmission are specified. Once polled, a client can send multiple data frames within the same TXOP. To accommodate requests for TXOPs, the AP advertises a Controlled Contention Interval (CCI) whereby clients can make resource requests (RRs) i.e. requests for TXOPs. Combined with EDCF, the HCF offers greatly improved QoS control at the wireless MAC protocol level.

In addition to differentiation and short-term medium reservation capabilities offered by the EDCF and HCF, the third 802.11e extension, namely the Traffic Specification (TSPEC) facility, offers flow-based traffic admission and reservation capabilities. When a client wants to reserve resources for a stream of frames (a flow), it sends a TSPEC (Action) frame with specification of the QoS requirements (e.g. data rate, delay) for the flow. The AP now decides whether it should admit the TSPEC, admit a modified TSPEC, or reject the TSPEC, before returning the result to the client. The admitted TSPEC is given one of 8 TSPEC IDs (TSIDs) identifying different flows associated with the client in question, and all frames using that TSID are classified for treatment according to the TSPEC registered for the specific flow. The TSPEC facility offers a means for MAC level admission control and reservation signalling between the wireless clients and the Access Points.

3.2.2 Admission Control Mechanism for 802.11 Wireless LANs

Admission Control Overview

Located at the Radio Resource Manager (RRM), the admission control functionality is responsible for admission control of session-based application streams. The admission decisions from FICC are made on the basis of stream QoS requirements and the current RRM’s knowledge of the resource usage (reservation) status in the WLAN. The admitted streams are registered with the edge router for the purpose of mapping between the 802.11e stream QoS descriptors (TSPEC), stream identifiers (TSID), the user priority levels on the Ethernet distribution network, and the DiffServ DSCPs visible at the edge of core/transport network. The QoS signalling between the wireless station and the Access Point is accomplished by means of MAC level TSPEC negotiation defined in 802.11e, even though the signalling is initiated at the application level as shown in Figure 8. The Service Differentiation performed at the 802.11e MAC layer level ensures that the high priority (session streams) frames have transmission opportunities (TXOP) satisfying their QoS requirements, as promised at the time of stream admission. Lower priority traffic is treated according to "best effort" principles, filling in the bandwidth available after the session-based streams admitted to the system have been satisfied. Details of the QoS architecture are described in the following figure.

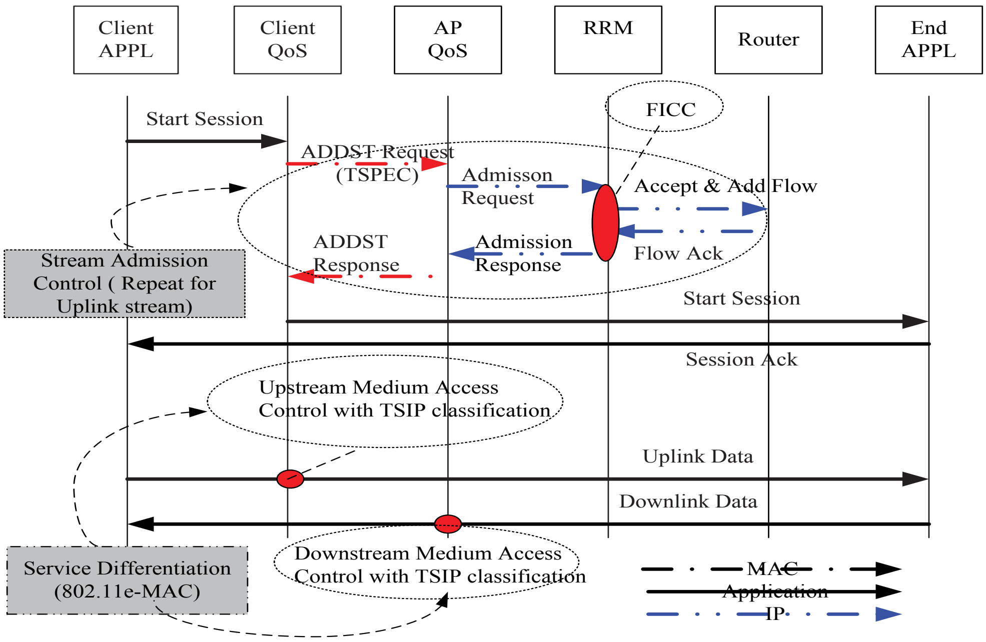


Figure 8 A Signalling Diagram for Flow Admission Procedure

When a wireless station (STA) initiates, or is invited to, a session-based application, a session set-up dialog is carried out (a Session Initiation Protocol (SIP) Invite dialog as an example). The QoS agent in the STA will capture the QoS requirements of media streams involved in the session, and map them to a MAC layer TSPEC description as defined in the 802.11e. In order to request admission and reserve radio resources for the stream, an 802.11e ADDTS-Request (add stream request) frame is sent to the AP. It carries a TSPEC element, which describes the source address (MAC), destination address, TSID, and QoS parameters of the stream. The QoS agent in the Access Point then forwards, in an IP packet, the admission request to the Admission Controller in the Radio Resource

Manager. The RRM has “global” knowledge of the WLAN resources and reservation status; and will either admit or reject the stream, taking into account the resource usage across the WLAN. If the stream is successfully admitted, the RRM registers the stream with the edge router (via IP level communication) and sends a positive reply to the AP’s QoS manager. Subsequently, a QoS ADDTS-Response frame is sent back to the wireless station, carrying a TSPEC element for the admitted stream. The admitted TSPEC can be as requested, or altered as a result of resource negotiation at the RRM.

Once resources are reserved for a stream, application data frames must be classified in order for the service differentiation mechanism to be applied in the AP at the MAC layer level. The task of the classifier (for downlink traffic, in the edge router) is to: give an IP datagram of a particular flow (identified by, for example, the source/destination IP addresses and port numbers; we refer to this as the flow ID), and allocate the stream identifier (TSID) to the corresponding MAC data frame. For user data traffic on the uplink, the classification is straightforward because the QoS agent in the wireless station has knowledge of the streams generated by this station and admitted to the system. As a result, the TSID can be inserted directly into the MAC frame when it is generated at the station. For downlink traffic, classification is more complex because the AP operates only at the MAC layer level, and has no knowledge of the traffic flows at the IP level. The process of classification and mapping of IP flows onto the TSIDs must begin at a layer 3 device, i.e. the edge router.

For downlink traffic, the edge router examines IP packets to detect flows and marks the distribution network (e.g. Ethernet) MAC frames with a priority level based on the TSID previously registered for the flow; (as part of admission control procedures, the RRM notifies the edge router of new flows). The priority information in the 802.3 MAC frame on the distribution network is carried in the additional 802.1p header. This additional header, which can be processed by most Ethernet products available today, carries a 3-bit user priority field. Therefore, when the router sends a distribution network MAC frame towards the AP, the frame contains the mobile host MAC address and the user priority value equal to the registered TSID. The AP’s QoS agent must then interpret the Ethernet user priority field as the TSID for this frame. Together with the identity of the destination station, it determines the service differentiation treatment the frame will receive at the AP.

In support of classifying downlink traffic using the Ethernet user priority field, the network should be configured as below. The WLAN distribution network should be separated from other parts of the LAN by an edge router where classification of flows is performed. Fixed hosts attached to the Ethernet (such as servers within the WLAN subnet) must be equipped with a QoS agent that ensures MAC frames sent by them are marked with the appropriate user priority that will be interpreted by the APs as TSID. These seem to be practical to most of Ethernet subnet today.

Procedure at Clients

Each traffic class that makes request for stream admission control shall maintain two variables. The first of these is the admitted time and the second is the used time. The admitted time shall be set at an association time to the value of default admitted time for that traffic class specified in the last received QoS parameter set element. The client may subsequently decide, using any heuristic algorithm it chooses, to explicitly request an amount of time per beacon interval for a specific class.

In order to make such a request, the client shall transmit a TSPEC request element contained in a management frame such as a generic management action request frame. That frame may also contain other elements. On receipt of a TSPEC response element, which is contained in a management frame, the client shall set the admitted time variable for the specified class to the value contained in that element. The client may choose to tear down the explicit request at any time. In order to tear down an explicit admission, the client shall transmit a TSPEC request element to the RRM indicating that the request is a teardown. If an explicit admission is torn down, either by the client or RRM, the client shall set the value of admitted time to the default value contained in the QoS parameter set element that it has just received.

If the used time reaches or exceeds the admitted time value, the corresponding class shall no longer transmit using the access parameters for that class as specified in the QoS parameter set element. However, a client may choose to temporarily replace the parameters for that class with those specified for best effort in order to continue transmissions at reduced priority.

If, for example, a client has made and accepted an explicit admission for a class, and the channel conditions subsequently worsen, possibly in conjunction with a change in the PHY layer data rate, such that it requires more time to send the same data, the client may make a request for more admitted time to the AP/RRM. At the same time it may choose to downgrade the parameters to best effort for short intervals in order to send some of the traffic of higher priority, whilst waiting for a response to the admission request.

Procedures at the RRM

The RRM shall respond to requests for time conveyed in TSPEC request elements. On receipt of a TSPEC request element conveyed in any management frame except when it is a probe request from an associated client, the RRM shall determinate whether to:

- Accept the request
- Deny the request
- Accept a modified request

The RRM may use any algorithm in making such a determination. Having made such a determination, the RRM shall transmit a TSPEC response element to the requesting AP/client contained in a generic management action frame. The RRM may choose at any time to modify or tear down an explicit admission. In that case, it shall transmit a new TSPEC response element to the AP/client contained in a generic management action frame.

The RRM shall make a determination, using any algorithm it chooses, as to the amount of time any client that has not made an explicit admission request for each traffic class may use for that class in a given beacon interval. The value for each class is designated a default admitted time. The values, contained in the QoS parameters element, shall be transmitted in every beacon from the AP.

A suitable algorithm for making this determination that the RRM may choose to use consists of these steps:

- Determine the total amount of time available in a beacon interval
- Subtract the total amount of time that has been explicitly admitted
- Divide the remainder between the non session-based classes
- Divide each value by the number of associated clients/stations in the BSS

However, this algorithm is simplistic at best. If it wishes, the RRM may choose to monitor the channel continuously in order to estimate the usage so that it may refine the default admitted times. This is what would be undertaken in FICC. We will give more details in the next section

Traffic Stream Signalling Procedures

As shown in Figure 8, the implementation for signalling procedure in the Hybrid structure largely relies on traffic stream operation between AP and client while the AC decision is made by the RRM. The traffic stream operation is implemented using the facilities of Traffic Stream (TS) and TSPEC description provided in IEEE 802.11e. A TSPEC describes the QoS characteristics of a traffic stream (TS). The main purpose of the TSPEC is to reserve resources within the AP/RRM and modify the AP/RRM scheduling behaviour. It also allows other parameters to be specified that are associated with the traffic stream such as ACK policy and use of FEC (Forward Error Correction).

A TSPEC is transported by the ADDTS (Add TS) and DELTS (Delete TS) QoS Action frames and across the MLME (MAC sublayer management entity) SAP (service access point) by the MLME-ADDTS and MLME-DELTS primitives. Following a successful negotiation, a traffic stream is created, identified within the client by its TSID and Direction, and identified within the AP by a combination of TSID, Direction and client address. Below, it outlines the complete TS lifecycle in a state transition diagram and further details their procedures of each lifecycle state in the time sequence diagrams.

- TS Lifecycle

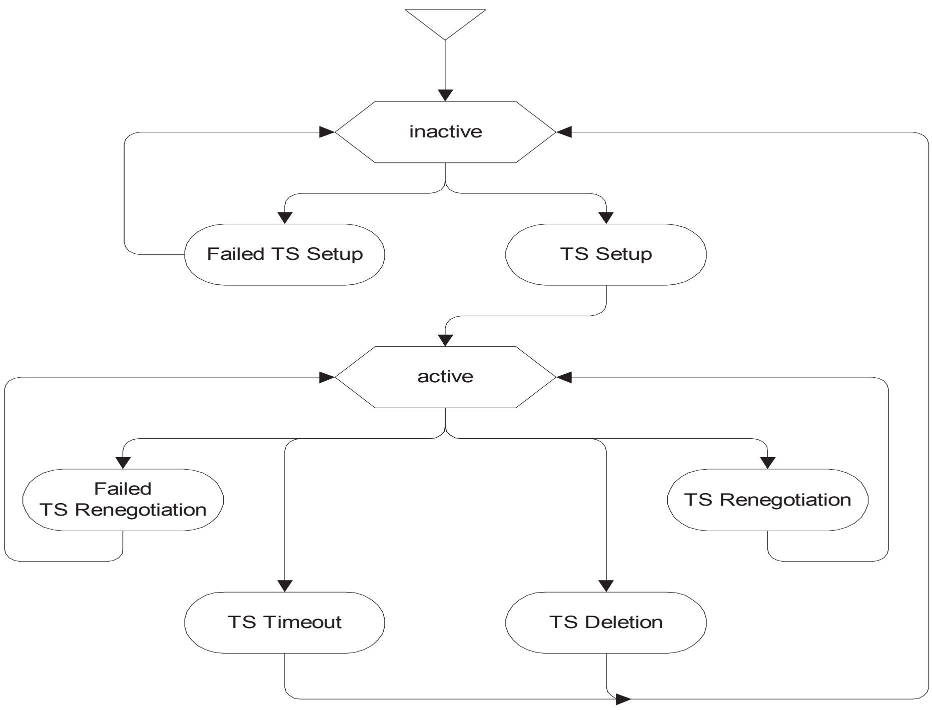


Figure 9 **State Transition Diagram of TS Lifecycle**

Figure 9 summarises the TS lifecycle. Initially TS is inactive. Following a successful TS Setup initiated by the client, the TS becomes active, and either the client or the AP may transmit MPDUs (MAC Protocol Data Unit) using this TSID. While the TS is active, the parameters of the TSPEC characterising the TS can be re-negotiated as initiated by the client. This negotiation can either succeed – resulting in a change to the TSPEC, or it can fail, resulting in no change to the TSPEC. Active TS becomes inactive following a TS deletion process initiated at either the client or the AP. It also becomes inactive following a TS timeout, which is detected at the AP.

- TS Setup

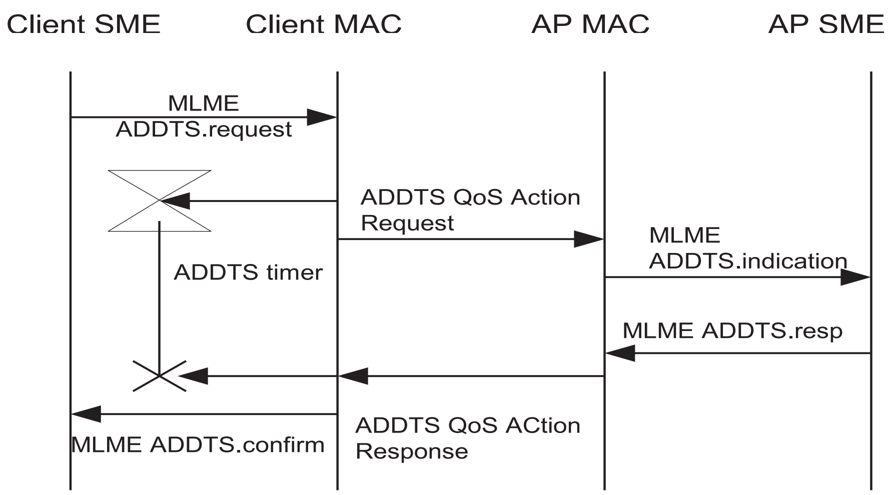


Figure 10 Time Sequence Diagram of TS Setup

Figure 10 shows the sequence of messages occurring at a TS setup. The client SME (Station Management Entity) decides that a TS needs to be created. How it does this and how it selects the TSPEC parameters are within the scope of application layer. It generates an MLME-ADDTS.request containing a TSPEC.

Condition	TSPEC Contents	Status
AP SME grants requested TXOP	Exactly as the requested TXOP	Success
AP SME grants an altered TXOP	TSID and Direction field the same as the requested TXOP. Other fields can be modified	Alternative
AP SME refuses TXOP	Exactly as the requested TXOP	Refused

Table 2 TSPEC and Status field contents in the MLME-ADDTS.response

The client MAC transmits the TSPEC in an ADDTS QoS action request to the AP and starts a response timer called ADDTS timer of duration. The AP MAC receives this MPDU and generates an MLME-ADDTS.indication primitive to its SME containing the TSPEC. The SME in the AP decides whether to admit the TSPEC as specified, admit the TSPEC with a counter proposal or refuse the TSPEC and generates an MLME-ADDTS.response primitive containing the TSPEC and a status value. An example of the contents of the TSPEC and status field containing values is specified in Table 2.

The AP MAC transmits an ADDTS QoS action response containing this TSPEC and status. The client MAC receives this MPDU and cancels its ADDTS timer. It generates an MLME-ADDTS.confirm to its SME containing the TSPEC and status. The client SME decides whether the response meets its needs. If the responses are "OK" and "ALTERNATIVE" Status, the TS is left in the active state. If an alternative grant is acceptable, the setup procedure ends here. The whole process can be repeated using the same TSID and Direction and a modified TSPEC until the client SME decides that the granted TXOP is adequate or inadequate and cannot be improved. If the client SME decides to terminate and an ALTERNATIVE is inadequate, it is the responsibility of the client SME to destroy the TS using the TS Deletion procedure.

There are three possible cases of failed TS setup:

1. An alternative grant is not acceptable to the client SME
2. The ADDTS MPDU transmission has failed
3. No ADDTS MPDU response is received from the AP (for example either because of delay due to congestion, or because the response frame cannot be transmitted)

The first case is indistinguishable from success within the MAC, and is not considered further in this discussion. Cases 2 and 3 are considered to be the same as the client cannot be accurately determine if its transmission had failed. Figure 11 summarises this case. The client MAC shall send a DELTS QoS Action request to the AP specifying the TSID and Direction of the failed request in case the AP had received the request and the response had been lost.

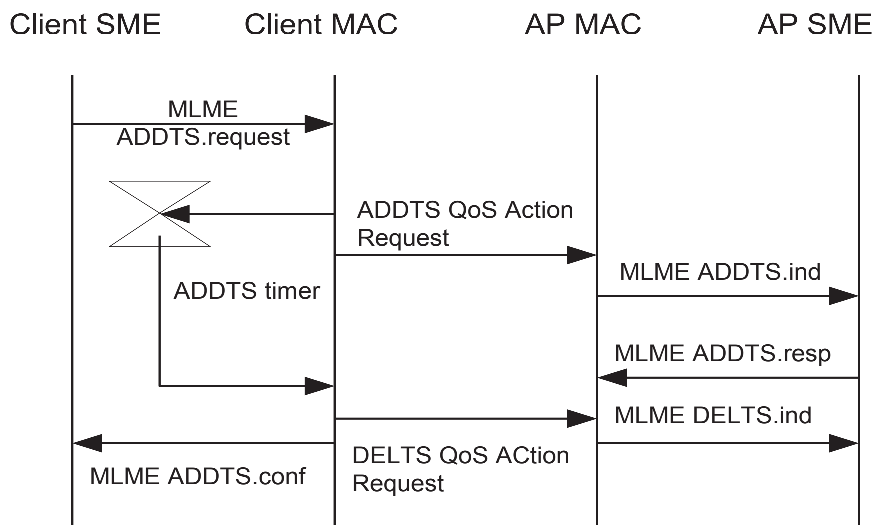


Figure 11 Time Sequence Diagram of Failed TS Setup detected within client MAC

- TS Deletion

There are two cases of TS deletion: client-initiated and AP-initiated. In both cases, the SME entity above the MAC generates an MLME-DELTS.request specifying the TSID and Direction of the TS to be deleted. This causes the MAC to send a DELTS management action frame.

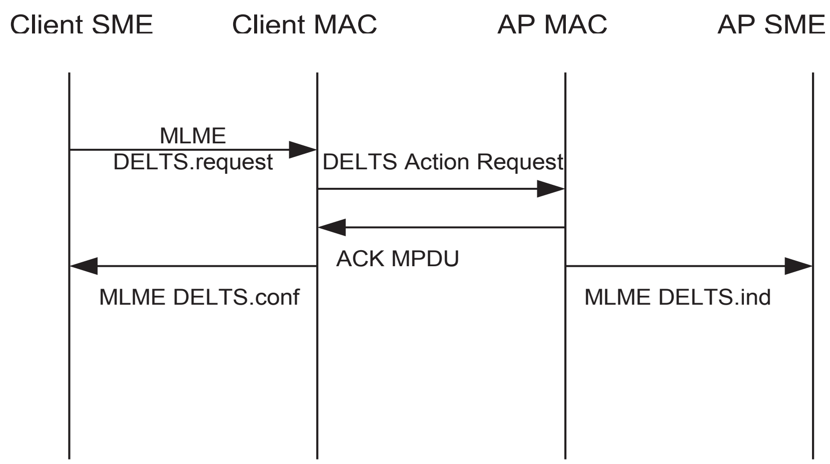


Figure 12 Time Sequence Diagram of TS Deletion (client-initiated)

The TS is considered inactive within the initiating MAC when the ACK MPDU to the management action frame is received. No management action frame response is generated. Figure 12 shows the schematics of client- initiated TS deletion. The case of AP-initiated TS deletion is the same with the client and AP labels swapped over.

- TS Timeout

TS timeout is detected within the AP MAC when no traffic is detected on the TS within the inactivity timeout specified when the TS was created. In response to an inactivity timeout, the AP shall send a DELTS QoS Action request to the client, and inform its SME using the MLME-DELTS.indication.

3.2.3 Fair Intelligence Congestion Control

The Admission Control mechanisms above are built to provide an effective way to convey traffic information in the TSPEC, between the QoS Agent in the client/applications and RRM/AP. Based on the available information; RRM makes admission decisions and informs the QoS agent with the same channel. Performances of the admission control rely on the decisions and its algorithms below. Here, we examine the effectiveness and efficiency of FICC, when applying it to our hybrid model. At the interface between wired and wireless networks, FICC should enable effective admission controls.

The Admission Control (AC) function plays a decisive role in QoS-oriented traffic management in a WLAN environment. Within the scope of the WLAN, it has to fulfil various tasks ranging from AC and QoS profile negotiation to radio resource status monitoring. The AC should represent the QoS management related functions of both the core network and the WLAN. Specifically, it is responsible for handling stream activation and deactivation requests, keeping track of traffic load and radio resource utilisation status within the radio BSS and performing QoS profile renegotiation. The RRM must provide methods and objects to store, update and evaluate radio resource utilisation data as well as traffic flow related QoS profile data (Traffic Flow Template). Furthermore, it has to perform admission control on the basis of this data and react to changing resource utilisation in a radio cell by engaging the QoS profile renegotiation procedures described above.

$$f(Q) = \frac{Buffer\_Size - Q}{Buffer\_Size - Q_0} \qquad \text{for } Q > Q_0; \quad (1)$$

$$f(Q) = \frac{(a-1)*(Q_0 - Q)}{Q_0} + 1 \qquad \text{for } Q \leq Q_0.$$

Equation 3 Linear Congestion Function

The purpose of the admission control functions is to prevent congestion at the edge as well as within a DiffServ domain, and to allocate resources fairly among traffic classes within the domain. It uses available up-to-date resource information to calculate an Explicit Rate (ER) for each class. The per-flow admission control guarantees fairness among individual flows within the same class. By doing this, the control algorithm must attempt to maintain the queue length at the bottlenecked router along the path of the session close to a target point to avoid router buffer overflow and underflow. The bottlenecked router always operates at the full capacity of the output link without interruption from traffic congestion or buffer starvation. Thus, the most efficient throughput can be achieved. In addition, variations on queue length and consequently queuing delays are reduced. The FICC also attempts to allocate the available bandwidth unbiasedly. Specifically, FICC tries to allocate bandwidth equally among aggregates (DSCPs) with equal status and to distribute the unused bandwidth (left over by constrained aggregates) fairly among the aggregates that can use an additional share. To achieve this objective, FICC oversells bandwidth when the network operates below the target point. Each sender is continuously informed about its current fair share of bandwidth based on the dynamic network traffic conditions by the feedback message.

Parameters

* β : The average ratio
* BUR : Buffer Utilization Ratio
* a : Congestion function parameter

Per Queue Variable

MACR: Mean Allowed Class Rate
DPF: Down Pressure Factor
Q_0: Target Queue Length

Initialization

Q_0 = BUR * BufferSize

At each router's network interface

```
if (receive RM (CCR, ER, DIR = forward))
  if (QueueLength > Q0)
    if (ACR < MACR)
        MACR = MACR + β * (CCR - MACR)
    else
        MACR = MACR + β * (CCR - MACR)
if (receive RM (CCR, ER, DIR = backward))
    if (QueueLength > Q0)
               BufferSize - QueueLength
      DPF = -------------------------
                     BufferSize - Q0

   else
           (a-1)*(Q0 - QueueLength)
       DPF = -----------------------+ 1
                       Q0
      ER = max (MCR, min (ER, DPF * MACR)
```

Algorithm: Fair Intelligence Congestion Control

In order to achieve this objective, it is thus essential to relate appropriately the buffer queue length to the degree of network congestion. We use Mean Allowed Class Rate (MACR) to measure the estimated fair share of the aggregate. This MACR, in turn, is based on the queue length at the router and will determine the explicit rate (ER) of an aggregate (the maximum rate at which the network informs the source of the aggregate that it can support). The “queue control function” is expressed using the Buffer Utilization Ratio (BUR) of an output queue as the target percentage of buffer capacity that should be occupied. When the target is met, the queue occupancy is calculated to be Buffer_Size *BUR. This target occupancy is designed to avoid link underutilization and the remaining buffer capacity Buffer_Size* (1-BUR) is available to absorb packets that might arrive in the queue when the network becomes highly loaded. While BUR defines the target buffer operating point, the corresponding target queue length Q0 (= BUR * Buffer_Size) is often referred to instead of BUR. Since the queue builds up and drains out continuously, the congestion function should be continuous to smoothly regulate the queue fluctuations through the computed ER values. A sophisticated but simple queue control function, the piecewise linear congestion function f(Q) is shown in Equation 3. It would fine-tune the performance of the congestion control algorithm, however, it should also be pointed out that BUR only indicates the desirable long-term operational level. The actual buffer utilization fluctuates around this level.

The actual algorithm has been described as above. The target is to estimate its available bandwidth and advise the traffic sources appropriately. Firstly, the current traffic rate of all aggregates passing through it are estimated and allocated the available bandwidth fairly among its aggregates. The MACR contained in TSPEC is updated with an exponential average factor, which is a true exponential running average of the current load from all aggregates only when the network operates below the target operating point. When the network exceeds the target operating point, FICC does not allow MACR to increase further, meaning that the MACR does not track any ACR value larger than the current MACR when the queue is congested. This rule prevents all those aggregates whose ACRs are already larger than the current MACR to increase their rates further, thereby preventing further loading of the network. Instead, all aggregates have to reduce their rates to the same explicit rate so that the throttling is performed fairly. However, when the network operates below the target operating point, all aggregates are allowed to increase their rate by a factor greater than 1 (also known as overselling), which enables aggregates that are capable of using the available bandwidth to take advantage of it. The explicit rate is calculated as above.

The Resource Discovery (RD) protocol is responsible for maintaining FICC communications among RRM and QoS Agents in clients. Its agent in AP monitors the available resources of wireless links (at the MAC level) and provides feedbacks to FICC, which include MACR and ER. With the supports of the signalling mechanism we discussed in Section 3.2.2, RD information captures the resources availability and is conveyed in a TSPEC frame among RRM/AP and client using the mechanism we proposed. In RRM, information is generated, calculated and updated, then sent to AP for execution of the determination to support clients' demand. With the support of RD and signalling mechanism discussed above, FICC manage the resources among traffic flows based on the algorithms and mechanism discussed above.

To perform an Admission Control, there is an AC policy that must be implemented specifically that all active traffic flows can be served according to the QoS profiles negotiated. Preferably, only a limited number of privileged connections would be allowed simultaneously. And any standard traffic should receive the resources necessary to meet its QoS requirements. Thus, it is be preferable to reject or downgrade a stream activation request rather than to endanger the quality of all sessions. On the other hand, it might be advantageous to displace background or even interactive traffic flows to allow for any additional conversational traffic flow (such as VoIP) to be admitted.

3.2.4 Simulation Analysis

Network simulator ns2 was used to evaluate the framework, where both last hops of DiffServ domain are wireless LAN. Several agents were designed in C++ to implement the schemes. The simulation topology is shown in Figure 6, where FICC is implemented particularly in AP and RRM, and the wireless link is the bottleneck of the network, which has available bandwidth of 1 Mpbs. The bandwidths and propagation delays of wired links are standard. The uniform error model was used to generate errors on the wireless link. The wired links were assumed to be error free. There are four classes of traffic, AF11 (Gold), AF21 (silver), AF31 (Bronze), and Best Effort, which claim 40%, 30%, 20%, and 10% respectively. They are comprise of a mixture of TCP and UDP traffic in which Best Effort has the UDP traffic and other three classes only have the TCP traffic (in Gold, Silver, and Bronze). In the simulation, UDP traffic has a constant bit rate of 1 Mbps so that it causes the bottleneck. The simulation results for FICC and regular DiffServ permit performance in terms of queue length, packet loss, end-to-end delay, throughput, goodput and fairness. The throughput is defined as the number of bits of all the TCP packets transmitted at the source (including RD packets if FICC is used) divided by the duration of the transmission. The goodput is defined as the number of bits of TCP packets transmitted at the source and successfully received at the destination divided by the duration of the transmission.

Violation Prevention and Congestion Control

We firstly look at how the FICC can prevent traffic violation from certain classes, in order to avoid traffic congestion. In Figure 13, with regular DiffServ schemes, silver class traffic does occupy most of bandwidth, exceeding the gold class indeed, as it has shorter RTT of 32ms. Obviously, sliver class violates the traffic condition stipulated for the class it has been assigned (sliver). The damage is that gold class and others cannot be allocated for the bandwidth fairly, as it has been agreed. Furthermore, the gold class actually experiences lower throughput to 0.28Mpbs on average, as shown in Figure 13 and longer delay with larger queue length.

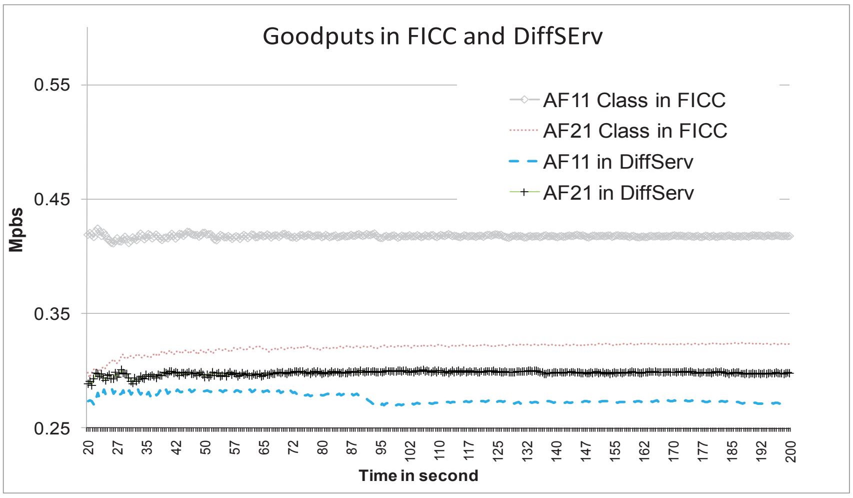


Figure 13 Goodput comparisons between FICC and DiffServ

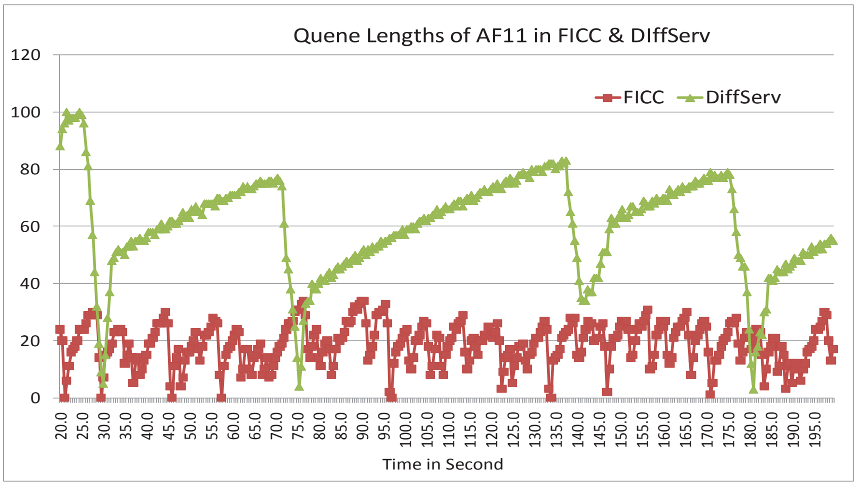


Figure 14 Queue Performance Comparisons between FICC and DiffServ

In FICC, the queue length around the target point is controlled. There is no congestion and no packet loss due to congestion during transmission. The queue length variation and average queue length with FICC are also smaller than those for regular DiffServ. In Figure 14, it has been shown that for the gold class traffic, its queue length has been largely improved from average 60 to 20 second.

Fair Allocation of Resources

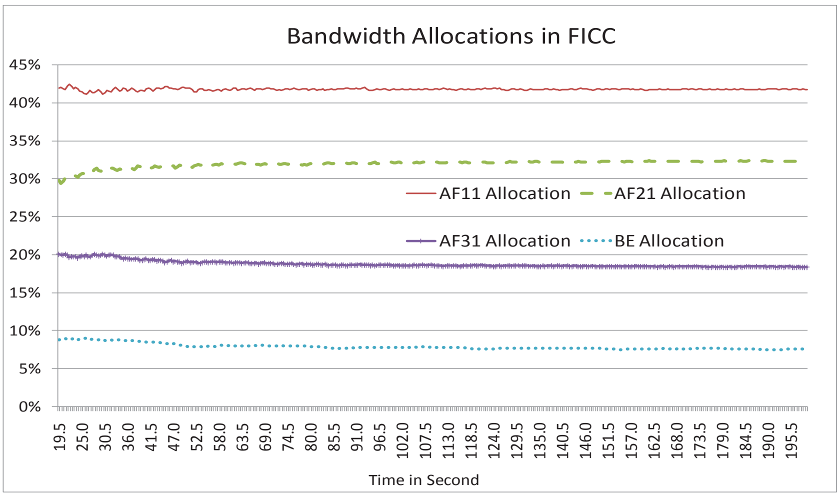


Figure 15 Resources Allocation of Bandwidth in FICC

As seen, DiffServ is unable to prevent any traffic violation, so it is important to investigate how FICC fairly allocate resources to different classes/PHB. FICC will always accurately estimate the fair share for each session at each router and will constantly convey the information to sender by RD and ACK packets. Based on these feedbacks, traffic is policed and sessions with FICC should share roughly the assigned amount of bandwidth. In Figure 15 below, all four classes are shown to be allocated with 10%, 20%, 30% and 40% bandwidth, as agreed in the Admission Control. In particular, in the BE class, the UDP is transmitted at a constant bit rate of 1 Mpbs, which would overload the system, particularly in the wireless section. However, FICC can minimise its impact and fairly distribute the resources. In Figure 16, the queue length comparison between AF11 and BE shows us such fair distribution, as AF11 queue constantly maintain at an average of 20 even when it is already allocated 40% bandwidth. In contrast, the BE queue performance fluctuates due to its large request for resources while it has only been allocated 10% of bandwidth.

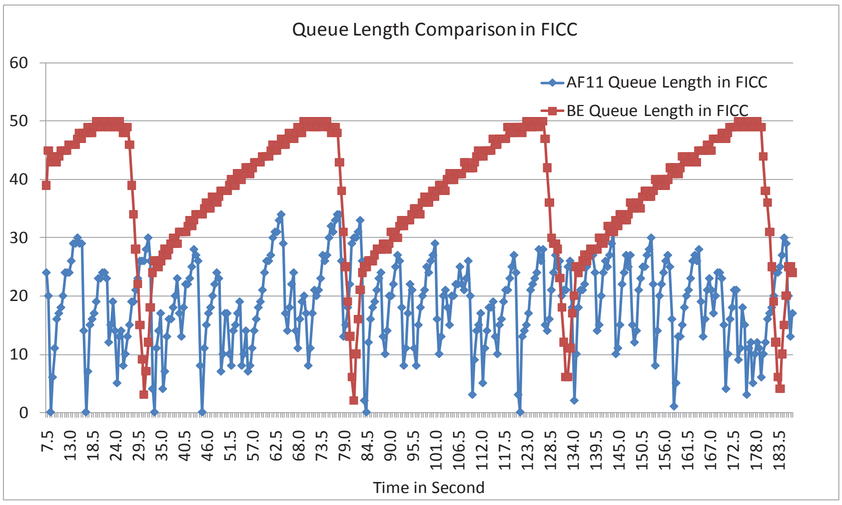


Figure 16 Resources Allocation of Queue Length in FICC

Delay Performance

The end-to-end delays performance comparison for regular DiffServ and FICC schemes are investigated and discussed in this section. The FICC is able to prevent traffic violation, and provide intelligent admission control, thus ensuring the delay performance. In Figure 17, the majorities of delays in FICC are within 0.5 second to 0.7 second, while the delay performance for DiffServ ranges mainly between 1.0 second to 3.7 second. Furthermore, the delay distribution analysis shows that in FICC, 44.11% occurs at 0.6 second, while delays in regular DiffServ is spread across from 0.5 and 3.7 with a maximum of 13.41% contribution at 2.8 second delay. This indicates the jitter, delay variation in the regular DiffServ is much worse than those in FICC. And this amount of delays has a great impact in the delay-sensitive applications, such as Voice over IP.

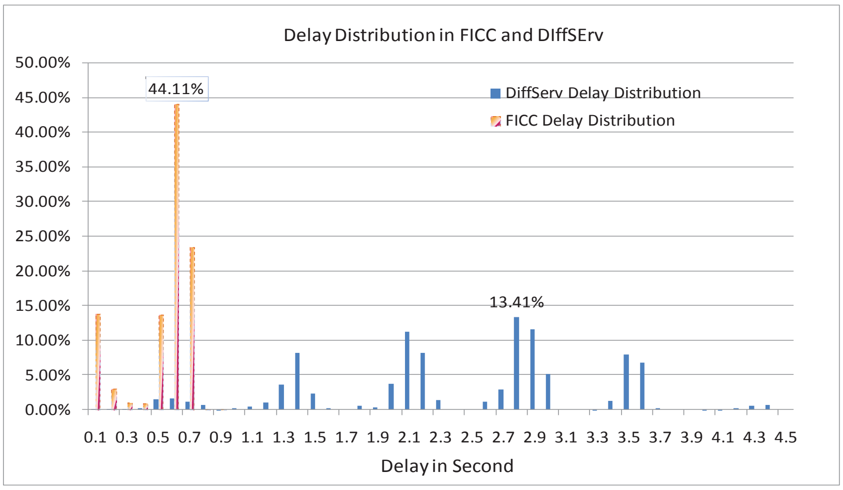


Figure 17 Delay Distributions in FICC and DiffServ

In summary, the Admission Control is implemented in a way that all active traffic flows can be served according to the QoS profiles negotiated. Preferably, there are only a limited number of privileged connections allowed simultaneously. Furthermore, any standard traffic should receive the resources necessary to meet its QoS requirements. Thus, it can be preferable to reject or downgrade a stream activation request rather than to endanger the quality of all sessions. However, it may be advantageous to displace background or even interactive traffic flows to allow for an additional conversational traffic flow (such as VoIP) to be admitted.

3.3 Summary

The chapter has presented a hybrid QoS architecture framework for next generation wireless networks. The framework deploys a DiffServ-enhanced QoS control model for the core/transport part of the end-to-end path, but applies IntServ principles of explicit admission control and resource reservation locally only in the wireless access network domain. In the context of IEEE 802.11 WLAN with 802.11e QoS extensions, a flow signalling mechanism has been designed to meet the proposed explicit admission control and resource reservation locally within the wireless access network. In particular, within the admission control, FICC is adopted as its algorithm in RRM. Simulation results have shown that FICC is able to effectively manage the overloading scenario in the edge section, which is the resources bottleneck of the wireless access domain. Particularly, FICC is able to intelligently predict per-queue fair share for all traffic aggregates, using feedback control to keep the Resources Manager operating at a desirable operating point at all times. It prevents traffic violation from uncontrolled traffic and provides guarantee to those priority traffic pertaining to guarantee bandwidth allocation and specified delay. It also allows overselling of bandwidth when the network is not congested thus making efficient use of the network resources. The results presented in this chapter demonstrate that this proposed hybrid framework with the Fair Intelligent Congestion Control incorporated could be realized for effective end-to-end QoS delivery.

In our proposed 802.11 WLAN QoS architecture, service differentiation plays another important role in keeping providing the promised QoS over wireless communication links in addition to admission control and resources reservation. They are the basic blocks of the whole QoS architecture, and are involved in enforcing the QoS goals for units of service as projected on the control path and QoS framework from the MAC layer. In a wireless environment, these basic blocks will couple tightly with individual link layer specification, where they should provide effective and efficient facilities for the construction of service differentiations because the last hop communication normally operates over link layer only. Importantly the principles of the QoS architecture framework, as described in this chapter, will still hold for other choices of service differentiation mechanisms at the local radio access network level and this will be discussed in the following chapters.

~~ I do not know anyone who has got to the top without hard work. That is the recipe. It will not always get you to the top, but should get you pretty near ~

(Margaret Thatcher)

Chapter 4 Service Differentiation in Contention Medium Access

As parts of standards for Wireless LANs, the IEEE 802.11 Medium Access Control (MAC) protocol [IEEE P802.11 1999] has been defined. Two coordination functions are specified in 802.11: the Distributed Coordination Function (DCF) and the Point Coordination Function (PCF). While the PCF is a polling scheme, providing more explicit control over access of wireless clients to transmission medium; DCF is a basic multiple access technique utilising Carrier Sense Multiple Access with Collision Avoidance (CSMA/CA). It is the mandatory, and most widely used, access scheme in the IEEE 802.11 standard.

DCF is the basic of all contention schemes and designed to provide standard and equal services. In fact, fairness and efficiency in utilising wireless medium, in respect to wireless clients and the traffic generated by them, have always been considered two main features of contention schemes. In previous years, many researches focus on the needs of multimedia application over DCF and other contention schemes. Most of DCF enhancement schemes propose the improved usage of backoff mechanism and interframe space (IFS) to provide service differentiation; some of them even introduce different scheduling schemes to provide more control.

This section gives a description on DCF, which consists of a basic access method and an optional RTS/CTS access method. And then we describe how the enhancement schemes are designed to provide service differentiation for multimedia traffic. While providing service differentiation, the features of fairness are further investigated with these enhancement schemes. Finally, we will point out some observation on fairness in respect to multimedia traffic, which has not been reported in the literature.

4.1 The Basic of Contention Medium Access Schemes

The Distributed Coordination Function (DCF) provides asynchronous type of service, based on the basic access method of the IEEE 802.11 MAC protocol, which is also known as the Carrier Sense Multiple Access with Collision Avoidance (CSMA/CA) protocol. It is the basic of all Contention Medium Access Schemes.

The medium access to the wireless medium is controlled by the use of inter-frame space (IFS) time between the transmissions of frames. In total, three IFS intervals have been specified by 802.11 standards namely Short IFS (SIFS), PCF (point coordination function) IFS (PIFS) and DCF IFS (DIFS). The SIFS is the smallest while the DIFS is the largest. The wireless client may proceed with its transmission if the medium is sensed to be idle for an interval larger than the Distributed Inter Frame Space (DIFS). If the medium is busy, the client defers until after a DIFS is detected and then generate a random back-off interval before transmitting. This backoff interval is then used to initialise the backoff timer. The back-off timer counter is decreased as long as the channel is sensed idle, frozen when the channel is sensed busy, and is resumed when the channel is sensed idle again for more than a DIFS. The decrement period is referred to as the slot-time, which corresponds to the maximum round-trip delay within the BSS and, hence, depends on the maximum BSS coverage. A client can initiate a transmission when the back-off timer reaches zero. The back-off interval (0, w-1) is uniformly distributed between zero and a maximum called Contention Window (CW), which is an integer with the range determined by the PHY characteristics CWmin and CWmax. At the first transmission attempt, w is set equal to the value of CWmin. After each unsuccessful transmission, w is doubled, up to value CWmax = 2m CWmin, where m is the maximum backoff stage.

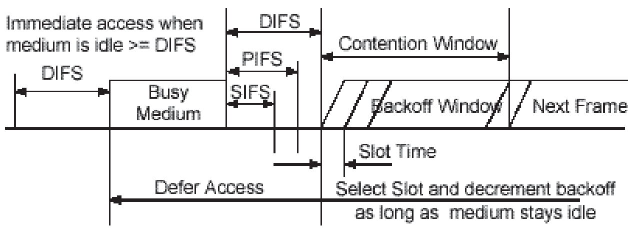


Figure 18 DCF Access Procedure

As soon as the backoff timer expires, the client is authorised to access the medium. Obviously, a collision occurs if two or more clients start transmission simultaneously. Unlike wired networks (e.g., with CSMA/CD), in a wireless environment, collision detection is not possible. Hence, upon received a packet correctly, the destination client waits for a SIFS interval immediately following the reception of the data frame and transmits a MAC ACK back to the source client, indicating that the data frame has been received correctly. Since the SIFS is, by definition, less than the DIFS, the receiving client does need to sense the medium before transmitting the acknowledgement. The procedure is shown in Figure 18 [IEEE P802.11 1999].

In case the source client does not receive an ACK, the data frame is assumed to be lost and the source client will schedule retransmission and enter the backoff process firstly again. However, to reduce the probability of collisions, after each unsuccessful transmission attempt, the Contention Window is doubled until a pre-defined maximum (CWmax) is reached. After a (successful or unsuccessful) frame transmission, if the client still has frames queued for transmissions; it must execute a new backoff process. When the data frame is transmitted, all the other clients hearing the data frame adjust their Network Allocation Vector (NAV), which is used for virtual CS at the MAC layer based on the duration field value in the data frame, which includes the SIFS and the ACK following the data frame.

In 802.11, DCF also provides an optional way of transmitting data frames that involve transmission of special short RTS and CTS frames prior to the transmission of actual data frame. When the destination receives the RTS frame, it will transmit a CTS frame after SIFS interval immediately following the reception of the RTS frame. The source client is allowed to transmit its packet only if it receives the CTS correctly. If a collision occurs with RTS frames, less bandwidth is wasted when compared with the situation where larger data frames are in collision. This access method is proposed to deal with hidden client problem, which arises when a client is able to successfully receive frames from two different transmitters but the two transmitters cannot receive signals from each other (hidden from each other). In this case a transmitter may sense the medium as being idle even if the other one is transmitting. This results in a collision at the receiving client.

4.2 Analysis in Service Differentiation

Enhancements over DCF have been studied for a long history since the introduction of DCF. These researches focus on the design of enhancement to meet the needs of service differentiation to service multimedia application over contention schemes. Most of DCF enhancement schemes propose the improved usage of backoff mechanism and interframe space (IFS) to provide service differentiation; some of them even introduce different scheduling schemes to provide more control. Among these forces, 802.11e plays an important role to define extension to enhance the access mechanisms of IEEE 802.11 to provide service differentiation. They have all been well studied and reported in the literature. We will give details of a few dominant enhanced schemes as below.

4.2.1 The Existing Approaches

Contention Windows and Interframe Space Schemes

As discussed above, setting the same values of CWmax and CWmin for each client, means that they have the same opportunities to access the medium under the DCF basic access method. And adaptation of CWmin and CWmax values has great impacts on the transmission probability. Therefore, it introduces priority to client simply by adopting different contention window in each client or services. In [Malone et al 2007][Alonso-Zárate et al 2012][Banchs and Perez 2002] and [Banchs et al. 2001], Banchs et al. proposed an extended DCF with a modified algorithm for the computation of the contention window. It supports real-time traffic and elastic traffic, which are also called Assured Rate Service and best-effort service in [Banchs and Perez 2002]. In the scheme of [Tinnirello et al 2010], similarly, high priority service always receives a higher throughput than a lower priority one; while delay of high priority service can be achieved sufficiently low. Virtual MAC is proposed in [Veres et al. 2001] to extend service differentiation for voice, video and best-effort TCP over the wireless hop. The core mechanism for providing service differentiation for these classes is to set different value of CW for each class in order to produce different backoff intervals.

Similarly, in [Kim et al. 2001][Vardakas et al 2008], DiffServ-like traffic services system are proposed and supported by legacy DCF and modified DCF with shorter contention window (DCF/SC). Kim et al. take further step to divide period II and I. In period I, only premium services are supported by DCF/SC while all services are allowed in period II. In period II, premium services and assured services have higher access priority than best-effort services, as premium services and assured services are supported by DCF/SC while best-effort services operate with DCF. The results show premium services and assured services perform much better than while best-effort services as expected.

Aad and Castelluccia present three service differentiation schemes for IEEE 802.11 in [Aad and Castelluccia 2001]. The first one is based on scaling the contention window according to the priority of each flow or user. The second one assigns different inter-frame spacings to different users. And the last one adopts different maximum frame lengths for different users, high priority traffic have larger frame limit and can be transmitted via fragmentation. TCP and UDP flows are simulated and their performance are analysed for each scheme. [Senthilkumar and Krishnan 2010] also carries similar studies.

Apart from using contention windows as above, another alternative solution is to use Interframe Space (IFS) to provide service differentiation. As described in 802.11 Specification, ACK packets get higher priority than RTS packets, simply by waiting SIFS, which is shorter than DIFS for RTS. The same idea is used to introduce priority for different data frames. [Deng and Chang 1999] exactly uses the interframe space (IFS) and the backoff mechanism to provide differentiation. If two clients use different IFS for data transmission, a client with shorter IFS will get higher priority than a client with longer IFS. Therefore, by using these two different interframe spaces, traffic can be differentiated and classified into two classes.

However, the more IFS are defined, the more overhead are heavily introduced. To further extend the number of available classes, the backoff mechanism is then used to differentiate among clients beside IFS. This is done by designing the backoff algorithm such that it generates backoff intervals in different intervals, depending on the priority of the client. The backoff algorithms chosen guarantees client that uses the low priority backoff algorithm always generates longer backoff intervals than clients with higher priority. In [Deng and Chang 1999], total four traffic classes are designed with combination usage of IFS and backoff algorithm.

Scheduling in Backoff Mechanism

In [Vaiday et al. 2000][Li et al 2009], they combine backoff mechanism with a Distributed Fair Scheduling (DFS) algorithm. It is based on the fair queuing mechanism known as Self-Clocked Fair Queuing (SCFQ) [Golestani 1993], and uses the backoff mechanism to determine which client should send first. Before transmitting a frame, the backoff process is always initiated, even if no previous frame has been transmitted. If the backoff interval is calculated longer, the lower the weight of the sending client is; therefore differentiation will be achieved. Furthermore, fairness is achieved by using the size of the packet to be sent in the calculation of the backoff interval. This means that it causes larger packets to get longer backoff intervals than small packet, allowing a client with small packets to send more often so that the same amount of data is sent. If a collision occurs, a new backoff interval is calculated using the backoff algorithm with the given CW as defined in the IEEE 802.11 standard. However, the backoff intervals of DFS are inversely proportional to the weight of a flow. The backoff interval is unnecessarily long if a client constantly has queued packets (also called backlogged) and the weights of its flow are small. Further researches are still being carried on to compress such backoff interval.

Other Alternatives

[Sobrinho and Krishnakumar 1996 & 1999] proposed a scheme called Blackburst. [Ruscelli et al 2012] propose more advanced approach based on Blackburst. It aims to minimise the delay for real time traffic, and it is somewhat different from the other schemes since it imposes certain requirements on the traffic to be prioritised. Firstly, all high priority clients access the medium with a shorter IFS comparing to low priority clients. Furthermore, Blackburst provides service differentiation among all high priority clients. It requires the ability to jam the wireless medium for a period of time. When a Blackburst client wants to send a frame, it senses the medium to see if it has been idle for a PIFS and then sends its frame. On the other hand, if the medium is found busy, the client waits until it has been idle for a PIFS and then enters a black burst contention period. The client now sends a so-called black burst by jamming the channel for a period of time. The length of the black burst is determined by the time the client has been waiting to access the medium, and is calculated as a number of black slots. After transmitting the black burst, the client listens to the medium for a short period of time (less than a black slot) to see if some other client is sending a longer black burst. That would imply that the other client has waited longer and thus should access the medium first. If the medium is idle, the client will send its frame, otherwise it will wait until the medium becomes idle again and enter another black burst contention period. In such way, it can be guaranteed that each black burst contention period will yield a unique winner.

4.2.2 The EDCF of IEEE 802.11e

As it can be seen that all these DCF-based proposals above have implied change in the basic medium access or modified the standard protocol behaviour of DCF in order to provide service differentiation with DCF. Therefore, IEEE 802.11 working group forms 802.11 Task Group E to conclude such schemes and define Enhanced DCF (EDCF) as a standard protocol of 802.11 to meet the QoS requirements by the method of contention. The EDCF is based on DCF and is designed to provide differentiated, distributed channel accesses for frames with 8 different priorities (from 0 to 7). This is an extension of the basic DCF access mechanism in the original standard. Since devices complying with the old standard are widely deployed, great care was taken to ensure that EDCF should be inter-operable with the old DCF.

QoS is realized with the introduction of Traffic Categories (TCs). In the context of the 802.11e, the priority value is called Traffic Category Identification (TCID). A single ESTA may implement up to eight TCs realized as virtual STAs inside an ESTA, with a parameter set that affects its behaviour under EDCF rules. Each frame from the higher layer arrives at the MAC along with a specific priority value and such value would be translated into one of TCs, carried in the MAC frame header if such frame is under EDCF transmission. By modifying the minimum contention window (CWmin) and the interframe space used for data transmissions, EDCF mechanism can provide service differentiation for 8 different traffic classes. As discussed above, choosing a smaller default contention window for a client will cause that client to generate shorter backoff intervals, thus gaining priority over a client with a larger CWmin which generates longer backoff intervals.

To be able to further differentiate between clients using the same contention window, different interframe spaces are used by different traffic classes. Instead of waiting a DIFS before trying to access the medium, or starting to decrement the backoff timer as in ordinary DCF, Arbitration Interframe Space (AIFS) is used. Each traffic class uses its own AIFS, which equals a DIFS plus a number of time slots (possibly zero). This means that traffic using large AIFS (many “extra” time slots) will have lower priority than traffic using a small AIFS, since they will wait longer before trying to access the medium or starting to decrement the backoff timer.

Therefore, an EDCF packet, which belongs to a TC, uses AIFS [TC], CWmin[TC], and CWmax[AC] instead of DIFS, CWmin, and CWmax, of the DCF, respectively, for the transmission contention. The values of AIFS [TC], CWmin [TC], and CWmax[TC] are determined according to its TC, and announced by the AP via beacon frames, which are transmitted periodically. The AP can adapt these parameters dynamically depending on network conditions. Each TC maintains a single queue and behaves as a single enhanced DCF contending entity, using its own AIFS and maintains its own Backoff Counter (BC).

During CP, each TC within the clients contends for a TXOP and starts an individual back-off after detecting the channel being idle for its AIFS; After waiting for AIFS each back-off sets BC to a random number from the CW. The back-off reduces the counter by one for every slot time. When the counter reaches zero, the contention is won. If the medium is sensed busy before the counter reaches zero, the back-off must wait for the medium being idle for AIFS again before continuing to reduce the counter. After any unsuccessful transmission attempt the next back-off is performed with a larger minimum size of the CW, increased by a Persistence factor (PF) to reduce the probability of a new collision on the medium. When there is more than one TC finishing the backoff at the same time, the collision is handled by a scheduler in the client, which gives the TXOP to the highest TC. That is, the highest priority frame among the colliding frames is chosen and transmitted, and the others perform a backoff with increased CW values.

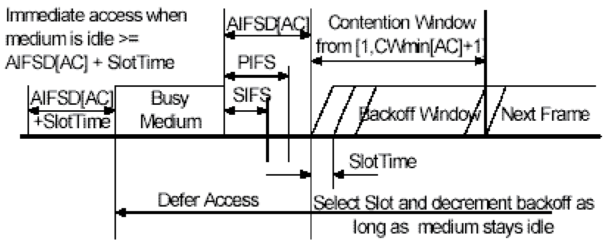


Figure 19 EDCF Access Procedure

In fact, EDCF is a contribution from the previous varied approaches discussed in Section 4.2.1. It provides standard way to modify DCF to offer service differentiation in CP. Performance analysis of such mechanisms is well studied not only in such previous proposal, but also in EDCF studies [Choi et al. 2003], [Lindgren et al. 2003] and [Grilo and Nunes 2002]. They all demonstrate that EDCF allows QoS service differentiation among different priority traffic.

However, EDCF can provide service differentiation at the cost of larger overhead, which comes from the longer backoff procedure and usage of different IFS for each TC. In [Choi et al. 2003b], performance comparison between DCF and EDCF is reported. In a condition of three TC creations for voice, video and data, EDCF over-performs than DCF in QoS guarantee of video and voice, but the total throughput drops approximately 15.6%. In fact, the more TC is used, the more overhead EDCF would produce as average backoff time experience longer and average IFS is also larger. In [Lindgren et al. 2003], medium utilisation of EDCF and DCF is also reported. On the other hand, [Lindgren et al.

2003] points out that EDCF also suffers a large number of collisions especially at the high loads. Collision not only further cuts off the medium utilisation, but also makes EDCF fail to provide hard QoS guarantee due to the date loss and delay caused by re-transmission.

4.2.3 Summary

As discussed above, the main constrains of all contention schemes are degree of QoS guarantee and the trade-off of overhead and the number of services it can provide. Also, all data transmission is based on contention. This may cause possibility of collision and disturbance by low priority traffic as the one of unavoidable natures of CSMA. Therefore, it cannot really provide hard QoS guarantee.

On the other hand, only limited numbers of service classes can be implemented due to the overhead impacts. If IFS and CW are set more values for more traffic classes, the overhead is significantly increased and heavily affects the efficiency of the contention schemes. Even the combination of IFS and CW are used, it does not improve too much. However, the contention schemes are traditionally efficient transmission methods. The enhancement of service differentiation over a contention scheme can provide priority-based QoS for a certain numbers of traffic classes, which do not really need hard QoS guarantee. And this is assumed that the admission control is properly applied. Therefore, we propose to remain contention schemes as options for service differentiation in our novel hybrid architecture. Such design not only provides priority-based QoS among all non-real-time application, but also achieves wireless network efficiency, which is also one of the crucial concerns on wireless communication.

The above analysis convinces that contention medium access is fundamental part of our novel architecture providing priority-based QoS among non-real-time applications while also achieving wireless network efficiency. However, fairness and efficiency in respect to wireless clients and the traffic generated by them have always been considered two main features of all contention schemes. Even though service differentiation provided for multimedia traffic as discussed above, the enhanced contention schemes should still provide equal services to the same traffic class in accordance with the foundation of their access methods - contention. We further examine the fairness issue with respect to multimedia traffic with assumption that the multimedia traffic is assigned the same traffic class in the service differentiation in the following sections.

4.3 Fairness Nature and Service Differentiation

Transmission of a packet involves contention over the joint neighbourhoods of the sender and the receiver, and the level of contention for the shared wireless channel in a geographical region is spatially dependent on the number of contending clients in the region. Furthermore, DCF is not a centralised control and no client is guaranteed to have accurate knowledge of the contention even in its own neighbourhood. These are fundamentally different from wireline and cellular channel models, wherein all flows perceive the same contention. Therefore, fairness study in a contention-based WLAN is a challenge, which has been addressed in a few existing works.

Many papers have studied the performance of DCF (e.g. [Alonso-Zárate et al 2012][Malone et al 2007][Crow 1997] [Bianchi, G. 2000] [Prasad 1999]). Issues of fairness have been investigated for bi-directional applications among different wireless clients. Among issues of interest in these studies are location-dependent contention, trade-off between optimising channel utilisation and achieving fairness, decentralised control and rate compatibility (also called Link Adaptation). Fairness has also been mentioned in reports on performance evaluation of IEEE 802.11 WLANs [Chhaya and Gupta 1995], [Chhaya and Gupta 1996] and [Crow 1997]. In [Nandagopal et al. 2000], an attempt is made to achieve fairness by deriving an appropriate contention resolution algorithm. In [Pagtzis et al. 2001], the authors concentrate on experimental evaluation of 802.11b and discuss fairness issues with regards to rate compatibility, namely the effect of the discrete set of supported signalling rates in 802.11b on the fairness in bandwidth allocation. Work reported in [Qiao and Shin 2002] proposes a priority-based MAC scheme (modified DCF) to achieve weighted fairness among multiple traffic flows while maximising the wireless channel utilisation. In [Chhaya and Gupta 1995], the authors study throughput and fairness properties of the IEEE802.11 DCF, specifically the impact of hidden terminals and the capture effect. In [Mangold et al. 2002b], issues of fairness are also addressed for the scenario of overlapping Basic Service Sets (BSSs), where geographically co-located Wireless LANs share the radio channel. However, none of the papers mentioned above have analysed fairness with respect to multimedia streams especially in a highly loaded and/or overloaded system.

4.3.1 Back-off Behaviour and Transmission Probability

The slotted binary exponential backoff mechanism is adopted in the classic DCF. Random backoff interval is selected in number of time slots; this random number is drawn from a uniform distribution over the interval [0, CW-1], where CW is the contention window size and its initial value is a CWmin. CW is doubled if there is an unsuccessful transmission and can reach a CWmax. If there is a successful transmission, CW is then reset to a CWmin before the random backoff interval is selected again.

Every time the wireless medium is sensed to be idle for at least a DIFS time, the client decrements its backoff counter. If the counter has not reached zero and the medium becomes busy again, the clients freezes its counter. Once the counter finally reaches zero, the client begins its queuing transmission. After successful transmission, this client has to waits for DIFS time and selects another random backoff interval before attempting to transmit the next frame. Another client, which has been waiting, has smaller backoff interval and its counter reaches zero earlier, it then can take up next opportunity for its transmission.

$$p = \frac{2(1-2\rho)}{(1-2\rho)(\mathrm{CW}_{\min}+1)+\rho\,\mathrm{CW}_{\min}(1-(2\rho)^{\mathrm{m}})}$$

Equation 4 Probability of Transmission

In [Bianchi, G. 2000], Bianchi, G. investigates the packets transmission probability of DCF basic access method in the assumption of ideal channel conditions (i.e. no hidden terminals). Let us consider a fixed number *n* of contending clients, $u, v \in f_i$. Each of them is under basic access method and always has packets to transmit. ρ is referred to as conditional collision probability, a probability of a collision seen by a packet being transmitted on the channel. Let m, “maximum backoff stage”, be the value such that $CW_{max} = 2^m CW_{min}$. According to the computation in [Bianchi, G. 2000], we can express the probability p that a client transmits in a randomly chosen slot time in Equation below. The probability p_s that a transmission is successful and is from client u can be calculated as: $p_{s,u} = p_u \cdot \prod_{i \neq u}^{n} (1 - p_i)$. If client u has the same successful probability p_s as client v, we must have

$$p_{s,u} = p_{s,v} \Leftrightarrow p_u(1-p_v) = p_v(1-p_u) \Leftrightarrow p_v = p_u$$

Equation 5 Probability of Successful Transmission

As seen in Equation 4, *p* depends on CWmin, CWmax and ρ. While we assume that each packet collides with constant and independent probability ρ, setting the same values of CWmax and CWmin for each client, means that they have the same opportunities to access the medium under the DCF basic access method. Therefore, if the network is busy and there are several clients with a frame ready to transmit, the time that each of the clients contending for access waits until transmission of a frame is, on average (over a number of transmissions), equal. This remains true whether the transmission is successful or not, as shown in Equation 5. This is where the notion of fairness has been derived from. In other words, when clients have frames in transmit buffers ready for transmission; each of these clients is given, on average, the same level of transmission opportunity. The numbers of frames sent by these clients during a certain period of time are (statistically) equal only if all involved clients have packets in transmit queues during the period of time in question.

4.3.2 Fairness in Transmission Opportunities

If all packets sent as described above are of the same size and have statistically comparable inter-arrival patterns, all clients can be expected to enjoy an equal allocation of bandwidth. This is often understood as equal allocation of bandwidth to all clients or, in some other cases, allocation of bandwidth to clients in proportion to the traffic they generate under lightly loaded conditions, and fair (equal) degradation of bandwidth available to each client under overload conditions. Simple analysis of the principles on which the 802.11 DCF is based reveals that neither of these popular beliefs is correct, and that the fairness of DCF should be understood as statistically equal opportunity to commence transmission of a frame among all clients that have a frame ready to transmit when the transmission medium has become available. As a result, the relative traffic patterns of different traffic streams have a profound impact on the relative allocation of bandwidth to clients and traffic streams. In addition, under overload conditions, transmit buffers of some clients may become full, further distorting the allocation of transmission opportunities (and bandwidth) to different traffic streams.

However, packets generated by multimedia applications may have different arrival patterns and sizes. Under such conditions, it is important to know if proportional fairness in bandwidth allocation can still be achieved among different multimedia streams sharing the same client and among clients with different traffic loads. Furthermore, for an overloaded system, the key question is whether all clients and/or application streams would suffer fairly the degradation of allocated resources.

Fairness is normally tested among two or more clients in a BSS, with identical offered load, and within the capacity boundaries of the MAC protocol. Under such conditions, fair (proportional) bandwidth allocation is expected. However, if clients transmit traffic streams generated by different applications, with more than one application serviced by a given client at a time, the packet generation rates and packet sizes differ from client to client and from one traffic stream to another. Under such conditions, the popular expectation of fair allocation of bandwidth to different traffic streams needs to be modified especially if it is possible for the entire system or selected clients to become overloaded. In wireless systems featuring QoS control, admission control would normally be implemented for multimedia (e.g. voice and video) sessions, thus preventing overload, as discussed in Chapter 3. However, the best effort traffic is normally not fully controlled in order to achieve good utilisation of available bandwidth, and therefore it will be possible for the network to experience high load conditions.

Apart from the emergence of multimedia application and multiple streams serviced by a single client, we further consider the scenario of overloading. When the system using a contention scheme is underloaded, clients and applications can always get the resources they require. However, when the system is overloaded, the key question is whether all clients or applications would suffer fairly the degradation in resources available. Observations resulting from simple experiments lead to a conclusion that under overload different traffic streams may suffer widely different levels of degradation in allocated resources. As an example, we can consider a client transmitting multiple streams of traffic. Let us assume that one of these is a stream of frequent but small frames (e.g. as in the case of Voice-over-IP traffic). We observe that such traffic streams, characterised by frequent requests for access to the transmission medium, tend to dominate the use of available radio resources. This may turn into nearly exclusive occupation of radio resources under severe overload; if such streams compete for available buffer space in overloaded transmit buffers against streams producing infrequent but large frames of data.

Effects such as the domination of some streams over others, as described above, can also severely impact on the symmetry of resources allocated to bi-directional streams. Also, in a typical infrastructure network, the Access Point is expected to transmit relatively high volume of downlink traffic (and many separate streams) to other wireless clients. Some of these streams may be severely disadvantaged due to the effects as mentioned above.

In the following sections, we present some observations related to the issues stated here. The simulation would further examine if proportional fairness in bandwidth allocation can still be achieved among different multimedia streams sharing the same client, as well as among clients with different traffic loads. Furthermore, it would answer the question whether all clients and/or application streams would suffer fairly the degradation of allocated resources for an overloaded system. Based on these results, we also discuss its effect on asymmetry in bi-directional multimedia sessions.

4.4 Performance Analysis on Fairness

4.4.1 Network Topology and Traffic Configuration

This section provides details about the simulation configuration. In our simulation environment, the simulation ignores the possible coexistence technologies may include: 802.15 (Bluetooth), 802.16a, 2.4 GHz Cordless Phone, 5.0 GHz Cordless Phone, 2.4 GHz Video Transmitter. 5.0 GHz non-OFDM Video Transmitter and 2.4 GHz Microwave oven. Only WLAN system is simulated by an OPNET model. Simulations are run mostly short period of time, as there would be no further change in any longer period as long as configuration remains. And Traffic re-acts in the unit of milli-second. Due to the large amount of scenarios, validation is done by control/incremental analysis method and network traces/markers and breaker are built to pause simulation in the middle of process and pull out data for analysis. The model is made up of process and state. Therefore, each state/process results were examined and validated. Therefore, traffic monitoring on packets is the main key step taken in the simulation plan to ensure a satisfactory level of reliability in the results.

The model includes traffic sources, the 802.11 MAC protocol, and characteristics of the 802.11 physical layer and several assumptions made for simulation experiments.

Simulation Layouts

Simulations cover the time duration of 30 seconds, including period I (the first 10 seconds) and II (the remaining 20 seconds). Data Client (sending the bulk data) becomes active only from the start of period II. This has been designed to allow observation of the effect of activation of high volume data stream on other streams. As mentioned earlier, we simulate only a single BSS, which is only equipped with the basic contention scheme, DCF. The study focuses on the fairness issue with respect to multimedia traffic, which are all assigned with the same class under the simulated DCF mechanism.

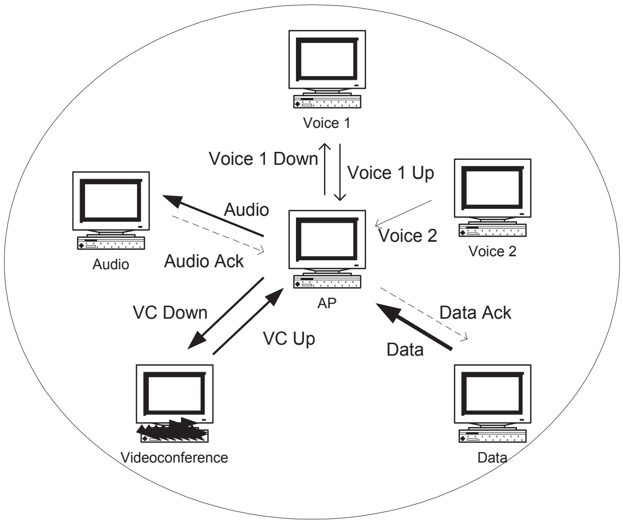


Figure 20 Simulated Network Topology

As shown in Figure 20, the BSS consists of one Access Point (AP) and the following clients:

- Voice 1 Client: two-way interactive voice, active from the start of simulation period I.
- Voice 2 Client: one-way voice (to Access Point), active from the start of simulation period I.
- Audio Client: audio download from Access Point, with Ack feedback returned up-link to the Access Point, active from the start of simulation period I.
- Videoconference (VC) Client: two-way videoconference application, active from the start of simulation period I.
- Data Client: Bulk data transfer up-link to Access Point, with Ack returned down-link; active from the 10th second of simulation (i.e. the start of period II).
- Access Point: In the simulation period I, it sends voice, audio and VC while also receiving the VC, voice and audio acknowledgments. In the simulation period II, it also receives the bulk data and sends bulk data acknowledgments.

Assumptions and Platform

The effect of 'hidden terminals' is not accounted for in the simulations, i.e. all nodes are located within hearing distance of every other node in the BSS. The RTS/CTS scheme is not used in our simulations.

We also assume that the transmit buffer available in each station is finite; when the transmission buffer fills; all newly generated packets are dropped. Each station maintains only one buffer, common to all application streams serviced by the station.

Multimedia Traffic Model

Voice has traditionally been the primary driver of the growth in cellular wireless networks. However, voice and other multimedia applications are also becoming the dominating type of traffic in wireless LANs. Unlike voice-only systems (e.g. 2nd generation cellular networks), wireless LANs must be able to handle heterogeneous traffic with varying characteristics and quality of service (QoS) requirements. Some of the traffic types may be delay-sensitive (e.g. voice, video conferencing) while others may be less sensitive to delay but more sensitive to errors (e.g. email, file transfers). In reality, Internet traffic is also generated by many kinds of traffic sources such as interactive voice and video, audio and video download, and interactive or bulk data. In our simulation, four different traffic sources are modelled: interactive voice, audio download, video-conferencing and bulk data transfer. They are assumed here to represent a range of multimedia applications with different inter-arrival rates and packet sizes. Traffic generated by each user is simulated by independent ON/OFF sources, as in [Kim et al. 2001] [Kim et al. 2003], with parameters listed below.

Application	*Offered Load (Kbps)*	*On/Off Time*		*Inter-arrival*	*Pkt Size*	*Max Delay (Ms)*
		On (ms)	*Off (ms)*	*Time (ms)*	*(Bytes)*	
VoIP	*36.8*	*Always on*		*20*	*92*	*30*
Audio	*52.0*	*Always on*		*Uni (8.3,250)*	*815*	*200*
Audio Ack	*2.5*	*Always on*		*Uni (8.3,250)*	*40*	*200*
Video	*1410.0*	*12*	*88*	*1*	*100*	*100*
Data	*3680.0*	*Always on*		*Exp (5)*	*2300*	
Data Ack	*102.0*	*Always on*		*Exp (5)*	*64*	

Table 3 Parameters of Multimedia Traffic Sources

As shown in Table 3, voice frames have fixed size of 92 bytes and arrive every 20ms (in each stream). Audio frames are much larger and arrive less frequently with a uniform distribution. The packet arrival process for a videoconference source consists of an ON-state and a silent OFF-state. In the ON-state, frames have constant inter-arrival interval and are large in size. The bulk data source is the most aggressive; it has an exponential distribution of inter-arrival times with high average frequency of arrivals and very large packet size. Therefore, it may consume large proportion of bandwidth. Another stream is also simulated, acknowledgements to the bulk data frames (Data Ack); it follows the same exponential distribution of inter-arrival times with the mean of 5ms and features very small frames of 64 bytes. The parameters are defined as follows:

- Packet size: MSDU size at the top of the MAC
- Maximum Delay: Maximum transport delay at the top of the MAC – i.e. between matching MA-UNITDATA.request and indication.
- ACK: Indicates the network-layer protocol running between the data source and the MAC. It takes one of two values: TCP or UDP. These are intended to represent a generic acknowledged and generic unacknowledged network-layer protocol.

Attributes	*Value*
Data Rate (Mbps)	*11*
Fragmentation Threshold (bytes)	*2304*
Physical Characteristics	*DSSS Long*
Buffer Size (bits)	*2048000*
DSSS preamble (bits)	*144*
DSSS header (bits)	*48*
slot_time (us)	*20*
SIFS_time (us)	*10*
Cw_min	*31*
Cw_max	*1023*

Table 4 Simulation Attributes

Applications identified as being carried by UDP are assumed to generate MSDUs at a fixed rate, as identified in the "Offered load" column. Traffic carried by TCP assume to be served on a best-effort basis, and applications using TCP are assumed to generate MSDUs at rates up to the value given in the "Offered load" column. Being an acknowledged protocol with a constrained window size, TCP responds to congestion in the BSS by reducing application throughput without losing MSDUs. This effect is reflected in simulation results by reporting achieved throughput for applications using TCP with possible large delay.

Test Parameters

The values of simulation attributes used in our experiments are provided in Table 4. In the table, the value of Physical Characteristics (DSSS Long) stands for Direct Sequence Spread Spectrum (DSSS) with long preamble; this scheme adopts 144 bits for the physical layer preamble and 48 bits for the physical layer header. The buffer size specifies the maximum length of the higher layer data arrival buffer. Each station maintains one buffer; once the buffer limit is reached, data arriving from the higher layer will be discarded until some packets are removed from the buffer.

4.4.2 Simulation Observation of Fairness Issues with Comparison Between Overloaded and underloaded Scenarios

Peak Throughput with Variable Packet Sizes

At first, we look at the peak throughput of the 802.11b wireless LAN in order to gain a clear understanding of the impact of system overload. As shown in Figure 21 below, during period I (the first 10 seconds) all clients and applications receive their required bandwidth (total of 2.982Mpbs). No data is dropped during this period and we consider the network is lightly loaded. During period II, the network becomes overloaded with the bulk data. The total bandwidth requirement reaches 6.764Mpbs while the system can only offer a maximum of approximately 5.6Mpbs. The 1.1Mpbs of data dropped is mainly due to the overflow of buffer(s) at one or more clients (note the delay in dropping as compared with the start of the bulk data transfer).

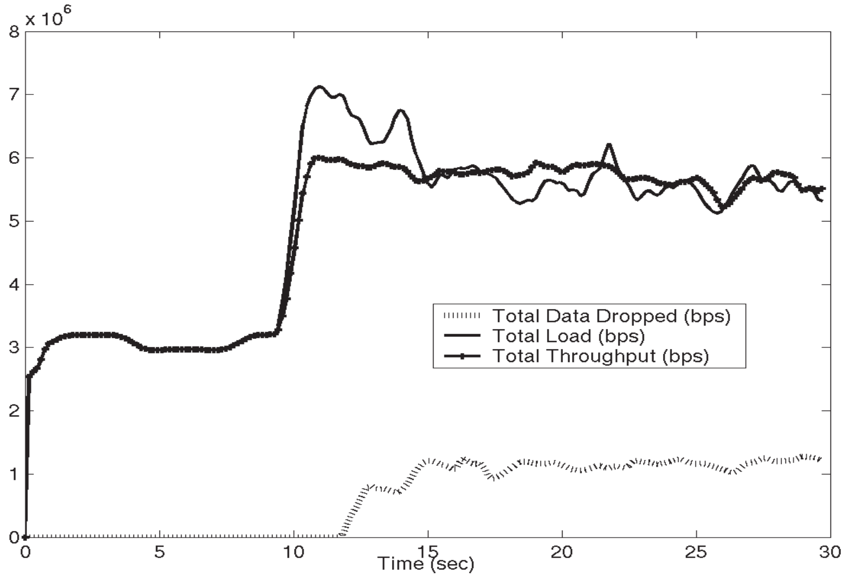


Figure 21 **Total Throughputs, Load and Data Dropped**

Fairness in Respect to Different Clients

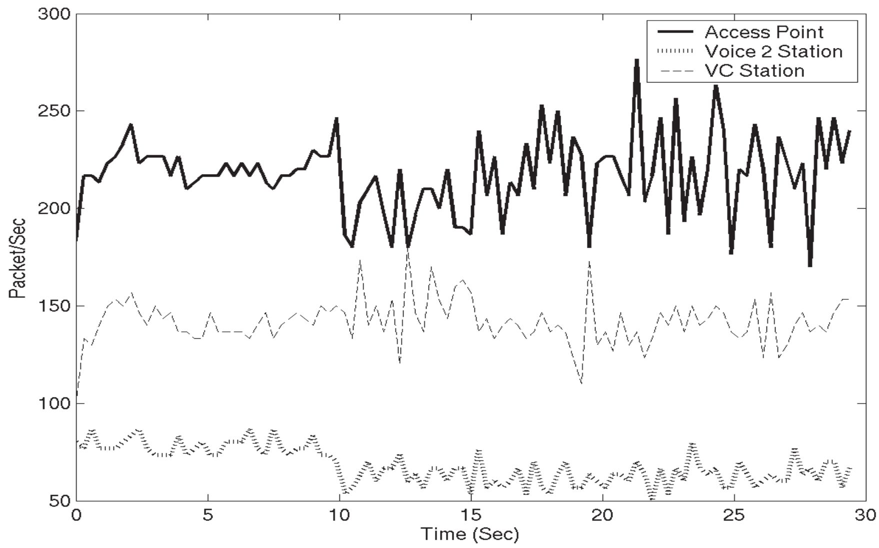


Figure 22 **Rate of Packets Sent by Clients**

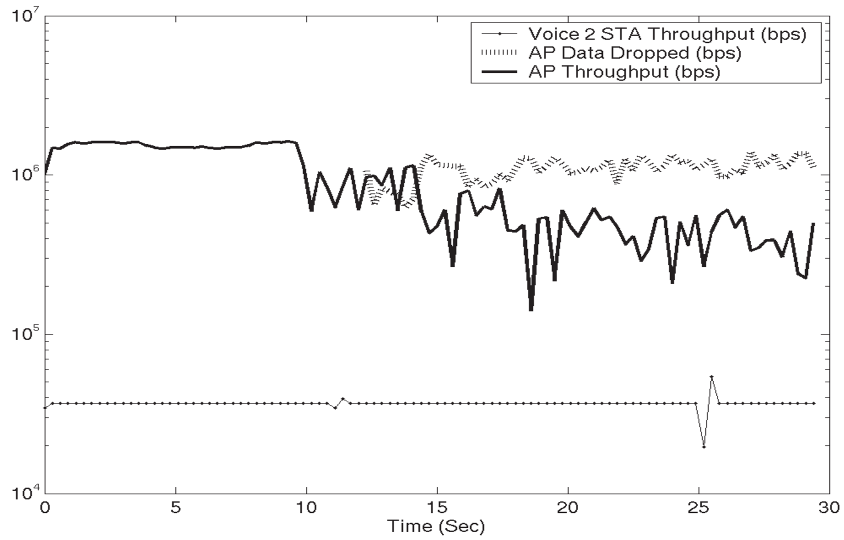


Figure 23 **Fairness between Voice 2 Client and AP**

In this section, we investigate the sharing of resources among different clients. As discussed in Section 4.3, each client that has a frame to transmit will enjoy the same level of opportunity to transfer when the medium becomes free. In an overloaded system, there must be buffer overflow in some clients, even when the bandwidth requirements of other clients are satisfied. Such clients have, on average, more frames to send than the transmission opportunities they receive allow for. In other words, the average frame arrival rate in these clients is greater than the average frame transmission rate. Such overloaded clients not only fully consume their 'fair' share of transmission opportunities, but also 'excess' opportunities left over by other, less busy clients (i.e. clients that may, at times, have their transmit buffer empty due to less frequent arrivals of frames from traffic sources).

We first look at simulation results in terms of rates of data packets sent. This may be interpreted as the number of transmission opportunities taken (used). In Figure 22, we show appropriate results for AP, Voice 2 and VC clients (other results were left out for the sake of clarity/readability of the graph). AP sends traffic contributed by one videoconference, one audio, one data Ack (only Period II) and one voice stream while Voice 2 client only needs to send one voice stream. According to the traffic patterns specified in Section III, frames of data Ack stream arrive with a much higher frequency than voice frames. At the same time, the frame sizes of audio and videoconference streams are much larger than the size of voice frames. Therefore, AP maintains a much larger buffer of frames ready to send than Voice 2 Client. Voice 2 Client uses its 'fair' share of transmission opportunities which are sufficient to achieve full required throughput with acceptable delay, while AP needs to consume more transmission opportunities to send its frames, thus using 'excess' opportunities left by other clients. In the simulation period II, when overload occurs, the AP cannot access transmission opportunities sufficient to satisfy its traffic, because the required 'excess' opportunities have to be shared with the bulk data client. As a result, AP experiences buffer overflow and starts dropping frames, while the Voice 2 Client can still have its requirement for transmission opportunities satisfied and thus maintain satisfactory throughput and delay as shown in Figure 23. Results above (in Figure 22) illustrate that each client receives fair treatment in terms of transmission opportunities, even in an overloaded system, and that busy clients may also be able to share the 'excess' opportunities left by other clients. In the next sub-section, we further study the issues of fairness among streams serviced by the same busy client.

Fairness among Application Streams in One Client

When the system is lightly loaded, the sharing of resources among different application streams in one multi-stream client follows the same principles as those applicable to sharing of resources among different clients. However, in case of overload, the application streams in a multi-stream client may interact with each other in a different manner, and the behaviours are highly dependent on packet arrival of the streams involved and the respective packet sizes.

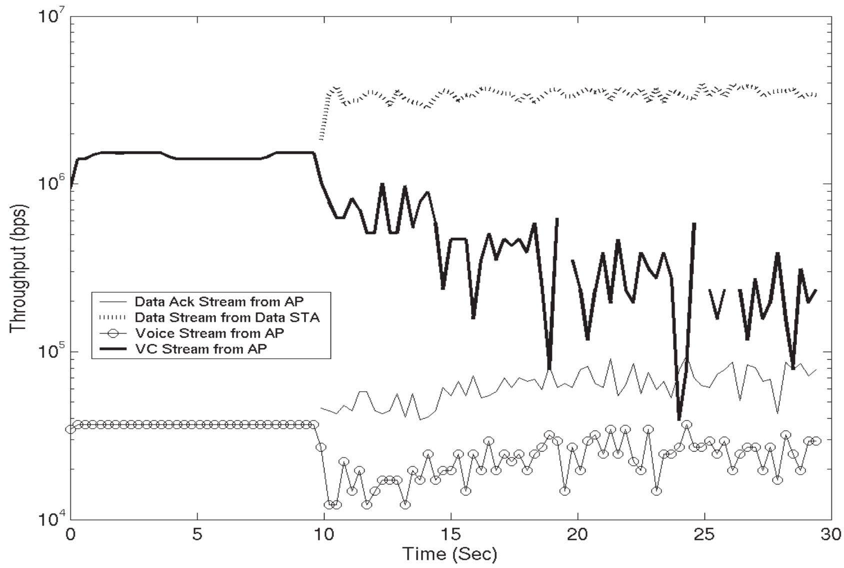


Figure 24 Multiple Streams Performance in AP

We will observe the interaction between streams in an overloaded client on the example of Access Point. For the purpose of our discussion, we further divide the simulation period II into period IIa and period IIb. Period IIa is defined as the period from the start of system overload to the time when the transmit buffer at AP fills up. Period IIb then follows. Streams serviced by the AP behave differently during periods IIa and IIb. During period IIa, the proportions between numbers of frames generated by different streams are reflected in the relative proportions of frames stored in the buffer. Because of the first-in-first-out service discipline, the sequence of frames in the buffer (and frames transmitted) still reflects the sequence of frames generated as in period I. Therefore each stream drops almost the same percentage of its packets (i.e. suffers the same degree of degradation in allocated bandwidth). In Figure 23, the period IIa can be identified as spanning between 10 and 14 seconds. The relative drop in throughput experienced by the VC and voice streams transmitted by the AP during the period IIa can be seen in Figure 22. Both streams experience a drop in throughput by approximately 50% (please note the logarithmic scale used in the graph).

However, during period IIb, the relative proportions between the numbers of frames in the buffer change. Because the buffer is full (or nearly full), most of the time it does not have enough room for the newly generated VC (large) frames, while the small voice frames can still fit in the buffer space freed up by a transmission of a frame. As a result, small but frequently generated frames will dominate the buffer, causing relatively larger packet (throughput) loss for streams with less frequent but large frames. In the case shown in Figure 24, the frames from Data Ack stream (64 bytes) are much smaller than videoconference frames (1504 bytes) and audio frames (815 bytes). Accordingly, we observe in Figure 24 that during the simulation period IIb the throughput of the videoconference stream decreases, while throughput of Data Ack and Voice streams increases. Not surprisingly, as shown in Figure 22, in period IIb (from 15th second approximately), Data Ack has almost full-required throughput while the videoconference stream can only maintain approximately 25% of the throughput it requires as compared to 50% in period IIa.

In order to verify our understanding of the buffer effect, we simulated the same scenario, this time with a much larger buffer (12144000bits); the results show that the larger buffer can only extend the duration of period IIa. Once the buffer fills up (or nearly fills up), clients and streams experience the same problems as described before.

We conclude that under overload conditions, traffic streams with short but frequent frames tend to dominate over other traffic streams within a multi-stream client. Moreover, a highly loaded client (typically, the Access Point will have more traffic to transmit than other clients) will, under overload conditions, tend to suffer more throughput degradation than other more lightly loaded clients. In the extreme, down-stream traffic transmitted by the AP may suffer unacceptably high degradation of throughput while up-stream traffic may still enjoy the required level of throughput.

A good example of this unfair and asymmetric allocation of bandwidth, as shown in Figure 24, is the throughput enjoyed by the data client, which requires 3.68Mpbs bandwidth for transmitting its bulk data stream. The client achieves nearly the required level of throughput while the much lighter VC stream from the AP (overloaded with packets to transmit) suffers significant degradation of throughput.

4.4.3 Discussion on Further Effects and Inheritance

The interaction of traffic streams described above has further effects on the traffic handling behaviour of the network. Those effects are stated as below.

Effect on Bi-directional Streaming

Bi-directional traffic streams (e.g. video or audio conference) often require symmetric throughput performance for both directions. In an “infrastructure” wireless LAN, these are the directions “to AP” and “from AP”. The AP is nearly always more loaded with frames to transmit than any other client in the network since it transmits to all other clients. It is therefore almost inevitable that under heavy load (or overload), the downlink direction of communication will suffer more than the uplink one. This can be illustrated by the example of videoconference streams between the AP and the VC client in Figure 25. The VC, as described before, generates large and infrequent frames. Downlink VC traffic shares the resources available to AP with many other traffic streams (some of them with short and frequent frames) and therefore in the case as illustrated in Figure 25, it drops to approximately 20% of the required throughput when the network is overloaded. The uplink VC traffic, however, is the only traffic transmitted by the VC client, and because this client can capture sufficient share of total opportunities to transmit, the uplink direction of the VC enjoys sufficient level of throughput.

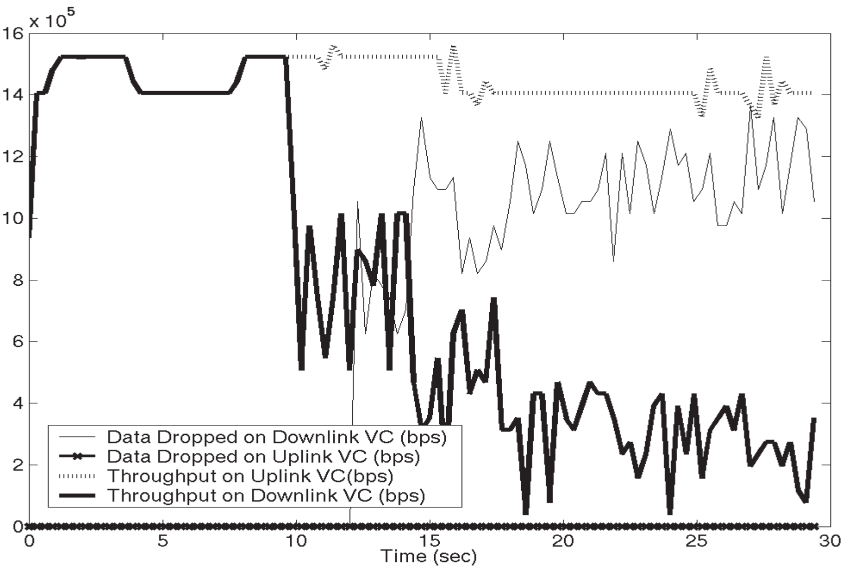


Figure 25 Videoconference Performance

Effect on Uplink and Downlink

Internet traffic usually exhibits asymmetry between the uplink (lightly loaded) and downlink (highly loaded) directions. Since the AP has no more transmission opportunities than other clients (AP is only considered by DCF a “normal” client), this creates a significant imbalance between the transmission load on the AP and transmission load on other clients. This creates conditions under which the phenomena described before can cause significant degradation of quality of service for the downlink traffic (e.g. downloads of data in response to uplink requests), while the uplink traffic (e.g. requests) enjoys sufficient throughput.

Further Inheritance in Enhanced Contention Schemes

One may question these fairness problem and effects would not be a concern in those enhanced contention schemes as multiple service classes are supported in these enhanced schemes unlike DCF. According to the discussion in Section 4.2, all enhanced contention schemes including EDCF are more realistic to support limited numbers of service classes, because the increase of supported service classes in number would have a clear trade-off of overhead produced by the additional usage of IFS and CW. In order to support a variety of applications, the obvious solution is that several applications share one service class in these enhanced schemes in order to avoid the overhead. However, the fairness issue among these applications (so the relevant clients) would still remain the same as those in DCF, as applications with same traffic class in enhanced schemes are exactly treated the same as that in DCF. Therefore, the usage of different service classes in enhanced schemes may only help ease the fairness problem among clients and application, as limitation on the numbers of service classes supported can not meet the requirement of the variety of application types.

In the worst case, the asymmetry effect on bi-directions applications (such as VoIP) and downlink/uplink performance seem remaining in the enhanced schemes. As it is known that even in the enhanced schemes, AP often services more traffic than any of its clients and it is given the same transmission opportunities as its clients for the same traffic class. When there are several applications simultaneously serviced by the same class in the AP and clients, those sent by AP are still at a disadvantage compared to those sent by clients. The scenario is the same as that in DCF. The downlink traffic would suffer compared to uplink even they are in the same traffic class; and those same class bi-direction applications would have uneven two-way communication performance. In short, the fairness problem among these applications (clients) remains and the same performance asymmetry problem would be observed in the enhanced contention schemes including EDCF.

4.5 Summary

Even though the enhanced contention schemes fail to provide hard QoS guarantee, it can provide priority-based service differentiation for non-real-time applications using IFS and CW mechanisms. However, all these enhanced schemes inherit the contention nature, where the notion of fairness has been derived from. We have presented simulation results to illustrate the issues of fairness in respect to multimedia traffic over the IEEE802.11 DCF WLAN. In an underloaded system, the DCF ensures fair (i.e. proportional to the number of frames to send) allocation of resources to clients and traffic streams. However, when it is overloaded, fair degradation of the level of service cannot be assumed due to behaviour of limited-size buffers and individual traffic arrival patterns. Traffic streams with light and frequent frames tend to dominate access to buffer space and radio resources available to a client at the expense of streams with larger and less frequent frames. Such problem is likely to affect AP performance in an infrasructure network, as well as create asymmetry of throughput enjoyed by bi-directional applications. Furthermore, as all enhanced contention schemes including EDCF inherit the natures of DCF, the fairness problem as discussed above, remains serious in these enhanced contention schemes. The usage of multiple service classes may only help ease the fairness problem among the applications, but the asymmetry effect on bi-directions application and downlink/uplink performance remains as an issue.

These results show that contention schemes could perform well in a fair way in lightly load network. In overloaded networks, using EDCF, it is suggested to assign non-session-based applications to assess DCF channels to maximise network efficiency.

However, under high load especially when one client (usually the AP) has to transmit more than other clients, session-based applications could not have any guarantee and they could not tolerate relatively obvious delays and packet loss. More sophisticated mechanisms are needed to control access of wireless clients to radio resources. Such mechanisms have to be capable of controlling bandwidth allocation of traffic streams to the resource-limited network, and to account for the greater transmission bandwidth requirements of some clients (e.g. APs) in the principles on which bandwidth allocation is based. In short, more explicit, sensitive to traffic characteristics, QoS control (service differentiation) is needed in resource-limited wireless LANs. These challenging problems lead us to further explore the options for service differentiation, which we will discuss in the next Chapter.

~ Success is never final and failure never fatal. It's the courage that counts ~

(George Tilton)

Chapter 5 Service Differentiation in Polling Medium Access

We have seen that even though the contention medium access schemes can provide a priority-based service differentiation with certain enhancements, it fails to maintain a fairness nature in respect to multimedia traffic. Provision of service differentiation cannot purely rely on the contention medium access schemes. In this section, we propose a novel solution for services differentiation with combination of polling and contention schemes. Firstly, varieties of polling schemes are reviewed, whereby polling-based Services Differentiation is introduced. Based on these foundations, the supportive specifications of IEEE 802.11e are then introduced. Owing to the shortcomings of these existing schemes, we present a novel Registered Multi-cycle-Polling scheme in a sequential way. Finally, we depict the entire Service Differentiation picture by multiplexing Registered Multi-cycle Polling method and contention method, based on the simulation performance observations.

5.1 The Basic Natures of Polling Medium Access

As it is well known that, the legacy IEEE 802.11 has two different access methods, the mandatory Distributed Coordinator Function (DCF) and the optional Point Coordinator Function (PCF). PCF is a centralised, polling-based access mechanism and requires the presence of a base station (client) that acts as Point Coordinator (PC).

In the PCF mode of operation, the AP alternates between Contention Periods (CP) where the DCF access rules are used and Contention Free Periods (CFP) where the AP explicitly allocates transmission opportunities to wireless clients by polling mechanism. The CFP is started by a broadcast of a special beacon frame from the AP; this forces all wireless clients to enter a mode where they can transmit only in response to a poll. The AP sends poll frames, possibly piggybacked on the downlink frames, to clients that have been placed on the polling list; these in turn respond with acknowledgement and uplink frames. At the end of the CFP (signalled by a CF-End frame), clients revert to their normal operation under the DCF rules (i.e. enter the Contention Period) until the next CFP.

In fact, PCF was originally developed to support delay-sensitive (time-bounded) services and it does offer a certain level of service differentiation (the AP can schedule transmissions to and from selected clients). To eliminate contention, only PCF-enable clients can allow transmission by being polled during CFP in a certain order. Therefore, PCF-enabled clients apparently have higher priority to assess the medium especially during CFP.

5.2 Service Differentiation for Session-based Applications Using Polling Schemes

Traffic Classifications

In Chapter 3, we have discussed that traffic on the WLAN can be divided into two major classes: traffic generated by session-based applications and traffic generated by non-session-based applications. The division criteria are actually based on the QoS requirements of such traffic. Non-session-based traffic does not need hard QoS guarantees while total throughputs possible are the only concern for them. Thus explicit resource reservation is not necessary. The traditional "best effort" service is sufficient for this class of traffic. Session-based applications normally start with session set-up procedure and require resource reservation as discussed. Applications such as Voice over IP and video streaming sessions are the examples. The QoS parameters need to be guaranteed for the duration of the session after the resources are reserved and the session is arranged between the application and network entities. Therefore, the two types of applications not only have different requirements on resource reservation, but also require different levels of service differentiation apart from the admission control. In brief, session-based applications require stricter and more guaranteed service differentiation while non-session-based applications would be satisfied with priority-based service differentiation.

Contention-based Schemes for Non-session-based Applications

In Chapter 4, we have already discussed the contention-based schemes in details. They provide certain levels of service differentiation and achieve medium utilisation. However, when the network is highly loaded, the collision rate is also high with usage of contention mechanisms. So it fails to provide effective traffic protection (hard guarantee) and may be only suitable for less critical (loss-tolerant) traffic such as Best

Effort. However, the usage of CW and IFS enables them offer certain level of service differentiation to meet the needs of different applications. But, as discussed earlier, the larger numbers of service classes it supports, the more overhead it produces from the increased usage of CW and IFS; and it then leads to medium utilisation decrease. Given that limited class number is supported, several different types of application may have to share one class. We then found out that the fairness issue among these applications in the same class remains a problem and the asymmetry effects have also been observed.

Based on the rationale above, contention-based schemes can not offer the full picture for service differentiation for a variety of applications. In order to avoid the sharing of the same traffic class for different applications, it is suggested to use contention-based schemes as part of service differentiation mechanisms to support a limited number of particular types of applications. In fact, according to the features of contention-based schemes, they are suitable to offer services for non-session-based traffic, as these traffic only require priority-based service differentiation while have a high demand for network resources. Furthermore, the contention-free period can end earlier before it reaches its per-defined period if there is no more data to be polled, contention period can then be extended. Therefore, non-session-based traffic is able to fill up the excess resources as desired to achieve better network efficiency during the extended contention period.

Polling-based Schemes for Session-based Applications

Comparing to contention-based schemes, polling-based schemes have more controls on medium access and provide more QoS guarantee by polling clients to access the medium in a certain order. By eliminating contention during its entire transmission period, it takes its traffic out from contenting with the contention traffic by using poll mechanisms. Particularly in an overloaded system, we emphasize that AP should take priority to allocate available resources to meet parameterised QoS for session-based applications; while other non-session-based applications are only offered limited opportunities in a short CP due to the network conditions. Hence, CFP dominates in the superframe.

Even after the CFP is allocated and the traffic stream is admitted, the frame loss and delay requirements of such traffic are still required to be met by intelligent polling mechanisms. Polling-based schemes normally work in a centralised method, which provides the full control of the medium during the contention free period. It enables the commitment of dynamic network resource allocation and Admission Control. Therefore, we consider session-based traffic should be treated by polling-based schemes to meet their parameterised QoS after such session-based traffic is admitted by Admission Control. Admitted session-based streams, like Voice over IP and video, should be offered transmission opportunities via polling in order to meet the strict delay requirement. In such situation, polling-based schemes not only offer session-based streams priority access over the non-session-based traffic, but also provide session-based streams hard QoS guarantees.

However, the efficiency of polling-based schemes to provide service differentiation mainly depends on how the Beacon Interval (superframe size), ratio of CFP to superframe and polling order are designed; whether client is required to send signal to PC to make request for time-bounded services and how these signals are transmitted. Therefore, the legacy polling scheme (PCF) fails to provide the level of service differentiation control necessary to deliver QoS guaranteed service. Several enhancements to legacy PCF for delivery of QoS guaranteed service are studied in the following section. Besides the design issues above, in [Mangold et al. 2003a], it was reported that QoS supports are also limited according to the problems of delayed beacon, unknown transmission time and beacon-missing clients.

Therefore, enhancements on polling scheme are necessary to provide parameterised service differentiation for session-based application. Below, we start with literature reviews of the proposed enhancement and then discuss the 802.11e enhancements as specified in the 802.11e standard. Based on such analysis, we propose the novel enhanced mechanisms for service differentiation support with optimal usage of 802.11e facilities and characterise their efficiency. A performance evaluation of the described mechanisms through simulation results is also presented after such analysis.

5.3 Analysis on Existing Polling Enhancements

5.3.1 Previous Enhancements

[Kabara and Calle 2012][Sheu S.T. and Sheu T.F. 2001] [Kim et al. 2005] tried to guarantee the quality of real-time traffic by using a distributed bandwidth allocation/sharing/extension (DBASE) protocol over the CSMA-based DCF and PCF. It is based on a contention process that only occurs before the first successful access and a reservation process after the successful contention. The real-time clients (having voice or video) contend for medium by RTS to join the reservation table and reserve needed bandwidth; while the non-real-time clients gain medium accesses based on conventional DCF. The actual allocation of reserved bandwidth is within CFP, supported by PCF. Obviously, this scheme requires reservation procedure and uses RTS/CTS for signalling.

In [Yeh and Chen 2002] [Niyato and Hossain 2006 & 2007], they proposed four polling schemes combined with the PCF to improve the utilisation of the wireless channel and support certain quality of service (QoS) of multimedia traffic. First In First Out (FIFO), Round Robin (RR), Priority and Priority with Effort-limited Fair schemes have been studied in their researches. The results show PCF with polling scheme generally performs better than DCF with respect to provide service differentiation. Within these four schemes, FIFO is good at providing max throughput while performance of most of these schemes can be affected by some traffic pattern. [Liu et al 2011] provides an experiment platform for such studies.

A previous study of [Coutras et al. 2000] [Inan et al. 2006] [Lam et al. 2006] [Liu and Guo 2011] gave suggestions for a better polling mechanism to achieve strict delay and loss guarantee. It firstly proposed that each client can only be polled once during each CFP and the maximum size of frame has to negotiate with PC/AP. These strategies ensure every PCF client has fair opportunity. Clients, which require opportunities during CFP for new data flow, are firstly a connection request; if accepted, it is placed at the end of the polling list. Therefore, data flows with existing connection have higher priority than new arriving flow. A new concept of 'time window' is also used to minimise the tail delay distribution of the difference between nominal and actual poll time.

'Superpoll' is proposed in [Ganz and Phonphoem 2001] as signalling scheme for reservation of time-bounded services. [Misra et al 2011] inherits such idea for more evaluation. In this scheme, AP broadcasts at the beginning of CFP a superpoll message, which includes the list of clients that will be polled during this CFP. In order to solve the problem of possible collisions due to erroneous reception, each client monitors the transmission on the channel and in its turn, it re-sends the superpoll message appended to its frame. Such a superpoll signalling actually forms a chaining mechanism to further improve the reliability of PCF transmission in a noisy environment. The results show it improves PCF support for multimedia application; while as a trade-off, it also produces certain amount of overhead, which is caused directly by superpoll messages.

5.3.2 802.11e Polling Enhancement

As discussed above, legacy IEEE 802.11 MAC (DCF and PCF) does not provide sufficient supports for QoS provisioning in MAC level. All those DCF-based and PCF-based proposals for providing service differentiation, which have been discussed above, have implied changes in the basic medium access or have modified the standard protocol behaviour of DCF and PCF in order to provide service differentiation with legacy 802.11 WLAN. Consequently, IEEE 802.11 working group forms 802.11 Task Group E to conclude such schemes and define certain supplements as a standard protocol of 802.11 to meet the QoS requirements by the method of either contention or polling.

The major enhancement in 802.11e is the new hybrid access mechanism, Hybrid Coordinator Function (HCF), which includes Enhanced DCF (EDCF) and Controlled Channel Access (CCA). An enhanced client, which may optionally work as the centralised controller for all other clients within the same QoS BSS (QBSS), is called the Hybrid Coordinator (HC) and typically resides within an 802.11e AP.

Same as that in legacy 802.11, there are still two phases of operation within a superframe, a CP and a CFP in 802.11e, which alternate over time continuously. HCF is proposed as a combination function of DCF and PCF with some enhanced QoS mechanisms. The EDCF is used in the CP only with a method of contention, while the HCF is used in both phases with a method of polling, which makes this new coordination function hybrid. The contention schemes have been well discussed in Chapter 4. Under the HCF, the AP can, in addition to polling clients during the CFP, poll clients (or itself) and allocate them a TXOP during the CP. Therefore, such access method is highly hybrid for being able to access the medium whenever it wants to.

To allow requests for TXOPs, the AP advertises a Controlled Contention Interval (CCI) in which clients can make resource requests (RRs) i.e. requests for TXOPs. As part of 802.11e, Controlled Contention / Resource Request (CC/RR) are defined an additional access protocol. It is firstly part of polling-based scheme. Only by receiving a QoS poll, the polled client can own the medium from the specified time up to its TXOP limit specified in the poll frame; while other client set their NAV and would not contend for medium during that period. For this, the HC requires information that has to be updated by the polled clients from time to time. The crucial design of CC/RR is the reservation channel for update of traffic stream on a frame basic. It provides a way for HC to learn which client needs to be polled, at which time and for which duration. Therefore, a request mechanism (RR) is designed for reservation of TXOP for individual frame of traffic stream (TS). An RR should not need to contend with other EDCF traffic, but controlled-contend with other RRs for other TSs. HC firstly broadcasts a specific control frame (CC frame), which defines a number of controlled contention opportunities (CCOP) and a filtering mask containing the traffic classes (TSs) in which requests may be placed. Each TS client with matching TS chooses one opportunity interval in a pre-defined order (contained in CC frame) and transmits such request to HC. A Resource Request frame may contain the requested TS and TXOP duration, or the queue size of the requested TS. Acknowledgements for reception of requests are broadcasted again in the following CC frame for collision resolution. This control frame also forces other clients (CC-ineligible) to set their NAV until the end of the controlled contention interval, thus they remain silent during the controlled contention interval.

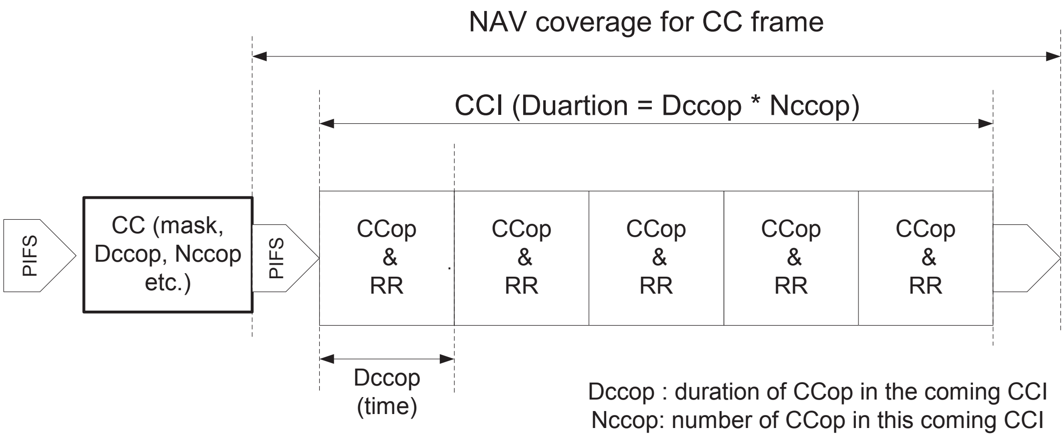


Figure 26 Controlled Contention Interval [IEEE 802.11e D3 2003]

However, if RR is used, the transmission of RR is Controlled Contention. Therefore, it has potential of RR collision. Within 802.11e working group, there are further proposals concerning on how to send Resource Request. If CC is used, the per-defined order is needed for TS client to find the right interval to transmit its RR while avoiding colliding with other RRs. Proposals on design of such per-define order mainly consider either backoff procedure, random, or compliance with polling order.

In the mean time, according to objective of 802.11, it only standardizes the specification of any possible options and facilities to be used for QoS enhancements.

Therefore, it leads to the challenges of how to optimally use these options and facilities to depict an entire QoS picture in an efficient and effective way.

5.3.3 Summary

Indeed, a great deal of research concerns actually has been drawn to polling-based channel access scheme. It is proposed to provide parameterised QoS based on the contract between clients and AP. Once a traffic stream is set up after the successful TSPEC operations, AP allocates the required/promised bandwidth by using controlled channel access method (either CFP or CAP). All these researches above contribute in different aspects for PCF to support time-bounded services; while at the same time, it also demonstrates that PCF has quite a few fundamental problems in supporting time-bounded services.

Firstly, PCF performance heavily depends on polling list scheduling, which should be implementation-dependent, while the specification of IEEE 802.11 only defines a simple round robin scheme. The scheduling is the key aspect to design an effective and efficient polling-based scheme. And such design has to give consideration to traffic pattern as well, as performance of standard PCF can be affected not only by polling list scheduling but also by the traffic patterns. For example, if a client with large frame in size or large numbers of frames is polled in the beginning of CFP, the duration of transmission of such client can be very long. Therefore, it may destroy any attempt to provide opportunities for other clients, which are listed for poll during the rest of CFP.

Secondly, our studies also show that designs of the maximum duration of CFP and superframe size are also crucial to PCF performance. Longer duration CFP provides more opportunities for time-bounded services while also treats normal services unfairly, which are operated with DCF. Smaller superframe can provide a short delay bound for time critical application while it also introduces a lot of overheads as trade-off.

Finally, in order to schedule polling list in an efficient way, it seems that AP should have information on traffic patterns and specifications on all demanding PCF-enabled clients. Therefore, signalling for such reservation seems necessary for this purpose while it also leads to certain overhead, which has not been addressed in 802.11e. Enhanced mechanisms for PCF or new coordination functions are required to solve these problems of PCF to achieve QoS efficiency and effectiveness.

In summary, overhead is one of the major concerns to polling-based schemes due to the additional overhead from establishment of CFP and CF-Poll comparing to contention-based schemes. Designs of polling method (either legacy polling or polling only by granted request), proportion of CFP, the scheduling of polling list and the beacon interval also heavily affect the performance of polling-based schemes. According to these technical challenges, we propose our Registered Multi-cycle Polling in the following sections.

5.4. Registered Multi-cycle Polling

As discussed in the review above, quite a few numbers of options and facilities are discussed to provide polling-based QoS supports over IEEE 802.11 WLAN. Based on these options and facilities, we propose a novel polling service differentiation for 802.11 WLAN under 802.11e supplements. The solution properly makes use of these options and facilities to achieve a better efficiency; while also effectively provides service differentiation for session-based application.

The solution focuses on tackling the design of the polling method and schedule, also given consideration to the performance factors of the proportion of CFP and the beacon interval. At first, we propose Registered Polling by discussing then examining the efficiency of Registered Polling comparing to polling method via Resources Request. As it is known that no single option or facility can solve the problem entirely; Registered Polling is designed as associated with the Hybrid QoS framework with coordination in 802.11e TSPEC operation discussed in Chapter 3. Furthermore, additional Multi-cycle polling schedule is enhanced to solve the dilemma of overhead and delay performance of those delay-sensitive applications such as VoIP.

In this section, we introduce our design of Registered Polling with performance analysis. Firstly, the necessity of traffic information update for an intelligent polling is discussed. Request polling method is then studied with its particular information update method. Owing to the shortcoming of Request polling method, we then propose Registered Polling. Performance of Registered Polling is also examined via simulation according to its design objectives of efficiency and effectiveness.

5.4.1 Analysis on Information Update and its mechanism

The Need of Information Update in Polling Methods

At first, we assume that all Traffic Streams (TS) discussed below are admitted by Admission Control in AP and eligible to be polled by one of the polling methods. Legacy polling method in 802.11 standard is designed originally to provide time bound service [IEEE 802.11 1999]. However, in legacy polling, a random polling order is used, which leads to its problem in efficiency and effectiveness. In order to achieve network efficiency and allocate network resources effectively, AP should implement an intelligent polling method to meet the individual QoS needs of varied applications. Such method should be based on the available information of individual QoS needs of varied applications. We recall that different TSs have different traffic characteristics, such as delay requirements, transmission duration and frame arrival pattern, therefore require different parameterised QoS. For example, for Voice over IP streams, frames arrive frequently in a small frame size. Once these voice frames arrive, they demand immediate attentions to maintain a strict delay bound; while video frames can tolerate certain period of delay but concern about the bandwidth allocation as they normally come in a large frame size. Therefore, information, especially on traffic arrival situation and resources request (for queuing traffic), across the network, are required in AP. And such information should be regularly updated. AP should schedule TXOP based on such information. Hence, it is admired for AP to keep updated on varied information across the network in order to be able to schedule an optimal polling order to meet individual QoS needs instead of a random order defined in legacy 802.11 standard.

According to the necessity of information update especially on traffic arrival situation, resource request (for queuing frames) and TS transmission duration demand, an information update channel is required. Such channel can be achieved either implicitly or explicitly. In an explicit information update, Controlled Contention (CC), defined in 802.11e supplements, may be used to send Resource Request (RR) message for TXOP allocation for queued frames before the actual transmission. Such polling methods can be categorised and defined as Request Polling. It indeed requires an individual physically signalling channel. On the other hand, we propose Registered Polling, where an implicit information update is applied. It intercepts information on available resources in AP and predicts the network resource request in a simple way. We analyse both schemes and make comparison in terms of their effectiveness and efficiency. Performance of Registered Polling is also examined via simulation according to its design objectives.

Request Polling and CC/RR

- CC/RR Option

As discussed above, CC/RR protocols are introduced as IEEE 802.11e supplements as a major option of request polling method. The AP only issues poll and TXOP after receiving resource request for individual frames. They are used for clients to make reservation for individual frames, which have arrived (or are expecting to arrive soon) in the client and require immediate attention for QoS treatment.

The operation of CC/RR has been described in the previous section. Such explicit CC/RR certainly works well on information update, as it defines the individual and separate signalling channel between eligible client and AP. For those time-critical frames, AP then takes sharp responses to allocate resource either in CFP or CP (CAP). The more frequent CC/RR is used, the more precise information the AP has. Additionally, CC/RR also tries to achieve efficiency in a large network. Since in a large network, the number of the eligible clients may not be small and they do not constantly have frames arrived in queue when they are polled. Therefore, these polls to the idle clients are waste of resources. Using CC/RR, AP can only poll those eligible clients, which have frames arrived in queue, based on the updated information from CC/RR. As a consequence, excess polls reduce.

- Other Request Polling Options

Beside the proposal of RR by EDCF above, other most dominant proposals can be summarised as Group Request Period (GRP) [Fischer & Ho 2003]. In these proposals, AP can access medium for GRP regularly or irregularly, in either CFP or CP, depending to its estimate of resource request across the network. During such GRP, a Controlled Contention frame is broadcasted and it includes information of an RR transmission pattern in next CC interval (CCI) and eligibility of RR transmission for each client. Eligible clients have chance to send RR in a certain pattern while ineligible clients have their NAV set so as the CCI is protected for interference. With CCI, RR is transmitted in a group, separated only SIFS, in a certain order, either random, per-defined or in a permission probability. Each RR transmission opportunity is defined as the same length and same format and they are only separated by SIFS in a group. Neither they are provided with collision avoidance mechanism, nor are they provided by polling. Even when there is a per-defined order in use, it certainly has a probability of collision among RRs themselves. Even when there is a per-defined order in use, it may only optimise the problem. In [Sherman 2002], a backoff procedure is proposed for only RR transmission to avoid request collision. This leads to further complexity and overhead.

Further Analysis and Concerns on CC/RR and Request Polling

Firstly, we must clarify that CC/RR reservation is different from the higher-level stream reservation that we define in the Hybrid QoS Framework. The higher-level reservation in Stream Admission Control is used to request resources for TS before they become active, so called stream-based reservation. CC/RR reservation is only applied after such TS is already active and is used to request recourses for frames of the active stream. It allows very frequent information updates in AP. AP can then fully be aware of the frame arrivals and their requests. Apparently, it is an explicit method of information update by direct information transmission over the medium. One may question that these two level reservations repeat themselves especially for those applications with regular arrival pattern and frame size. Explicit information update for these application streams may not be that necessary as their frames arrive regularly in the same size even though CC/RR may provide information more accurately.

The second concern is the overhead CC/RR produces. Because medium is used for CC/RR transmission, it certainly introduces signalling overhead caused by those CC/RR packets, which travel frequently. It is actually similar to RTS/CTS overhead problem in DCF. As it is known that, a frame, transmitted via RTS/CTS, has related overhead of 178 bytes, comparing to 96 bytes for a frame, which is transmitted by normal contention under DCF. Therefore, CC/RR overhead should only be counted and shared by the frames, which gain TXOP via the CC/RR channel. If there are more requests in CCI, the more transmitted frames share the overhead, the less overhead per frame it produces. Furthermore, the less frequently CC/RR packets travel, the less overhead it produces. CC/RR overhead problem becomes a major concern when considering the usage of CC/RR.

The frequency of CCI usage in each superframe has further affects on the CC/RR efficiency. The more frequently CC/RR is used, the more overhead it produces. If packets arrive more frequently and/or the wireless medium is fully/overloaded with a large number of applications, AP is certainly desired to be updated more frequently to gain precise information. While more frequently CCI is used, overhead problem would be even worse.

Therefore, the question is how frequent of CC/RR is appropriate. It is a question of how the trigger mechanism for CC is designed. For the newly established TS, AP certainly has information about it as such TS is actually admitted by AP and RRM. If AP broadcasts a CCI regularly (regularly triggered according to a timer), efficiency may be a concern. Others suggest a prediction method where previous request information from the historical CCI is used for estimate of near future demand. The trigger mechanism seems to be the dilemma between the precision of information and efficiency, which is still under discussion. Fundamentally, the question is also the choice of implicit information update or explicit information update.

The third concern on CC/RR is the collision problem in RR transmission within the CC interval. In fact, within 802.11e working group, there are a number of parties working on this problem and proposed a number of solutions. The first proposal is to use contention method (EDCF) to send RR, giving the highest priority to RR [Ho 2002] over other EDCF data traffic. In such way, RR does not necessarily wait for a defined period for transmission and can be sent at the time it wants. It enables a highly precise information update according to the minimum RR transmission delay. However, as already known, EDCF experiences certain rate of collision itself and has possibility of dropping or re-transmission of the RR, which would cause higher latency or larger Jitter. The usage of contention for RR transmission seems against the reliability principle of polling method. Therefore, the RR collision problem seems unavoidable.

The CC/RR Usage Limitation

In fact, CC/RR protocols are defined as QoS enhancement for 802.11e, the usage of CC/RR becomes crucial to its functionality and efficiency according to the problem we pointed out earlier. We thus further recommend the appropriate usage of CC/RR usage as below.

Firstly, CC/RR should be used more suitably on unpredictable situations, such as unpredictable traffic arrival pattern and unpredictable network conditions (such as congestion and interferences in core network). In such unpredictable situations, packets arrive in an irregular and unpredictable way, so AP is not able to predict the time of requests and the amount of requests as no information of such requests are available before packets actually arrive in the clients. Therefore, CC/RR is a specific tool to keep such unpredictable information updated with AP. For example, TS set-up signalling (or re-modification) is time-critical and also unpredictable from time to time. AP can response quickly with CC/RR usage for these signalling. However, such situations are rare.

Furthermore, CC/RR should be used in the case that the system has large proportion of CP in a long superframe while delay-sensitive applications are also present. During CP in such situation, delay-sensitive applications cannot really be guaranteed a strict delay bound due to uncontrolled contention methods. These applications can not wait for the CFP neither as there is a long superframe. The longer superframe and the larger proportion of CP there are, the worse delays of these delay-sensitive applications have. Therefore, AP should be informed the arrival of these application frames even during the CP and offer transmission opportunity. If AP is updated with request information for these arriving frames during CP using CC/RR and gain medium for CAP (polling method during CP) to allocate resource for these applications during CP. The problem is solved. However, compared to the EDCF that gives contention priority to these applications during CP, CC/RR methods can only improve the delay performance at a higher cost (with more overhead).

Summary

In summary, we conclude that design of CC/RR and the usage of CC/RR need to be carefully considered according to the efficiency while RR collision problem, overhead problem and design complexity need to draw further attentions. In some cases, it may not be necessary according to the trade-offs it has. In fact, most TS traffic has certain arrival patterns and arrives on a periodical basis. And these are admitted via TSPEC in AP and RRM, AP has their arrival patterns in record and can predict the traffic arrival in a certain level of precision. Those applications may not necessarily need CC/RR. For TC applications, we only consider contention-based service differentiation for them via contention methods. Hence, the amount of applications, which requires CC/RR, may not be majority at all.

5.4.2 Registered Polling

We have well discussed the explicit request polling mechanisms, particularly on CC/RR; as well as several concerns and limitations have been identified. In this section, we propose an implicit method, Registered Polling, to achieve a better efficiency and simplicity. We firstly recall our Hybrid QoS Architecture in Chapter 3, which would offer the facility for AP to gain implicit information on arriving packets in clients. The main components of such Registered Polling are then depicted. Finally, a detailed analysis on this mechanism is presented with comparison to the request polling mechanism.

The Implicit Information in the Agreed TSPEC

Traffic Specification (TSPEC) is introduced as a crucial facility in use for QoS supports. Real-time traffic demands strict performance guarantees from the network. For example, a client wanting to transmit real-time voice to another client would require that maximum delay for data frames be bounded. Before certain performance level can be guaranteed to a client, characteristics of the traffic generated by it have to be known, so that appropriate resources (transmission time) can be allocated to it. Therefore, clients that require performance guarantees are required to set up a 'contract' with AP. Henceforth, such contract is established between the AP and the client(s), and the AP should guarantee the performance desired as long as the client does not violate the declared traffic characteristics. Such establishment can be considered as QoS reservation means of special QoS management action frames. These mechanisms are well discussed in Chapter 3.

Such contract (agreed TSpec) between clients and AP can also provide description of agreed traffic specification. The information elements in TSPEC include Mean data rate, Delay bound, nominal MSDU size, User priority, Maximum MSDU size, Maximum Burst Size, Minimum PHY rate and Peak data rate. These parameters are very similar to those included in the IP FlowSpecs defined in RFC 1363 and used by IntServ and DiffServ. They also facilitate QoS mapping between the IP MAC layers. The TSPEC facility not only offers a means for MAC level admission control and reservation signalling between the wireless clients and the AP, but also provides traffic pattern information of the admitted TS. More importantly, they are registered at AP and available to be used for AP to gain traffic arrival information of the current admitted TS in the Registered Polling method.

Using Implicit Information in TSPEC

Clearly, implicit information makes most use of the available information located in AP. We now look at how this information to be used to predict the resource request from the queuing packets in the active streams. We recall that RR messages in the explicit method include information of requesting TS ID, duration of requesting TXOP and possibly the queue size of current TS in the client. With the complex usage of CC/RR, AP only achieves a certain level of precision of 'current' information about the arrived traffic because the arrival of individual packets varies from time to time. Compared to the explicit information, the implicit information in agreed TSPEC offers simplicity while providing sufficient information to predict the traffic arrival as discussed below.

Firstly, the TSPEC provides all information about the traffic arrival pattern and the starting time of the admitted TS. Most TSs, like VoIP and Video, have constant arrival pattern. For example, VoIP packet arrives from the higher layer at constant every 20ms and keeps such pattern until stream of VoIP finishes. Therefore, it is not necessary to update the arrival of each VoIP packet once it is set up as AP can predict its arrival within certain level of precision based on its agreed traffic pattern in TSPEC. Therefore, the traffic arrival information can be predicted based on the TSPEC information elements in most cases especially for those periodic applications such as VoIP and Video Conferencing.

Secondly, duration of the stream communication is also known in AP by monitoring the TSPEC operation. When the TS ends/establishes communication, a TS deletion/establishment action as part of TSPEC operations is taken and AP is informed of its completion/commencement. Hence, we suggest making the most use of the information, including the agreed TSPEC, TS deletion and establishment information to predict the valid duration of the resource request from the queuing packets in the active streams. TXOP would be granted for the active stream packets only during the duration of the stream communication; and the duration is precisely recorded in AP based on the stream deletion/establishment information.

Also, such resource request prediction is relatively reliable. In fact, once an agreed TSPEC is negotiated, such TS would not normally violate its traffic pattern, otherwise, it would violate the entire network allocation. In case of violation, enhancements are required as the network experiences unpredicted situation. In most cases, network is under control with admission function and is prevented from most of violating traffic.

Determination of Polling Order

Based on the information, AP can then maintain a table of active TSs for a proper polling schedule. The guidelines below are followed in scheduling the poling order. Firstly, only active admitted TS is allowed to enter such polling table with its traffic pattern once it is admitted. Any TS deletion and establishment should be informed and such table is updated conveniently with these deletion and establishment in AP. All completed TS would be taken out from the table. As a consequence, excess polling overhead in legacy PCF can be avoided.

We also adopt part of the polling rules of 802.11 specifications that once a client is polled, AP would keep polling such a client until the end of current traffic queue of such client [IEEE 802.11 1998]. Therefore, once the TS is polled, it would have opportunity to transmit all the arrived frames in the queue.

Most importantly, AP further distinguishes traffic pattern of individual TS according to their time (delay) bound requirements in order to schedules a correspondent polling order and allocates resource properly to meet individual needs. As it is known that Delay-sensitive applications, like Voice over IP, have strict delay requirements and need absolute service differentiation; therefore AP should keep their arrival information regularly updated in order to give them high-level QoS treatment. For example, Voice over IP normally requires 25ms delay bound if delay echo canceller technique is not used. Therefore, along the entire data path, it should aim to provide as less delay as possible in each section path. Certain On/OFF traffic is also time-critical, such as Internet chatting. Other session-based traffic may be able tolerate certain delay, such video according to their traffic characteristics. However, such delay-tolerant applications may require higher bandwidth allocation comparing to delay-sensitive application such as Voice over IP.

Therefore, great cares should draw to meet delay bound for delay-sensitive applications while also providing efficient bandwidth usage for delay-tolerant applications such as video. Therefore, according to the delay bound requirement of delay-sensitive and bursty applications, we place these applications on the top of polling list while delay-tolerant applications come after.

In such way, applications like VoIP gain medium access in higher priority than others, and are transmitted with the possible minimum delay (less queuing delay). Additionally, all queued frames are treated in a group in a superframe and transmitted in a group with similar delay performance according to the polling rules we adopted above. Provided if the superframe is not long, the queuing delay variation between the current transmitted frames and previous frames, which are transmitted in the previous superframe, should not be large. For example, if the VoIP frames arrive in every 20ms, and we adopt a 20ms superframe, such polling method should meet the strict delay bound for VoIP, offering minimum queuing delay in the wireless channel.

We also recall that such VoIP applications are normally in a very small packet size such as 92 bytes of VoIP comparing to 1464 bytes of Video packet. As a consequence, VoIP packets would not occupy the medium for long in each superframe. Therefore, other applications, such as video, which usually starve for bandwidth, can occupy the remaining CFP, which is large proportion of CFP. In other words, applications like video can be allocated large portion of bandwidth even they are not placed on the top of the polling list. Certainly, they experience longer delay than VoIP, but as discussed, they are not so delay-sensitive while bandwidth allocation is more crucial to them. As a consequence of such order, each type of applications gets what they admires; video applications are allocated with large proportion of available bandwidth while VoIP applications can also be guaranteed delay bound.

Efficiency and Simplicity of Registered Polling

In this section, we firstly discuss the efficiency and simplicity of Registered Polling with comparison to Request Polling. With implicit information update in AP, the problem of excess polling on idle clients can be minimised. It is achieved largely by updates of completed or newly established TSs and prediction of frame arriving based on its constant arrival pattern. If Registered Polling can ensure that the AP poll clients effectively and provide parameterised QoS, especially for delay-sensitive applications in an overloaded system, we consider that implicit information update is sufficient enough and further additional explicit signalling may not be necessary.

Request polling uses CC/RR for explicit information update. It defines a separate physical signalling channel and has the ability to provide very precise information update. However, as discussed in the previous sections, the precision of information largely depends on the implementation details, such as CCI frequency, CCI trigger and eligibility of clients for CCI, which all introduce overhead. The more advanced design of CC/RR is implemented, the more precise information it provides, but the more overhead it produced. We believe that networks with different levels of traffic load require different complexity levels of CC/RR implementation. Therefore, in an overloaded system, complex implementation of CC/RR cannot be avoided. It is known that such implementations are yet not to be well studied in the literature, but large overheads are expected in an advanced CC/RR implementation.

Compared to CC/RR, Registered Polling does not cause transmission overheads as no additional physical mechanism is used within the implicit information update. Gaining traffic information for prediction and maintaining a polling table in AP are also relatively simple. As all TSPEC operation is performed mainly in AP, gaining information from TSPEC operations is straightforward. Such information update is certainly implicit as AP only predicts the packet arrival based on its agreed TSPEC contract. However, providing entire explicit information update is at the high cost of efficiency and complexity as discussed above. If AP takes a role of positively polling described above with the implicit (predicted) request information and can also achieve the same objective, registered polling would over-perform request polling, which relies on CC/RR, according to the bandwidth efficiency and implementation simplicity.

We recall that the objectives of a smart QoS-enhanced polling method should achieve delay bound guarantee to delay-sensitive applications while providing large portion of bandwidth allocated to other applications.. If Registered Polling can meet such design principle, the implicit information update is considered sufficient. Registered Polling can achieve the same objective as Request Polling can, but at the lower cost of complexity and overheads. Especially in overloaded situations, where CC/RR implementation is more complex, the efficiency and simplicity benefits of Registered Polling are more apparent as Registered Polling would only handle more information processing at AP.

Furthermore, we carry out simulation below to examine if Registered Polling is able to achieve delay bound guarantee for delay-sensitive applications while providing bandwidth efficiency in an overloaded system. To validate the performance, we firstly describe the simulation environment and network configuration to ensure all scenarios have sensible parameters setting that do not affect the performance of the schemes adversely.

5.4.3 Performance Analysis of Registered Polling

Simulation Assumptions

In order to analyse service differentiation schemes in respect to multimedia streams, we have developed a simple simulation model of an IEEE 802.11b network using the OPNET discrete event simulation package. The OPNET model includes traffic sources, the 802.11 MAC protocol, and characteristics of the 802.11 physical layer and radio medium (e.g. BPSK bit error rate model, free space path loss model). The traffic model remains the same as those in Chapter 4. Several assumptions have been made for simulation experiments.

- The effect of 'hidden terminals' is not accounted for in the simulations, i.e. all nodes are located within hearing distance of every other node.
- The RTS/CTS scheme is not used in our simulations. It is not necessary for experiments aiming at observing fairness of DCF in an environment free of hidden terminals.
- Only one BSS is simulated, therefore no interference from other BSSs is considered.
- Each client maintains only two transmit buffer, PCF traffic buffer and DCF traffic buffer. They are finite; when transmit buffer fills; all newly generated frames are dropped.

Attributes	Value
Data Rate (Mbps)	11
Fragmentation Threshold (bytes)	2304
Physical Characteristics	DSSS Long
Buffer Size (bits)	2048000
DSSS preamble (bits)	144
DSSS header (bits)	48
slot_time (us)	20
SIFS_time (us)	10
Cw_min	31
Cw_max	1023
Beacon Interval (superframe size)	20ms
CFP	19ms

Table 5 Simulation attributes used in Registered Polling experiments

The values of simulation attributes used in our experiments are provided in Table 5. In the table, the value of Physical Characteristics (DSSS Long) stands for Direct Sequence Spread Spectrum (DSSS) with long preamble; this scheme adopts 144 bits for the physical layer preamble and 48 bits for the physical layer header. The buffer size specifies the maximum length of the higher layer data arrival buffer. Each client maintains one buffer; once the buffer limit is reached, data arriving from the higher layer will be discarded until some frames are removed from the buffer.

The performance measurements of interest in our analysis include:

- Throughput: in our simulations, it represents the total number of bits (in bits/sec) handed up from 802.11 MAC layer to higher layers;
- Load: it represents the total number of bits (in bits/sec) submitted to 802.11 MAC layer by higher layers.
- Data dropped: it is the total size of higher layer data frames (in bits/sec) dropped by the 802.11 MAC; it is either due to overflow of MAC layer buffer, or failure of transmission repeated 7 times.

Network Topology

Simulations cover the time span of 30 seconds. Data Client is only non-PCF-enabled client, which only sending the bulk data. Other clients are all PCF-enabled and have multimedia traffic including VoIP, Audio and Video Conference. As mentioned earlier, we simulate only a single BSS with one Access Point (AP). As shown in Figure 27, the BSS includes the following clients:

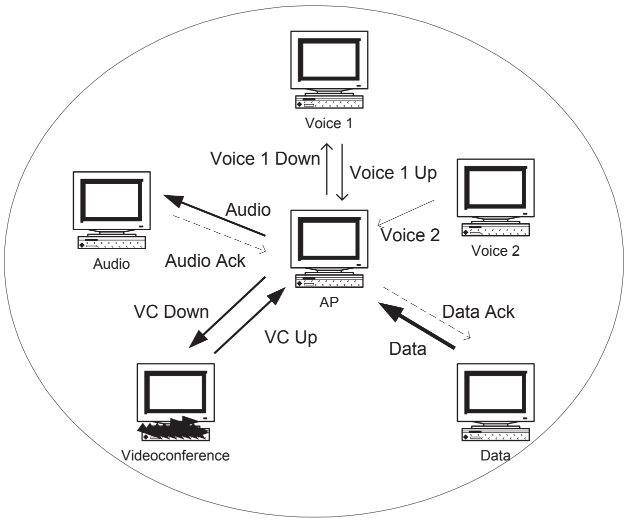


Figure 27 **Simulated Network Topology for Registered Polling**

- Voice 1 Client: PCF-enabled; two-way interactive voice.
- Voice 2 Client: PCF-enabled; one-way voice (to Access Point).
- Audio Client: PCF-enabled; audio download from Access Point, with ACK feedback returned up-link to the Access Point.
- Videoconference (VC) Client: PCF-enabled; two-way videoconference application.

- Data Client: Non-PCF-enabled; Bulk data transfer up-link to Access Point, with ACK returned downlink.
- Access Point: PCF-enabled; it sends voice, audio, bulk data acknowledgments and VC while also receiving the VC, bulk data, voice and audio acknowledgments.

Performance Analysis of Registered Polling

The scenario describes the overloaded situation in the network, where the polling-based service differentiation can be examined. Only data client is ineligible for polling and access the medium via contention as it generates bulk data and only best-effort service is offered. However, it does request large bandwidth allocation at 3.68Mpbs and becomes one of major forces to create overloaded situation. Other clients generate VoIP, Audio and Video Conference; and request parameterised QoS. We assume that they are all admitted in AP and are already active at the beginning of simulation. The service differentiation of these multimedia applications is guaranteed by polling schemes. In our Registered Polling, such polling table includes two VoIP and one VC in order; while downlink VC, VoIP and audio take advantages as downlink data traffic can piggyback on poll and ACK messages.

We also assume that AP has full knowledge of admitted TS streams (VoIP, Audio and Video Conference) from the stream admission procedures. Hence, AP schedules the packet transmission based on the current arrival patterns of the traffic streams. As the system is overloaded with large bulk data, TS streams are protected from the aggressive bulk data and their servicing clients are all PCF-enabled. Within all TS streams, voices streams have higher priority for transmission than other as they are delay-sensitive, so are placed on the top of polling list as described before. Therefore, VoIP would be expected to have the best delay performance compared to other TS streams.

The performance of polling also largely depends on the polling method parameters used in the simulation such as Superframe size (Beacon Interval) and the ratio of CFP to superframe. Even though the throughput of polling traffic is relatively lower than the throughput of contention traffic, we set the ratio of CFP up to 95% in order to guarantee parameterised QoS in CFP rather than prioritised QoS. In fact, our scenario consists of large portions of TS traffic, which requires parameterised QoS. On the other side, as discussed above, if superframe size is large, it would increase the queuing delay variation between current transmitted VoIP frames and previous VoIP frames transmitted in previous superframe. Further analysis will be described in the next sections. At the beginning, we simply adopt 20ms superframe, the same value of the VoIP arrival interval.

- Throughputs and Voice Delay

We firstly look at the network performance in term of throughput to examine the overloaded situation. The total network throughput achieves approximately 2Mpbs as shown in Figure 28. Comparing DCF performance results in the previous chapter; such total throughput is relatively low. However, the system basically ensures communication for VoIP, audio and mostly Video Conferencing (VC), where two VC streams occupy 900Kbps. They are protected from aggressive bulk data traffic, which accesses medium by contention in CP only and experiences large data drop. These two VC streams actually are allocated most of available bandwidth after VoIP and Audio are satisfied with sufficient delay bound. In short, the system is seriously overloaded with large data drop occurred only in bulk data client; and the multimedia applications streams are guaranteed with bandwidths required.

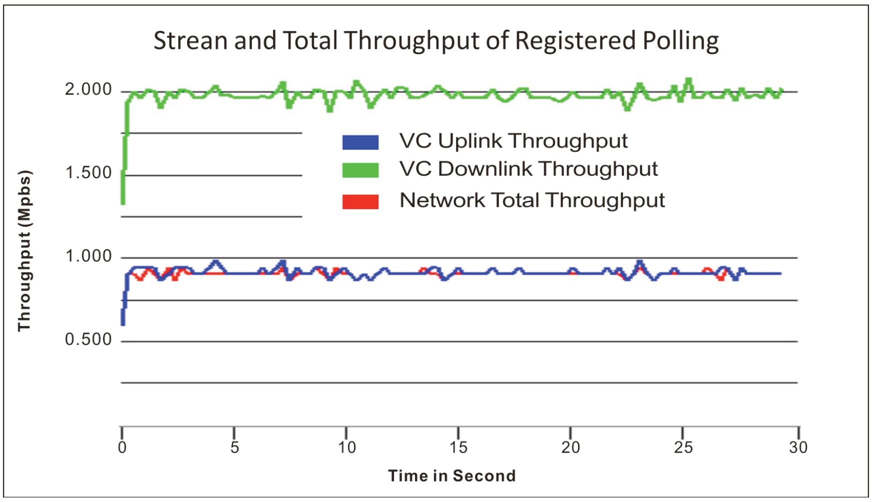


Figure 28 Streams and Total Throughput of Registered Polling

We then further examine whether registered-polling can maintain delay bound for delay-sensitive applications (VoIP and Audio streams) with implicit information update. As seen in Figure 29 below, three VoIP streams perform well and achieve a good delay bound within 3ms - 4ms. VoIP uplink from voice client 2 has slightly higher delay as it is scheduled after the VoIP uplink of voice client 1. Both VoIP uplink are uplink traffic and transmitted when polled; while VoIP from AP (to voice client 1) is downlink traffic and piggybacked on poll message to Voice Client 1 and enjoys the same performance as VoIP uplink from voice client 1. The two-way VoIP performance seems balanced in both ways compared to those in the DCF scenario as discussed in Chapter 4. The audio stream transmission just comes after VoIP streams and can achieve approximately 6ms - 7ms in average. As all VoIP packets are relatively small, the AP can poll audio client after a short period of VoIP transmission. Therefore, audio does not experience long delay either.

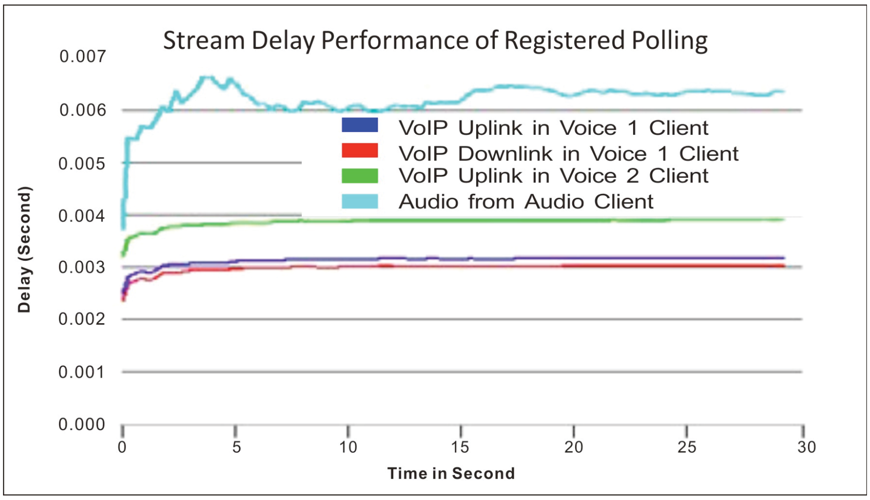


Figure 29 Stream Delay Performance of Registered Polling

As expected, Registered Polling can provide delay bound for VoIP in general and the individual performance of voice also depends on their polling order. The delay of VoIP mainly consists of large portion of medium access delay instead of queuing delay. To examine this, we adjust the superframe size for further simulations. We carried out simulation with the same conditions except that a 10ms superframe is used. The results show that delay performance of VoIP remains at 3ms - 4ms. Even the superframe size is further decreased; VoIP delay performance cannot be further improved. In fact, the delay performance of Registered Polling has already minimised the queuing delay in the clients. Medium access delay is a nature of the medium capability and does not depend on the usage of either implicit information update or explicit information update. Therefore, Registered Polling has well performed in respect to delay bound guarantee as expected.

- Further Analysis on Throughput Performance and Overhead Effects

Clearly, the total network throughput of Registered Polling only achieves approximately 2Mpbs, compared to a total throughput of 4.5Mbps as found in our DCF study in Chapter 4, with similar simulation environment and network configuration. This can be explained as polling schemes generally introduce more transmission overhead than contention schemes do.

The classic contention scheme produces overhead mainly by the InterFrame Space and backoff period. If RTS/CTS is used to secure the transmission, it can be up to 178 bytes per data frame. However, polling itself introduces 174 bytes overhead per data frame during the contention-free period (CFP); and this excludes the CFP set-up overhead (CFPini). Even though CFPini is shared by all the date frames within one CFP, it cannot be ignored as parts of total CFP overhead, as it can be up to 347 bytes, including beacon management frames. Table 6 below shows the general overhead of different schemes. These increased overhead produced in the polling schemes obviously results in throughput degrade with comparison to contention schemes. It is the cost of providing parameterised QoS in the polling schemes.

CP (RTS/CTS)	178 bytes
CP (per data frame)	96 bytes
CFP (per data frame)	174 bytes
CFPini (shared)	Up to 347 bytes

Table 6 General Overheads of Service Differentiation Schemes

Among all overhead, CFP (per data frame) overhead are fixed for each frame as long as polling scheme is applied; while CFP set-up overhead (CFPini) varies mainly according to length of superframe and data frames number in one CFP. Therefore, we mainly draw our attentions on CFPini overhead. In fact, our Registered Polling prefers a short superframe to meet the delay bound for delay-sensitive applications; it may suffer more on CFP set-up overhead (CFPini) as beacons are transmitted more often. This overhead is shared by the frames, which are transmitted in that CFP. Therefore, more frames are transmitted in CFP; the CFP setup overhead effect is less obvious. In other words, a longer CFP (longer superframe in our case) amortizes the setup cost over more data frames, provided that the network traffic configuration is unchanged [Pagtizs et al 2001]. In [Lindgren et al. 2003], such CFPini effect (superframe size) on medium utilisation is also reported. In fact, the results in the simulation also show that the network throughput would further decrease when the superframe is shorter at the value of 10ms. These figures raise the questions whether medium efficiency of Registered Polling can be improved with alternative design on superframe while also achieving delay bound guarantee for delay-sensitive applications.

Obviously, the problem does exist both in Registered Polling and Request Polling as long as polling mechanism is applied. We recall that when using a long superframe in Request Polling, it requires more frequent information update and more CCIs are used in superframe in order to have the same level of information precision. For example, if one CCI is used in a 20ms superframe, five CCIs may be required in a 100ms superframe. Therefore, the usage of long superframe in Request Polling increases more additional CC/RR signalling overhead while CFP set-up cost may be minimised. Therefore, increasing Superframe Length does not necessarily result in better efficiency in Request Polling. Furthermore, as discussed, if a same value of superframe is used for both Registered Polling and Request Polling, it is obvious that Request Polling has more overheads due to the heavy CC/RR signalling than Registered Polling. Therefore, Request Polling cannot necessarily avoid the CFPini overhead even a long superframe is applied.

Owing to such shortcoming of Registered Polling, we now question that whether Registered Polling has to stick with a short superframe in order to achieve delay bound guarantee for delay-sensitive applications. If Registered Polling can adopt longer superframe to minimise the CFP set-up overhead, it may be able to have less CFP set-up overhead effect. According to this challenge, we further provide observations and analysis, develop the solution and enhance our Registered Polling with Multi-Cycle scheme to solve the problem, which is described in the following sections.

5.4.4 Further Performance Observations for Registered Polling

As discussed above, polling methods generally suffer from medium utilisation mainly due to the overhead comparing to contention methods. In order to achieve better efficiency in Registered Polling, we further investigate the (superframe length) CFPini overhead impact on polling schemes performance. We firstly outline the CFPini overhead in details and then provide performance observations in different superframe scenarios. These analyses and observations would be the solid ground for us to propose the enhancement for Registered Polling to achieve better efficiency in the next sections.

Beacon Functionality and CFPini Overhead

CFP is initialized by a beacon transmission. At the very beginning of each superframe/CFP, a beacon is transmitted from AP as a management frame. Normally, in each superframe, there is one beacon. Therefore, in this case, superframe size is the same value of beacon interval. The frame body of a management frame of beacon contains the information of Beacon Interval, time stamp, SSID (basic service set identification), Rates, FH/DS (Frequency Hopping/Direct Sequence) set, CF set and TIM (traffic indication map). The functionality of beacon transmission is to provide the information to clients to initialize the incoming CFP. After the beacon frame is broadcasted and received by clients, all clients read such beacon and have knowledge about the coming CFP and CF from the beacon. Pollable clients would prepare to enter CFP and do not attempt to access medium by contention during CFP. Non-pollable clients set their NAV and do not interfere with polling medium access until the end of CFP. AP would then reset the polling list and start to poll clients from the top of the polling list. The initialization of CFP then completes.

Obviously, a short superframe may not be necessary for the purpose of informing clients about CFP, even though it provides slightly better delay performance for delay-sensitive applications in Registered Polling. Instead, if the beacons are transmitted repeatedly and more frequently; it then incurs more CFPini overhead. Another concern on short superframe is problem of delayed beacon. The beacon is scheduled at each target beacon transmission time. Because the medium may be in busy condition due to DCF traffic at the nominal beacon transmission time, beacon is then only transmitted after medium becomes free again. As defined in IEEE 802.11, in such case, the following CFP may be foreshortened by the amount of the delay, which is the time required to complete the current DCF frame exchange. In [Mangold et al. 2002], the worst case of 4.9ms delay is reported for delayed beacon. The delayed beacon problem seems more serious in a short superframe scenario.

However, if superframe is too long, clients may not be updated with precise information available at AP such as synchronization and CF information set. Particularly, hidden client may cause serious problem. In a long superframe, if they miss the previous beacon, before the next beacon is received, they would keep operating under DCF contention rules. Therefore, it may interfere with Contention Free access. It is against the design principle of polling schemes. The interfering period cannot be ignored as beacon interval is too long and the hidden clients may take long time to receive the next beacon. Such problems would be revisited when we present the enhancement to Registered Polling. Below, simulation results of different superframe scenarios are provided to further confirm its effect and assist in finding the optimal size.

Observation of Superframe Size Effect

As discussed above, in Registered Polling, a short superframe would allow clients to be polled more frequently especially for those clients on the top of polling list. Such high priority clients can send data more frequently. However, it increases the number of management frames sent, which might waste resources. Information contained in Beacon may not be necessarily needed for update so often. On the other hand, using longer superframe would reduce the amount of management frames sent. However, it causes the problem of high delay for delay-sensitive applications even they are placed on the top of polling list as they are polled less frequently and experience longer queuing delay waiting for the next CFP.

To validate the comparison between scenarios of different superframe sizes in Registered Polling, it is important to ensure that all scenarios have reasonable parameters setting that do not affect the performance of the scenarios adversely. We carry out simulations with same environments and network configuration as we did in Registered Polling above; and adopt different superframe sizes to observe the throughput effect of superframe size. Particularly, we also maintain the same ratio of CFP to superframe, i.e. 95%. Even though our network scenarios are overloaded, AP maximises available resources for providing parameterised QoS to voice and video rather than prioritised QoS to bulk data.

Maximising the Throughput Improvement with Long Superframe

As discussed above, adopting longer superframe can improve network total throughput. In this section, we try to find out the size range of the superframe that can support maximum throughput. We first investigate the throughput performance of two superframe sizes namely 100ms and 400ms. As shown in Figure 30, in the scenario of 410ms, the throughput experiences more peak variation than that of scenarios with 100ms superframes. As known, contention method generally achieves more network total throughput than polling method does. And in 400ms superframe scenario, its CP is a relatively longer period than another and its throughput is higher in this period. Therefore, the 'peak' throughput period is actually CP period in each superframe. In case of 100ms, even though it has the same ratio of CFP to Superframe; its CP period is shorter, appears more frequently and its peak is lower. These explain that the longer superframe is, the more obvious throughput variation it shows.

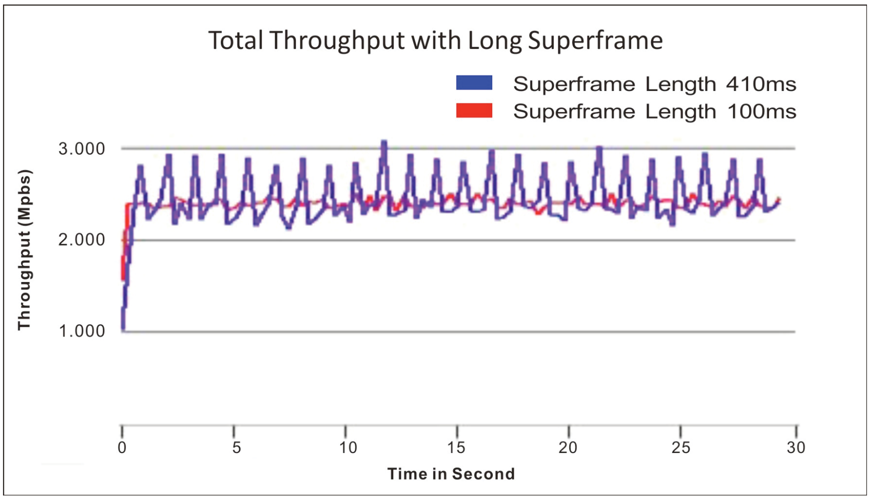


Figure 30 Total Network Throughputs with Long Superframes

However, their average throughput is very similar and achieves 2.4Mbps approximately, as shown in Figure 31. It is 20% improvement comparing to our result of 20ms superframe scenario. Although the improvement is as expected, it also indicates that even further increase superframe up to 800ms or more, the total network throughput would not increase any more. As shown in Figure 31, this indicates that there is no need to further increase the superframe length once it reaches 100ms. It is because when the CFP set-up overhead is shared by up to certain amount of transmitted frames, the CFPini overhead effect on network throughput becomes less obvious than other CFP overheads such as overheads caused poll and ACK messages.

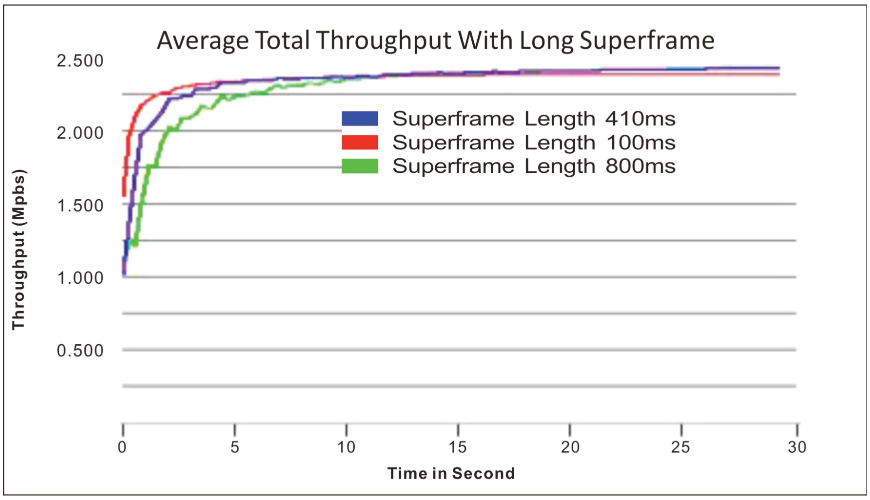


Figure 31 Average Total Throughputs of Three Long Superframes

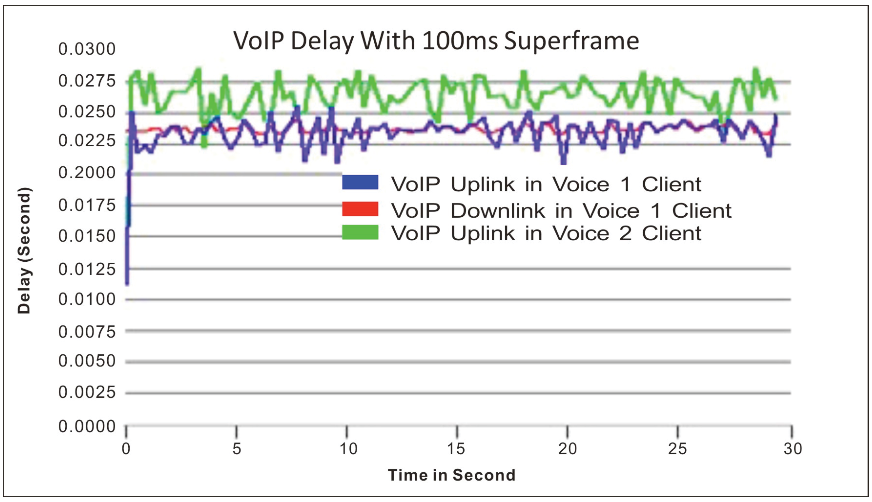


Figure 32 Voice Delay with 100ms Superframe

On the other hand, as we expect, all these scenarios experience large delay. Figure 32 shows average 23ms - 24ms delays of all three voice streams in the scenario of 100ms superframe. It is relatively higher than our initial result in 20ms superframe Registered Polling scenario. Such delay may not be good enough for a quality VoIP, considering that the voice packets may also be transferred in a long data path and each section in such data path contributes certain amount of delay as well.

Furthermore, long superframe not only causes the large delay as expected, but also leads to large delay variation as shown in Figure 33. Clearly, the Delay Cumulative Distribution Function of three Voice streams in the scenario of a 400ms superframe is presented. Up to only 5% of packets can have less than 50ms delay, while 80% of packets experience 100ms delay. Even echo canceller technique is used for voice transmission; such delay and delay variation are not acceptable.

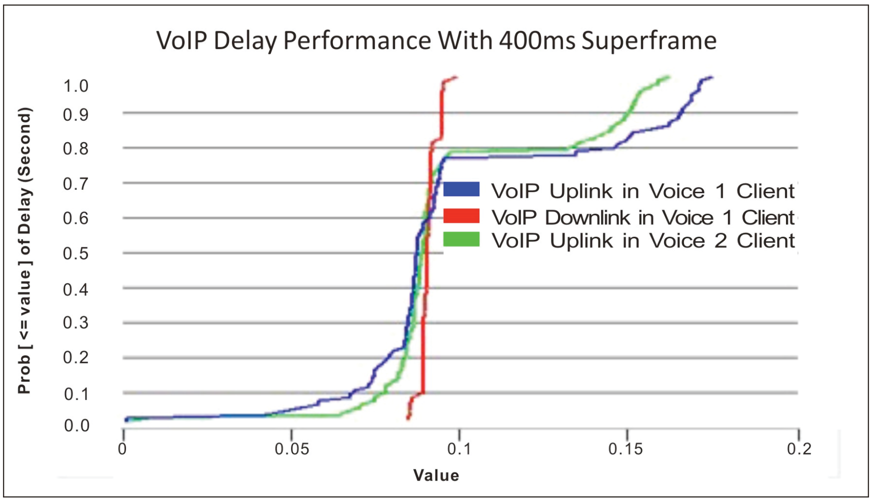


Figure 33 **VoIP Delay Cumulative Distribution Function With 400ms Superframe**

Maximising the Delay Performance for Voice Stream

As discussed in Section 5.4.3, the voice delay bound can be guaranteed with Registered Polling with short superframe of 20ms. Implementation of 20ms superframe is based on the reference to the inter-arrival time of Voice frame. It is known that delay mainly consists of medium access delay and queuing delay especially in our cases. While medium access delay largely depends on the physical characteristics of the wireless channel, the queuing delay can be minimised using different mechanisms such as short superframe. In this section, we investigate if the queuing delay can be even minimised with shorter superframe.

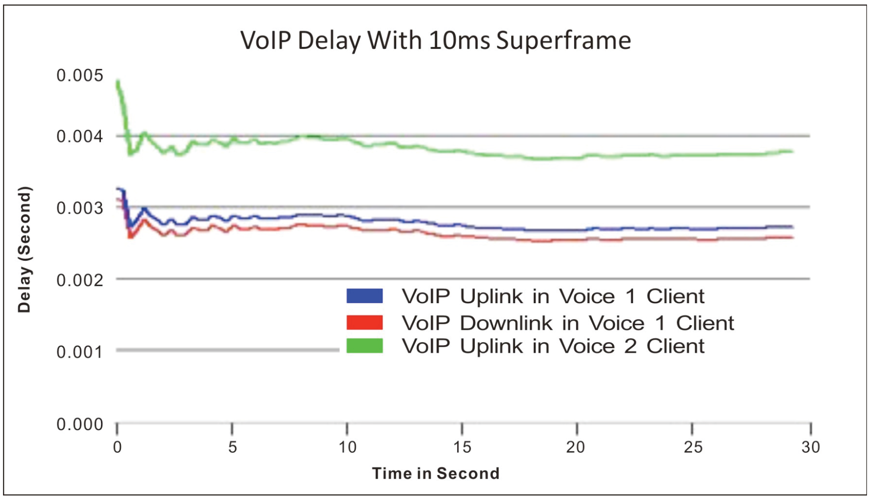


Figure 34 VoIP Delay with 10ms Superframe

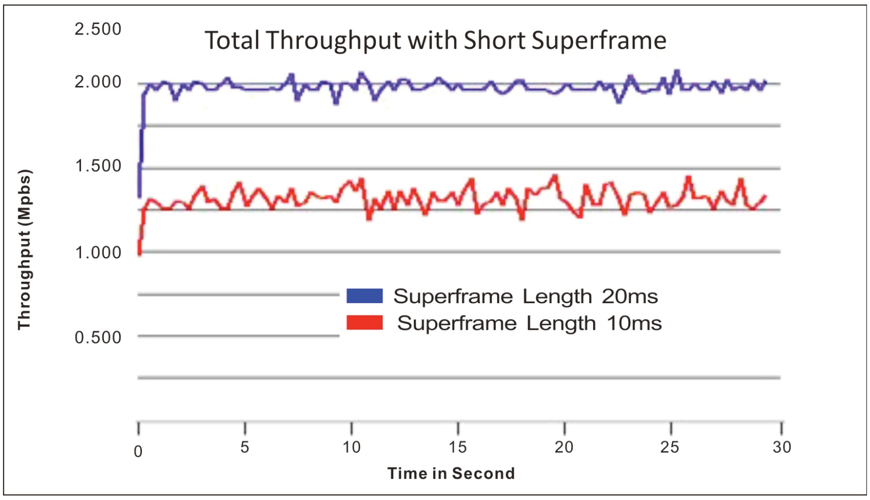


Figure 35 Total Throughputs with 10ms Superframe

Figure 29 has shown the Voice Streams Delay Performance of 20ms Superframe Scenario. The average delay of all streams is approximately within range of 3ms and 4ms. In order to find out if the delay bound can be further improved with shorter superframe, we carry out scenario with 10ms superframe and results are shown in Figure 34. Apparently, it does not have much improvement. It indicates that the delay performance is almost the best when the superframe reaches 20ms. In the scenario of 20ms superframe, queuing delay has been minimised and medium access delay contributes large portion of total delay. While further decreasing superframe size from 20ms, it cannot further improve the voice delay performance. However, the network seriously suffers utilisation downgrade in the mean time. In Figure 35, a large decrease of total network throughput in 10ms scenario is shown. Comparing to the throughput in 20ms superframe scenario, there is almost 35% decrease in throughput.

Finding the Optimal Point

Based on the observations above, we firstly conclude that scenario with a superframe, which is greater than 100ms, can achieve the maximum network total throughput, regardless of the delay effects. Secondly, delay performance cannot be further improved even with decrease of the superframe size when it has reached 20ms.

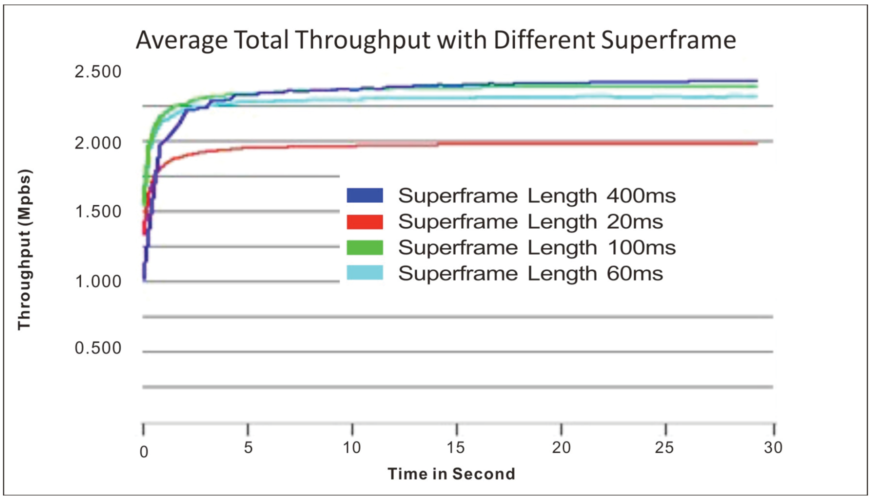


Figure 36 Average Total Throughputs with Different Superframes

In the following, we try to find out an optimal superframe size between 20ms and 100ms to balance the dilemma of throughput and delay bound guarantee in legacy Registered Polling. The delay and throughput performance and the superframe size, seem to have a approximately liner relationship as seen above. Therefore, we carry out the simulation with a 60ms superframe. As shown in Figure 36 and Figure 37, it maintains 2.3Mpbs throughput, which is slightly less than 2.4Mpbs of 100ms superframe scenario, and achieve average 13.5ms delay for voice streams. Obviously, such delay consists of certain portion of queuing delay. However, the overall performances of such scenarios are acceptable in comparison to the scenario of 100ms superframe and scenario of 20ms as expected.

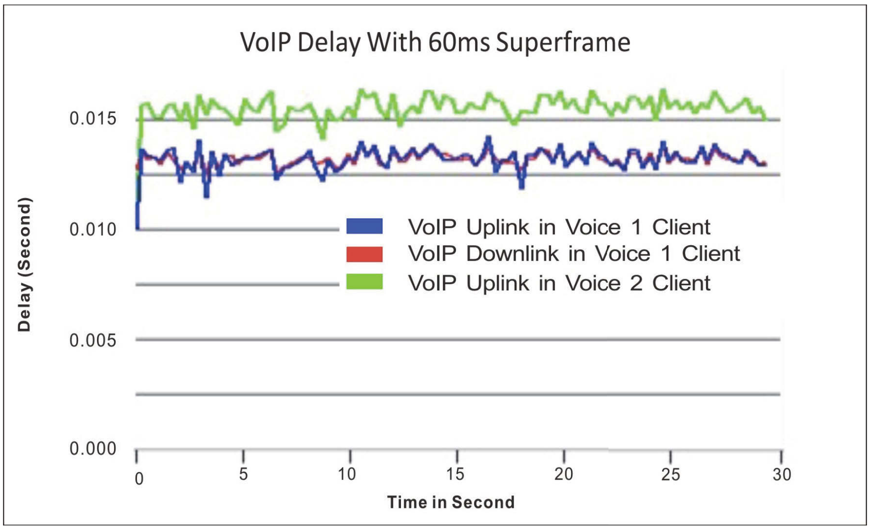


Figure 37 VoIP Delay with 60ms Superframe

It seems that a 60ms superframe is the optimal point for the balance of throughput and delay performance in such network configuration provided that there is no further enhancement on the Registered Polling method. However, 13.5ms delay performance in 60ms superframe scenario seems still not good enough for VoIP communication. In fact, voice delay bound guarantee should be taken as first priority in order to maintain a quality voice communication as promised at AP. And this motivates us to further design enhancement on the Registered Polling method to achieve better network utilisation while still providing delay bound guarantee at the range of 3ms and 4ms for voice streams. The design of the enhancement to Registered Polling below would be largely based on such observation.

5.4.5 Multi-cycle Enhancement

In this section, we propose Multi-cycle enhancement to Registered Polling to improve network utilisation while maintaining parameterised QoS for delay-sensitive applications in an overloaded WLAN system. Such enhancement itself becomes part of Registered Polling method and solves the problem of efficiency caused by using short superframe in legacy Registered Polling. Simulation of such Multi-cycle Registered Polling would be carried out to prove its efficiency and effectiveness afterwards.

The objective of such enhancement is to improve network utilisation based on Registered Polling method. However, such enhancement should not conflict with the QoS principle for delay-sensitive applications. In other words, delay bound guarantee must be achieved for delay-sensitive applications in enhanced Registered Polling. And network utilisation should not pay a cost of lower delay bound performance for delay-sensitive applications.

Determining the Superframe Size

To design a proper superframe size (Beacon Interval), we firstly consider the beacon functionality. Clearly, the information contained in each beacon should be regularly updated, such as time stamp and SSID; therefore, a superframe should not be relatively too long. In case of hidden clients, regular beacon broadcast would create more chances for hidden client to synchronise with AP and stop interfering with other clients. Furthermore, a short superframe would further downgrade the quality in the situation of Delayed Beacon. We consider 4.9ms delay as the worst case of delayed beacon [Mangold et al. 2002]. If a 20ms superframe is used and even CFP occupies large portions, a foreshortened CFP would be much shorter than normal CFP. The delayed beacon may be too obvious for such a short superframe and may cause problem in next CFP. A longer superframe may be easier to accommodate foreshortened CFP than a short superframe. Therefore, using a short superframe (Beacon Internal) would actually damage the functionality, which the beacon is initially designed for.

While avoiding the short superframe, we now consider the reasonable size. We firstly refer back to the classic 802.11 WLAN performances in [Crow 1997], where 410ms is used as beacon Interval. However, 410ms superframe would certainly have large impact on delay performance for those delay-sensitive applications. The results are not acceptable and we need to further investigate the possibility to reduce the size from 410ms to a lower value.

We recall the performance observations in respect of network efficiency in previous Section 5.4.3. As seen in Figure 31, when the superframe size reaches approximately 100ms, regardless of further size increase, the network utilisation improvement becomes less obvious as the CFP set-up overhead is shared by certain amount of frames transmitted during that CFP. Therefore, this provides us the foundation to reduce the size up to 100ms with respect to network efficiency.

As seen, in the scenario of 100ms superframe, the beacon functionality can be reasonably maintained and the network efficiency can be satisfied. However, its delay performance of VoIP applications (delay-sensitive) are still not good enough at 23ms -

25ms as shown in Figure 32. Further enhancement to Registered Polling is needed to satisfy the delay performance requirement; while a 100ms superframe is used to maintain the principle beacon functionality and network efficiency.

Enhancing with Multiple Polling Cycles

The objective of enhancement to Registered Polling is to maintain delay bound guarantee for delay-sensitive applications while a reasonably long superframe (100ms) remains necessary. The key to provide delay bound guarantee for voice-based delay-sensitive application is to keep granting TXOP to those voice streams repeatedly at every 20ms, which is almost the same value as the inter-arrival time of voice frames. However, in most polling scheme, polling list is reset only at the beginning of CFP and most of voice streams have to wait for an entire beacon interval to get next TXOP for transmission (even though they are scheduled on the top of polling list) after such client has been served. According to this, we create 'virtual polling cycle' within CFP to reset the polling list and repeat the polling every pre-defined period, which should be a portion of beacon interval. Therefore there are multiple cycles in each CFP and polling is repeated at each cycle instead of each CFP.

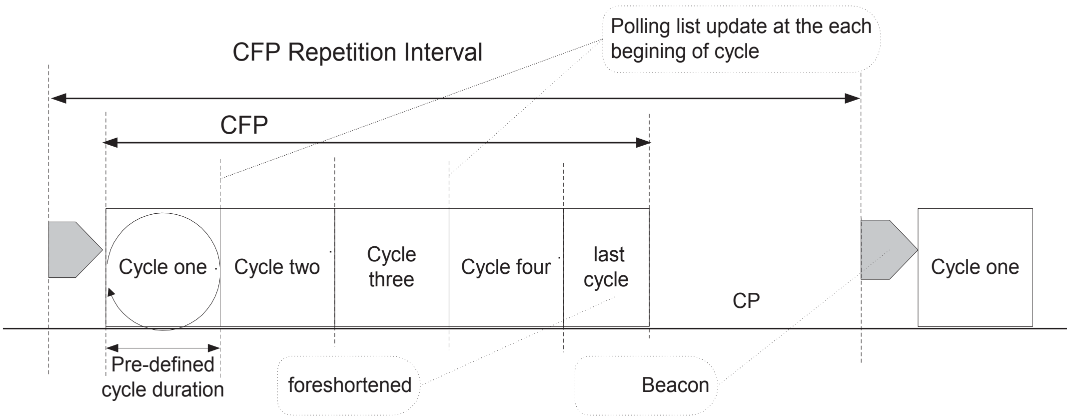


Figure 38 **Multi-cycle Enhancement**

We propose that individual cycle duration is 20ms referring to the inter-arrival time of VoIP frames and each cycle try to maintain the same duration. During the entire polling period, CFP consists of a number of such cycles. As the CFP may not be the exact multiple of a number of equal (approximately) cycles, the last cycle may experience shorter period than others. For example, within a 150ms CFP in a 170ms superframe, there are 8 cycles and the last one is about 10ms. At the end of each CFP, it complies with all the CFP ending rules. When time is up for next cycle, the ongoing transmission would continue until it is completed, therefore, such cycle may be a bit shorter than its pre-defined period. Hence, most cycles have only approximately equal duration. As it is only timing for polling list reset/update, it does not affect any other operations of PCF.

During each cycle, AP performs polling as the same as that in Registered Polling and at the beginning of each cycle, additionally there is regular polling table update and determination of polling order. As the cycle is a shorter period than a CFP, once the AP gains stream information update implicitly and recognise the need for change in the polling list, it can implement such change in time when entering the next cycle. Therefore, the updated polling order would meet the needs of delay-sensitive streams more precisely in time.

The implementation of Multi-cycle enhancement is also straightforward. We recall that in Registered Polling, AP polls clients following a polling order with a pointer. The pointer starts from the top of the list and the client on the top of list get polled firstly. After such client completes all transmission of its queued frames, pointer points to the next client, which is second in the list. Pointer moves to the bottom of the list at the end of CFP. In our multi-cycle mechanism, polling scheduling is enhanced with a pointer reset at the end of each cycle. An interrupt process is used as a timer to reset the polling order. The reset enables the polling pointer move back to the top of the polling list at the beginning of next cycle. Therefore, Voice streams are able to access medium by being polled again repeatedly at the beginning of each cycle within a single long superframe as long as it has not completed its session transmission. Other Pollable applications also enjoy such multi-cycle enhancement as it reduces their queuing delay in a long superframe in the same principle.

With such multi-cycle enhancement, Registered Polling is able to provide delay bound guarantee for delay-sensitive applications, within a relatively large superframe, which ensures better efficiency. Even if there is no active delay-sensitive TS in the network, Registered Multi-cycle Polling also optimises other application delay performance and improves total network throughput more significantly. We prove such judgments with the simulations as demonstrated below.

5.4.6 Performance Analysis of Registered Multi-cycle Polling

To validate the improvement of Registered Polling with Multi-cycle enhancement, it is important to ensure that legacy Registered Polling scheme and Registered Multi-cycle

Polling (RMP) scheme have reasonable parameters setting that do not affect the performance of the schemes adversely. We carry out simulations with the similar environments and network configuration as that in legacy Registered Polling scheme. A 100ms superframe with a number of 20ms cycles is implemented in the simulation. Again, we maintain the same ratio of CFP to superframe as before.

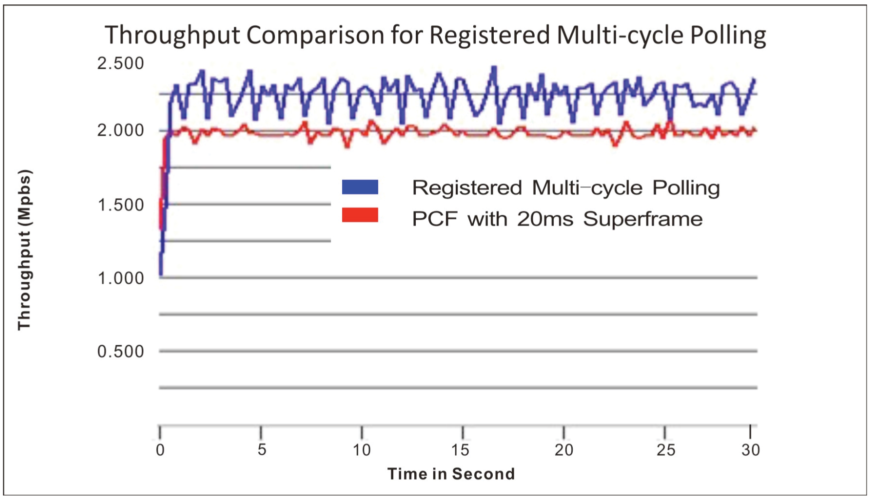


Figure 39 Throughput Comparison for Registered Multi-cycle Polling

Firstly, the throughput comparison between Registered Multi-cycle Polling and legacy Registered Polling is shown in Figure 39. The throughput improvement of Registered Multi-cycle Polling against to the legacy Registered Polling is obvious. While legacy Registered Polling reaches 2Mpbs total throughput, Registered Multi-cycle Polling achieves an average of 2.25Mpbs total throughput. It provides average improvement up to approximately 12.5% in such network configuration. The reason for such improvement is the usage of long 100ms superframe in Registered Multi-cycle Polling compared to 20ms in legacy Registered Polling. The analysis has been well studied in the above sections. The throughput variation is also observed as one of by-products of the usage of longer superframe. This has also been well explained in previous Sections (The network has higher network throughput during CP compared to those during CFP).

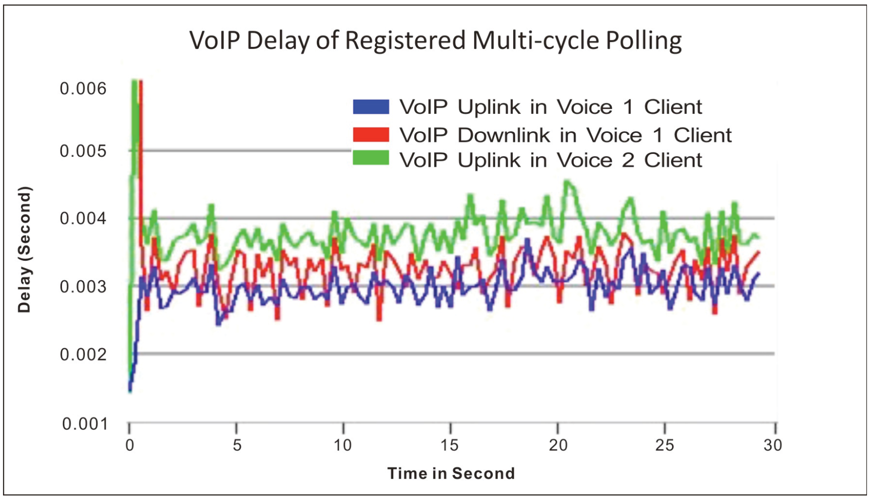


Figure 40 VoIP Delay of Registered Multi-cycle Polling

More importantly, we examine if Registered Multi-cycle Polling can meet the delay principle while the network efficiency has been improved as above. As shown in Figure 40, Registered Multi-cycle Polling is still able to achieve delay bound in the range of 3ms to 4ms for all voice streams. Comparing to the delay performance (average 23ms-24ms) in legacy Registered Polling as shown in Figure 34, it has large improvement and has minimised the queuing delay for voice streams transmission over the wireless channel to meet the delay requirement. In addition, the effect of ordering in the polling list has also been shown. Voice stream 0 of voice client 1 over-performs voice stream 2 of voice client 2, as voice client 1 is placed on the top of the polling list and polled firstly. In summary, Registered Multi-cycle Polling does improve network utilisation while maintaining parameterised QoS for delay-sensitive applications in an overloaded WLAN system. It achieves its design objectives indeed.

5.4.7 Performance Analysis of Registered Multi-cycle Polling in a Scenario of Large Data Applications

Furthermore, we examine the network efficiency improvement in situations that the network is only overloaded with large delay-tolerant application such as Video Conferencing and bulk data, instead of mixture of delay-sensitive and delay-tolerant applications. Therefore, we carry out simulations for Registered Polling scheme and Registered Multi-cycle Polling scheme; and disable all voice streams and audio stream in both schemes. Other network configuration and environment remain the same as before, where 20ms and 60ms superframes are configured for Registered Polling scheme and Registered Multi-cycle Polling respectively.

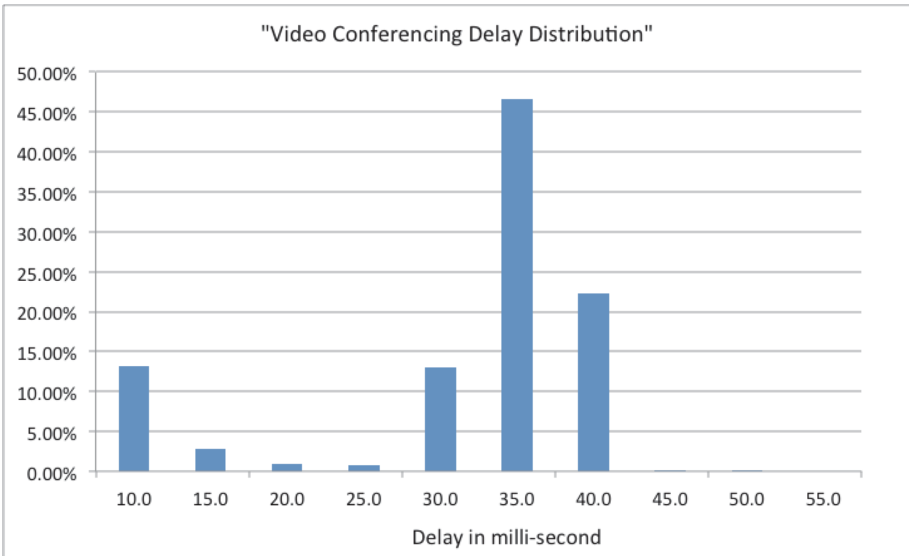


Figure 41 Video Conferencing Delay Distributions

Firstly, delay performances of Video Conferencing are studied in Registered Multi-cycle Polling. As seen in Figure 41, the majority of Video Conferencing delays are within 30ms and 40ms, while the Maximum of 47% of collections is at a value of 35ms approximately. Such delay performance of Videoconference streams in the enhanced schemes is acceptable given that 100ms is generally considered from end to end communication for Video conferencing.

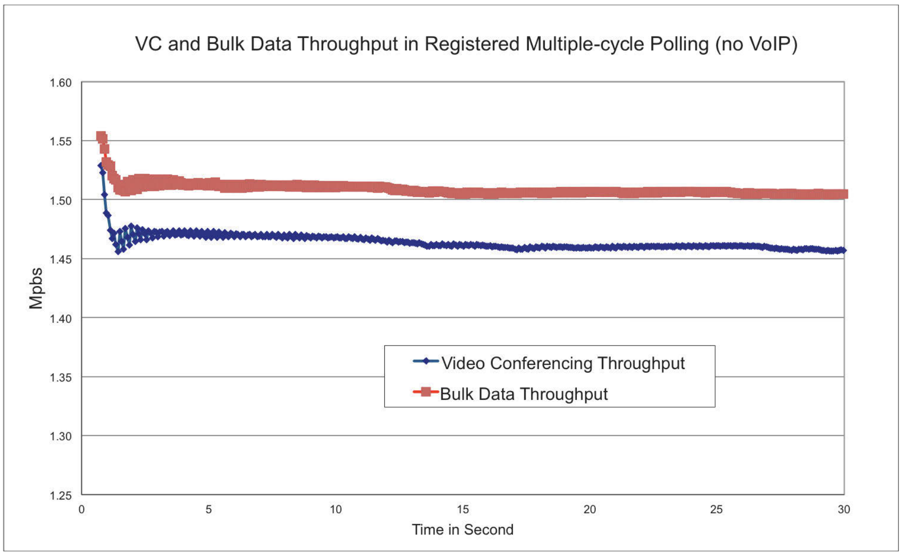


Figure 42 VC and Bulk Data Throughput in Registered Multi-cycle Polling (no VoIP)

Figure 42 shows us that video conferencing could gain full throughput at its 1.46Mpbs in average. This is in line with its satisfactory delay performance we have seen in Figure 41. Obviously, the performance is guaranteed by placing the video conferencing on the top of the polling list, which enabling its gaining parameterised opportunities for transmission. The remaining opportunities are occupied by the aggressive bulk data at approximately 1.52Mpbs, which is also shown in Figure 42.

We repeat such experiment with PCF that polls stations in 20ms superframe, also with Bulk data and Video conferencing only. In Figure 43, as by placing the video conferencing on the top of the polling list, video conferencing could still gain full throughput at its 1.46Mpbs in average. However, the aggressive bulk data could only have 1.22Mpbs in average. That is a reduction of 300Kpbs compared to that in 60ms superframes Registered Multi-cycle Polling. 24% throughput loss is considered caused by the frequent overhead of 20ms superframe. As discussed in Section 5.4.2, the maximum CFP overhead could be as much as 347bytes every 20ms. Therefore, the legacy PCF scheme has 2.68Mbps total network throughput, while our enhanced scheme over-performs at 2.98Mbps. As the Bulk data application is not given any priority, it experiences most of the packet loss. In contrast, Registered Multi-cycle Polling scheme not only mostly satisfies two Videoconference streams in bandwidth allocation, but also allows more bulk data share the channel effectively.

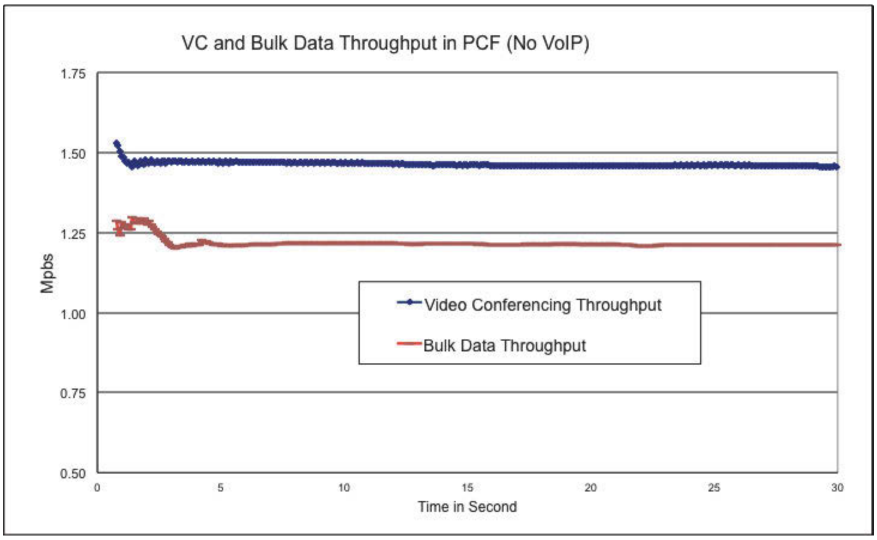


Figure 43 VC and Bulk Data Application in PCF (no VoIP)

The results also show that the throughput improvement is even more significant when there is in a scenario of only large data application. Referring to result in Figure 36, we have seen throughput of 2.30Mpbs in 60ms Registered Multi-cycle Polling where there are three VoIP and Audio streams. Without three VoIP and Audio streams, Figure 44 shows us a throughput of 2.98Mpbs, with improvement of 30%. It is understandable that small and frequent packets like VoIP indeed requires relatively large transmission overhead, even they are small in packet sizes. Therefore, network operators should always monitor network traffic profile in order to adjust network configuration for the best performance of their network.

In conclusion, Registered Polling can achieve parameterised QoS guarantee, particularly in an overloaded system, using implicit information update method. Queuing delay has been minimised for delay-sensitive applications on transmission over the wireless channel without any additional signalling overhead. Due to its shortcoming of network efficiency caused by the usage of short superframe, we propose Multi-cycle enhancement in Registered Polling and prove that Registered Multi-cycle Polling can over-perform legacy Registered Polling in respect of total network throughput, while it retains the original principle of Registered Polling to provide delay bound guarantee for delay-sensitive applications.

5.5 Summary

In the last few chapters, we have described an entire picture of service differentiation for IEEE 802.11 WLAN above. Firstly, session-based applications and non-session-based applications are distinguished according to their session set-up requirements. A hybrid QoS framework is then outlined at the level of Admission Control to ensure the QoS guarantee commitments for admitted session-based applications and protect traffic violation. Concept of two levels Service Differentiation is introduced, where session-based applications are provide with parameterised/hard QoS with polling schemes and non-session-based applications are provided with prioritised/soft QoS with contention schemes. The medium still alternates between CP and CFP. During CP, non-session-based applications access medium via contention methods such as EDCF. The contention method helps improve network utilisation while also provides prioritised QoS for non-session-based applications.

In this chapter, we then focus on providing parameterised QoS using polling schemes and propose Registered Multi-cycle Polling. It not only meets the delay bound guarantee for delay-sensitive applications effectively, but also improves network utilisation due to the usage of implicit network information update for scheduling of polling order. In Registered Polling, polling table is updated in an implicit method, rather than an explicit method where separate request transmission is used and therefore additional overhead is introduced. Multi-cycle enhancement further minimises the effect of CFP set-up overhead caused by using longer superframe while still maintains strict delay bound for delay-sensitive applications.

We have examined the performance of Registered Multi-cycle polling. It works well especially in a situation of overloaded system, where QoS guarantee for session-based applications is crucial to the network performance. The scheme not only provides parameterised QoS for these session-based applications, but also improves the total network utilisation. These objectives are achieved with a couple of novel QoS mechanisms, such as Multi-cycle enhancement and implicit information update.

However, there is a need of the co-ordination between the Registered Multi-cycle Polling method and contention method for allocating resources for session-based application and non-session-based applications either with an underloaded system or with an overloaded system. In fact, it also implies the needs of dynamic QoS control over varied network scenarios, which may be caused by different traffic load and physical channel conditions. Such dynamic controls may associate with policy management and extend to a scope of internetworking between WLAN themselves, between WLAN and Access Network, or even between 802.11 WLAN and HiperLAN. In this case, the proposed hybrid QoS framework would play an important role in it, which leads to meaningful future works.

~ The secret to success is to do the common things uncommonly well ~

(John D. Rockerfeller Jr.)

Chapter 6 Service Differentiation and Resources Coordination in Multi-cell Overlapping BSS Environments

6.1 Introduction of OBSS

In previous Chapters, we have investigated the MAC protocol enhancement in order to support QoS. It was found that the contention schemes based on principles of the DCF have certain shortcomings in supporting QoS in 802.11 WLAN, such as fairness and imbalance of uplink and downlink. Therefore, contention-free schemes, based on a polling mechanism, are suggested for resource allocation with more central controls. In order to improve efficiency and provide time-bound service for delay-sensitive applications, Registered Multi-cycle enhancements have been designed and evaluated. However, these studies are all based on the scenario of one single BSS.

In reality, hot spots, like airport, university campus and office building, are normally equipped with Multiple BSSs. As the user in these hot spots may be in a high density while the capacity of single BSS in user number is quite limited, these BSSs may be located in a close distance. Therefore, their transmission range may overlap fully or partially and this overlapping causes some interference among each other. In such case, it is considered as Overlapping Basic Service Sets (OBSS). Generally OBSS is referred as two or more BSSs running at the same channel, where the transmissions by some clients belonging to one BSS affect some stations (clients) in the other BSSs.

One of the obvious effects on QoS from OBSS may be collision on polling mechanism. In the situation of overlapping BSSs, the contention-free transmissions are not guaranteed due to the possible channel contentions from clients in the co-located and/or adjacent BSSs. That means, consequently, it makes very difficult to support QoS in the situation of Overlapping BSSs (OBSS). However, supporting QoS is an essential and important problem, which should be solved in order to make 802.11 WLAN more realistic in the era of multimedia.

In this Chapter, we first consider the probability of having OBSS in the reality and define the scope of the problem. Then we point out the essential of overlap management. An

OBSS coordination scheme to support QoS is presented for both downlink and uplink transmission. The design details of OBSS resource allocation system are then given, including message collection mechanisms and the downlink channel assignment, using graph colouring technique.

6.2 Literature Review and Observations on OBSS Interference

Wireless LANs operate in the unlicensed portions of the spectrum, where they provide interference-free simultaneous transmissions on multiple channels; each cell (BSS) transmits on a single channel. The number of channels available varies with the spectrum allocation and the physical layer technologies. (In IEEE 802.11b & g, it provides 3 channels for duplex data transmission in the 2.4 GHz ISM band; while IEEE 802.11a standard provides 12 channels in the 5 GHz ISM band.) In particular, the interference–free channels are very limited. Considering when several BSS locate closely and some of them would then be assigned the same frequency channel, co-channel interference would occur accordingly [Shan et al 2012].

A number of researches, such as [Dunat et al. 2004][Prabhu et al 2012], points out the inter-cell co-channel interference between overlapping cells, and concludes that it largely reduces the network performance. Some simulation results are provided in [Balachandran et al 2009][Mangold 2003][Nguyen et al 2011], regarding to throughput and delay performance under OBSS. It shows that a variety of delay and throughput are observed, in OBSS, depending on the degree of overlapping, the number of BSS overlapped together. Even though throughput decrease is partially caused by the limited physical channel, while the user capacity increases, the additional collisions from OBSS lead to further reduction in throughput and significant increase in delivery delay; Furthermore, such delay and throughput reduction cannot be forecasted and are really out of control [Nethi et al 2011]. The neighbourhood capture also has been pointed out in [Benvensite 2002]. Its effect can be understood that besides they experience the normal co-channel interference, medium can be captured unfairly by certain BSSs. In the following section, we will further give analysis on how co-channel interference is occurred and affected by BSS distance, medium access scheme and neighbourhood capture.

Inter-BSS Distance Effect

OBSS only happens if two or more BSS operate on the same channel, while they are located within certain 'close' distance. According to the distance between these two APs (which correspondently controlling two relevant BSSs), we categorise the types of overlapping between two BSS, AP-overlapped and non-AP-overlapped. In non-AP-overlapped scenario, none of these tow APs is overlapped by any of foreign stations or foreign AP. Even though the transmission range of two AP stations does not overlap each other, those foreign stations, which locate between two AP stations, may locate close enough to interfere the AP. Figure 44 shows an example of two non-AP-overlapped BSSs. A circle around each client (and AP) represents the transmission range of the client. $client_{x,1}$ belongs to the first STA of BSS_x and is controlled by AP_x. This example represents a very problematic situation, since $client_{2,1}$ is not in the coverage area of AP_1. $client_{2,1}$ cannot receive beacons transmitted by AP_1, and hence it would not set up its NAV during the CFP of BSS_1. Therefore, $client_{2,1}$ can transmit a frame while $client_{1,2}$ is receiving a data frame from AP_1 during a CFP, thus resulting in a collision.

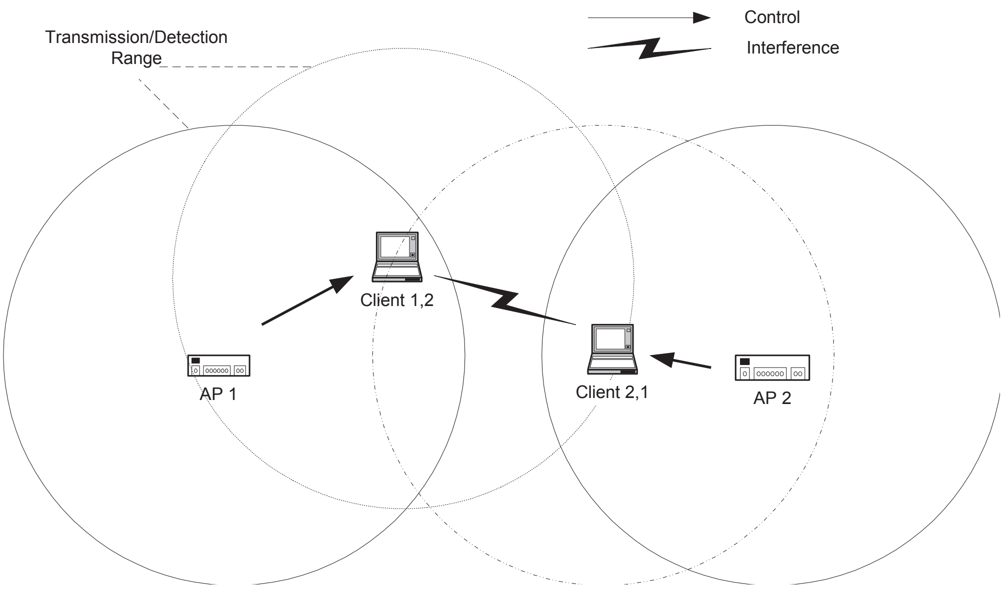


Figure 44 Non-AP-Overlapped: $client_{2,1}$ is not reachable from AP_1.

The AP-overlapped scenario happens when two AP_x move closely and their transmission ranges overlap each other. In this case, at least one of AP_x would be overlapped by some of foreign clients or AP. The example in Figure 45 shows such situation, where $client_{2,1}$ will not initiate a data transmission during the CFP of BSS_1 after hearing the beacon frame from AP_1

at the beginning of a CFP. This leads to large delay and reduced throughput in $client_{2,1}$ additionally. The same co-channel interference between $client_{2,1}$ and $client_{1,2}$ still exists here as discussed above. As the distance of two BSSs is getting closer, more stations may actually hear signals from other BSS, the interference effect is even worst and collision rate may increase.

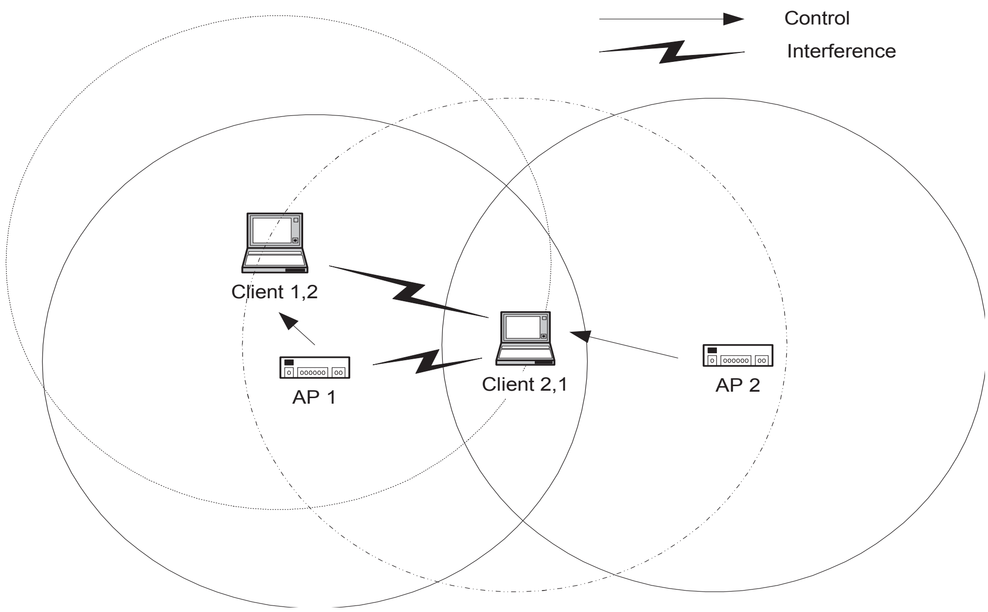


Figure 45 **AP-overlapped Scenario: $client_{2,1}$ is reachable from AP_1.**

One of the distinctions of these two types of overlapping between two BSS is distance between these two BSS. Obviously, the closer the distance of two BSSs is, the higher the potential of interference would be. In other words, AP-overlapped situation may experience worst interference than non-AP-overlapped situation. Therefore, more attention should be paid to the case of AP-overlapped situation. In addition, as AP normally plays as the gateway of all uplink and downlink traffic to the distribution system and the coordinator for contention-free medium access, if AP is overlapped, the impact normally is more serious.

Overlapping Effect on Medium Access Scheme

As discussed above, location distance and frequency distance would make impact on the levels of interference. Besides that, the medium access schemes operating in the interfering parties can also make an impact. The co-channel interference at least involves two parties, and these parties may operate either on contention-based medium access scheme or polling-based medium access scheme. As each BSS has full choices of using different schemes, we analyse this problem between two BSSs, in three categories, when single access scheme (either polling or contention) is used or when both schemes are used simultaneously by two parties.

At first, we look at the scenario if both two parties adopt contention scheme when they overlaps each other. The fundamental of contention scheme is Carrier Sense Multiple Assess /Collision Avoidance. Both DCF and EDCF define a Collision Avoidance mechanism to reduce the probability of collision. Therefore, even two BSSs overlap each other, station can still use such Collision Avoidance mechanism to avoid any potential collision, not only with stations in its own BSS, but also with stations in other BSS, which it can detect. However, hidden terminal problem becomes a bit worse as the entire system has more station and covers wider areas. However, the medium can be still fully utilised in such a random access scheme while collision rate may be acceptable. When more than two parties overlap altogether and they are all operating on contention scheme, neighbour capture effect is also needed to be considered. In [Benvensite 2002], channel slotting and synchronization are suggested to eliminate the neighbourhood capture effect.

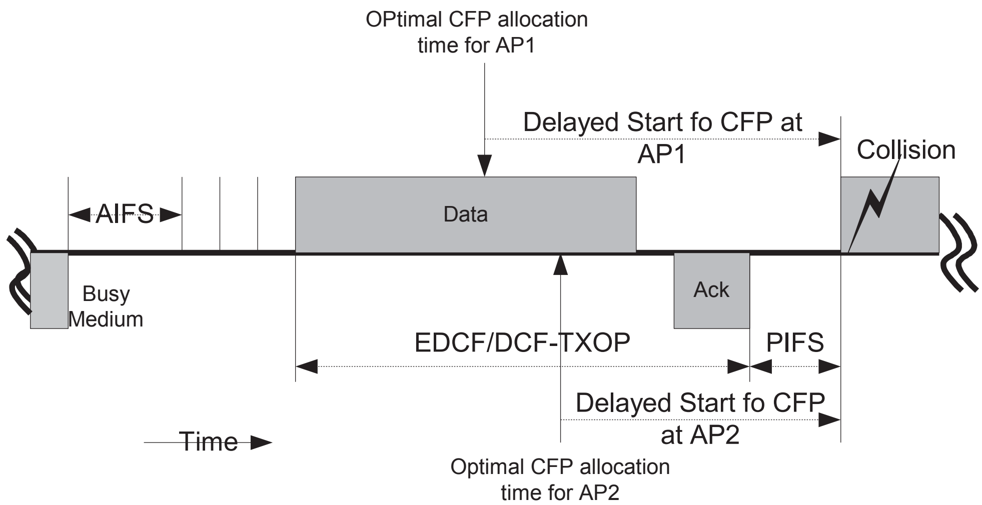


Figure 46 **Frame Collisions of Two Delayed CFPs**

In the case that both two parties adopt polling scheme, the interference effect may be even worse. The main reason is that when the polling scheme operates without the CSMA/CA contention window randomization and backoff operations in the contention scheme, there is a risk of repeated collisions if multiple, overlapping, point-coordinated

BSSs are operating on the same PHY channel, and their CFP Rates and beacon intervals are approximately equal. Figure 46 shows the beginning of collision of two CFP in such situation. In [IEEE802.11 1999], to eliminate such problem, a random backoff delay and a DIFS are used to start a CFP when the initial beacon is delayed because of deferral due to a busy medium. However, it seems to be not sufficient enough to eliminate all possible interference, because both AP may not hear from each other. Even two AP can hear from each other, the medium can also be grabbed by foreign station, which is hidden to another AP and performs additional backoff and DIFS. In the case of these repeated collision, contention free transmissions are not guaranteed at all. That means, consequently, it makes very difficult to support QoS, which is supposed to be promised in polling scheme as its design fundamental. OBSS becomes one of critical challenges for polling scheme.

In the case of combination of polling scheme and contention scheme for two parties, which overlap each other, interference becomes complex. Note that polling scheme always has priority over contention scheme. Therefore, BSS, which operates on polling scheme, may capture the medium over another BSS, which operates on contention scheme. And such capture may not be fair to the BSS, operating on contention scheme, even though collision may not be serious. Furthermore, in the case of non-AP-overlapped scenario, all stations may not be in the detection range of a foreign AP, which operates on polling scheme. Those stations, which are out of range of detection of such AP, become hidden to this AP. They perform channel contention regardless of polling period in other BSS and cause serious collision.

In summary, QoS supports appear to become problematic when BSSs overlap, regardless of how access schemes are used when there is no co-ordination in place among BSSs, even though the interference severity is different according to adoption of schemes among BSSs. As we propose resource management on MAC protocol to allocate resource among BSSs in order to re-use the frequency channel while avoiding co-channel interference, it would be very useful to keep this analysis in mind and see how these problems are solved.

Neighbourhood Capture Effect

The neighbourhood capture effect can be understood by considering three BSSs locate closely and are assigned the same frequency channel. Besides that they experience the normal co-channel interference, medium can be captured unfairly by certain BSSs. As illustrated in Figure 47, each BSS can only detect the transmission of its respective neighbour BSS. So BSS 1 and BSS 3 are hidden to each other and they operate independently at the same time. While all stations in three BSSs use CSMA protocol to access medium, a station or BSS may retain from transmitting while it detect the channel busy and transmission is deferred. Once a station in BSS 1 starts a frame exchange, stations in BSS 2 would defer transmission as they can detect the current BSS1 transmission. However, it implies that the stations in BSS 3, which is hidden to BSS 1 and do not detect the BSS 1 transmission, can start its frame exchange independently during the BSS 1 transmission. This can be repeated as long as BSS 1 and BSS 3 have data to transmit. As a result, stations in BSS2 may have no opportunity for transmission for long period, while the medium is captured by the neighbourhood of BSS 2. This leads to large delay and reduced throughput in BSS 2 with such unfair treatment.

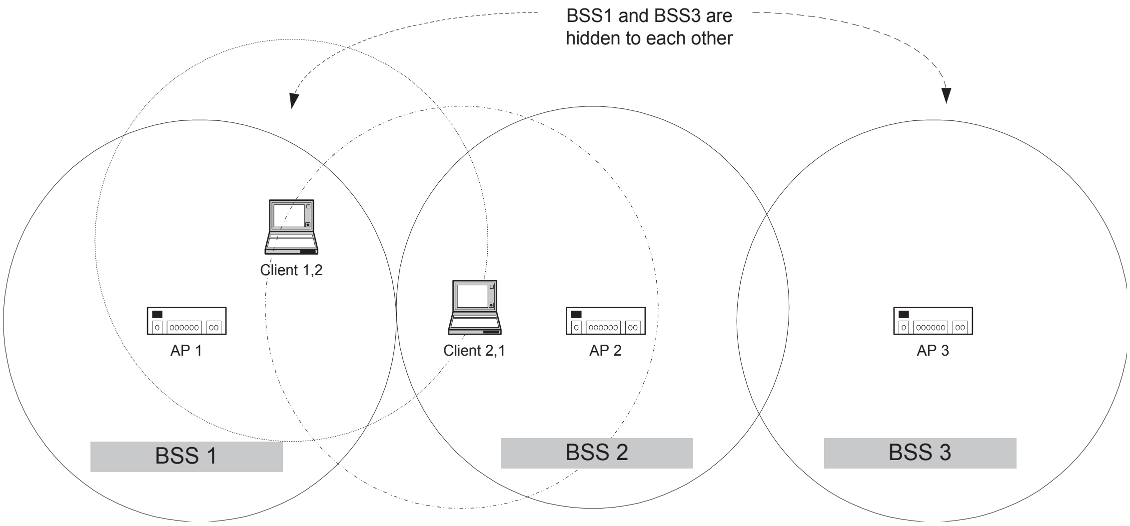


Figure 47 Neighbourhood Capture

6.3 OBSS Coordination Essential

Avoiding OBSS

The best way one can do is to avoid the BSS overlapping situations. Most of researchers approach the problems based on the Dynamic Frequency Selection principle by assigning different frequency channels to co-located and/or neighbouring BSSs. This can be done manually by the system administrator in some WLAN environments. However, for example, in the home environment where a professional system administrator is not available, this kind of manual channel selection may not be a feasible solution. Therefore, the new MAC should have a function to select the channel dynamically and adaptively by observing the channel condition and scanning the status of other possible channels. IEEE 802.11 committee has such scheme, named Dynamic Frequency Selection (DFS), as part of the 802.11h supplement standard. [Collotta et al 2012] [Matsunaga and Katz 2004] and [Hui and Shankaranarayanan 2004] show such implementation and propose resource allocation on the limited number of available channels across WLAN domains by introducing network-initiated load balancing based on workload measurements. Their researches aim to result in the best channel for an overlapping BSS.

However, as discussed previously, channels for WLAN are very limited in an unlicensed band. The limited availability of channels implies that they must be re-used, much like in cellular communication network, especially it comes to deploy in a large areas. On the other hand, when these channels are re-used, two BSS having the same physical channel must have a certain physical distance. Such distance is considered reuse distance and the optimal selection of reuse distance is a complex issue being under discussion at the moment. However, in order to have interference free, the re-use distance at least may be considered as 2-3 times of the normal transmission range, which is already an issue for a highly-populated area. Therefore, when WLAN comes to service a highly populated area, which has higher communication demand with high traffic load, reuse distance and limited availability of channels make OBSS avoidance impractical [Ding et al 2008][Liu et al 2010]. In time of high offered traffic and a large number of BSSs, it is desirable that BSSs can share a single frequency channel without interference and at the same time, QoS is supported as if they do not have to share a frequency channel with competing BSS.

Consequently, multiple-cell WLAN must rely on certain coexistence/ coordination on MAC protocol to allocate resources among BSSs in order to re-use the frequency channel while avoiding co-channel interference between BSSs. Some management schemes, such as reduction of transmission power and selection of robust PHY mode may help to further reduce OBSS effect, but it is out of the scope in this research.

Existing Coexistence Schemes

Most of early researches suggest that decentralized approaches could fix the problems. In [Mangold et al. 2002b], Mangold uses EDCF-TXOP bursting mechanism to gain improvement at the cost of potential increased MSDU delivery delay of other streams. He also proposes game theory to allocate resources among overlapping WLAN and Bluetooth, with assumption of absence of central coordination in [Mangold et al 2003c]. [Prabhu et al

2012] also discussed similar decentralized clustering algorithms for dense wireless sensor networks. In [Cheng et al 2012] [Chiasserini and Rao 2002] [Golmie and Rebala 2003] [Chevrollier et al. 2003] [Xhafa et al. 2007 & 2008] [Erceg et al. 2008][Sherman et al 2008], similar coexistence mechanisms based on Clear Channel Assessment and traffic scheduling techniques, are introduced to mitigate interference between systems.

However, in recent years, we have seen much higher density in real deployment. The decentralized approaches become impractical obvisouly. Centralised Control is introduced in [Vergados and Vergados 2004] [Rastin et al. 2006] [Han et al. 2009] and [Eladly and Chen 2004]. Vergados proposes 'Super Point Coordination' to work out possible collision sets, which is only useful in cases where there is only a small proportion of the connections or in small networks. Eladly studies similar disciplines (virtual access points) on top of 802.11 for the interaction among overlapping IBSSs.

Most recent studies of [Collotta et al 2012] [Wang 2012] [Liu et al 2010] studied load balancing among BSS to avoid congestion and so then overlapping at a higher cost of large signalling traffic. [Diepstraten and Haagh 2012] [Nakajima and Zhang 2010] [Prabhu and Sophia 2012] [Diepstraten and Haagh 2012] proposed technique to adjust antenna power and gain to optimize overlapping situations, where at the same time, limiting the total throughput performance. [Cui et al 2011][Hou et al 2012][Nethi et al 2011] present good ideas on channel assignment and bandwidth segmentation in certain particular dense environments. All of these studies either focus on coexistence between two standards when there is no central coordination, or use PHY technique to minimize interference as the first priority, or lack of scalability of the problem even when central coordination is proposed.

Essential of Coordination in OBSS WLAN

We have seen the problems of OBSS in aspects of medium access schemes, distance and neighbourhood capture and recognised the needs for coexistence scheme for overlapping situations. Compared to these coexistence schemes, we further emphasize the coordination essential to address interference problems, particularly in the common scenario that a WLAN domain is deployed in a high-density environment with central resources management function. Such coordination should also provide scalability to the problem compared to the coexistence schemes above.

Firstly, QoS supports can only be provided with OBSS coordination when under OBSS interference in a multi-cell WLAN domain. As known, QoS streams are highly time-repetitive in nature. Under the polling scheme, collision due to OBSS interference will be highly repetitive as well, which cause high failure rates for these QoS streams. Even worse, such failures are usually not hidden to AP and out of control of AP. These all mean no QoS guarantee at all. Even if under contention scheme, collision due to overlap are more random than that in a single cell, so it produces a higher retry rate, and it translates into more delay and lower throughput.

Secondly, Coordination for the usage of polling scheme and contention scheme in OBSS environment has to be considered for bandwidth reuse purpose. Bandwidth reuse is one of critical issues for multiple-cell WLAN deployment due to the limited availability of frequency channel. The management should also be able to allow substantial bandwidth reuse among nearby BSSs where both contention-based traffic and polling-based traffic are present.

Also, even though additional backoff mechanism is designed to avoid CFP and CFP collision with overlapping BSS, as stated in [IEE802.11 1999], further enhanced coordination is needed to reduce delay in allocating CFP, to avoid hidden terminal issue and to allow efficient sharing of the medium with CFPs by multiple BSSs. According to the nature of Neighbourhood Capture, such coordination should also be able to avoid the unfairness issue among co-located BSSs using the central management. Without the CFP collision and Neighbourhood Capture among BSSs, QoS supports would then be provided for QoS streams as that in an isolated single BSS.

In summary, coordination in OBSS is essential to address the problems discussed above. Based on these rationales, we propose central coordination to avoid interference among co-located WLAN cells and further improve channel utilisation with re-use strategies.

6.4 Concept of a Novel OBSS Coordination

General Principles

As discussed above, certain OBSS overlap management on MAC protocol is essential to allocate resources / channel time among BSSs in order to re-use the frequency channel while avoiding co-channel interference between BSSs. The intention of resource management for overlapping scenario is to define a set of configurable interfaces and services that formalize the share of resources and quality of service among all individual BSS, providing the coordination framework for the integration of control and management mechanisms. The following principles govern the construction of an overlap management framework.

- Coordination Principle: It states that all individual BSS should be willing to participate in overlap management and follows the guidelines, which are designed and determined in a Resource Manager (RM). The essential of overlap management has been discussed previously in order to efficiently utilise the limited spectrum resources, while avoiding OBSS interference. And such overlap management has to be united among the participating BSSs.
- Fairness Principle: The RM should be selected and built independently to each individual BSS. It considers each participating BSS equal to each other while making resource management policies.
- Simplicity Principles: The coordination framework should be built as simple as possible. The design addition and change to IEEE 802.11 specification should be minimum to avoid additional complex implementation especially on wireless client.
- Scalability Principle: The coordination should accommodate not only area-limited deployment, but also large scope of network deployment. The larger network is co-ordinated; the resources re-use should be more outstanding.
- Compatibility Principles: As the overlap management itself is parts of QoS control issue; the overlap management framework should be compatible to existing QoS framework and collaborate with QoS framework for overall radio resources management.
- Dynamic Adaptation Principles: Based on the updated OBSS overlap and traffic information through the control and management mechanisms, RM should respond dynamically to meet the changes on the traffic and overlapping topology, with proper updated channel assignment scheme for efficient medium utilisation periodically.
- Resource Re-use Principles: The schemes should maximise medium utilisation with certain spectrum re-use strategies, while also avoiding interference.

Coordination Assumptions

- Assumption of a Fully-Coordinated OBSS Environment

The Overlapping study ignores the possible coexistence technologies may include: 802.15 (Bluetooth), 802.16a, 2.4 GHz Cordless Phone, 5.0 GHz Cordless Phone, 2.4 GHz Video Transmitter. 5.0 GHz non-OFDM Video Transmitter and 2.4 GHz Microwave oven.

We assume that the WLAN network is deployed in an infrastructure network configuration, where several wireless clients and AP (access point) form a BSS, and such access points (AP) are connected to Internet cloud via access network. All transmission in WLAN is omni-directional. AP locates at the centre of BSS with a transmission range of R and BSS is represented as circle with a centre of AP. Hence, all the possible transmissions occur only within this circle. If we further consider the power control issue, each circle has different size and it means overlapping possibilities with others.

The foundation of OBSS coordination is that, central resources management for each BSS is already in place. As discussed in Chapter 3, any QoS flow is admitted and managed by Radio Resource Manager. Within the OBSS environment, such overlap management should be collaborated as parts of Radio Resource Management. While performing the admission control for each BSS, the RM should also give consideration on the overlapping effects, such as that the BSS, which is under severe overlapping, should not be admitted large amount of QoS flow.

Each BSS within the OBSS environment is also willing to participate in OBSS coordination and follows the coordination managements and strategies, which are computed and defined in the Resource Manager. In case of non-participating BSSs, different frequency channels should be used in order to avoid interference between coordinated OBSS and uncoordinated OBSS. If two or more network operators deploy network closely and have potential OBSS interference, they should either adopt different frequency channels or establish such resource management.

- Assumption of a Supporting Distribution System

To create a wireless network of arbitrary size and complexity, IEEE defines Extended Service Set (ESS) network. From IEEE 802.11-1999 (R2003), it is defined as a set of one or more inter-connected basic service sets (BSSs) and integrated local area networks (LANs) that appears as a single BSS to the logical link control (LLC) layer at any station associated with one of those BSSs. It provides traffic segmentation and range extension.

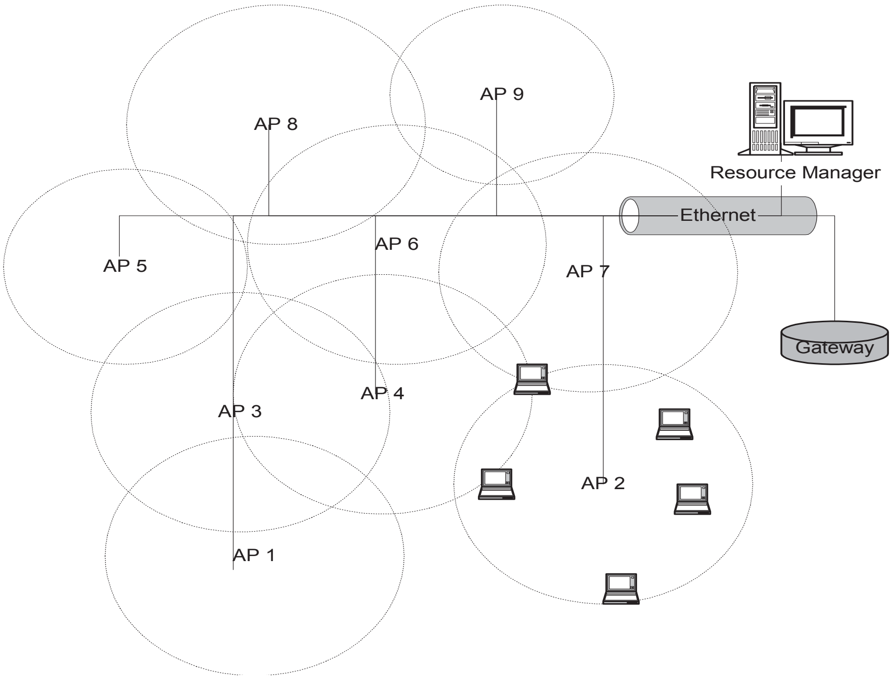


Figure 48 Supporting Distribution Systems

The ESS architecture consists of BSSs and the Distribution System (DS), which is normally created from the current IEEE 802 wired LANs. A collection of BSSs, each controlled by a single ESS Controller under one [E] SSID. Clients within an ESS may communicate and mobile clients may move from one BSS to another (within the same ESS) transparently to LLC.

As shown in the Figure 48, the DS provides a reliable and simple communication structure among APs in an ESS, as each participating BSS is part of ESS and connected via the DS. As the communication is transparent to LLC, the actual message transfer among APs in an ESS is a bridge operation. MAC Bridges interconnect the separate IEEE 802 LANs that comprise a Bridged LAN by relaying and filtering frames among the separate MAC of the Bridged LAN. The details of bridge operation refer to ANSI/IEEE Std 802.1D, “Media Access Control (MAC) Bridges”. All MAC frames of 802.11 WLAN are supported in bridge operation.

Most of ESS-related functions are also implemented fully or partially in the ESS Controller, such as Radio Resources Management (RRM), Roaming, and Security. The central controller of the overlapping coordination, Resource Manager (RM), is designed as part of RRM, which was discussed in Chapter Three. According to Chapter Three, RRM is located in ESS Controller within the wired DS and so is RM as well. Therefore, the supporting DS provides the effective channels for communication for central overlapping coordination between AP and RM.

Coordination Concepts

- Separation on Uplink and Downlink Transmission

In OBSS environment, comparing to contention scheme, polling scheme has more serious problem in OBSS interference issues. The contention scheme seems to be a better option than polling scheme in OBSS environment. However, unfairness issue between uplink and downlink in the contention scheme has been observed in previous study. Even in EDCF scheme, such disadvantage cannot be improved much as it is caused by the nature of contention scheme. On the contrast, in polling scheme, downlink is piggybacked on the poll frame for broadcast while uplink traffic is transmitted after poll frame by individual stations. Therefore, uplink and downlink traffic are generally given equal opportunities in polling scheme.

In order to maximise the benefits of using contention scheme in OBSS environment, unfairness issue between uplink and downlink has to be solved. Nevertheless, as known, either polling access scheme or contention access scheme in the standard MAC operation does not separate uplink and downlink transmission, which may be one of reasons of unfairness. Therefore, one of the options is to separate uplink and downlink transmission while allocating separate resources dynamically and individually for downlink and uplink transmission. Such treatment should be used to guarantee the equal opportunity for downlink. In such separation, the medium alternates between uplink period and downlink period.

- Coordination for Uplink Transmission

For uplink transmission in OBSS, contention-based scheme is suggested to use and only contention scheme is used among all BSS. As discussed, if all BSSs operate in contention scheme, the co-channel interference is minimum as the collision avoidance mechanism is used. On the other hand, due to the random access nature and finite detention range, the medium is fully utilised within the relevant the distance. Once the stations are out of the detention of another station, which is current occupying the medium, they are free to contend for the medium without any further coordination during the uplink period. In uplink period, the entire OBSS is considered as a 'super cell' in bigger scope, operating in contention-based scheme. It benefits from the natures of contention, being simplicity, self-organization and efficiency. In such case, priority-based QoS mechanism would be supported as parts of EDCF functions, and the QoS provisioning (traffic class mapping etc) should be unique among each participating BSS in order to provide service differentiation.

- Coordination for Downlink Transmission

For downlink transmission in OBSS, we use polling scheme to provide transmission opportunity. During downlink period, as polling scheme is used, AP can only poll itself for all downlink traffic while all uplink traffic are scheduled during the uplink period. However, interference is obvious if polling schemes are used among OBSS without coordination as discussed. Therefore, channel assignment, as the key part of the downlink coordination, is essential in order to avoid OBSS interference and maximise resources utilisation.

The scarcity of spectrum necessitates efficient channel assignment mechanisms. Whether the channel sharing is based on upon time (Time Division Multiple Access), frequency (Frequency Division Multiple Access), code (Code Division Multiple Access), or a combination thereof, there exists the fundamental limit on the number of users sharing the same channel assignment simultaneously. This has motivated the need for spatial reuse of the channel. In our OBSS downlink channel assignment solution, the channel sharing is based upon time, spatial reuse is to have users (BSSs) sufficiently far apart (in a reuse group) use the same frequency band. Members in the same reuse group are generally interference-free to each other and have certain physical distance sufficiently far apart from each other. Neighbouring BSSs, which are in different groups, cannot be assigned the same channel simultaneously. Therefore, the downlink period is divided into several limited time slot/span for these reuse groups. Apparently, the less number of time slots are divided, the more efficient the channel assignment solution is. We apply graph colouring technique to solve this problem with the best solution.

Control & Management Mechanisms for Coordination

These mechanisms provide real-time dynamic control and management to allow RM to finely tune the overall system performance.

- Overlapping Message Collection

Each WLAN client has relatively mobility and ON-OFF alternation for services. Hence, the overlapping topology may change and this leads to new requirement to channel assignment. RM should have periodically updated information on concurrent OBSS overlap situation in order to update the new channel assignment scheme to meet the new situation to maintain good medium efficiency. Hence, each BSS should perform OBSS message collection periodically for this purpose.

- Traffic demand Update and slot re-adjustment

Traffic demands vary from time to time. Dynamic management should respond to the balance of downlink and uplink traffic demands and the existence of delay-sensitive applications in the system. It should be supported by a traffic update mechanism.

- Allocation Message Transfer

Such messages include resources allocation decision/assignment and synchronisation; and they are conveyed between each BSS and Resource Manager. And they are transmitted in the supporting distribution system, which collaborates with existing Radio Resource Management (RRM) framework.

- Synchronisation

It is required to control the event ordering and the precise timing of BSS interaction. Synchronisation guarantees the implementation of resource allocation (such as time slot division). These are detailed in Appendix B.

6.5 OBSS Coordinated Resource Allocation Scheme and Protocol

Based on the overlapping information collected in RM (as detailed in Appendix B), RM would then follow the resources allocation protocol (sub-process) to execute the co-ordinated resource allocation scheme. Below, we describe resource allocation scheme before we present the allocation protocol.

The resources allocation scheme within one common frequency channel is a timeslot assignment topic; therefore, the scheme can be expressed in the unique superframe structure. The superframe design is the key component of the resource allocation scheme in the MAC layer. We explain the resource allocation scheme collaborated with the formation of the superframe structure.

Based on the current overlapping topology and traffic condition, the RM defines the scheme for all BSSs and executes it following the resource allocation protocol. It firstly divide all BSSs into several re-use groups, then design the length of each timeslot and assign each re-use group certain timeslots for transmission within such superframe. As the timeslot assignment is also dynamically adjusted by the current traffic situations, the slot duration re-adjustment protocol is also defined to re-adjust the timeslot length and so to keep the scheme dynamical to the current traffic requirements. The allocation message transfer protocol is then defined for transport of the allocation scheme from RM to AP in the BSS MAC layer. Finally, an example implementation of all messages transfers in the context of IEEE 802.11 is also described in the Appendix C.

6.5.1 OBSS Coordinated Resource Allocation Scheme

Coordinated Superframe

The OBSS-coordinated medium access functions can be described as two parts, Downlink Polling Access and Uplink Contention Access. So the scheme separates the contention and polling schemes. All BSSs in such coordinated OBSS environment would be synchronised and have a unique and coordinated superframe structure, which consists of downlink period and uplink period. Each coordinated BSS simultaneously operates in either downlink period or uplink period. During the downlink period, only the algorithm of polling access mandates; while during the uplink period, only Uplink Contention Access is applied to all stations.

During uplink period, all client stations in all BSSs become active for uplink transmission. And beacon frames are regularly broadcasted in every beacon interval from AP stations to all client stations. The Beacon frame contains necessary information for synchronisation and contention procedure, including the time length of uplink period and QoS service differentiations.

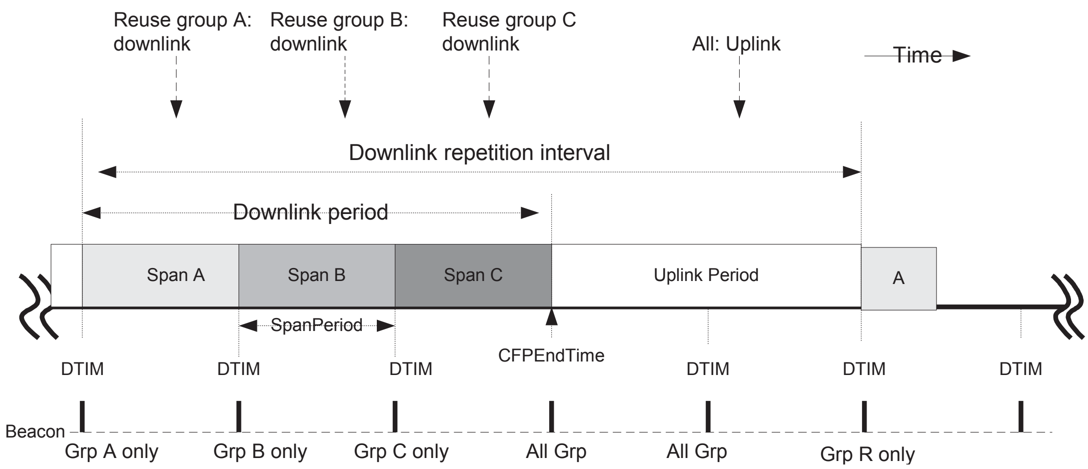


Figure 49 an Example Superframe

The downlink period alternates with an uplink period, in which the contention method controls the frame transfers. The downlink period would occur at a repletion rate. The RM generates downlink period at the downlink repetition rate, which is defined as a number of DTIM intervals (DTIM is a basic time unit defined in the IEEE 802.11). The repetition rate is also called Superframe size. All BSSs are synchronised for downlink period (the synchronisation system is described in the later sections). The initial proportion of superframe length allocated to uplink (/downlink) transmission is based on well-justified traffic assumptions.

The downlink period itself consists of several time spans for different reuse group of BSS. Each re-use group is at least assigned one individual time span. The reuse group is defined as a collection of BSSs, such that any two of BSSs in the group would not overlap each other. When a certain reuse group is assigned active during an assigned time span, its member BSSs are allowed to be active at that time span. Their AP stations start transmission with the Beacon frames. The beacon frame would firstly set NAV in all client stations in this active BSS and also carry the necessary information for polling period, such as polling order, and polling duration. Other groups should just remain silent by setting their NAV until they are assigned active. They should not even transmit Beacon frame during their silent periods to avoid any possible Beacon collision with the current active group.

Service Differentiation in Uplink Access

The uplink access method in a coordinated OBSS environment can be an EDCF. The fundamental of EDCF remains as CSMA/CD, in which collision avoidance is adopted. The collision avoidance mechanism minimises potential collisions in the OBSS environment, including those collisions between clients, which belong to different BSSs. Therefore, the nature of collision avoidance helps minimise the potential OBSS interference in the uplink transmission. RTS/CTS may be used as refinement under various circumstances to further minimise collisions. The details of EDCF access procedures are given in IEEE 802.11e specifications and other published articles, while we only address the concerns on the QoS service differentiation for the OBSS environment as below.

QoS service differentiation is provided for the BSSs. According to the principle of fairness among BSSs, the access to the QoS service differentiation should be equal and fair among BSSs. Therefore, all the clients are treated ‘equally’ in the entire OBSS system according to the EDCF access rules.

The QoS facility of EDCF supports 8 user priorities, which are identical to the IEEE 802.1D priority tags. One or more user priorities shall be assigned to one of four EDCF access categories (AC). The QoS mapping table is the guidelines for mapping traffic types to different EDCF priorities as below in Figure 50. And the admitted TS (such as VoIP and video) should be assigned higher ACs to meet the delay requirements.

User priority	802.1D Designation	Access Category (AC)	Designation (Informative)
0	Best Effort	0	Best Effort
1	Background	0	Best Effort
2	-	0	Best Effort
3	Excellent Effort	1	Video Probe
4	Controlled Load	2	Video
5	Video (<100ms delay & jitter	2	Video
6	Voice (<10ms delay & jitter)	3	Voice
7	Net Control	3	Voice

Figure 50 **User priority to Access Category mappings**

Each AC is an enhanced variant of the DCF that contends for TXOPs using one set of EDCF QoS parameters element in Beacon frames. The implementation values for EDCF QoS parameters set and the mapping between user priorities and AC should be unique to all BSSs and controlled by the Resource Manager. In such way, all BSSs are guaranteed equal for accessing prioritised treatments for delay-sensitive / critical message and data.

Octets: 1	1	2	1 * 4	2 * 4	1 * 4	2 * 4
Element ID (12)	**Length** (44)	**EDCF Parameter Set Count**	**CWmin[AC] Values** CWmin[0] … CWmin[3]	**CWmax[AC] values** CWmax[0] … CWmax[3]	**AIFSN[AC] values** AIFSN[0] … AIFSN[3]	**TXOPLimit[AC] Values** TXOP[0] … TXOP[3]

Table 7 **EDCF QoS Parameter Set element format**

Open Options of Service Differentiation for Downlink Polling Access

All BSS also incorporate polling access method for downlink traffic in order to avoid OBSS interference. This polling access method is controlled by the point coordinator in the AP. However, the AP would firstly follow the access rules from the OBSS Resource Manager, for accessing the assigned timeslot/timeslots for its downlink traffic transmission. The RM performs the role of the master and tries to avoid the simultaneous transmission from the BSSs, that belong to different reuse group, as discussed previously.

Each AP has full controls and options for varied service differentiation for its own traffic only during its assigned downlink polling transmission period. The adoption of different service differentiation schemes in each BSS would not make any effect to transmissions in its neighbouring BSSs. The AP is equipped with a QoS scheduler and should always satisfy (offer certain TXOPs) the admitted TS under the controlled polling mechanism based on the accepted TSPEC information of the admitted TS. The implementation has been discussed in IEEE 802.11e as part of Traffic Stream transmission.

However, even the client is polled for transmission, as data of admitted TS are not always pending in the queue; the client may miss the polled opportunity to transmit and have to wait for the next polling period. And it leads to degrade on delay performance. In such case, we consider Multiple-Cycle Polling, which is particularly designed for delay-sensitive applications during Contention Free Period. Further implementation details and performance analysis are referred to previous Chapters.

Slot Duration

The Superframe length (downlink period repetition rate) is a multiple of DTIM; but it should not be constrained to a certain multiple of DTIM and dynamically controlled by the Resource Manager according to the traffic demands. It is one of parameters for the performance tuning. A common example of the superframe is illustrated in Figure 49 above, where a beacon interval is a DTIM. The superframe consists of five DTIM and each time span is a DTIM period for each re-use group of total three. It is not allowed to have delayed beacon and foreshortened downlink period by the usage of strict synchronisation within the entire OBSS environment (we gives details on synchronisation mechanism in the later section.).

RM allocates certain proportion of superframe duration to uplink transmission based on the demands of uplink/downlink traffic and the admitted multimedia traffic (further re-adjustment details are described in the following sections.).

Each downlink time span has equal/approximately equal length according to the fairness principle among each reuse group. BSSs with high traffic demands may deserve an uneven long span due to the equal time span allocation. However, longer span should not be used for them, because they are randomly deployed, may overlap each other and not always able to assign into the same group to 'enjoy' a longer span. Therefore, uneven span is inappropriate and additional second time span is considered with second grouping assignment.

First Grouping and Second Grouping

The number of time span (re-use group) is calculated in RM using certain grouping (first and second) algorithm. The purpose of channel (group) assignment computation is to produce guidelines for dividing all participating BSSs into several re-use groups, namely Group R (Red), B (Blue), G (Green), Y (Yellow), etc. The constraint of grouping is that every two overlapping BSSs cannot be assigned into the same group. The objective of grouping is to find the minimum of group number it needs to satisfy the grouping purpose. The decision purely depends on the number of BSSs in the coordinated environment and the overlapping severity.

The result of the first grouping exercise gives each BSS a group and an approximately equal time span. However, in order to maximise the spatial reuse, we propose second grouping assignment, which BSSs with high traffic (downlink) demands may be additionally assigned its second group. Therefore, these assigned BSSs with high traffic demands can still be satisfied with the additional second group assignment strategy. The second grouping neither increases the number of groups used nor changes the result of the first grouping assignment. Instead, it only assigns additional group to certain BSSs. The constraint of the second grouping remains the same as that in the first grouping, which is every two overlapping BSSs, cannot be assigned into the same group. In such way, each BSS is at least assigned one group and such group contains BSSs members, which do not overlap and interfere with each other. We would describe the grouping algorithm using colouring theory in the following section.

6.5.2 OBSS Resource Allocation Protocol

The resource allocation protocol consists of foreign ESS detection, justification of grouping re-arrangement, first group assignment, second group assignment, timeslot assignment, traffic condition adjustment and allocation message transfer.

We recall the BSS reports handling process in the overlapping messages collection section, which we have already discussed previously. The foreign ESS detection process is carried after RM record update. The RM would firstly examine if the record contains any SSID, which is different form its SSID of the home ESS. By finding different SSIDs, RM is aware of existing of other neighbouring foreign ESSs. And at least one of BSSs in the foreign ESS overlaps with a BSS in the home ESS. Upon the awareness of such situation, the RM should inform ESS Central Controller to make arrangements on resources allocation together with the foreign ESS (such as the use of dynamic frequency selection technique).

After the ESS detection process, RM would justify whether there is need for the grouping re-arrangement according to overlapping changes, as part of the BSS report handling process. Only overlapping topology change triggers group re-assignment, and consequently the resources allocation re-arrangement. As known, the topology is made of the RM records (overlapping pairs). A change in overlapping topology occurs only if:

- When a record is removed and its counterpart record does not exist in the database.
- When a record is added and its counterpart record does not exist in the database.

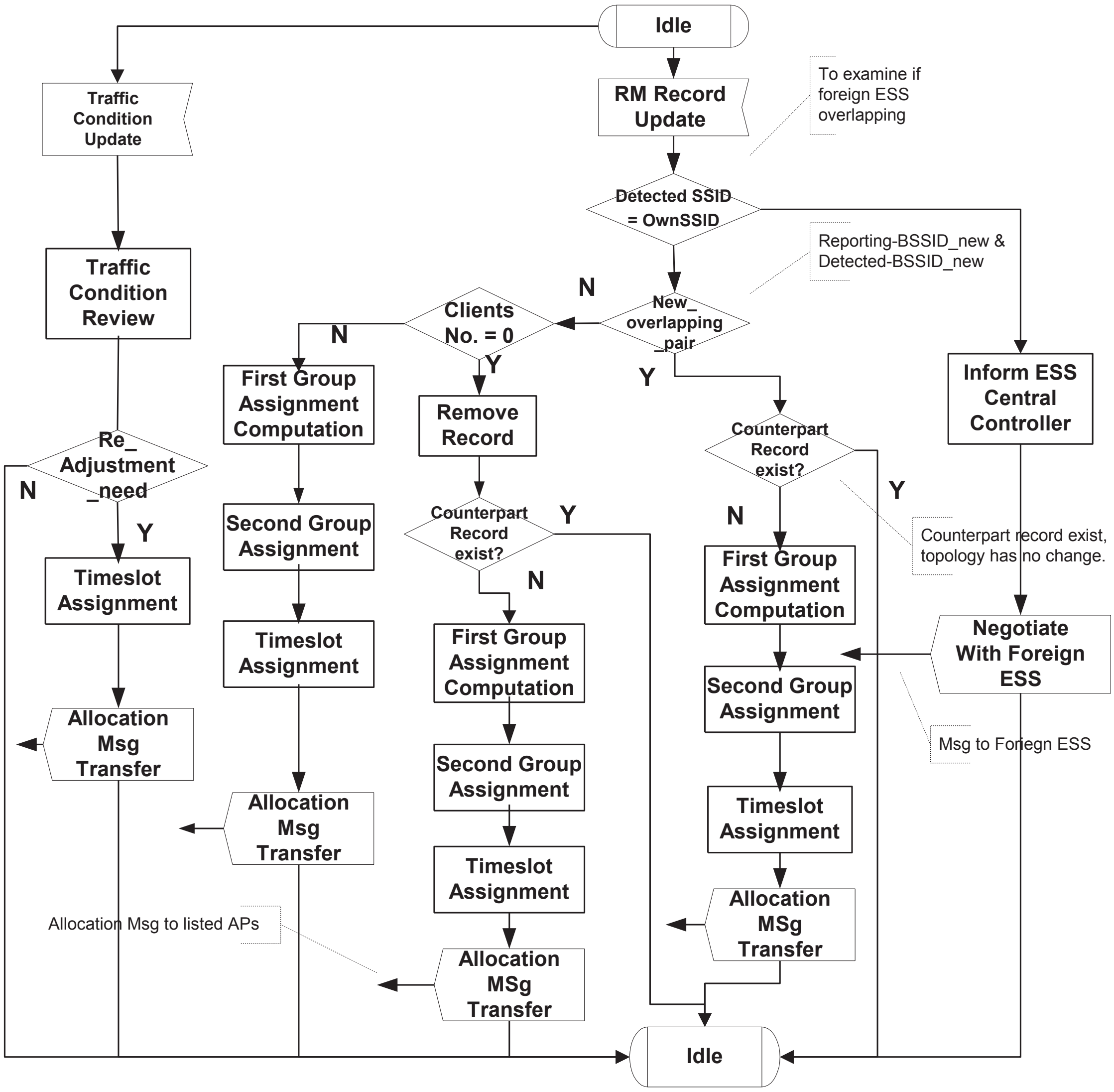


Figure 51 Resources Allocation Protocol

The group assignment consists of first group assignment and the second group assignment. RM assigns BSSs into the first-level groups, such that any two of BSSs in the group would not overlap each other. The BSS is considered to assign its second group, based on the first group assignment and the existing traffic demand reports. When assigning the second group, RM gives first/higher priority to BSSs, which have higher downlink traffic demands. The demands are reported via the traffic re-adjustment mechanism as discussed in the Appendix B.

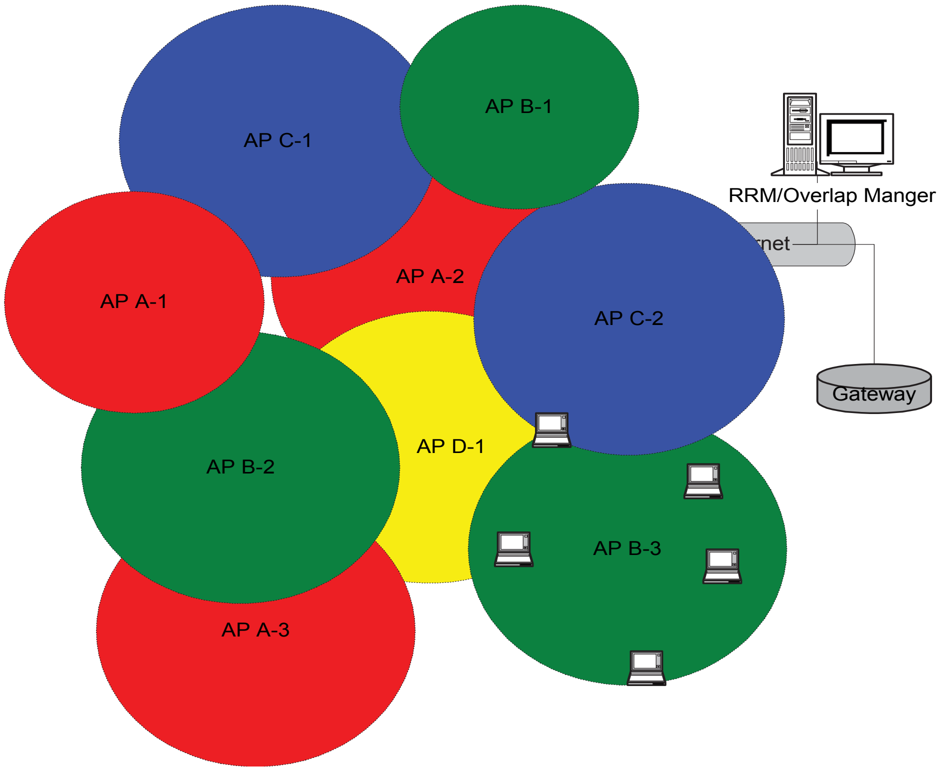


Figure 52 an Example of the OBSS First Grouping

The constraint of both the first group assignment and second group assignment is that any of two overlapping BSSs cannot be assigned into the same group. The objective of the first grouping is to find the minimum of group number it needs to satisfy the grouping purpose/constraint. The objective of the second grouping is to maintain the results if the first group assignment including group number used; while trying to additionally assign the second group to those BSSs with high downlink traffic demand. The information needed for group computations is:

- The total amount of the participating BSSs in the ESS
- The neighbouring/overlapping relationship among all participating BSS.
- Downlink traffic demands in each BSS.

The algorithms used in the group assignments are discussed separately in Section 6.7 with the design justifications and formal specification. An example of grouping is given below in Figure 52 and Figure 53, where four groups are used and BSS B3, A3 and C1 are assigned their second groups.

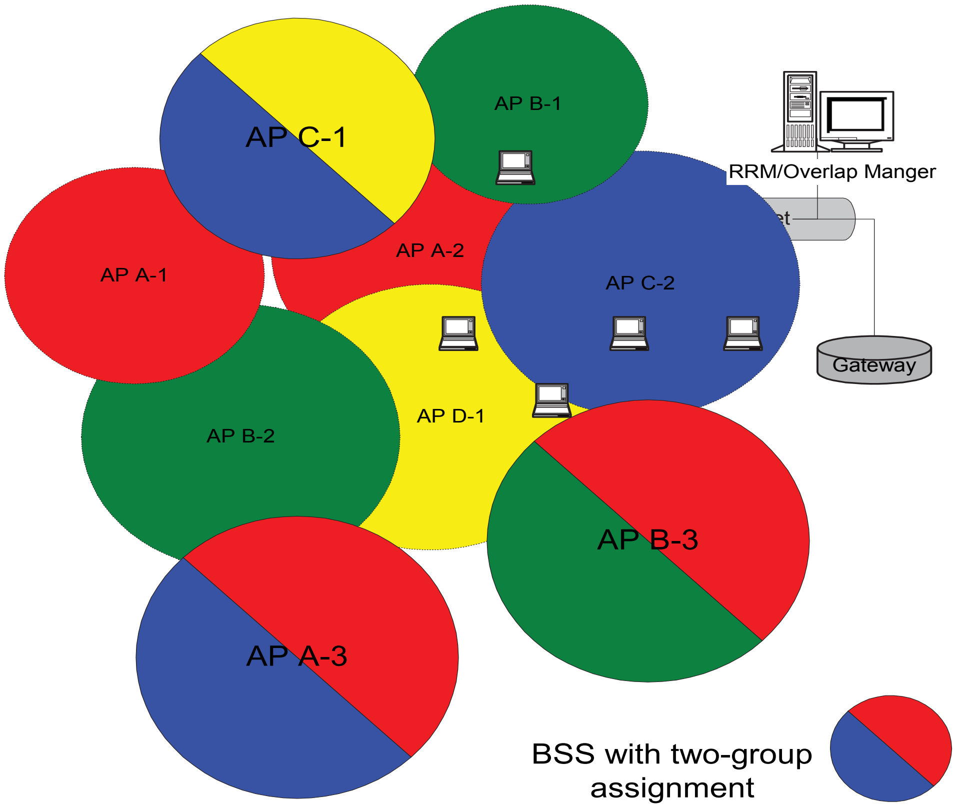


Figure 53 an Example of the OBSS Second Grouping

The timeslot assignment then defines all the slots duration based on the traffic demand reports, the first group and second group assignment. It firstly assigns one uplink timeslot for all BSSs to share the medium together; while the downlink slot is divided into several time spans. Within each time span, it would only allow downlink transmission in the BSSs, which belong to one assigned group. It results in a unique superframe structure specifying the timeslot allocation for the respective re-use groups. The superframe structure and slot duration design is as discussed in the previous section.

The slot duration can be re-adjusted. The entire protocol can also be triggered by the receipt of traffic condition report apart from the overlapping record update. RM would then review the traffic condition with new changes and justifies the need for slot duration re-adjustment. If the traffic condition changes meet certain criteria for resources allocation re-arrangement, only timeslot assignment would be involved for the resource allocation re-arrangement. As the overlapping topology may not change at the time, the traffic condition change may not necessarily lead to group re-assignments. The timeslot re-assignment would address the changes on proportions of uplink and downlink traffic in each BSS. The details of slot duration re-adjustments on traffic conditions will be presented in Appendix B.

After the timeslot assignment completes, RM would produce a list of BSSs, which require update with the new allocation scheme. RM would then respectively send allocation messages to the listed BSSs (AP). And AP takes further responsibility to inform its clients about the relevant allocation scheme/superframe structure. The entire allocation message transfer procedures would be described in Appendix B.

6.6 OBSS Grouping Assignment Using Graph Colouring

6.6.1 Graph Technique Background

A packet radio WLAN network is a collection of radio transmitters and receivers (transceivers) located in a geographical region. Associated with each transceiver is a transmission range, which depends on its transmission power. A transceiver $client_{ap1,\ 1}$ located within the transmission range of another transceiver $client_{ap2,\ 1}$ can receive the messages from transceiver $client_{ap2,\ 1}$. The transceivers of the network share the communication medium for message transfer. This gives rise to conflict situations where two or more of the transceivers may want the same shared medium simultaneously. In a coordinated OBSS environment, it means the sharing of a common radio channel for all BSSs (and all its clients) without causing garbling and eventual loss of messages.

The spatial distribution of the transceivers permits sharing or reuse of the channel. Most popular techniques for this sharing includes time slot division; frequency division and code division multiple access schemes. The medium division or scheduling are often modelled and solved by the graph colouring technique, where the radio network is modelled using graph. In our overlapping resource coordination scheme, the medium is scheduled for sharing among all participating BSSs by time slot division technique. Collaborated with the natures of overlapping WLAN environment, we model the overlapping WLAN network, using graphs and solve the graph colouring problems as below.

We recall that problem discussed in previous Sections. The purpose of channel (group) assignment computation are to produce guidelines for dividing all participating BSSs into several re-use groups, namely Group R (Red), B (Blue), G (Green), Y (Yellow), etc. The constraint of grouping is that every two overlapping BSS cannot be assigned into the same group. The objective of grouping is to find the minimum of group number it needs to satisfy the grouping purpose. The information needed for computation is:

- The total amount of the participating BSSs in the ESS
- The neighbouring/overlapping relationship among all participating BSS

The overlapping pair as discussed previously indicates the neighbouring relationship between a pair of BSSs. The entire overlapping relationships are made up of all these pairs in the topology database, which are regularly updated in the RM.

6.6.2 Graphical Model

The channel assignment problem can be mathematically modelled in a few ways. The most famous models are multiple interference model and binary constraint model [Jensen and Toft 1995]. The multiple interference modelling studies the signal propagation, defines attenuation factors for adjacent channel interference and evaluates Signal-to-Interference Radio (SIR) at certain test point in order to assign channel properly without occurrence of channel interference.

The binary constraint model has a higher level of abstraction and requires less computational complexity. It can be seen as a generalisation of the well-known Graph Colouring Problem [Jensen and Toft 1995]. In this thesis, we adopt this model and use it to solve the channel assignment problem in OBSS.

Our presentation of channel assignment is based on a standard representation of a radio network by a directed graph G = ***(V, E).*** We limit that graph only refers to a network of edges and vertices. Here, ***V*** is a set of vertex denoting the clients in the radio network, and ***E*** is a set of directed edges between vertices such that for any two distinct vertices ***u, v*** $\in$ ***V, (u, v)*** $\in$ ***E*** if and only if v can receive ***u***'s transmission. In such way, G denotes an interference graph, where the node set denotes cells or base stations (BSS) that require communication service, and the edge set represents geographical approximation of cells and therefore the possibility of co-channel interference.

Above, the standard definition of the edges is with directedness. That is, ***(u, v)*** $\in$***E*** does not imply that (***(v, u)*** $\in$ ***E***. In fact, within WLAN environment, even though the interference between cells may not be bi-directional, the information necessary to make assignment decision is the fact that interference occurs, no matter which transceiver of two adjacent transceivers causes the interference to occur. And such information is sufficient to schedule the channel assignment to prevent interference. Therefore, from the assignment point of view, the network can be represented by an undirected graph. That is, if there is/are ***(u, v)*** $\in$***E*** or/ and (***(v, u)*** $\in$ ***E***, we establish the undirected edge between ***u*** and ***v***.

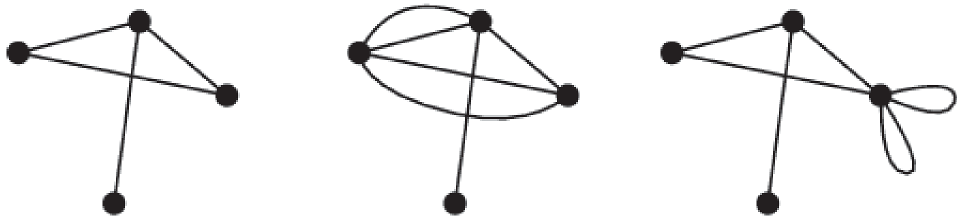


Figure 54 Graph Varieties

Graphs come in a wide variety of different sorts [Garden 1984] as shown in Figure 54. The most common type is graphs in which at most one edge (i.e., one edge or no edge) may connect any two vertices are said to be simple graphs. If multiple edges are allowed between vertices, the graph is known as a multi-graph. Vertices are usually not allowed to be self-connected, but this restriction is sometimes relaxed to allow such "graph loops." A graph that may contain multiple edges and graph loops is called a pseudo-graph. A simple graph is enough to represent the wireless network model. Therefore, in our graph, there are no multiple edges and graph loops.

The edges, vertices, or both of a graph may be assigned specific values, labels, or colours (properties), in which case the graph is called a labelled graph. A vertex colouring is an assignment of labels or colours to each vertex of a graph such that no edge connects two identically coloured vertices. Similarly, an edge colouring is an assignment of labels or colours to each edge of a graph such that adjacent edges (or the edges bounding different regions) must receive different colours. The assignment of labels or colours to the edges or vertices of a graph based on a set of specified criteria is known as graph colouring.

Further formal definitions of terms we use through this section are as follows. If $(u, v) \in E$, then v is a neighbour of u. As our graph is undirected, u is also a neighbour of v. A neighbour must be a vertex. The degree of a vertex v is the total number of neighbours of v. The maximum degree of a graph is the maximum of degrees taken over all vertices in such graph.

6.6.3 Mapping to Graph Theoretic Problem

Due to the similarity between scheduling and some graph theoretic problem, we solve our OBSS group assignment problem using graph theory; therefore, the problem can be considered in a simpler way and certain theorems and techniques in graph theories can be used to solve the problem. From the problem domain, two BSS cells cannot be scheduled in the same time slot for downlink transmission when they interfere. If modelled as a graph with a vertex for each cell and edges to represent interference, this group assignment / scheduling problem can be solved using graph vertex colouring technique according to its constraint below.

Constraints Definition

The OBSS group assignment problem models the task of assigning time slot to a set of OBSS cells. A feasible group assignment must additionally satisfy certain interference constraints. An optimal assignment globally minimises a cost function, which is the minimum of time slots it uses in the OBSS system.

It is important to understand that interference can only be defined with respect to a receiver. That is, any signal, occurring within the same channel that a desired signal is to be received in, is considered to interfere with the reception of the desired signal if it is of sufficient power (as measured at the designated receiver). Interference power is a function of transmitter power, receiver sensitivity, antenna gains, patterns and polarizations, and channel loss. Channel loss is a function primarily of distance, frequency, and weather and is quantified by minimum acceptable signal to noise power ratio, or maximum permissible interference to noise power ratio as measured at a receiver. We can assume that most of these factors are already determined or beyond our ability to influence. Consequently, we will define interference directly as a function of distance in the case of OBSS problem.

We now study the problem by defining constraint, which is a symmetric relation between two vertices or two edges in a graph. A constraint imposes a restriction on colouring: two vertices or edges that are mutually constrained (related by some constraint) must receive different colour for an assignment to be legal. The constraint is normally described by whether they are between vertices or edges, the separation between them, and whether it is a transmitter and/or a receiver based constraint. It is either a vertex based or an edge based colouring. An edge colouring of a graph G is a colouring of the edges of G such that adjacent edges (or the edges bounding different regions) receive different colours. A vertex colouring is an assignment of labels or colours to each vertex of a graph such that no edge connects two identically collared vertices.

In our OBSS group assignment problem, the constraint is relatively simple. During downlink transmission period, only those cells (vertexes), which can hear other (BSS is detected by either or both sides), would cause interference between themselves. Therefore, such constraint for OBSS group assignment is vertex based. Furthermore, these cells are very often co-located and adjacent, so called neighbours. The interference will not be carried out further to neighbour's neighbour as long as there is no overlapping detention among them. The constraint is so called Distance-1. And it makes no difference whether constraint is transmitter and/or a receiver based, with the same reason that undirected graph is applied here. Distance-1 graph vertex colouring of an undirected graph is defined for the OBSS group assignment problem in the following way:

Given an undirected graph G = ***(V, E)***, a Distance-1 graph vertex colouring of an undirected graph, is a mapping of

$$\langle f(v) \to N \mid f(v_i) = f(v_j) \Leftrightarrow (v_i, v_j) \notin E \rangle$$

Equation 6 Graph Colouring Technique

A vertex colouring, one of the graph colouring techniques, is an assignment of labels or colours to each vertex of a graph such that no edge connects two identically coloured vertices. The most common type of vertex colouring seeks to minimise the number of colours for a given graph. Finding a minimal colouring can be done possibly using brute-force search [Christofides 1971; Wilf 1984; Skiena 1990]. The minimum number of colours, which with the vertices of a graph G may be coloured is called the chromatic number, denoted X (G). A graph with chromatic number three is said to be three-colourable. In general, a graph with chromatic number n is said to be an n-chromatic graph, and a graph with chromatic number, which is less than or equal to n, is said to be n-colourable. Calculating the chromatic number of a graph is an NP-complete problem [Skiena 1990]. If a problem is known to be NP-complete (verifiable in Nondeterministic Polynomial time), and a solution to the problem is somehow known and such chromatic number can be found.

Previous graph studies [Christofides 1971; Wilf 1984; Skiena 1990] have proved the NP-completeness according to most different types of graph. The only one-colourable graphs are empty graphs, and two-colourable graphs are exactly bipartite graphs. The four-colour theorem establishes that all planar graphs are 4-colourable.

Mapping to the OBSS group assignment, the Distance-1 graph vertex colouring technique is applied to find the minimum colours/ re-use group needed for the OBSS group division. After finding out the minimum number, the procedure on how to form the re-use groups can also be considered as the procedure of the graph labelling and colouring. In order to use the results of these graph studies, we firstly investigate the types of graph representing our network model to determine the minimum number.

6.6.4 Network Model and Properties

OBSS Environment Characteristics

The selection of graph representation for a network should give consideration of the characteristics of the network scenario, which is modelled by the particular type of graphs. In most assignment studies of radio networks, nodes are often considered as transmitter or transceiver station. Not only is the direction of interference considered, but also the geometric node (network) configuration is carefully concerned. Most of these scenario, the graphs used in modelling must be directed and/or poor planar (high thickness) [Ramanathan and Lioyd 1992].

Unlike most scenarios in these studies, the OBSS environment is defined by the location of each BSSs and the respective detection of other BSSs from such BSS. Therefore, BSS is considered as the node/vertex. Considering the practical deployment scenario, the amount of BSS deployed should not be as large as that of transmitter or transceiver station in a normal radio network and the distance among neighbouring BSSs (nodes) should not be as close as those in a normal radio network. The OBSS environment should therefore be relatively planar, with a possible light network density.

Existing Model Properties and Problems

In all prior researches [Arikan 1984] [Ephremides et al. 1990] [Even et al 1984] [Ramanathan and Lioyd 1992] [Ramanathan and Lioyd 1993] [Ramanathan and Lioyd 1989], a packet radio network is modelled as a graph, for the purpose of constructing an optimal channel schedule. Models are designed to encapsulate the essential features of the system. The discussion in the previous study clearly shows that, almost all previous researches in this area used an arbitrary graph (either directed or undirected) as the model of the radio network. Many scheduling problems were shown by the researchers to be completed using this model.

Arbitrary graphs have the advantage of being able to represent all possible network configurations. However, some arbitrary graphs have no real physical counterpart in radio network. Certain restricted graphs can give an accurate representation to certain radio network or network scenario; and illuminate some aspects of the problem structure, which might help in solving the problem, and find the optimal solution, such as finding a chromatic number. Tree and planar graphs are most famous restricted arbitrary graphs studied used in modelling radio networks.

Tree is the simplest graphical representation used in modelling networks. Problem such as message routing and propagation can be well addressed using tree models. However, according to the nature of tree structure, it fails to be flexible enough to represent many possible network configurations.

The limitation of trees led Ramanathan to propose planar graphs as the model of choice [Ramanathan and Lioyd 1989, 1992, and 1993]. A planar graph is a graph if it can be drawn in a plane without graph edges crossing. The use of a planar graph (directed or undirected) as the model of a radio network can capture some essential features of the WLAN network. This mechanism of constructing the graph model for a radio network was used by many other researchers. At the first level of abstraction of the radio networks, we think of the BSSs as some points on a two-dimensional plane. Associated with each node/BSS is a range value, representing the transmission range of the BSS. At the second level of abstraction, we construct a graph where each node/BSS represents a point on the plane and there is an undirected edge between node Vi and node Vj, on the two-dimensional plane if the BSSi detects the existing of the BSSj or the BSSj detects the existing of the BSSi. The planar graph constructed in this manner captures all the pertinent features of the WLAN radio network.

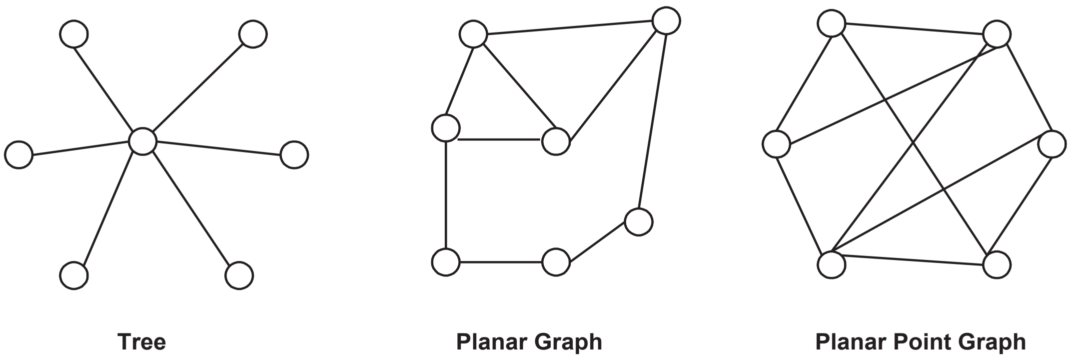


Figure 55 **Graphs Used for Modelling**

However, the planar graph is incomplete in representing all possible network configurations, particularly those complex scenarios. For example, when there are many nodes deployed closely and overlapped heavily with each other, these nodes potentially cause dense interference and these are reflected as large number of edges within a limited area on the graph. Therefore, such radio network cannot be represented by a planar graph, as the edge crossing can not be avoided. In [Arunabha and Huson 1997], a planar point graph is further introduced for the purpose of modelling such network configuration of heavily-overlapped radio network and has shown its ability of capturing these features of such networks.

In conclusion, a planar point graph may be appropriate in modelling most possible network configurations of the OBSS environment. However, in practice, a high-density deployment of BSSs is unlikely, because the extreme limitation on resource within such small physical scope would not allow efficient communication for all these closely-located BSSs. Therefore, a planar graph may be more appropriate to use to represent OBSS environment while also illuminating some aspects of the problem structure and helping in solving the problem (finding the chromatic number).

6.6.5 The Problem Solution and its Completeness

NP-completeness of Chromatic Number Problem

In order to find the chromatic number, NP-completeness problem needs to be studied to ensure that such number is verifiable. It is very important to realize that if a problem is NP-complete in a certain domain, it does not imply that the problem will continue to be NP-complete when the domain is substantially restricted [Arunabha and Huson 1997]. More specifically, NP-completeness of a graph problem does not necessarily imply NP-completeness of the problem for a restricted class of graphs. Therefore, we study this NP-completeness problem according to different types of graph used in modelling the OBSS environments.

In most studies [Arikan 1984] [Ephremides et al. 1990] [Even et al 1984] [Ramanathan and Lioyd 1992] [Ramanathan and Lioyd 1993] [Ramanathan and Lioyd 1989], most scheduling problems are shown to be NP-complete using the arbitrary graph. In [Arunabha and Huson 1997], the problems were shown to be NP-complete using the planar point graph, which has been discussed.

If a planar graph is used for modelling the network, the four-colour theorem can be applied. The four-colour theorem states that any map in a plane can be coloured using four colours in such a way that regions sharing a common boundary (other than a single point) do not share the same colour. More specifically, NP-completeness of a planar graph problem is proved [Wagon 1998] [Wagon 1999], and the chromatic number for the assignment problem where the network can be modelled using the planar graph is 4. The chromatic number should remain 4 as long as the network can be represented by a planar graph regardless of number of nodes/BSSs.

Recommendation for Deploying OBSS

As seen, if an OBSS environment is modelled by a planar graph, maximum four time spans can be enough for use in scheduling the downlink traffic without collision. Obviously, as network density increases, planar graphs may not be used, most likely more colours need to be applied, and so the network efficiency may be concerned. This should be considered particularly before deploying OBSS. An efficient deploying should indicate guideline of a 'maximum' network density, which is the limitation for application of four-colour theorem.

In [Bondy and Murty 1976], it is proved that as a fundamental property of a planar graph, the minimum degree vertex has at most five neighbours. Therefore, it is suggested to use this fact to examine the 'maximum' network density while deploying the OBSS and try to avoid over-dense situation. The guideline for not being over-dense is that:

The BSS, which has minimum neighbours, has at most five neighbours.

Therefore, when deploying OBSS, RM should always keep an eye on those BSS, which only overlaps minimum numbers of other BSSs, and ensure the number is not over five. If the number is over five, RM should make suggestions on changing overlapping topology, such as relocation of these BSSs or limiting their transmission range.

6.6.6 The Colouring Algorithm

The Algorithms Varieties

The above sections have shown that the common OBSS environment mostly can be modelled by a planar graph and its channel assignment problem can be considered as a vertex colouring problem, where such planar graph can be colourable by 4 colours. We now consider how to perform such vertex colouring (colouring / labelling each vertex) in a planar graph.

The algorithms used in performing the colouring have also been well studied. The selection of these algorithms relies on the graphs used in modelling the network, or whether it is vertex colouring or edge colouring. There are mainly two classes of algorithms: geometric algorithms and greedy algorithms [Ramanathan 1999]. The geometric algorithms make use of the geometric arrangements in the network. In particular, for a 'linear' network, where all nodes are in a single line, a geometric algorithm, called linear projection, is efficient. The greedy algorithms mainly assign the first non-conflicting colour based on the colours already assigned to nodes. As it has wider applicability to most graphs, including planar graph and planar point graph, we give further details on the greedy algorithm below, in the case of the vertex colouring.

The greedy algorithm consists of two phases, a labelling phase and a colouring phase. In the labelling phase, each vertex in the graph is firstly assigned a unique label between 1 and n, where n is the total number of the vertices. The vertices are then considered in decreased order of labels to be coloured in the colouring phase. The algorithm completes until the vertex labelled 1 is coloured. The colour assignment follows the greedy principles that no constrains is violated. As seen, the labelling determines the colouring order and largely determines the final results. There are a few possible ordering in the literature. The random ordering is the simplest and straightforward, so called the Pure Greedy. This is an algorithm in which a vertex is chosen at random and coloured with the first available, non-conflicting colour. Most of the algorithms described in previous works [Chlamtac and Kutten 1985] are essentially this method. Other greedy algorithms using different ordering, such as Maximum Degree First (MDF) and Maximum Degree Last, present slightly complex ordering technique, which results in higher time complex. However, in [Sen and Huson 1996], based on the experiment results, it pointed out that the performances of these algorithms begin to diverge as the density (node population) of network increase. On average, most greedy algorithms produce similar results closer to optimal.

Maximum Degree First Algorithm

After having discussed the background of several algorithms, it is shown that Maximum Degree First is one of the best choices among other algorithms to perform a vertex colouring in a planar graph, where the OBSS channel assignment problem is modelled. The section below depicts the procedure of Maximum Degree First with an example.

The well-known Maximum Degree First ordering (MDF) is also called Minimum Neighbour First. Here, degree is defined as the number of neighbours for this vertex. In this method we take a maximal mutually conflicting clique of edges around the maximum degree vertex first, colour it, and then progressively do the same for the remainder of the graph. This method is based on the intuitive notion that it is better to colour the more 'crowded' areas first. Heuristics with this philosophy were first examined in [Matula et al. 1972] and were found to do quite well for vertex colouring in a planar graph.

This ordering heuristic is based on the degree of each node. The determination of the degree is dynamic and the order is not determined, prior to any node being coloured. In fact, a node with maximum degree is selected and coloured and the degree of all its adjacent nodes are decreased before the next node is selected and coloured. In the colouring phase, it should follow the principle of 'first-available-colour', where colours are given priorities to be selected without collision. Figure 56 below depicts the MDF procedure of an example planar graph. It has also shown that the planar graph is colourable by 4 colours only.

The first group assignment technique, Maximum Degree First algorithm, is given in formal specification as below:

Input: **(1) A graph G =** *(V, E),*
(2) *V* **to be coloured with constraint of**
$\langle colour(v) \to N \mid colour(v_i) = colour(v_j) \Leftrightarrow (v_i, v_j) \notin E \rangle$.
(3) Ordering of MDF
Output: **A colouring c:** *V* **← C{1,2,3,…}**

Begin
1. for all vertex v of *V* do
2. colour(v) ← 0
3. Assign-Label (G)
4. for *j* from largest down to smallest label do
5 for each v with label *j*
6. colour(v) ← First-Available-Colour (G, v, C)
7 Assign-Label (G)
End

Procedure Assign-Label (G)
1. for labels *l* from 1 through | *V* | do
2. *v* ← unlabelled Max. degree vertex in G
3. for each uncoloured vertex *u do*
4. if (*u, v*) ∈ *E*, degree of *u* decrease 1
5. label (*v*) = *l*

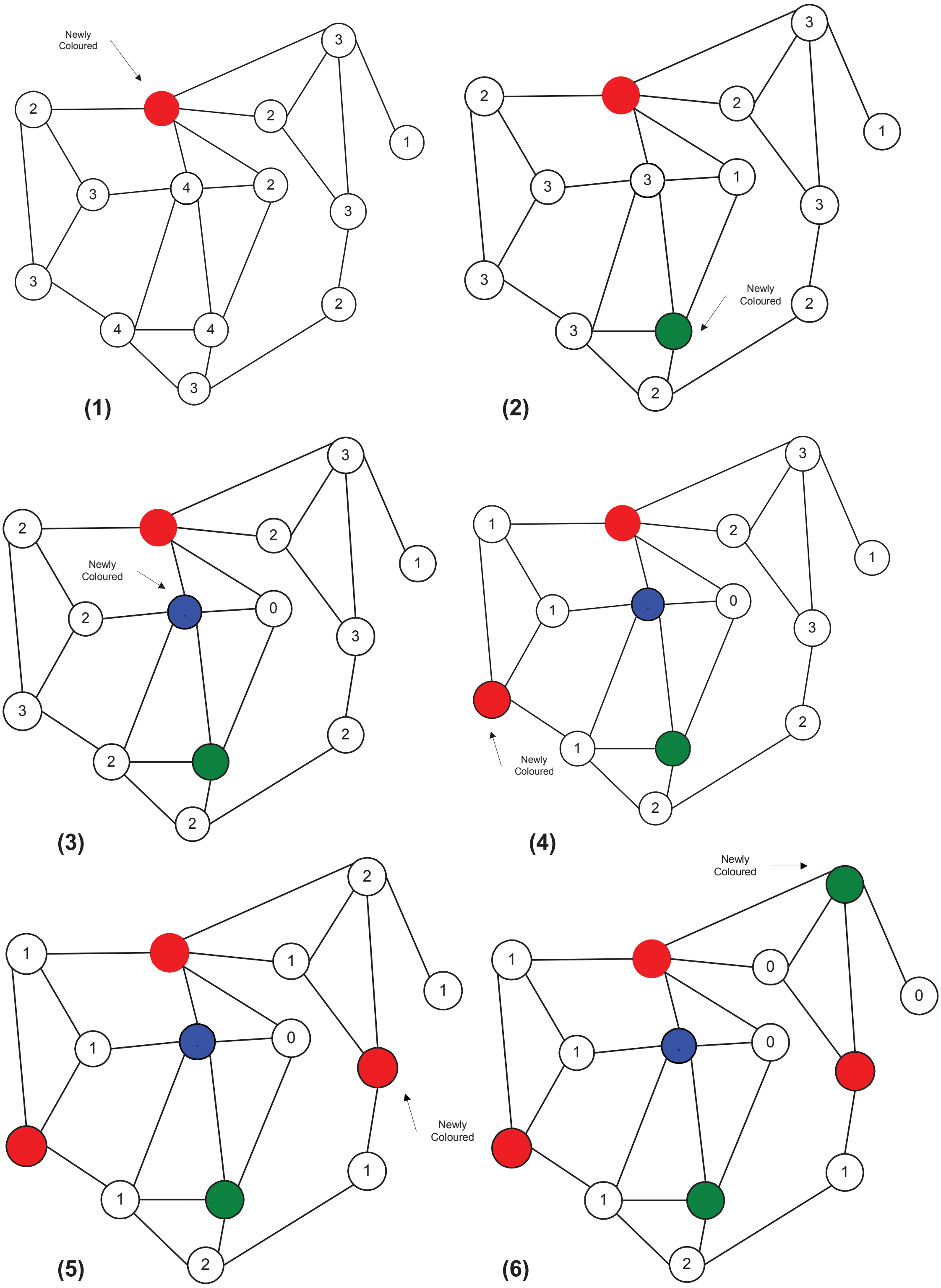
Procedure First-Available-Colour (G, v, C)
1. Taken ← { }
2. for each coloured vertex *u do*
3. if (*u, v*) ∈ *E*, Taken ← Taken ∪ colour (*u*)
4 Return (smallest colour ∉ Taken)

The second group assignment technique, Maximum Degree First algorithm, is given in formal specification as below:

Input: **(1) A graph G =** *(V, E),*
(2) Every vertex v of *V* **is assigned one colour (***V* **← Colour{1,2,3,…}**
and to be assigned second colour with constraint of
$\langle colour(v) \to N \mid colour(v_i) = colour(v_j) \Leftrightarrow (v_i, v_j) \notin E \rangle$.
(3) *V* **← Downlink Demand{1,2,3,…}**
Output: **A colouring c:** *V* **← C{1,2,3,…}**

Begin
1. for *j* from largest down to smallest Demand do
2. for each v with label *j*
3. Second-Colour(v) ← First-Available-Second-Colour (G, v , C)
End

Procedure First-Available-Second-Colour (G, v , C)
1. Taken ← { }
2. for each coloured vertex *u do*
3. if (*u, v*) $\in$ *E,* **Taken ← Taken $\cup$ colour (*u*)**
4 Return (smallest colour $\notin$ Taken)


Newly Coloured
(1)
Newly Coloured
(2)
Newly Coloured
(3)
Newly Coloured
(4)
Newly Coloured
(5)
Newly Coloured
(6)

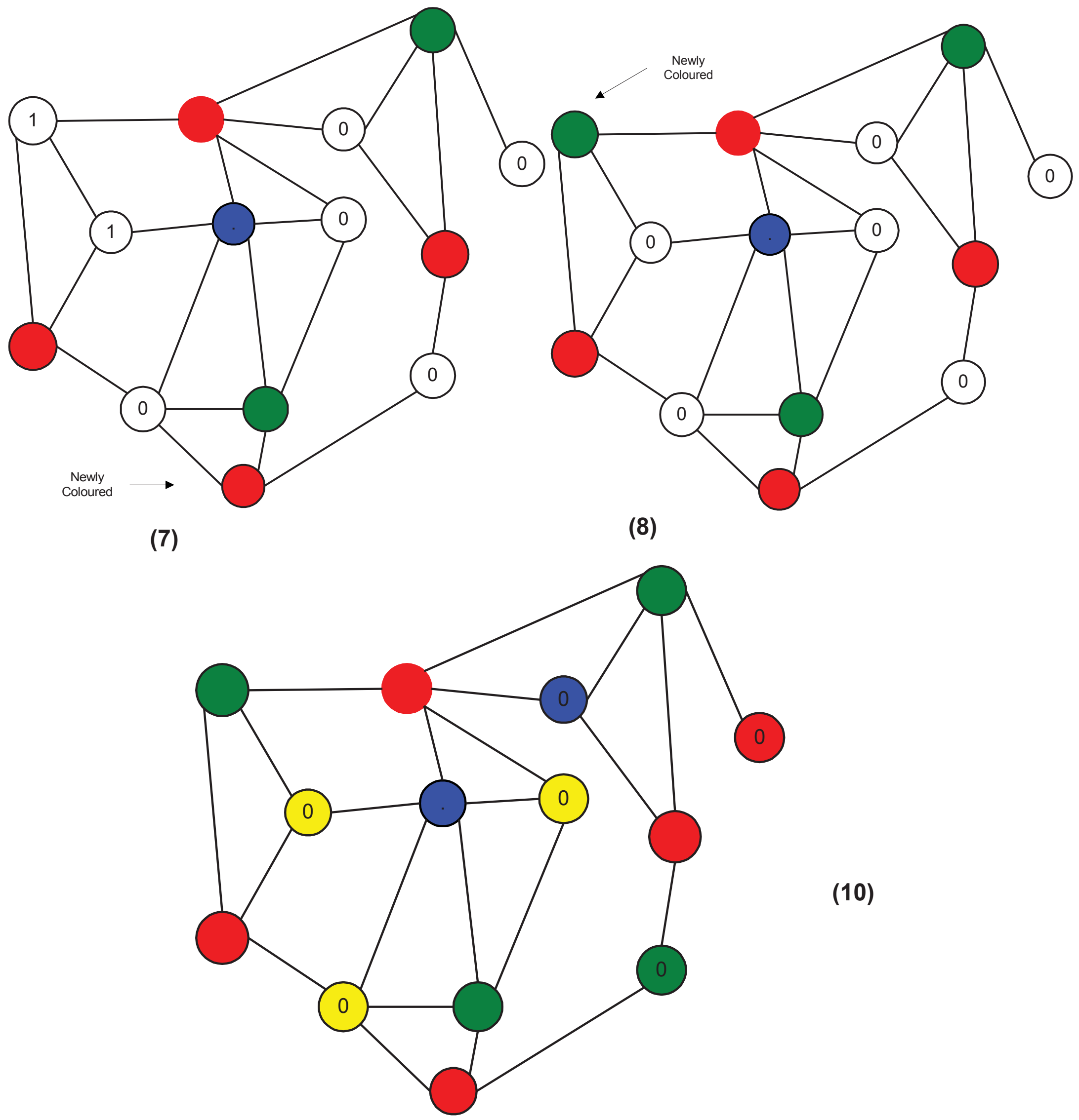


Figure 56 **Example of Maximum Degree First on a Planer Graph**

6.6.7 Summary

We have seen that an OBSS environment can be modelled as a graph, particularly a possible planar graph in most practical cases; its corresponding channel assignment is considered as a vertex colouring problem on such graphs. This vertex colouring problem has been proved NP-completeness either in a planar point graph or in a common planar graph. Therefore, the number of colours (time slots) we need for the assignment problem can always be found.

According to the characteristic of the OBSS deployment, the common OBSS environment can always be represented by a planar graph. In such case, 4-colour/time span is enough to solve the problem. This also provides us the guideline for deploying OBSS, which is used to avoid over-density situation and network inefficiency. We finally have given an example on performing the colouring on a plan graph with 4 colours, using a heuristics of Maximum Degree First.

6.7 Performance Analysis

In order to analyse coordination schemes for the OBSS environment, we firstly analyse effectiveness of message transfer mechanisms, which are the major control and management function for coordination schemes. We then present its performance using a simple simulation model built on an IEEE 802.11b network in the OPNET discrete event simulation package.

6.7.1 Analysis on Messages Transfer and Its Overhead

Message that are transferred as control and management information for coordination includes overlapping information, traffic information and allocation messages. It involves two-way communications among entities of clients, AP and RM. We firstly study the client record lifetime, which would have large influence on the message distribution pattern. The message overhead can then be mainly determined by the study of its distribution pattern and size. We also outline the options of its implementation. Besides the distribution patterns and sizes of the messages, the choice of implementation also makes impact on its delivery performances.

Lifetime for client records

As discussed previously, timer of the record is used to keep the track of the current status of the overlapping. If timer expires as its lifetime, overlapping record is removed in the client. If the same overlapping is heard in the client, its timer of this overlapping record resets. In such way, client can keep track of the existence of overlapping. And it avoids sending too much report traffic to the AP every time it hears the overlapping. Therefore, it becomes important to have a good design of the timer and the lifetime for the timer is the critical parameter for its effectiveness and performance. The design refers to the mobility pattern research as below.

[Jardosh et al. 2003] [Hsu et al. 2005] [Camp et al. 2002] [Jardosh et al. 2005] investigate the mobility modelling in WLAN. [Hsu et al. 2005] carries further survey on mobility patterns for normal human walking scenario and [Camp et al. 2002] summarises mobility models, which have been proposed and used in the performance evaluation of ad hoc network protocols.

BSS coverage, mobile client speed and pause time are the main aspects to characterize the user mobility and the stopovers in other BSS. In most models, client-moving speed is found as uniformly distributed between speed 0.5m/s to 1.25m/s with consideration to a normal pause time. And the time step is defined between 0.1 to 2 second for location update. If the range of 0.1 to 2 second is selected, movement distance is between 0.5 to 2.5 m, and such distance is considered needed for location update in a normal BSS with 300 meters range. If the BSS has less range, the time step should be defined smaller to reflect the sensitive change in the short distance. Therefore, we adopt the lifetime for client records between 0.1 to 2 second as the same as the time step used in most mobility modelling.

Discussion on message size, distribution pattern and its overhead

According to the design in Section 6.5 and 6.6, we summarize the size of overlapping information, traffic information and allocation information messages as below. The message sizes are between 4 to 66 bytes and relatively small compared to a general management/control frame in an 802.11 WLAN system.

- Overlapping information: 4 - 66 octets [2 octet (status) + 2-64(BSSID & SSID)].
- Traffic information: 8 octets (4 octets for 2 status and 4 octets for two real).
- Allocation messages: 18 octets (standard size for information element of Timestamp, beacon interval and CF parameter set) on the wired link. (A standard Beacon is 33- 324 octets; it is part of standard overhead for 802.11 and should not be counted as an additional overhead to the network.).

As described in the design, the update of client overlapping record triggers the overlapping message transfer and then the allocation message transfer. Therefore, inter-arrival for messages transfer largely depends on the distribution pattern of client record updates. An update only occurs when a new overlapping is detected or when an existing overlapping disappears (the timer of this client record expires).

[Guerin 1987] [Hsu et al. 2005] [Jardosh et al. 2005] found out that handover inter-arrival process is an exponential distribution (memoryless random Poisson distribution). In [Xhafa and Tonguz 2004], the mean outcome of the exponential distribution in a wireless network is calculated and 30 second is suggested and used in their performance analysis.

Given consideration to these references, the inter-arrival pattern of message transfers is an exponential distribution with mean outcome of 30 second. As the messages sizes range from 4 to 66 bytes as discussed before, therefore, the average overheads that are caused by the message transfer as the addition to the network are relatively very small.

Implementation options for message transfers

- Option 1: Transferred as files in the application layer:

The message can be contained in a simple text file and transferred as the same way that most of the common system configuration files are transferred from the server to the client. The advantage of this option is that it is independent to any wired and wireless network standard, and therefore provides better scalability and applicability to the network. The disadvantage is that the file transfer requires supporting protocols, such as FTP or TFTP, and every message requires certain header in the TCP/IP layer for management and control, compared to those in Option 2 (MAC management) and it produces additional overheads. Based on the analysis of the message distribution pattern and size, the application profile for message transfer largely matches the profile of standard database access application.

- Option 2: Transferred in the pre-defined management frames in the MAC layer of IEEE 802 networks.

This option is preferred because the bridge operation within 802 sub-network groups brings simplicity and unique for system management; and the existing IEEE 802 MAC service facilities also provide ease of implementation. Additionally, the allocation message transfer between client and AP is strongly advised to implement as parts of beacon operation in the 802.11 MAC, which would further reduce the implementation complexity and overhead of message transfer. The implementation details in the context of 802 standards are given in the Appendix B.

Secondly, option 2 provides better transfer quality and less delivery time as compared to option 1. As known, management and control frames normally have higher priority or better treatment for transmission than application data traffic, either in wireless network or in wired network. Therefore, message transfers are assured for better performance if they are implemented as parts of the MAC Layer management.

Discussion on delivery time and quality across wireless and wired link

It is desired that the RM receives the messages timely, so the report messages are able to report the current overlapping and traffic conditions precisely. The RM is then able to work out the scheme to meet the current situation. More importantly, the allocation messages also need to be timely transferred. On the other hand, the message transfer should have minimum drop rate.

We firstly study the worst scenario that the messages are transfer in the application layer. We carry out a simple simulation in an environment, consisting of a wireless subnet, a wired distribution network, and file servers. The system carries the message transfer application with other simple applications of HTTP/FTP. The message transfer is modelled as an exponential distribution with mean outcome of 30 second with constant average size of 35 bytes. The results show that the message delivery time is less than 10ms with a very minor dropped rate, compared to 20-30ms for HTTP/FTP application with assumption that network is normally loaded.

As seen, the performance of message transfer in the application layer meets the delivery requirements in terms of time and quality if the message transfer is implemented as parts of management in the MAC layer, the performance of delivery time and quality would be only better than that if it is implemented in the application layer, as discussed above.

6.7.2 Analysis on Resources Coordination

The analysis above shows that the overlapping information, traffic information and allocation information can be transferred effectively between AP, RM and clients. After the RM is updated with the current overlapping information and traffic information, if necessarily, it figures out the coordination scheme and sends the scheme information in the allocation message to the clients and AP.

Traffic information update in RM would make impacts on balance of uplink and downlink. Unlike the contention scheme in a single BSS, uplink and downlink are separately in two periods supported by contention access and polling access in the coordination scheme. After receiving the traffic information in RM, RM may dynamically control the allocation of contention (uplink) and polling (downlink) period. The medium alternates between uplink period and downlink period. Therefore, balance of uplink and downlink throughput can be achieved by the dynamic control. Unfairness issue between uplink and downlink can be minimised by dynamically controlling the resources allocation for downlink/uplink according to the traffic demand. Therefore, even the contention scheme is used; we would not have the fairness problem for uplink and downlink balance, which we observed previously when a contention scheme is used in a single BSS.

Based on the overlapping information, RM would then be able to work out the grouping assignment for coordination and consequently the superframe structure for each BSS, following the group assignment technique discussed in Section 6.6. Regardless of the size of the system and the numbers of BSS involved, all BSSs can be always assigned into several groups for coordination using the grouping assignment. Therefore, coordination scheme provides scalability for the OBSS environment. The performance analysis of group assignment technique has been discussed in Section 6.6. In the following discussions, we assume the result of assignment is 3-colourable, which is the common and typical scenario.

We now further analyse how the unique superframe is implemented in each BSS and how the coordination scheme performs at the levels of MAC in each BSS. With the assumption of 3-colourable assignment, three groups of A, B, and C are used for BSS grouping and its superframe structure is as shown in Figure 49. As seen, the coordination separates uplink and downlink transmission and the medium alternates between uplink period and downlink period. Therefore, we analyse the coordination both in uplink period and downlink period subsequently. The analyses mainly follow the metric of interference probability, fairness among BSSs, medium utility, and QoS support (Service differentiation).

6.7.3 Performance Analysis on Uplink Period

Compared to the analysis on a single BSS, the analysis on Uplink Resources Coordination for the OBSS environment additionally addresses the interference possibility across co-located BSS and also its medium utility. These are the fundamental functions to be achieved by a coordination scheme. Its performance is then further examined according the fairness and sharing issue across co-located BSS, and its QoS supports for multimedia traffic.

Medium Utility and Interference across Co-located BSSs

Within a single BSS, clients that can hear each other, shares the medium together and there is no spatial reuse. However, in OBSS environment, clients (in different BSSs) that are far away from each hearing ranges and do not interfere each other, do not share the medium and can transmit data at the same time. Therefore, during uplink period, spatial reuse can be achieved naturally when using contention scheme and its performance largely depends on the location density of BSSs in the OBSS environment.

However, for those BSSs that can hear each other partially or fully, the possible interference issue among them needs to be paid attention to. And discussion of interference issues is limited among these co-located BSSs, which can hear each other partially or fully. As known, when clients operate in the contention scheme, they can all avoid the interferences with the collision avoidance mechanism provided as the nature of the contention scheme. Also the coordination scheme would only allow the contention scheme to be used in every BSS for the uplink transmission during the uplink period. Therefore, during the uplink period, even several BSSs may be co-located; they can avoid transmission collision from each other with the collision avoidance mechanism as parts of contention scheme. The interference problem becomes the general collision and hidden terminal issue of the contention schemes in a single BSS, which has been well studied and solved [Crow 1997][Bianchi, G. 2000][Prasad 1999].

Fairness across Co-located BSSs

When several BSS co-locates and share the medium together, fairness among these BSSs should be ensured. If each BSS is equally loaded with same traffic pattern and configuration and service differentiation for all BSS is unique, the performances are expected approximately equal under the contention scheme. A simple simulation (case-one: even-loaded scenario) is carried out to further confirm this as below.

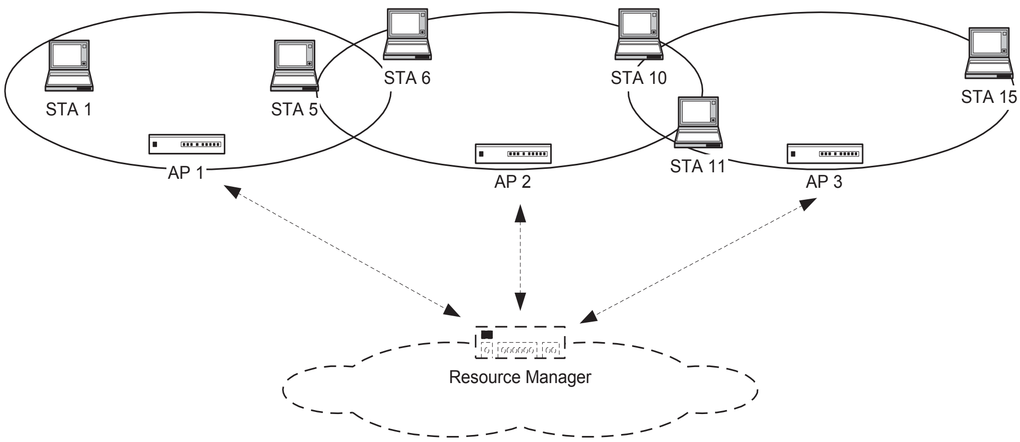


Figure 57 Simulation Layouts

Network Configuration consists of BSS A, B, and C representing group A, B and C as shown in the Figure 57. They co-locate and overlap each other in the worst scenario that all clients hear each other. Each BSS have two to three active clients transmitting traffic at all time. All clients are equipped with the basic contention scheme, based on the OPNET WLAN model. We assume that not only neighbouring clients can experience interferences but also all clients can hear each other within range of 300m. Three BSSs are virtually connected by fixed network and communicated with the Resource Manager. The channel assignments (time slot assignments for coordination) are pre-defined for each AP. All other assumptions, traffic model and test parameters remain the same as those in Chapter 4.

BSS A

client Name: A1		Role: VoIP upload	Location:		
Data Sources					
Destination for client	Mean Rate	Inter-arrival (ms)	MSDU Size	MAX Delay ms	Notes
AP_A	36.8 Kbps	20	92	30	

client Name: A2		Role: VC upload	Location		
Data Sources					
Destination client	Mean Rate	Inter-arrival (ms)	MSDU Size	MAX Delay ms	Notes
AP_A	1,41Mbps	1	1464	100	

BSS B

client Name: B1		Role: VoIP upload	Location:		
Data Sources					
Destination for client	Mean Rate	Inter-arrival (ms)	MSDU Size	MAX Delay ms	Notes
AP_B	36.8 Kbps	20	92	30	

client Name: A2		Role: VC upload	Location		
Data Sources					
Destination client	Mean Rate	Inter-arrival (ms)	MSDU Size	MAX Delay ms	Notes
AP_B	1,41Mbps	1	1464	100	

BSS C

client Name: C1		Role: VoIP upload	Location:		
Data Sources					
Destination for client	Mean Rate	Inter-arrival (ms)	MSDU Size	MAX Delay ms	Notes
AP_C	36.8 Kbps	20	92	30	

client Name: C2	Role: VC upload	Location	
Data Sources			

Destination client	Mean Rate	Inter-arrival (ms)	MSDU Size	MAX Delay ms	Notes
AP_C	1,41Mbps	1	1464	100	

client Name: C3	Role: data upload	Location:			
Data Sources (only when overloaded in case 2&3)					
Destination client	Mean Rate	Inter-arrival (ms)	MSDU Size	Delay Max	Notes
AP_C	3.68Mbps	Exp(5)	2300		

Table 8 Traffic Configurations for Uplink OBSS Scenarios

Only uplink traffics are simulated from clients to three APs. Each client has one of three applications (VoIP, Video conference (VC) and bulk data) with single service differentiation. Case one scenario represents even-loaded situation in each BSS, where three BSSs have the same active VC and VoIP transmission; Case two and three scenarios represent over-loaded situation (for discussions in the following sections.) where BSS C is additionally overloaded with large bulk data. The detailed traffic configurations are listed as in Table 8.

In this even-loaded scenario, network has relatively good total throughput at 4.3Mpbs and each BSS does not experience any data drop due to interference, compared to scenario that system lacks of coordination, which have been reported in [Dunat et al. 2004]. The use of collision avoidance mechanism in the contention scheme, as parts of coordination can avoid potential interferences between co-located BSSs during the uplink period.

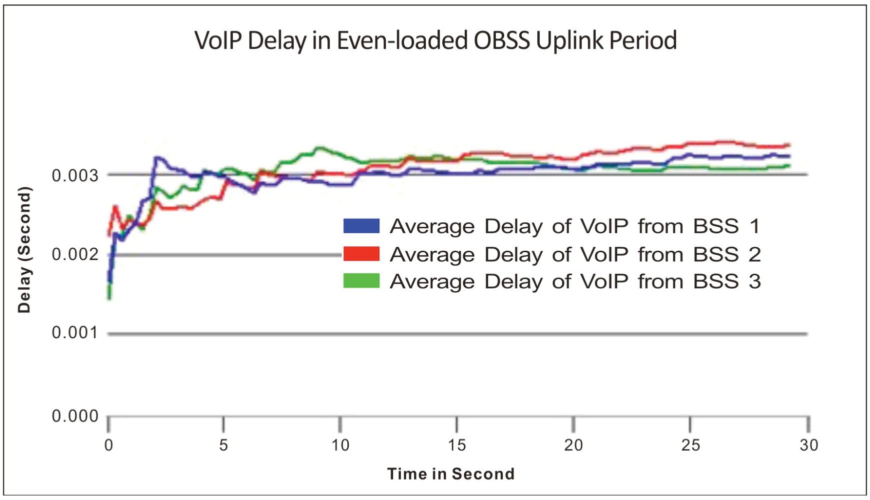


Figure 58 VoIP Delay Performances in Even-loaded Uplink Scenario

As also shown in Figure 58, the delays of VoIP application in each BSS are all approximately less than 5ms and the throughput of VC in each BSS is close at 1.43Mpbs. VC and VoIP performance in each BSS are treated fairly and approximately the same in an even-loaded scenario. Therefore, when the co-located BSSs are even-loaded with same traffic configuration and unique service differentiation, they can fairly share the medium together following the coordination.

Resource Sharing across Co-located BSSs when Overloading

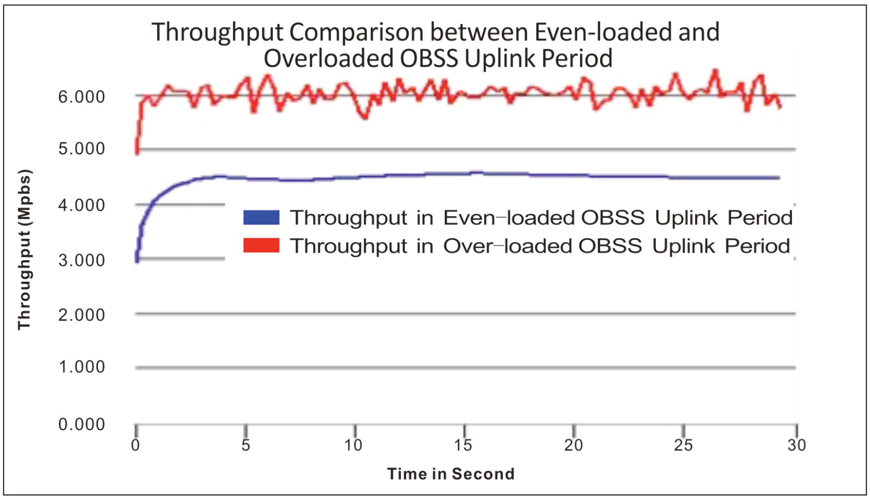


Figure 59 Total Throughput Performances in Over-loaded Uplink Scenario

However, network is unlikely often even-loaded in the reality. In the worst case where co-located BSSs are fully overlapped each other, resources are limited and shared by all co-located BSSs. These limited resources should be maximally utilized if there are BSSs with overloading traffic. If these BSSs are uneven-loaded, the overloaded BSS should fill the remaining capacity left over from other BSSs. A simple simulation is carried out to investigate this in an overloading scenario (Case-two), where three BSSs have the same active VC and VoIP, while the third BSS is overloaded with one more additional bulk data. The traffic configurations of Case-two are detailed in Table 8.

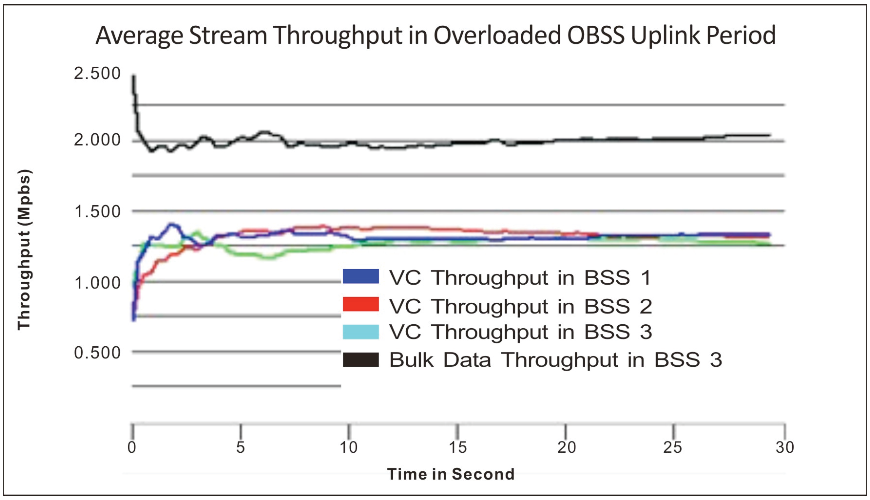


Figure 60 Throughput Performances in Over-loaded Uplink Scenario

Figure 60 shows that bulk data stream can fill up the remaining capacity and have throughput of 2Mbps. The total throughput of the system for three BSSs increases from 4.3Mpbs to 6Mbps as shown in Figure 59, when the third BSS is overloaded with a bulk data stream. The contention scheme supports the resource sharing when system is overloading during the uplink period.

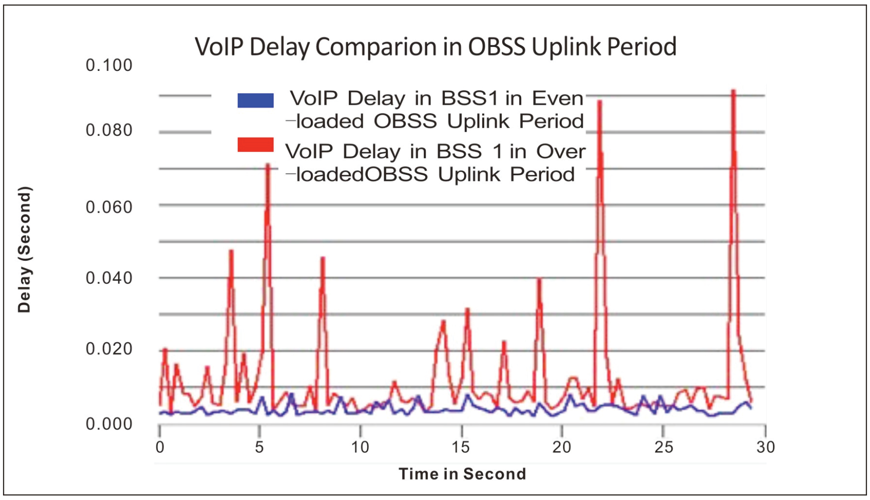


Figure 61 VoIP Delay Performances in Over-loaded Uplink Scenario

However, even thought the throughputs of VC in each BSS can still have 1.30Mbps close to those in even-loaded, we find that delays of VoIP application in each BSS all have large jitter and their delays increase as shown in Figure 61. The delay performances are seriously affected by the overloading from the bulk data. According to the fairness principle among the co-located BSS, even when any of the co-located BSSs are overloaded with their aggressive applications, such as bulk data, they should not affect the performance of applications in other BSSs, particularly for those delay-sensitive applications, like VoIP. The resource sharing should be provided with condition that the medium is not violated by any aggressive applications from overloading BSSs.

These can be achieved by implementing unique service differentiation among all BSS. The aggressive applications (such as bulk data) in the overloaded BSS should be regulated by being assigned a lower Access Category (AC) for the medium access. In such way, it would not affect other applications assigned with higher AC, either in its own BSS or neighbouring BSS. And often, applications of VoIP and Video are considered to have higher AC. Therefore, VoIP and Video are ensured for quality transmission while the bulk data can still fill the remaining capacity for a better medium utility.

These performances of service differentiations in contention schemes have been well studied. The performance of EDCF, as one of main modified contention scheme, has also been fully reported in [Choi et al. 2003] [Mangold et al. 2002a] [Kim et al. 2003] [Qiao and Shin 2002]. They mainly prove the service differentiation effectiveness and efficiency using sets of EDCF QoS AC parameters and its well-accepted frame collision rate. The key of using these contention-based service differentiations (such as EDCF) for OBSS is unique QoS AC assignment as discussed in the design. According to the fairness principle, same applications in different BSSs should be assigned the same AC, and individual BSS is not allowed to change such assignment. Therefore, the same applications belonging to different BSSs are treated fairly.

In the overloading scenario above, all VoIP and VC in all BSSs should be assigned higher AC while bulk data in BSS C should be assigned a lower AC. We carry out simulation (Case-three) with such unique AC assignments. A simple Service differentiation is simulated, where Transmission Attempt is implemented as one of QoS parameters. The bulk data is assigned a lower AC with a larger Transmission Attempt.

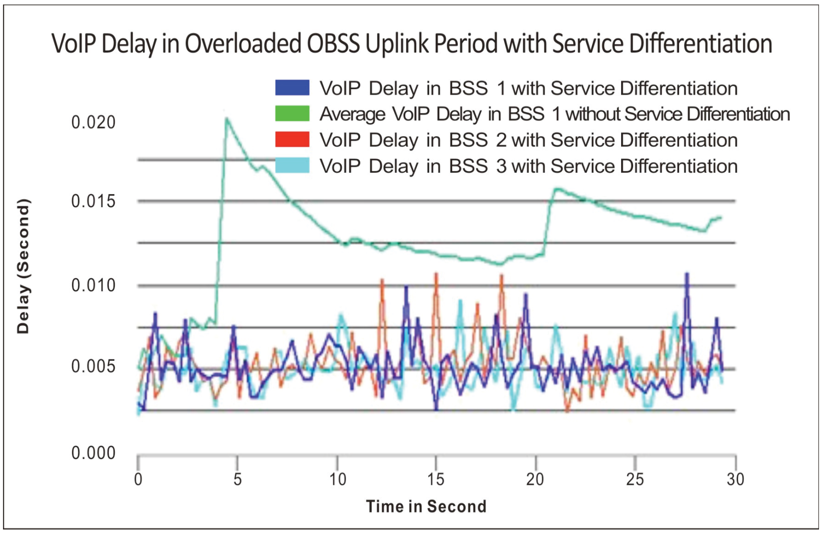


Figure 62 VoIP Delay Performances in Over-loaded Uplink Scenario with SD

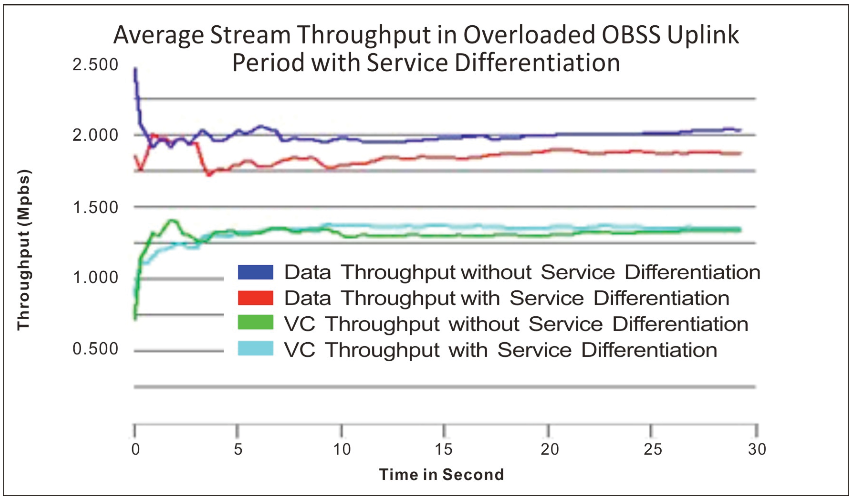


Figure 63 Throughput Performances in Over-loaded Uplink Scenario with SD

As seen in Figure 62, delays of all VoIP applications are guaranteed with satisfied performance. The delay performances of All VoIP are also approximately the same as the VoIP streams are all assigned the same AC according to the unique service differentiation assignment principle. Particularly, VoIP in BSS A and B are not affected by the aggressive bulk data application in BSS C, compared to those performances in the Case-two, where AC assignment is not in place. On the other hand, the bulk data is regulated by Service differentiation with a lower AC. However, it can still be able to fill up the remaining capacity left over from other application. As shown in Figure 63, it achieves throughput at 1.75Mbps, slightly dropping from 2.00Mbps of Case-two. As discussed, at the cost of slight drop in throughput for the aggressive application, unique service differentiation ensures the performance of the delay-sensitive application. The contention scheme with such unique service differentiation performs well in providing efficient resource sharing among co-located BSS, while in support of QoS for delay-sensitive applications.

6.7.4 Performance Analysis on Downlink Period

Spatial Reuse

As discussed, the coordination scheme divides all BSSs in several groups for downlink transmission. BSSs belonging to the same group do not need to share the medium and can transmit traffic simultaneously. Therefore, the downlink coordination with the grouping assignment achieves spatial reuse.

For those co-located BSSs, which belong to different groups, have to share the medium in the single channel. Obviously, the fewer groups we use, the better and further spatial reuse the scheme can achieve. Analysis on OBSS grouping has shown that the minimum number of groups can always be found using the algorithm discussed in the previous sections 6.6. In most case, four is the maximum group number it need for the OBSS grouping. The network deployment should only avoid the extreme case, in which the BSS with 'lightly' overlapping (having minimum overlapping neighbours) has no more than five neighbouring BSSs overlapped and more than 4 groups would be applied. The analysis given in the previous sections 6.6 confirms that coordination scheme can achieve the further spatial reuse with a maximum of 4 groups used in the OBSS grouping.

Interference Probability

The polling-based medium access mechanism is only allowed for the downlink transmission in the coordination scheme. As discussed in the interference observation section 6.2, a polling-based BSS may experience interferences from co-located BSSs, which operates contention or polling mechanism. As all BSSs only operate polling mechanism during the downlink period, we only need to consider the possibility of serious collisions between two polling-based BSS.

As known, grouping is used in the coordination scheme to separate co-located BSSs, which may interfere each other and each group is allocated individual time span for transmission. Therefore, the overlapping BSSs are assigned in different groups and then given different time spans, they transmit in separate time in the polling scheme. It effectively prevents interference between two polling BSSs that overlap each other. The possible collision should be reduced to the minimum level. The acceptable dropped data rate should be expected as the same as that in a normal polling BSS.

In short, the coordination scheme firstly avoids simultaneous use of the contention and polling schemes in different BSSs, to prevent any possible collisions between polling-based BSS and contention-based BSS. And the use of grouping and time division during the downlink period further enhance the prevention of collisions between two overlapping BSS, which both operate polling mechanism.

Fairness across Each Group

Member BSSs in the same group have been allocated the same time slot for transmission. The fairness issue among these member BSSs in the same group is obvious. However, those co-located BSSs are often assigned in several groups for different time slots to prevent interference from each other. The fairness issue among these co-located BSSs is the major discussion for the fairness problem across groups.

As known, each group is assigned the approximately equal span length, and these time spans do not overlap each other. Therefore, each group does not affect other and have the same duration to gain transmission opportunities. The fairness across each group is achieved in term of duration that they occupy the medium for. In fact, each BSS is given approximately equal duration to gain transmission opportunities except that it is also given a second group assignment (the details would be discussed in the later sections). Therefore, if each BSS is even-loaded with the similar traffic patterns and configuration, they all have similar performance either in term of throughput or in term of delay. If BSSs are loaded with different traffic pattern and configurations, the performance of each BSS purely depends on its own traffic pattern and configurations. For example, if a BSS has applications, which produces large and frequent packets, it mostly achieves high throughput. More importantly, these performances do not affect each other. A simulation is carried out to present an example.

Network Configuration consists of group A, B and C. Any BSS is treated and performs as the same as other members in the same group, so it is sufficient to simulate BSS A, B, and C to represent three groups. In such, BSS A, B, and C co-locate and overlap each other, so they are assigned in three separate groups. The superframe is implemented with three equal time spans as a typical case, according to the design. All BSS only become active when it comes to its span; otherwise, they are set NAV for idle at other time. All BSSs are only equipped with simple polling scheme. All WLAN parameters are as the same as those in previous simulations.

As described in Table 9, Each BSS would only have downlink traffic and its AP simultaneously transmits two or three applications (VoIP, light Video conference (VC), Audio and bulk data) to its clients at all time. Each BSS is loaded with different traffic. BSS A has three VoIP transmissions and in BSS B, there are light VC and Audio, while BSS C is overloaded with large bulk data.

BSS A

Client Name: AP_A	Role: VoIP download	Location:			
Data Sources					
Destination for client	Mean Rate	Inter-arrival (ms)	MSDU Size	MAX Delay ms	Notes
A1	36.8 Kbps	20	92	30	

Client Name: AP_A	Role: VoIP download	Location			
Data Sources					
Destination client	Mean Rate	Inter-arrival (ms)	MSDU Size	MAX Delay ms	Notes
A2	36.8 Kbps	20	92	30	

Client Name: AP_A	Role: VoIP download	Location			
Data Sources					
Destination client	Mean Rate	Inter-arrival (ms)	MSDU Size	MAX Delay ms	Notes
A3	36.8 Kbps	20	92	30	

BSS B

Client Name: AP_B	Role: VC download	Location			
Data Sources					
Destination for client	Mean Rate	Inter-arrival (ms)	MSDU Size	MAX Delay ms	Notes
B1	555Kbps	1	560	100	

Client Name: AP_B	Role: Audio download	Location			
Data Sources					
Destination client	Mean Rate	Inter-arrival (ms)	MSDU Size	MAX Delay ms	Notes
B2	36.8 Kbps	Uniform(0.0813-0.25	815	100	

BSS C

Client Name: AP_C		Role: bulk data download	Location:		
Data Sources					
Destination client	Mean Rate	Inter-arrival (ms)	MSDU Size	Delay Max	Notes
C1	3.68Mbps	Exp(5)	2300		

Table 9 Traffic Configurations for Downlink Scenarios

We now firstly look at performance in individual BSS. As seen in Figure 64, three VoIPs in BSS A have good delay performances, which all are less than 20ms. BSS B also maintains a satisfied performance for an audio and light VC, as its audio stream performs at less than 25ms delay as shown in Figure 64 and light VC achieves approximately 600Kbits/second.

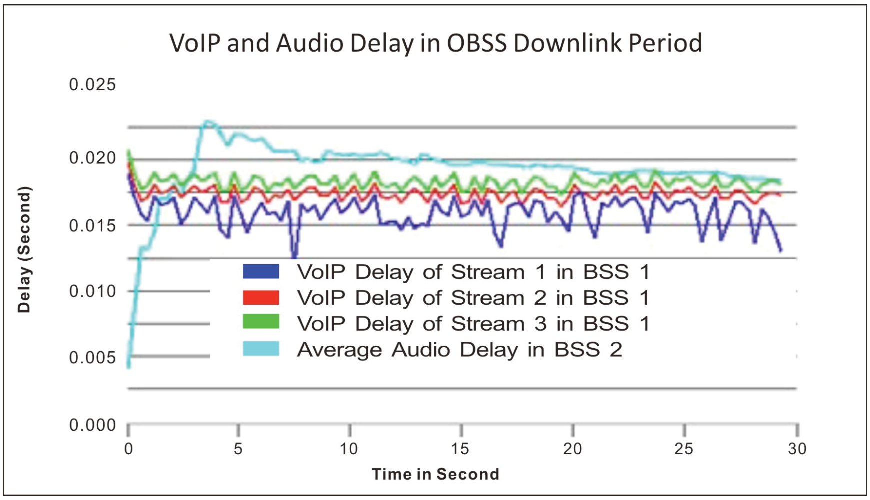


Figure 64 Delay Performances for VoIP and Audio in Downlink Scenario

However, Bulk data in BSS C experiences data drop at 1.75Mbits/second as shown in Figure 65. The application reaches the capacity of BSS3 during the assigned duration and achieved throughput of 1.8Mbits/second. While BSS C is overloaded with such aggressive bulk data and has consequent data drop, it does not gain further transmission opportunities at the cost of any possible performance degrade in other applications in BSS A and B at all. In fact, the coordination scheme solves the overloading problem with its second group assignment, which will be discussed in later section.

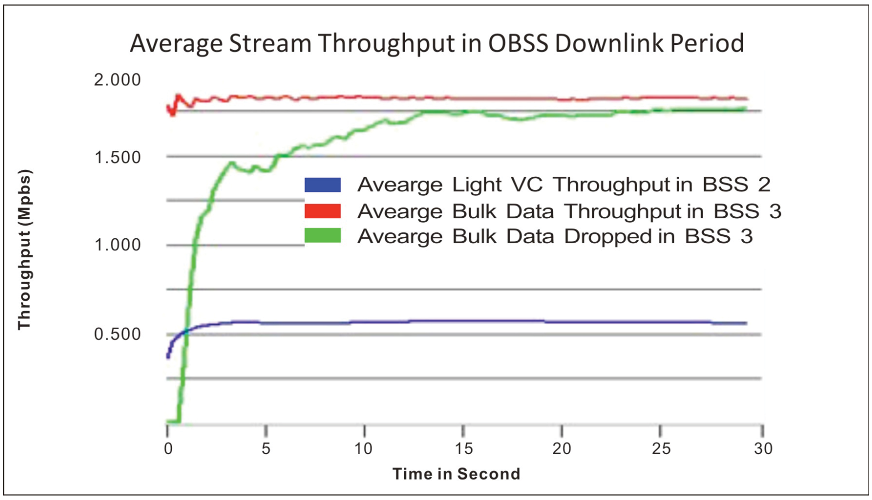


Figure 65 Throughput Performances in Downlink Scenario

Superframe Length and Delay Performance

As discussed in the Chapter 5, in a single BSS, if a simple prioritised polling mechanism is implemented, even VoIP application is always on the top of the polling list, its delay performance largely depends on the duration/length of the superframe. In our coordinated OBSS environment, regardless what polling mechanisms are used in the BSS, VoIP application has to wait for certain duration before it gets chance for transmission in its next span. The duration consists of uplink period and other time spans, which are assigned to other BSSs/groups. It is in line with superframe length. This causes the additional challenge for the system to ensure delay performance for VoIP applications. The loner the superframe length is, the challenge is more serious. Therefore, superframe length should be also considered less than certain duration in order to have an acceptable delay performance for VoIP applications.

However, if the superframe were assigned too short, as there are multiple Beacons (3-4 normally) in each superframe, it would produce relatively large overhead in a short superframe. Consider that a large Beacon is 324 octets and it is transmitted in a BSS_Basic_Rate (normally 1Mbits/s for management and control frames), it takes approximately 2.4ms to complete its transmission. And overhead for 3-4 beacons may be considered relatively large in a 20ms superframe. Therefore, the medium efficiency is better when a longer superframe is used.

In order to maintain a minimum waiting period for VoIP to achieve an acceptable delay performance, while also minimizing the multiple beacon overhead impact, we need to find an optimal range for superframe length. The simulation below is carried out to find the optimal solution. The network and traffic are configured as same as above, while different superframe lengths are used.

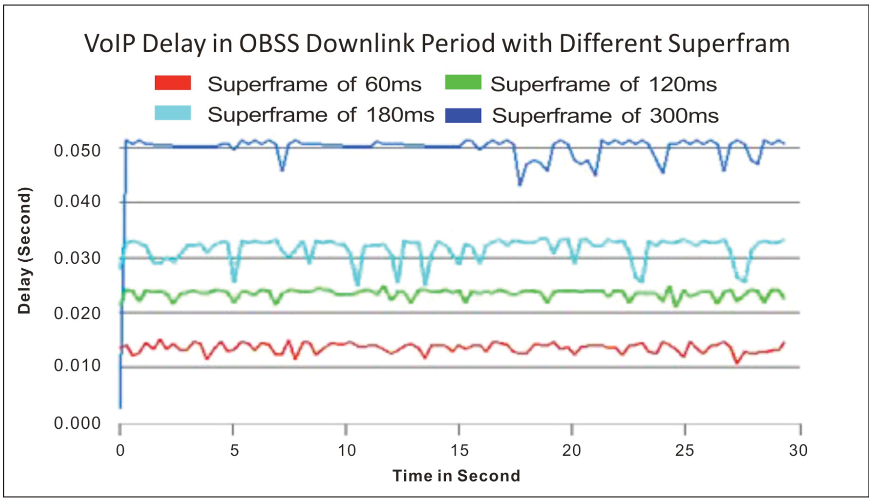


Figure 66 VoIP Delay in Downlink Scenarios with Different Superframe Length

The main focus of investigation is to find out how VoIP performs in term of delay in line with the changes of the superframe length. Therefore, we firstly look at the VoIP performance in BSS A, which is even-loaded as a typical case. As seen in Figure 66, as the superframe length increases, the VoIP delay performance in BSS A gets worst. When the superframe reaches 300ms duration, the VoIP has 50ms delay. As a good delay performance for VoIP should be less than 30ms; therefore, the recommendation for the superframe length should be no more than 180ms-120ms as shown in Figure 66.

Further simulation below is carried out to find out the lower threshold for the superframe length, with the objective of achieving medium utility. In the simulation, a 30ms superframe is implemented and the traffic and network configurations are as same as above. The results show that the system capacity has degrades particularly in those BSSs with multiple streams with frequent packet arrival. In BSS A, all three VoIP application experience large delay and jitter. In other words, BSS A cannot maintain three VoIP applications any more, when superframe is less than 30ms. In BSS B, VC throughput has obvious decrease as shown in Figure 67. Compared to the scenario that a 90ms superframe is implemented, the VC throughput drops from 550Kbits.second to 500Kbits/second. At the same time, Audio stream in the same BSS performs worst with large delay as well.

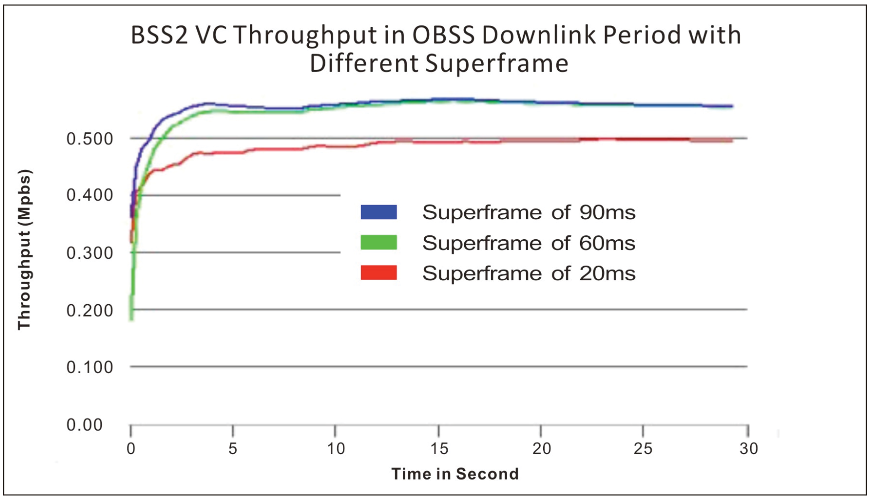


Figure 67 VC Throughputs in Downlink Scenarios with Different Superframe Length

As seen, the superframe length is a key performance parameter for polling mechanism. RM should consider the implementation recommendations above for decision of the superframe length when defining its coordination scheme. In order to provide service differentiation, particularly for delay-sensitive applications, superframe length is recommended with an upper threshold of 180ms-120ms. On the other hand, it should also have lower threshold, to prevent the system degrades as shown above. A range of 30ms-50ms is suggested based on the simulation implemented with common traffic configuration.

6.7.5 Performance Analysis on Second Group Assignment

In order to maximize the medium utility and satisfy transmission needs of BSS with high traffic demands, longer transmission period may be considered for these BSSs. One solution is to define an uneven longer span and assign all these busy BSSs into the same group for the longer span. However, they are not always able to assign into the same group/span as they are randomly deployed and may overlap each other. Therefore, uneven span is inappropriate. Instead of uneven span, a second group/span assignment is considered for these busy BSSs.

As the BSS with the overloading traffic can be assigned its second group if the environment is allowed (the BSS is only be assigned its second group, if none of its second group member overlaps with this BSS; details see Section 6.6.2), it would have two spans for transmissions. Compared to one span for most of BSSs. the BSS with its second group assignment has longer transmission period. Therefore, these busy BSS would have a better performance with more transmission opportunities and the overall system performance improves as well. Such improvements do not make any effects on any other BSSs, as the second group assignment would strictly follow the grouping principle that none of any two overlapping BSS would be assigned in one group.

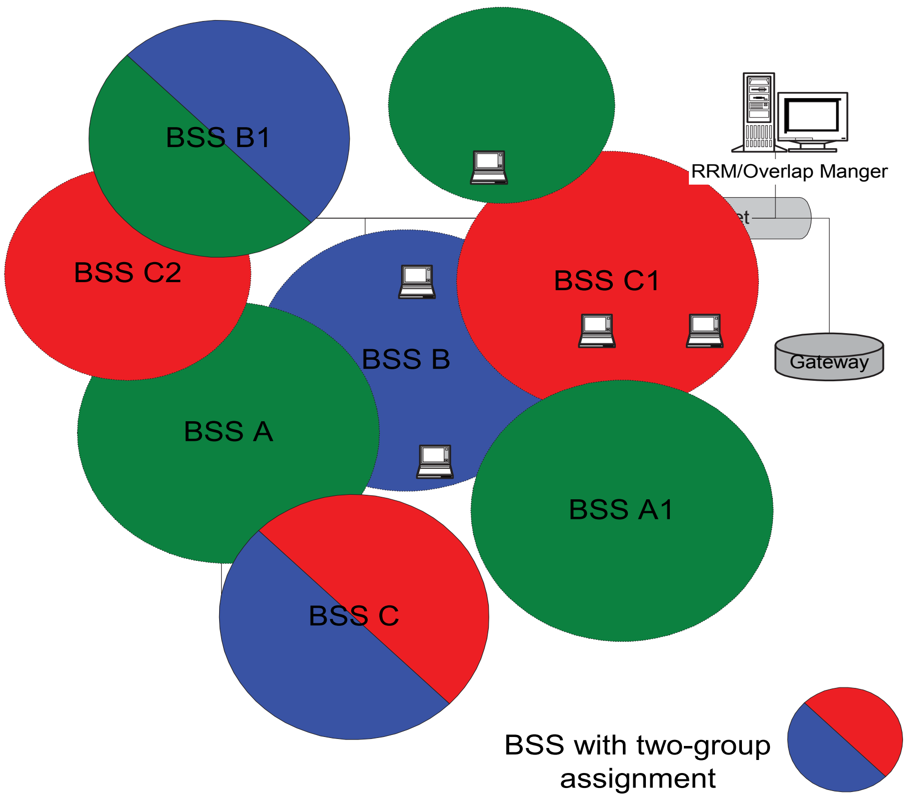


Figure 68 An Example of Second Group Assignment

We give an example here. In previous scenario, we have seen that BSS C is overloaded with aggressive bulk data and experiences data drop. It is admired that BSS C can be assigned its second group to satisfy the bulk data transmission. If BSS C does not overlap with any members of group Blue/B (such as BSS B), it can be assigned its second group and allocation additional time span B. An example of final grouping scenario is shown in Figure 68 when BSS C is given its second time span additionally in a three-group environment.

Firstly, we should examine if such additional arrangement on BSS C have any negative impact on its neighbouring BSS, such as BSS B and A, by looking at performances of their applications, such as the VoIP/Audio Delay and VC throughput. Figure 69 shows both VoIP applications in neighbouring BSS A perform well within 15-18ms, this is in line with what we have seen in Section 6.7.4. BSS A does not have any side impact from the second group assignment in its neighbouring BSS C.

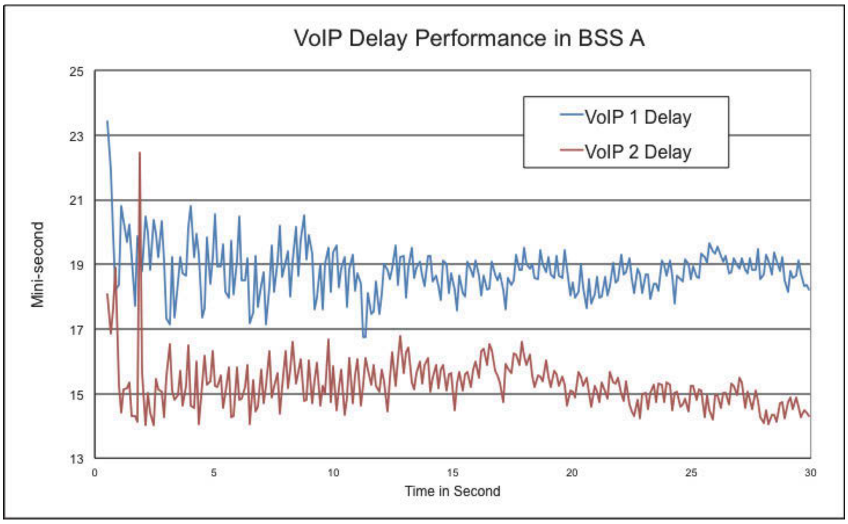


Figure 69 VoIP Delay Performance in Neighbouring BSS A

Another neighbour BSS B of BSS C has network loads of applications of Audio stream and Video conferencing application. The main measurements of BSS B performance are delay and throughput outcome of Audio stream and Video conferencing application respectively. Figure 70 below tells Audio performs well with delay in average within 19ms and 22ms. Secondly, we exams if downlink Video Conferencing throughput in BSS B could maintain in average 550Kbits/second in the scenario of the second group assignment. And we have seen such positive evident in Figure 71 below.

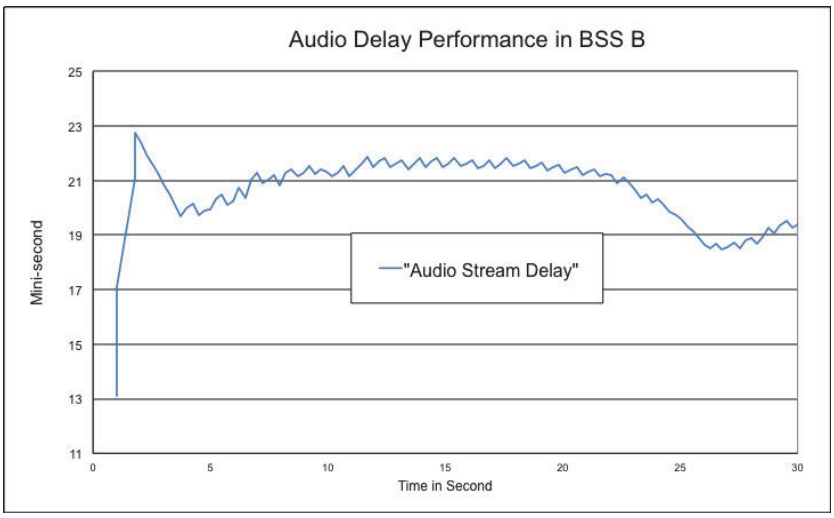


Figure 70 Audio Stream Delay Performance n BSS B

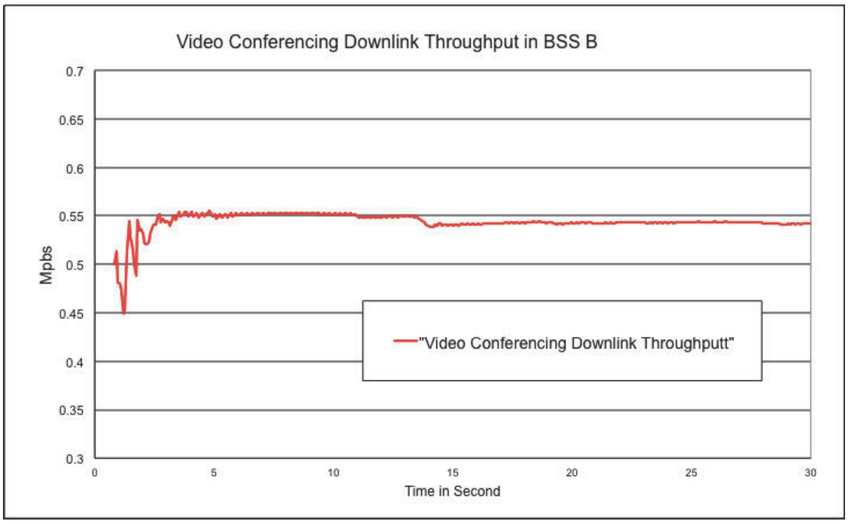


Figure 71 Video Conferencing Downlink Throughputs in BSS B

Finally, we verify if the expected performance improvements occur in the respective BSS, such as BSS C. As an example, BSS C has the second assignment, so as longer period, for transmission as admired and its throughput should improve accordingly. Consequently, the throughput improvement on the bulk data has been shown in Figure 72, from 1.80Mbits/second to 3.55Mbits/second. That is almost 100% improvement.

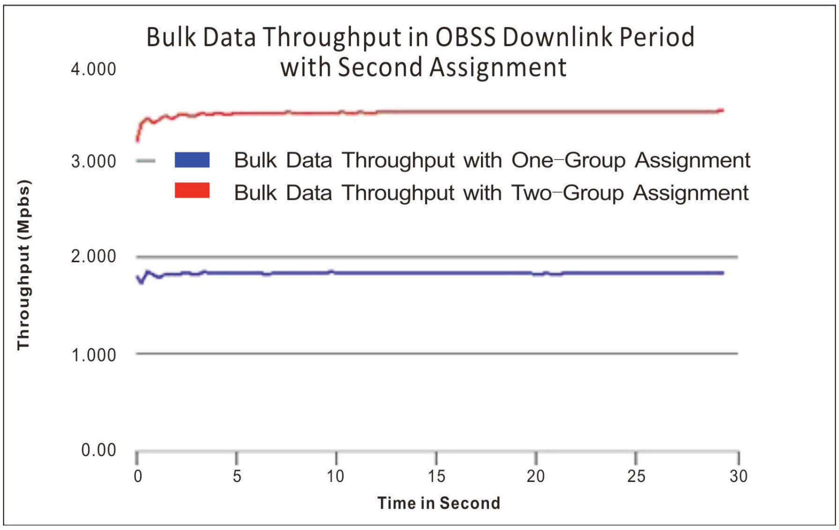


Figure 72 Bulk Data Throughput in Downlink Scenario with Second Assignment

However, not all the busy BSSs are always can be assigned second group. For example scenario in Figure 68, even if BSS B is additionally overloaded with another bulk data, it would be able to have second time span for transmission to meet the new demand, as it overlaps other BSSs belonging to both group A and C. If more groups are used in order to allow BSS B have second time span, it would only largely degrade the overall system performance instead because it is against the grouping objective of using minimum group for grouping. However, second group assignment maximise the medium utility by allowing busy BSSs have longer transmission period, whenever possible.

6.8 Conclusion

We have observed the unavoidable problem of overlapping BSS and pointed out that a central coordination should be in place to solve the OBSS interference problem, particularly in the common scenario when a WLAN domain is deployed with central resources management function in a high-density environment. The coordination separates uplink and downlink transmission into two periods, which are supported by contention and polling mechanisms respectively. The overlapping information reports inform RM the current overlapping topology. The message transfer/collection mechanisms are built to achieve reporting timely for RM to work out the coordination scheme. Traffic information update enables the dynamic control on the balance of uplink and downlink period. The grouping assignments are carried out based on the topology to divide BSSs into groups and time span to avoid downlink collision. The grouping analysis on 4-colourable scenario using graph colouring technique shows the medium efficiency and scalability of the coordination. The allocation message transfer mechanisms are also built to convey coordination scheme from RM to each BSS with minimum additional overhead introduced.

Our performance analyses show that the novel coordination not only prevents the overlapping collision but also achieves medium utility and fairness among the participated BSSs. Most importantly, it maintains a reasonable service differentiation supports particularly for delay-sensitive applications as the same levels as those in a single isolated BSS. The implementation recommendations, such as superframe length and deployment topology, are also investigated to ensure an effective and efficient performance for the proposed OBSS coordination scheme.

~ Our greatest glory in life consists not in never falling but in rising each time we fall ~

(Silken Laumen)

Chapter 7 Conclusion and Future Works

7.1 Summary of Results

In this dissertation, we have addressed the issue of delivering diverse QoS/QoE to mobile users in wireless/mobile networks. We firstly proposed a hybrid QoS architecture that follows the principles of Differentiated Service (DiffServ) model over the core part of the network and the principles of Integrated Services (IntServ explicit control) model locally over the wireless access segment. The model was then discussed in detail with an example solution in the context of WLAN exploiting the 802.11e QoS extensions, supported by an admission control core of Fair Intelligent Congestion Control (FICC).

Within our proposed hybrid architecture, the importance of service differentiation mechanisms was identified. As the basic option of the MAC service differentiation scheme, contention was studied and its QoS fairness problem was addressed, with respect to multimedia streams, particularly in a highly loaded and/or overloaded system. Such a problem would likely to affect performance of an infrastructure network as well as create asymmetry of throughput by bi-directional applications.

Polling schemes, as more sophisticated mechanisms, was then studied. While it provided more control access of wireless stations to radio resources via an intelligent polling, a sophisticated mechanism for traffic information update would be required. Consequently, we then raised a question on the necessity of explicit traffic information update for an intelligent polling, which formed the fundamental question for its effectiveness and efficiency. Based on these analyses, we proposed a novel Registered Multi-cycle Polling Scheme with an implicit information update mechanism achieved by stream registration, which is a component of our proposed Hybrid QoS architecture studied in Chapter 3. Not only was the design consistent with the whole QoS architecture, the simulation results also further confirmed improvements both on delay performance for real-time applications and medium utilisation with its Registered Multi-cycle Polling mechanism.

Finally, the study investigated the QoS problem in the situation of overlapping WLAN after the single cell scenario was studied in the previous sections. The necessity of overlapping coordination to support Service Differentiation was identified and the correspondent resource coordination scheme was then proposed. Initially, it was able to separate uplink from downlink transmission by assigning two individual periods in a superframe. Then each member BSS was assigned one group or more using graph colouring assignment technique. Each group was only allowed for communication within its assigned timeslot in order to avoid potential interference. The timeslot and group assignment not only offered collision avoidance but also achieved the efficiency and balance of uplink and downlink period by meeting the topology and traffic changes in the environment.

7.2 Further Development and Motivation

The studies have proven that wireless communication, particularly in unlicensed frequency bands, is a challenging task when QoS is required to support the multimedia Internet traffic. The necessity of spectrum management/coordination for acceptable QoE has been observed. The sharing of frequency spectrum seems to be a common topic of forming a society. Information exchange and management among society members is the key for the coordination.

In this dissertation, we have presented Information exchange and the coordination scheme in a centralised pattern given that the member subsystems form an infrastructure environment and are relatively stable in term of mobility. However, when the members not only interact each other but also are very mobile, the sharing coordination among the society members becomes even more challenging. In such scenarios, distributed coordination pattern should be considered instead of centralised infrastructure pattern. In particular, if these members operate in Ad Hoc mode where the distributed network is not available to provide an infrastructure environment for overlapping information exchange and management in a centralised mode, distributed coordination pattern seems to be an effective option.

Further motivation comes from the scenarios when members are from different types of network. Clearly, ETSI BRAN HiperLAN/2 (referred as HiperLAN/2) and WiMAX are radio system for Wireless LAN & MAN and popular standards nowadays. They define their own layer 1 and layer 2 protocols like 802.11 WLAN and aims to support QoS. The possible coexistence of 802.11 WLAN and HiperLAN/2 & WiMAX requires interwoking between these two as both operate on the 5GHz unlicensed band. However, different radio systems (802.11 WLAN and HiperLAN/2 & WiMAX) in an unlicensed spectrum are generally not capable of exchanging information because of the lack of a common protocol. In this circumstance, the prediction method may be used to allow coexisting entities to estimate the demands of other coexisting entities to avoid collision and share the limited media. From this point of view, new solutions such as prediction have to be found and verified to enable 'communication' between these coexisting entities, which do not share a common and obvious protocol that allows simplistic and dependable information exchange.

More Importantly, WLAN has been become so widely used almost in every smart-handset and notebook today. It is a really legacy technology, and its shortcoming in spectrum efficiency and interference remain for long time as long as we have all these notebooks and smart-handset in use on the markets. There is more and more serious interference incidents reported particularly in Metro transport systems in 2013. Co-existence among WLAN itself and other systems would become even more challenging because there would be more smart-handset and notebook in use and higher density deployments.

Another interesting topic would be how to combine the latest multimedia adaptation, QoE technique and the existing QoS mechanisms we discussed above. Obviously, there is a close relationship between QoS parameters and QoE parameters. The degree of importance and priority of each QoS parameter are influenced by the QoE and hence providing guidance to where and what to focus in adapting QoS framework. Only all these components are considered in a hierarchical structure, the problem of service quality in the packet switching Internet today, particularly over wireless/mobile segments, could be solved.

~ The world is moving so fast these days that a person who says it can't be done is generally interrupted by someone doing it ~

(Harry Emerson)

References and Bibliography

[3GPP 2003] 3GPP Web site: http://www.3gpp.org

[3GPP2 2003] 3GPP2 Web site: http://www.3gpp2.org

[Aad and Castelluccia 2001] Aad, I.; Castelluccia, C., "Differentiation mechanisms for IEEE 802.11", INFOCOM 2001, Twentieth Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE, Volume: 1, 22-26 April 2001,

[Agboma and Liotta 2012] F Agboma, A Liotta; "Quality of experience management in mobile content delivery systems"; Telecommunication Systems, 2012 – Springer

[Akhshabi et al 2011] Saamer Akhshabi, Ali C. Begen, Constantine Dovrolis; "An Experimental Evaluation of Rate-Adaptation Algorithms in Adaptive Streaming over HTTP"; In Proceedings of the second annual ACM conference on Multimedia systems (MMSys 2011)

[Alonso-Zárate et al 2012] J Alonso-Zárate, C Crespo, C Skianis, L Alonso; "Distributed point coordination function for IEEE 802.11 wireless ad hoc networks"; Ad Hoc Networks, 2012 – Elsevier

[Anastasi and Lenzini 1999] G. Anastasi, L. Lenzini, "QoS provided by the IEEE 802.11 wireless LAN to advanced data applications: a simulation analysis", Wireless Networks, Volume 6, Issue 2 1999.

[Antonio et al. 2001] J. Antonio García-Macías, Franck Rousseau, Gilles Berger-Sabbatel, Leyla Toumi and Andrzej Duda, "Quality of service and mobility for the wireless internet", Pages 34 – 42, Proceedings of the first workshop on Wireless mobile internet, July 21, 2001, Rome Italy.

[Arikan 1984] E. Arikan, Some complexity results about packet radio networks, IEEE Trans. Inform. Theory 30 (July 1984) 681–685.

[Arunabha and Huson 1997] Arunabha Sen, Mark L. Huson, A new model for scheduling packet radio networks, Wireless Networks, v.3 n.1, p.71-82, March 1997

[Aurrecoechea et al. 1998] Cristina Aurrecoechea, Andrew T. Campbell, and Linda Hauw, "A survey of QoS architectures", Multimedia Systems (1998) 6: 138–151.

[Awduche et al. 1999] D. Awduche, J. Malcolm, J. Agogbua, M. O'Dell, and J. McManus, "Requirements for Traffic Engineering Over MPLS", RFC 2702, Network Working Group, September 1999.

[Balachandran et al 2009] Balachandran, K., Q. B. Rudrapatna, J. Seymour, R. Soni, and A. Weber, "Performance assessment of next-generation wireless mobile systems," Bell Labs Technical J., Vol. 13, No. 4, 35-58, 2009

[Bianchi 2000] G. Bianchi, Performance Analysis of the IEEE 802.11 Distributed Coordination Function, IEEE JSAC, vol. 19, no. 3, March 2000.

[Banchs and Perez 2002] Banchs, A.; Perez, X; "Providing throughput guarantees in IEEE 802.11 wireless LAN", Wireless Communications and Networking Conference, 2002. WCNC2002. Volume: 1, Mar. 17-21, P130 –138

[Banchs et al 2001] Banchs, A.; Perez, X.; Radimirsch, M.; Stuttgen, H.J. "Service differentiation extensions for elastic and real-time traffic in 802.11 wireless LAN", High Performance Switching and Routing, 2001 IEEE Workshop on, 2001 P245 –249

[Banerjea et al. 1996] Anindo Banerjea, Domenico Ferrari, Bruce A. Mah, Mark Moran, Dinesh C. Verma, and Hui Zhang, "The Tenet Real-time Protocol Suite: Design, Implementation, and Experiences", IEEE/ACM Transactions on Networking, 4(1) 1–10, February 1996.

[Benslimane et al 2008] Benslimane, Djamal; Schahram Dustdar, and Amit Sheth (2008), "Services Mashups: The New Generation of Web Applications". IEEE Internet Computing, vol. 12, no. 5.

[Bernet et al. 2000] Y. Bernet; P. Ford; R. Yavatkar; F. Baker; L. Zhang; M. Speer; R. Braden; B. Davie; J. Wroclawski; E. Felstaine, "A Framework for Integrated Services Operation over Diffserv Networks", Request for Comments (RFC) 2998, IETF, November 2000.

[Bettermann and Rong 2011] Stephan Bettermann and Yue Rong, "Estimating the Deliverable Quality of a Fully Redundant Dispersity Routing System", 2011 17th Asia-Pacific Conference on Communications (APCC)

[Bhatti, and Knight 1999] Sallem N. Bhatti, and Graham Knight, "Enabling QoS adaptation decisions for Internet applications", London, UK, 1999.

[Bloch 2008] Bloch, Joshua (2008). "Effective Java (2nd edition)". Addison-Wesley. pp. 259–312. ISBN 978-0-321-35668-0.

[Bianchi, G. 2000] Bianchi, G., "Performance analysis of the IEEE 802.11 distributed coordination function", Selected Areas in Communications, IEEE Journal on Volume: 18 Issue: 3, March 2000, Page(s): 535 –547

[Biedl and Kant 1994] T. Biedl and G. Kant, A better heuristic for orthogonal graph drawing, in: Proc. 2nd Ann. European Sympos. Algorithms (1994).

[Blake et al. 1998] S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang and W. Weiss, "An architecture for differentiated services", RFC 2475, December 1998.

[Bonald et al. 2000] T. Bonald, M. May, and J-C. Bolot, “Analytic Evaluation of RED Performance”, in Proc. IEEE INFOCOM'2000, Tel-Aviv, Israel, March 2000.

[Bondy and Murty 1976] J. A. Bondy and U. S. R. Murty; “Graph Theory with Application”, American Elsevier Publishing Co. Inc., New York, 1976.

[Brand and Aghvami 1998] A. Brand and A. H. Aghvami, “Multidimensional PRMA with Prioritized Bayesian Broadcast – A MAC Strategy for Multiservice Traffic over UMTS”, IEEE Transaction on Vehicular Technology, vol. 47, no. 4, November 1998.

[Braden et al. 1994] Robert Braden, David Clark, and Scott Shenker, “Integrated Services in the Internet Architecture: an Overview”, Informational RFC 1633, June 1994.

[Braden et al. 1997] Robert Braden, Lixia Zhang, Steven Berson, Shai Herzog, and Sugih Jamin, “Resource Reservation Protocol (RSVP) - Version 1 Functional Specification”, Proposed Standard RFC 2205, September 1997.

[BRAIN D1.2 2001] IST-1999-10050 BRAIN, Deliverable D1.2, “Concepts for Service Adaptation, Scalability and QoS Handling on mobility enabled IP networks ”, March 2001. Web Site: http://www.ist-brain.org/

[BRAIN D2.2 2001] IST-1999-10050 BRAIN, Deliverable D2.2, “BRAIN architecture specifications and models, BRAIN functionality and protocol specification”, March 2001, Web Site: http://www.ist-brain.org/

[Callon et al. 1999] R. Callon, P. Doolan, N. Feldman, A. Fredette, G. Swallow, and A. Viswanathan, “A Framework for MPLS”, IETF draft, Network Working Group, September 1999.

[Camp et al. 2002] T. Camp, J. Boleng, and V. Davies, "A Survey of Mobility Models for Ad Hoc Network Research," Wireless Communication & Mobile Computing (WCMC): Special Issue on Mobile Ad Hoc Networking Research, Trends and Applications, Vol. 2, No. 5, pp. 483-502, 2002.

[Campbell et al. 1994] Andrew T. Campbell, Geoff Coulson, and David Hutchison, “A Quality of Service Architecture”, ACM Computer Communications Review, 24(2):6–27, April, 1994.

[Campbell et al. 1996] A. Campbell, C. Auurecoechea, L. Hauw, “A Review of QoS Architectures”, Proc. Of 4th IFIP International Workshop on Quality of Service (IWQS'96), Paris, France, March 1996.

[Cheng et al 2012] H Cheng, N Xiong, AV Vasilakos, L Tianruo Yang; “Nodes organization for channel assignment with topology preservation in multi-radio wireless mesh networks”; Ad Hoc Networks, 2012 – Elsevier

[Chevrollier et al. 2003] Chevrollier, N.; Rebala, O.; Golmie, N.; “Bluetooth and WLAN coexistence: challenges and solutions” Wireless Communications, IEEE, Dec. 2003

[Chhaya and Gupta 1995] Chhaya, H.S. and Gupta, S., "Throughput and fairness properties of asynchronous data transfer methods in the IEEE 802.11 MAC protocol," in Proc. Int. Symp. PIMRC'95, Vol: 2, Sep 1995.

[Chhaya and Gupta 1996] Chhaya, H.S.; Gupta, S.; "Performance of asynchronous data transfer methods of IEEE 802.11 MAC protocol", IEEE Personal Communications, Volume: 3 Issue: 5, Oct 1996 Page(s): 8 -15.

[Chiasserini and Rao 2002] Chiasserini, C.F.; Rao, R.R.; "Coexistence mechanisms for interference mitigation between IEEE 802.11 WLANs and Bluetooth"; INFOCOM 2002.

[Chlamtac and Kutten 1985] I. Chlamtac and S. Kutten, "A spatial reuse tdma/fdma for mobile multihop radio networks", in: INFOCOM (IEEE, March 1985).

[Choi et al. 2003] Sunghyun Choi, del Prado, J., Sai Shankar N, and Mangold, S., "IEEE 802.11e contention-based channel access (EDCF) performance evaluation", Communications, 2003, ICC '03. IEEE International Conference on, Volume: 2, 11-15 May 2003, Page(s): 1151 –1156.

[Cisco 2010] Cisco, "Cisco visual networking index: Global mobile data traffic forecast update," February 2010

[Clark et al. 1992] D. Clark, S. Shenker, and L. Zhang, "Supporting Real-Time Applications in an Integrated Services Packet Network: Architecture and Mechanism", Proc. SIGCOMM '92, August 1992, p. 14-26.

[Clark et al. 1994] Clark, D. et al.. "Integrated Services in the Internet Architecture: an Overview" RFC1633, June 1994

[Clark 2001] A. D. Clark, "Modeling the Effects of Burst Packet Loss and Recency on Subjective Voice", IPtel 2001 Workshop

[Coda 2010] Coda Research Consultancy, "Mobile broadband and portable computers: Revenue, user and traffic forecasts 2009-2017," July 2009.

[Collotta et al 2012] M Collotta, G Pau, VM Salerno, Scata, G; "A distributed load balancing approach for industrial IEEE 802.11 wireless networks", Emerging Technologies & Factory Automation (ETFA), 2012 IEEE 17th Conference

[Conforto et al. 2002] Conforto, P.; Tocci, C.; Losquadro, G.; Sheriff, R.E.; Chan, P.M.L.; Fun Hu, Y., "Ubiquitous Internet in an integrated satellite-terrestrial environment: the SUITED solution", IEEE Communications Magazine, Volume: 40 Issue: 1, Jan. 2002 P98 –107

[Coutras et al. 2000] Constantine Coutras, Sanjay Gupta, and Ness B. Shroff, "Scheduling of real-time traffic in IEEE 802.11 wireless LANs", Wireless Networks Volume 6, Issue 6 (December 2000), Pages: Pages: 457 - 466

[Crow 1997] Brian P.Crow, "IEEE 802.11 Wireless Local Area Network", IEEE Communication Magazine, 1997.

[Cui et al 2011] Y Cui, W Li, X Cheng; “Partially overlapping channel assignment based on “node orthogonality” for 802.11 wireless networks”; INFOCOM, 2011 Proceedings IEEE

[Davin and Heybey 1990] J. Davin and A. Heybey, “A Simulation Study of Fair Queueing and Policy Enforcement”, ACM Computer Communication Review, Vol. 20 No. 5, pp. 23-29, Oct. 1990.

[Deering and Hinden 1995] S. Deering and R. Hinden, “Internet Protocol, Version 6 (IPv6) Specification” RFC 1883, December 1995.

[Demers et al. 1990] A. Demers, S. Keshav, and S. Shenker, “Analysis and Simulation of a Fair Queuing Algorithm”, Internetworking: Research and Experience, Vol. 1, 1990, pp.3-26.

[Deng and Chang 1999] D-J. Deng and R-S. Chang, “A priority scheme for IEEE 802.11 DCF access method”, IEICE Transactions on Communications, E82-B (1), January 1999.

[Diepstraten and Haagh 2012] WJM Diepstraten, JPN Haagh; “Power control method in a wireless communication system”; European Patent EP0579372 Kind Code: B1, 2012

[Ding et al 2008] Y Ding, Y Huang, G Zeng, L Xiao; “Channel assignment with partially overlapping channels in wireless mesh networks”; Proceedings of the 4th Annual International Conference on Wireless Internet 2008

[Dixit et al. 2001] Dixit, S.; Yile Guo; Antoniou, Z., “Resource management and quality of service in third generation wireless networks”, IEEE Communications Magazine, Volume: 39 - 2, Feb. 2001; P125 –133

[Dohoon 2010] Dohoon Kim, “Application of the HoQ framework to improving QoE of broadband internet services”, Network, IEEE, Volume: 24, Issue: 2, 2010

[Dunat et al. 2004] Dunat, J.-C.; Elicegui, L.; Bonnet, C.; “Impact of inter-cell interference in a IEEE 802.11a network with overlapping cells” Personal, Indoor and Mobile Radio Communications, 2004.

[Durham et al. 2000] David Durham, Jim Boyle, Ron Cohen, Shai Herzog, Raju Rajan, and Arun Sastry, “The COPS (Common Open Policy Service) Protocol”, Proposed Standard RFC 2748, January 2000.

[Dutta et al.2012] Dutta, P.; Seetharam, A.; Arya, V.; Chetlur, M.; Kalyanaraman, S.; Kurose, J., “On managing quality of experience of multiple video streams in wireless networks”, INFOCOM, 2012 Proceedings IEEE

[Eladly and Chen 2004] T.M.Eladly, H.A.; Chen, T.M.; “An overlay algorithm to improve the support of multi-hopping in the IEEE 802.11 WLANs”; Communications, 2004 IEEE International Conference.

[Ephremides et al. 1990] A. Ephremides and T. V. Truong, Scheduling broadcasts in multihop radio networks, IEEE Trans. Communications 38 (April 1990).

[ERC 1999] ERC/DEC/(99)23. ERC Decision of 29 November 1999 on the harmonised frequency bands to be designated for the introduction of High Performance Radio Local Area Networks (HiperLANs)

[ERC 2002] ERC, "Study of the frequency sharing between HiperLANs and MSS feeder links in the 5 GHz band," ERC Report 67.

[ERC 2003] ERC, "Compatibility Studies Related To The Possible Extension Band For HiperLAN at 5 GHz," ERC Report 72.

[Erceg et al. 2008] Erceg, Vinko; Fischer, Trachewsky, Jason A.; "Matthew James A method and system for coexistence between 20 MHZ and 40 MHZ overlapping basic service sets (OBSS) in wireless local area networks", European Patent Application EP1931086, US.

[Eriksson 1994] H. Eriksson, "MBone: The Multicast Backbone," Communications of the ACM, August 1994, Vol.37 No.8, pp.54-60.

[ETSI 123 107 2000] ETSI TS 123 107 V3, "Universal Mobile Telecommunication System (UMTS); QoS Concept and Architecture", version 3.3.0 Release 1999, June 2000.

[ETSI 23.101 2000] ETSI 3GPP TS 23.101, "Universal Mobile Telecommunication System (UMTS); General UMTS Architecture", version 3.3.0 Release 1999, June 2000.

[ETSI 2003-09] ETSI, "Quality of Service (QoS) measurement methodologies", ETSI TS 102 024-5 V4.1.1 (2003-09)

[Evans and Filsfils 2007] John Evans, Clarence Filsfils; "Deploying IP and MPLS QoS for Multiservice Networks: Theory and Practice", Morgan Kaufmann, 2007, ISBN 0-12-370549-5

[Even et al 1984] S. Even, O. Goldreich, S. Moran and P. Tong, On the Npcompleteness of certain network testing problems, Networks 14 (1984) 1–24.

[Ferguson & Huston 1998] Paul Ferguson & Geoff Huston, 'Quality of Service: Delivering QoS on the Internet and in Corporate Networks', Wiley Computer Books, ISBN 0-471-24358-2288 pages, January 1998.

[Fiedler et al. 2010] Fiedler, M.; Hossfeld, T.; Phuoc Tran-Gia, "A generic quantitative relationship between quality of experience and quality of service", Network, IEEE, Volume: 24, Issue: 2, 2010

[Fischer & Ho 2003] Michael Fischer & Jin-Meng Ho, "Group Polling for DCF Based Reservation Request", IEEE 802.11-02/408r0, April 2003.

[Fings 2001] T. Fingscheidt et al., "RTP Payload Format for AMR" IETF draft, draft-fingscheidt-avt-rtp-amr-00.txt, January 2001.

[Floyd 1991] S. Floyd, Connections with Multiple Congested Gateways in Packet-Switched Networks Part 1: One-way Traffic, Computer Communication Review, V.21 N.5, October 1991, pp. 30-47.

[Floyd and Jacobson 1993] S. Floyd and V. Jacobson, Random Early Detection Gateways for Congestion Avoidance, IEEE/ACM Transactions on Networking, V.1 N.4, August 1993, p. 397-413.

[Floyd and Jacobson 1995] S. Floyd and V. Jacobson, Link-sharing and resource management models for packet networks. IEEE/ACM Transactions on Networking, 3(4): 365–386, August 1995.

[Fourmaux and Fdida 1999] Olivier Fourmaux and Serge Fdida, "Multicast for RSVP switching: An extended Multicast Model with QoS for Label Swapping in an IP over ATM environment", Telecommunication Systems Journal, 11(1-2): 85–104, January 1999, Special Issue on European Activities in Interactive Distributed Multimedia Systems and Telecommunication Services.

[Ganz and Phonphoem 2001] Aura Ganz and Anan Phonphoem, "Robust SuperPoll with chaining protocol for IEEE 802.11 wireless LANs in support of multimedia applications", Wireless Networks, Volume 7, Issue 1 2001, Pages: Pages: 65 – 73.

[Garden 1984] Gardner, M. "The Sixth Book of Mathematical Games from Scientific America", Chicago, IL: University of Chicago Press, p. 91, 1984.
[Gerard et al 200] Gerard G. Cervello and Sunghyun Choi, "Protecting QoS-Enabled BSSs in Situations of Overlapping BSSs in IEEE 802.11 WLAN," IEEE 802.11-00/108, May 2000.

[Ghanwani et al. 1998] Ghanwani, A., Pace, W., Srinivasan, V., Smith, A., Seaman, M., "A Framework for Providing Integrated Services Over Shared and Switched LAN Technologies", Internet Draft, May 1998 <draft-ietf-issll-is802-framework-05>

[Golestani 1990] S. Golestani. "A stop-and-go queuing framework for congestion management", In Proceedings of ACM SIG-COMM'90, pages 8–18, Philadelphia Pennsylvania, September 1990.

[Golestani 1993] S.J. Golestani, "A Self-Clocked Fair Queuing Scheme for Broadband Applications", in Proceedings, INFOCOM, June 1993.

[Golmie and Rebala 2003] Golmie, N.; Rebala, O.; "Bluetooth adaptive techniques to mitigate interference", Global Telecommunications Conference, 2003. GLOBECOM '03. IEEE

[Grilo and Nunes 2002] Grilo, A.; Nunes, M., "Performance evaluation of IEEE 802.11e", Personal, Indoor and Mobile Radio Communications, 2002, The 13th IEEE International Symposium on, Volume: 1 , 15-18 Sept. 2002 Page(s): 511 -517 vol.1

[GSM 03.64 2000] GSM 03.64: "Digital cellular telecommunication system (Phase 2+); Overall Description of the General Packet Radio Service (GPRS) Radio Interface; Stage 2".

[Guerin 1987] Guerin, R., "Channel occupancy time distribution in a cellular radio system", IEEE Transactions on Vehicular Technology 36, August 1987.

[Guerin and Peris 1999] Roch Guerin and Vinod Peris, “Quality of Service in Packet Networks – Basic Mechanisms and Directions”, Computer Networks, 31(3): 169–179, February 1999.

[IEC13818-3 2002] “Generic Coding of Moving Pictures and Associated Audio Information: Video”, ISO/IEC 13818-2: International Standard.

[Hajek et al 1988] B. Hajek and G. Sasaki, Link scheduling in polynomial time, IEEE Trans. Inform. Theory 34 (September 1988) 910–917.

[Han et al. 2009] Bo Han; Lusheng Ji; Seungjoon Lee; Miller, R.R.; Bhattacharjee, B.; “Channel Access Throttling for Overlapping BSS Management”, IEEE International Conference on Communications, 2009, Dresden.

[Handley et al. 1999] M. Handley, H. Schulzrinne, E. Schooler, J. Rosenberg, “SIP: Session Initiation Protocol“, IETF Request for Comments: 2543, March 1999.

[Hassan et al] Hassan, J.A.; Hassan, M. ; Das, S.K. ; Ramer, A., “Managing quality of experience for wireless VoIP using noncooperative games”, IEEE Journal on Selected Areas of Communications, 30(7), pp. 1193- 1204, August 2012.

[Hatauchi et al. 2008]Hatauchi, Takaaki; Fukuyama, Yoshikazu; Ishii, Misato; Shikura, Tatuyuki, “A Power Efficient Access Method by Polling for Wireless Mesh Networks”, IEEE Transactions on Electronics, Information and Systems, Volume 128, Issue 12, 2008.

[Heinanen et al. 1999] J. Heinanen, F. Baker, W. Weiss, and J. Wroclawski, “Assured Forwarding PHB Group”, RFC 2597, June 1999.

[Hiertz and Habetha 2003] G. Hiertz, J. Habetha, “A new MAC Protocol for a wireless multi-hop broadband system beyond IEEE 802.11”, Wireless World Research Forum, 9th Meeting in Zurich, Switzerland, 2003, July.

[HiperLAN 2003] HiperLan Web site: http://www.etsi.org/technicalactiv/hiperlan2.htm
[Hoang and Phan, 2007] Hoang, D.B., Phan, H. T, "End-Diff: An End-to-End QoS Architecture," International Symposium on Communications and Information Technologies (ISCIT 2007), Sydney, October 2007.

[Hoang 2006] D. B. Hoang, "Fair Intelligent Congestion Control," in Australian Telecommunications and Applications Conference Melbourne, Australia, 2006, pp. 153-157.

[Hou et al 2012] J Hou, D Feng, S Cai, M Chion, J Wang; “Techniques for providing efficient transmit diversity and bandwidth segmentation in a wireless communication system”, US Patent 8,185,160, 2012

[Hsu et al. 2005] W. Hsu, K. Merchant, H. Shu, C. Hsu, and A. Helmy, “Preference-based mobility model and the case for congestion relief in WLANs using ad hoc networks,” USC technical report.

[Hui and Shankaranarayanan 2004] Hui Luo; Shankaranarayanan, N.K.; “A distributed dynamic channel allocation technique for throughput improvement in a dense WLAN environment, Acoustics, Speech, and Signal Processing, 2004. Proceedings.

[Ickin et al 2012] S Ickin, K Wac, M Fiedler, L Janowski; “Factors influencing quality of experience of commonly used mobile applications”; IEEE Communications Magazine, 2012

[IEC13818-3 2002] “Generic Coding of Moving Pictures and Associated Audio Information: Video”, ISO/IEC 13818-2: International Standard.

[IEEE P802.11 1999] IEEE 802.11 WG, Standard for Telecommunications and Information Exchange Between Systems -LAN/MAN Specific Requirements - Part 11: Wireless Medium Access Control (MAC) and physical layer (PHY) specifications. IEEE 802.11 std, 1999.

[IEEE P802.11e 2002] IEEE 802.11 Task E WG, “Supplement to standard for Telecommunications and Information Exchange Between Systems -LAN/MAN Specific Requirements - Part 11: Wireless Medium Access Control (MAC) and physical layer (PHY) specifications: Medium Access Control (MAC) Enhancements for Quality of Service (QoS)”; IEEE 802.11e/D3.0, Jul. 2002.

[IEEE P802.1D 1998] IEEE P802.1D/D16, "Information technology - Telecommunications and information exchange between systems - Local and metropolitan area networks Common specifications - Part 3: Media Access Control (MAC) Bridges: Revision (Incorporating IEEE P802.1p: Traffic Class Expediting and Dynamic Multicast Filtering", ISO/IEC Final CD 15802-3 IEEE P802.1D/D16, March 1998

[IEEE P802.1Q 1998] IEEE P802.1Q/D11, "IEEE Standards for Local and Metropolitan Area Networks: for Virtual Bridged Local Area Networks", July 1998, IEEE Draft Standard P802.1Q/D11.

[Inan et al. 2006] Inan, I.; Keceli, F.; Ayanoglu, E.; “An Adaptive Multimedia QoS Scheduler for 802.11e Wireless LANs”, IEEE International Conference on Communications, 2006, Istanbul.

[ITU-T 1996] ITU-T - Recommendation I.371 (Draft): Traffic Control and Congestion Control in B-ISDN, March 1996. ITU-T Study Group 13.

[ITU-T 2003] ITU-T, “The E-model, a computational model for use in transmission planning,” ITU-T Recommendation, 2003.

[ITU-T 2006] International Telecommunication Union, “Vocabulary and effects of transmission parameters on customer opinion of transmission quality, amendment 2”, ITU-T Recommendation G.100, 2006

[ITU-T 07/2006] International Telecommunication Union, “Performance parameter definitions for quality of speech and other voiceband applications utilizing IP networks”, ITU-T Recommendation G.1020, July 2006

[ITU-T 2008] International Telecommunication Union, "Subjective video quality assessment methods for multimedia applications", ITU-T Recommendation P.910, April 2008

[Jacobson et al. 1999] V. Jacobson, K. Nichols, K. Poduri, "An Expedited Forwarding PHB", RFC 2598, June 1999.

[Jardosh et al. 2005] Jardosh, A.P.; Belding-Royer, E.M.; Almeroth, K.C.; Suri, S.; "Real-world environment models for mobile network evaluation", Selected Areas in Communications, IEEE Journal on Volume 23, Issue 3, March 2005 Page(s): 622 - 632

[Jardosh et al. 2003] A. Jardosh, E. M. Belding-Royer, K. C. Almeroth, and S. Suri, "Towards realistic mobility models for mobile ad hoc networks," MobiCom 2003, pp.217-229, Sep. 2003.

[Jelassi et al. 2012] Jelassi, S.; Rubino, G.; Melvin, H.; Youssef, H.; Pujolle, G., "Quality of Experience of VoIP Service: A Survey of Assessment Approaches and Open Issues", Communications Surveys & Tutorials, IEEE, Volume: 14, Issue: 2, 2012

[Jensen and Toft 1995] T.R. Jensen and B. Toft; "Graph Colouring Problem", Wiley InterScience, 1995.

[Johnson 1982] D. S. Johnson, The NP-completeness column: An ongoing guide,
J. Algorithms 3 (1982) 182–195

[Kabara and Calle 2012] [HTML] J Kabara, M Calle; "MAC protocols used by wireless sensor networks and a general method of performance evaluation"; International Journal of Distributed Sensor Network, 2012

[Kassler et al. 2000] A. Kassler, L. Burness, P. Khengar, E. Kovacs, D. Mandato, J. Manner, G. Neureiter, T. Robles, H. Velayos, "BRENTA - Supporting Mobility and Quality of Service for Adaptable Multimedia Communication", in Proceeding of the IST Mobile Communications Summit 2000, Galway, Ireland, October 2000, pp. 403 - 408.

[Katevenis et al. 1991] M. Katevenis, S. Sidiropoulos, and C. Courcoubetis. "Weighted round-robin cell multiplexing in a general-purpose atm switch chip," IEEE Journal on Selected Areas in Communications, vol. 9, no. 8, pp. 1265-1279, Oct. 1991.

[Kelly 1997] F.P. Kelly, "Charging and rate control for elastic traffic," in European
Trans. Telecomm., 1997

[Ketykó, et al. 2010] Ketykó, I.; De Moor, K.; Joseph, W.; Martens, L.; De Marez, L., "Performing QoE-measurements in an actual 3G network", 2010 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB).

[Khorsandroo et al. 2012] Khorsandroo, S.; Noor, R.M.; Khorsandroo, S., "A generic quantitative relationship between quality of experience and packet loss in video streaming services", Fourth International Conference on Ubiquitous and Future Networks (ICUFN), 2012

[Khorsandroo et al. 2012 Symposium] Khorsandroo, S.; Noor, R.M.; Khorsandroo, S., "Stimulus-centric versus perception-centric relations in quality of experience assessment", Wireless Telecommunications Symposium (WTS), 2012

[Kim et al. 2001] Kanghee Kim; Seokjoo Shin; Kiseon Kim, "A novel MAC scheme for prioritized services in IEEE 802.11a wireless LAN", ATM (ICATM 2001) and High Speed Intelligent Internet Symposium, 2001, Joint 4th IEEE International Conference on, 2001 P196 –199

[Kim et al. 2003] Kanghee Kim; Ahmad, A.; and Kiseon Kim, "A wireless multimedia LAN architecture using DCF with shortened contention window for QoS provisioning", IEEE Communications Letters, Volume: 7 Issue: 2, Feb 2003; Page(s): 97 –99.

[Kim et al. 2005] Byung-Seo Kim; Sung Won Kim; Yuguang Fang; Wong, T.F.; Two-step multipolling MAC protocol for wireless LANs", Motorola Inc., USA, IEEE Journal on Selected Areas in Communications, June 2005

[Kleinrock 1976] L. Kleinrock. "Queueing Systems: Volumes I and II". J. Wiley and Sons, New York, NY, 1976.

[Koodli and Puuskari 2001] Koodli, R.; Puuskari, M., "Supporting packet-data QoS in next generation cellular networks", IEEE Communications Magazine, Volume: 39; Issue: 2, Feb. 2001; P180-188

[Lam et al. 2006] Lam, R.Y.W.; Leung, V.C.M.; Chan, H.C.B.; "Polling-based protocols for packet voice transport over IEEE 802.11 wireless local area networks", Wireless Communications, IEEE, Feb. 2006

[Lederer et al 2013] S. Lederer, C. Mueller, C. Timmerer, C. Concolato, J. Le Feuvre and K. Fliegel, Distributed DASH Dataset, In Proceedings of the ACM Conference on Multimedia Systems (ACM MMSys) 2013, Oslo, Norway, 2013.

[Li, et al 2001] T. Li, et al. , Capacity of Ad Hoc Wireless Networks, 7th ACM MobiCom, 2001.

[Li et al 2009] T Li, T Tang, C Chang; "A new backoff algorithm for IEEE 802.11 Distributed Coordination Function"; Sixth International Conference on Fuzzy Systems and Knowledge Discovery, 2009.

[Lindgren et al. 2003] Anders Lindgren, Andreas Almquist, Olov Schelén, "Quality of service schemes for IEEE 802.11 wireless LANs: an evaluation", Mobile Networks and Applications, Volume 8, Issue 3 (June 2003).

[Liu, et. Al 2004] C. Liu, et. al., Adaptive Contention Control for IEEE 802.11 Wireless LANs, ACM MobiCom 2004

[Liu et al 2010] Y Liu, R Venkatesan, C Li; "Load-Aware channel assignment exploiting partially overlapping channels for wireless mesh networks"; GLOBECOM 2010

[Liu et al 2011] Q Liu, D Zhao, D Zhou; “An analytic model for enhancing IEEE 802.11 point coordination function media access control protocol”; European Transactions on Telecommunications; Volume 22, Issue 6, October 2011

[Liu et al 2013] Y. Liu, J. Geurts, J.-P. Point, S. Lederer, B. Rainer, C. Mueller, C. Timmerer and H. Hellwagner, “Dynamic Adaptive Streaming over CCN: A Caching and Overhead Analysis”, In Proceedings of the IEEE International Conference on Communication (ICC) 2013 – Next-Generation Networking Symposium, Budapest, Hungary, June, 2013

[Liu and Guo 2011] LIU Kai, GUO Jinhua; “RAMC: A RSU-assisted multi-channel coordination MAC protocol for VANET”; IEICE TRANSACTIONS on Communications Vol.E94-B No.1 , 2011

[Mah 1997] B. A. Mah, “An empirical model of HTTP network traffic,” in Proceedings of the IEEE INFOCOM ’97, April 1997, pp. 592–600.

[Mahadevan and Sivalingam 2000a] Indu Mahadevan; K. M. Sivalingam, “Architecture and Experimental Results for Quality of Service in Mobile Networks using RSVP and CBQ", in ACM/Baltzer Wireless Networks Journal, Vol.6, No.3, pp.221-234, 2000

[Mahadevan and Sivalingam 2000b] Indu Mahadevan and K. M. Sivalingam, "Architecture and Experimental Framework for Supporting QoS in Wireless Networks Using Differentiated Services", Accepted for ACM/Baltzer Mobile Networks and Applications Journal, May 2000.

[Malone et al 2007] D Malone, K Duffy, D Leith; “Modeling the 802.11 distributed coordination function in nonsaturated heterogeneous conditions”; Networking, IEEE/ACM transaction on 2007.

[Mangold et al. 2002a] S. Mangold; S. Choi; P. May; O. Klein; G. Hiertz; L. Stibor, “IEEE 802.11e Wireless LAN for Quality of Service”, Proceedings of the European Wireless, 2002, February, Volume 1, P32-39, Florence, Italy.

[Mangold et al. 2002b] Mangold, S., Sunghyun Choi, May, P. and Hiertz, G, “IEEE 802.11e - fair resource sharing between overlapping basic service sets”, Personal, Indoor and Mobile Radio Communications, 2002, The 13th IEEE International Symposium on, Volume: 1 , Sep 15-18, 2002, Page(s): 166 –171.

[Mangold et al 2003a] S. Mangold, L. Berlemann, B. Walke, “Radio Resource Sharing Model for Coexisting IEEE 802.11e Wireless LANs”, In Proc. of ICCT 2003 International Conference on Communication Technology, 2003, April.

[Mangold et al 2003b] S. Mangold, L. Berlemann, B. Walke, “Radio Resource Sharing Model for Coexisting IEEE 802.11e Wireless LANs”, In Proc. of ICCT 2003 International Conference on Communication Technology, 2003, April.

[Mangold et al 2003c] S. Mangold, L. Berlemann, B. Walke, “Equilibrium Analysis of Coexisting IEEE 802.11e Wireless LANs”, 14th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications, 2003, September.

[Maniatis et al. 2000] Maniatis, S.; Grecas, C.; Venieris, I., "End-to-end quality of service issues over next generation mobile Internet", Communications and Vehicular Technology, 2000. Symposium on, P150 –154

[Matsunaga and Katz 2004] Matsunaga, Y.; Katz, R.H. "Inter-domain radio resource management for wireless LANs"; Wireless Communications and Networking Conference, 2004.

[Matula et al. 1972] D, W. Matula, G. Marble, and J. F. Issacson. "Graph coloring algorithms". in Graph Theory and Computing, Academic Press NY, 1972.

[Misra et al 2011] S Misra, PV Krishna, KI Abraham; "An adaptive learning scheme for medium access with channel reservation in wireless networks"; Wireless Personal Communications, 2011 - Springer

[Müller et al 2011] C. Müller and C. Timmerer, "A VLC Media Player Plugin enabling Dynamic Adaptive Streaming over HTTP", In Proceedings of the ACM Multimedia 2011, Scottsdale, Arizona, November 28, 2011.

[Nandagopal et al. 2000] T. Nandagopal, T. Kim, X. Gao, and V. Bharghavan, "Achieving MAC Layer Fairness in Wireless Packet Network", MOBICOM 2000 Boston, ACM 2000.

[Nakajima and Zhang 2010] Ni, W., N. Nakajima, and S. Zhang, "A broadband compact folded monopole antenna for WLAN/WIMAX communication applications," Journal of Electromagnetic Waves and Applications, Vol. 24, No. 7, 921-930, 2010

[Nethi et al 2011] S Nethi, J Nieminen, R Jantti; "Exploitation of multi-channel communications in industrial wireless sensor applications: Avoiding interference and enabling coexistence"; Wireless Communications, IEEE 2011

[Nguyen et al 2011] NP Nguyen, TN Dinh, S Tokala, MT Thai; "Overlapping communities in dynamic networks: their detection and mobile applications"; Wireless Communications and Networking Conference (WCNC), 2011 IEEE

[Nichols et al. 1998] Kathleen Nichols, Stephen Blake, Fred Baker, and David Black, "Definition of the Differentiated Services Field (DS Field) in the IPv4 and IPv6 Headers", Proposed Standard RFC 2474, December 1998.

[Nichols et al. 1999] Kathleen Nichols, Van Jacobson, and Lixia Zhang, "A Two-bit Differentiated Services Architecture for the Internet", Informational RFC 2638, July 1999.

[Niyato and Hossain 2006] Dusit Niyato, Ekram Hossain, "Queue-Aware Uplink Bandwidth Allocation and Rate Control for Polling Service", IEEE 802.16 Broadband Wireless Networks, IEEE Transactions on Mobile Computing, June 2006.

[Niyato and Hossain 2007] Dusit Niyato; Ekram Hossain; "Radio resource management games in wireless networks: an approach to bandwidth allocation and admission control for polling service in IEEE 802.16, Wireless Communications, IEEE, Feb. 2007.

[Nyandoro et al] Nyandoro, A., Libman, L., Hassan, M, "Service Differentiation Using the Capture Effect in 802.11 Wireless LANs", IEEE Transactions on Wireless Communications, Volume:6, Issue: 8, August 2007.

[OPNET 1998] OPNET Modeller Tutorial, 1989–1998, MIL 3, Inc., 3400 International Drive, NW, Washington DC, 20008 USA, http://www.mil3.com.

[Phan and Hoang, 2006] H. T. Phan, Hoang, D. B, "Extension of BGP to support multi-domain FICC-Diffserv architecture," presented at IEEE Conference on Advanced Information Networking and Applications (AINA06), Vienna, April 2006.

[Pagtzis et al. 2001] Pagtzis, T.; Kirstein, P.; Hailes, S., "Operational and fairness issues with connection-less traffic over IEEE802.11b", in Proc. ICC'2001, Volume: 6, 2001 Page(s): 1905 -1913.

[Piamrat et al. 2009] Piamrat, K.; Viho, C.; Bonnin, J.-M.; Ksentini, A., "Quality of Experience Measurements for Video Streaming over Wireless Networks", Sixth International Conference on information Technology: New Generations, 2009. ITNG '09.

[Pitts 2008] Pitts, Jonathan, "Quality of experience: the route to scalable performance evaluation of networked systems?", International Symposium on Performance Evaluation of Computer and Telecommunication Systems, 2008. SPECTS 2008.

[Prabhu et al 2012] SRB Prabhu, S Sophia, PD Manivannan; "A research on decentralized clustering algorithms for dense wireless sensor networks"; International Journal of Computer Applications (0975 – 8887), Volume 57– No.20, November 2012

[Prasad 1999] A. R. Prasad, "Performance comparison of voice over {IEEE} 802.11 schemes", Proceedings of the 50th IEEE Vehicular Technology Conference, Amsterdam, Netherlands, Sep. 1999.

[Prabhu and Sophia 2012] SRB Prabhu, S Sophia; "A Review Of Energy Efficient Clustering Algorithm For Connecting Wireless Sensor Network Fields"; International Journal of Engineering, 2013

[Qiao 2011] Zizhi Qiao, "Smarter Phone based Live QoE Measurement", 2011 15th International Conference on Intelligence in Next Generation Networks (ICIN), 2011

[Qiao and Shin 2002] Daji Qiao and Shin, K.G., "Achieving efficient channel utilization and weighted fairness for data communications in IEEE 802.11 WLAN under the DCF", in Proc. 10th IEEE Int. Symp. on QoS, 2002.

[QoS Forum, 1999] QoS Forum "Introduction to QoS Policies", QoS Forum White Paper, July 1999. (http://www.qosforum.com).

[Rajan et al. 1999] Raju Rajan, Dinesh Verma, Sanjay Kamat, Eyal Fel-staine, and Shai Herzog, "A policy framework integrated and differentiated services in the internet", IEEE Network, 13(5):36–41, September/October 1999.

[Ramanathan and Lioyd 1989] R. Ramaswami and K. K. Parhi, Distributed scheduling of broadcasts in a radio network, in: INFOCOM (IEEE, 1989) pp. 497–504.

[Ramanathan and Lioyd 1992] S. Ramanathan and E. L. Lloyd, Scheduling algorithms for multi-hop radio networks, in: SIGCOM (ACM, 1992) pp. 211–222.

[Ramanathan and Lioyd 1993] S. Ramanathan and E. L. Lloyd, Scheduling algorithms for multihop radio networks, IEEE/ACM Trans. Networking 1 (April 1993)
166–172.

[Ramanathan 1999] S. Ramanathan, "A unified framework and algorithm for channel assignment in wireless networks", Wireless Networks, v.5 n.2, p.81-94, March 1999

[Rastin et al. 2006] Pries, Rastin; Tran-Gia, Phuoc; Wirth, Thomas, "QoS Traffic in Wireless LAN Overlapping Cells", European Wireless - 12th European Wireless Conference 2006, Athens, Greece.

[Ray et al. 2003] S. Ray, et al, RTS/CTS-Induced Congestion in Ad Hoc Wireless LAN, WCNC 2003

[Reis et al. 2010] Reis, A.B.; Chakareski, J.; Kassler, A.; Sargento, S., "Quality of experience optimized scheduling in multi-service wireless mesh networks", 17th IEEE International Conference on Image Processing (ICIP), 2010

[Robles et al. 2001] Robles, T.; Kadelka, A.; Velayos, H.; Lappetelainen, A.; Kassler, A.; Hui Li; Mandato, D.; Ojala, J.; Wegmann, B., "QoS support for an all IP system beyond 3G", IEEE Communications Magazine, Volume: 39 Issue: 8, Aug. 2001 P64 -72

[Ruscelli et al 2012] AL Ruscelli, G Cecchetti, A Alifano, G Lipari; "Enhancement of QoS support of HCCA schedulers using EDCA function in IEEE 802.11 e networks"; Ad Hoc Networks, 2012 – Elsevier

[Sadeghi, et al. 2002] B. Sadeghi, et al. , Opportunistic media access for multirate ad hoc networks, 8th ACM MobiCom, 2002.

[Saha et al. 1996] D. Saha, S. Mukherjee, and S. K. Tripathi, "Carry-over Round Robin: A Simple Cell Scheduling Mechanism for ATM Networks", in Proceedings, INFOCOM. Extended version available as Technical Report CS-TR-3658/UMIACS-TR-96-45, University of Maryland, March 1996.

[Schatz et al. 2011] Schatz, R.; Egger, S.; Platzer, A., "Poor, Good Enough or Even Better? Bridging the Gap between Acceptability and QoE of Mobile Broadband Data Services", IEEE International Conference on Communications (ICC), 2011

[Schenker et al. 1997] Schenker, S., Partridge, C., Guerin, R., "Specification of Guaranteed Quality of Service", RFC 2212 September 1997

[Schulzrinne et al. 1996] H. Schulzrinne et al.., "RTP: A Transport Protocol for Real-Time Applications", IETF RFC 1889, January 1996.

[Seaman et al. 1999] H. Seaman, A. Smith, E. Crawley, J. Wroclawski, “Integrated Services Mappings on IEEE 802 Networks”, June 1999, <draft-issll-is802-svc-mapping-04.txt>, Work in Progress.

[Sen and Huson 1996] A. Sen and M.L. Huson, A new model for scheduling packet radio networks, in: Proceedings of INFOCOM (1996) pp. 1116-1124.

[Senthilkumar and Krishnan 2010] D Senthilkumar, A Krishnan Nonsaturation; “Throughput enhancement of IEEE 802.11 b distributed coordination function for heterogeneous traffic under noisy environment”; International Journal of Automation and Computing 2010, Springer

[Shah et al 2011] D Shah, DNC Tse, JN Tsitsiklis; “Hardness of low delay network scheduling”; Information Theory, IEEE Magazine, 2011

[Shamsan et al 2012] ZA Shamsan, TA Rahman, AM Al-Hetar; “Point-Point Fixed Wireless and Broadcasting Services Coexistence with IMT-Advanced System”, Progress In Electromagnetics Research, 2012

[Shan et al 2012] S Shan, W Wu, W Wang, H Du, X Gao; “Constructing minimum interference connected dominating set for multi–channel multi–radio multi–hop wireless network”; Proceedings of the 17th annual international conference on Mobile computing and networking, 2012

[Sherman et al 2008] Sherman, M., A. N. Mody, R. Martinez, C. Rodriguez, and R. Reddy; "IEEE standards supporting cognitive radio and networks, dynamic spectrum access, and coexistence," IEEE Communications Mag., Vol. 46, No. 7, 72-79, 2008

[Shenker et al. 1993] S. Shenker, D. Clark, and L. Zhang, “A Scheduling Service Model and a Scheduling Architecture for an Integrated Services Packet Network”, working paper, Xerox PARC, August 1993.

[Shenker et al. 1997] Scott Shenker, Craig Partridge, and Roch Guerin, “Specification of Guaranteed Quality of Service”, Proposed Standard RFC 2212, September 1997.

[Sheu S.T. and Sheu T.F. 2001] Shiann-Tsong Sheu and Tzu-Fang Sheu, “DBASE: a distributed bandwidth allocation/sharing/extension protocol for multimedia over IEEE 802.11 ad hoc wireless LAN”, INFOCOM 2001, Twentieth Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE, Volume: 3 , 22-26 April 2001, Page(s): 1558 -1567 vol.3.

[Shreedhar and Varghese 1996] M. Shreedhar and George Varghese, “Efficient Fair Queuing using Deficit Round Robin”, IEEE/ACM Transactions on Networking, 4(3), June 1996

[Singh and Schulzrinne 2000] K. Singh, and H.Schulzrinne, "Interworking Between SIP/SDP and H.323". Proceedings of the 1st IP-Telephony Workshop (IPTel'2000), April 2000.

[Smith et al. 1999] A. Smith, E. Crawley,and J. Wroclawski, "Integrated Service Mappings on IEEE 802 Networks", Internet Draft, May 1999, <draft-ietf-issll-is802-svc-mapping-03>.

[Sobrinho and Krishnakumar 1996] J. L. Sobrinho and A. S. Krishnakumar, "Real-time traffic over the IEEE 802.11 medium access control layer", Bell Labs Technical Journal, pages 172–187, Autumn 1996.

[Sobrinho and Krishnakumar 1996] J. L. Sobrinho and A. S. Krishnakumar, "Quality-of-Service in ad hoc carrier sense multiple access networks", IEEE Journal on Selected Areas in Communications, 17(8): 1353–1368, August 1999.

[Stiller 1996] Burkhard Stiller, "Quality of Service", Dienstgüte in Hochleistungsnetzen.TAT No. 21, International Thompson Publishing, Bonn, Germany, 1996.

[Stoica and Zhang 1999] Ion Stoica and Hui Zhang, "Providing Guaranteed Services Without Per Flow Management", ACM Computer Communication Review, 29(4): 81–94, September 1999. Proceedings of SIGCOMM'99 Conference.

[Stoica et al. 1997] Ion Stoica, Hui Zhang, and T. S. Eugene Ng, "Hierarchical Fair Service Curve Algorithm for Link-Sharing, Real-Time and Priority Service", ACM Computer Communication Review, 27(4), October 1997.

[Tao et l. 2008] S. Tao, J. Apostolopoulos, and R. Guerin, "Real-Time Monitoring of Video Quality in IP Networks," in IEEE/ACM Trans. Network, 2008.

[Teitelbaum and Shalunov 2011] Teitelbaum, Ben & Stanislav Shalunov. "Why Premium IP Service Has Not Deployed (and Probably Never Will)". Internet2 QoS Working Group. Retrieved 17 October 2011

[Tinnirello et al 2010] I Tinnirello, G Bianchi, Y Xiao; "Refinements on IEEE 802.11 distributed coordination function modeling approaches", Vehicular Technology, IEEE Transactions, Volume: 59 , Issue: 3, 2010

[Vaiday et al 2000] N. H. Vaidya, P. Bahl, and S. Gupta, "Distributed fair scheduling in a wireless LAN", In Sixth Annual International Conference on Mobile Computing and Networking, Boston, August 2000.

[Vardakas et al 2008] JS Vardakas, MK Sidiropoulos, MD Logothetis; "Performance behaviour of IEEE 802.11 distributed coordination function";
- IET circuits, devices & systems, 2008

[Veres et al 2001] Veres, A.; Campbell, A.T.; Barry, M.; Li-Hsiang Sun, "Supporting service differentiation in wireless packet networks using distributed control", Selected Areas in Communications, IEEE Journal on, Volume: 19 Issue: 10 , Oct. 2001, Page(s): 2081 –2093

[Vergados and Vergados 2004] Vergados, D.D. Vergados, D.J., "Synchronization of multiple access points in the IEEE 802.11 point coordination function", Vehicular Technology Conference, 2004.

[Vogel et al. 1995] A.Vogel et al.., “Distributed Multimedia and QoS: A survey”, in: IEEE Multimedia, Vol.2, No. 2, 1995, pp.10-19.

[Vogt et al. 1998] Carsten Vogt, Lars C. Wolf, Ralf Guido Herrtwich, and Hartmut Wittig, “Quality-of-Service Management for Distributed Multimedia Systems”, ACM Multimedia Systems Journal, 6(3):152–166, May 1998. Special Issue on QoS Systems.

[Wagon 1998] Wagon, S. "An April Fool's Hoax." Mathematica in Educ. Res. 7, 46-52, 1998.

[Wagon 1999] Wagon, S. Mathematica in Action, 2nd ed. New York: Springer-Verlag, pp. 535-536, 1999.

[Wang 2012] JY Wang; “Set-based broadcast scheduling for minimizing the worst access time of multiple data items in wireless environments”; Information Sciences, 2012, Elsevier

[Watanabe 2008] Watanabe, K.; Yamagishi, K.; Okamoto, J.; Takahashi, A., “Proposal of new QoE assessment approach for quality management of IPTV services”, 15th IEEE International Conference on Image Processing, 2008. ICIP 2008.

[Wetherall et al. 1998] David J. Wetherall, Ulana Legedza and John Guttag, “Introducing New Internet Services: Why and How”, IEEE Network Magazine July/August 1998.

[Willinger et al. 1994] W. E. Willinger, W.E. Leland, M.S. Taqqu, and D.V. Wilson. On the self-similar nature of Ethernet traffic (extended version). IEEE/ACM Trans. Networking, 2(1): 1–15, February 1994.

Wong et al. 2012] Wong, Y.W.; Lau, W.C.; Chan, K.M.; Yang, Y.; Tang, C.Y.; Lam, F.; Lo, K.M., “An Extensible Distributed Measurement Platform for Analyzing Quality-of-Experience (QoE) of Multimedia Applications over Wireless Networks”, Vehicular Technology Conference (VTC Fall), 2012 IEEE

[Wroclawski 1997a] John Wroclawski, “Specification of the Controlled Load Network Element Ser-vice”, Proposed Standard RFC 2211, September 1997.

[Wroclawski 1997b] John Wroclawski, “The Use of RSVP with IETF Integrated Services” Proposed Standard RFC 2210, September 1997.

[Xhafa et al. 2007] Xhafa, A.E.; Batra, A.; Zaks, A.; “On the Coexistence of Overlapping BSSs in WLANs”, Vehicular Technology Conference, IEEE 66th, October 2007

[Xhafa et al. 2008] Ariton E. Xhafa, Anuj Batra, Artur Zaks, “On the 20/40 MHz Coexistence of Overlapping BSSs in WLANs”, Journal of Networks, Vol 3, No 7 (2008), 56-63, Jul 2008.

[Xhafa and Tonguz 2004] A.E. Xhafa and O.K. Tonguz, "Dynamic Priority Queuing of Handover Calls in Wireless Networks: An Analytical Framework", IEEE J. on Selected Areas in Communications, Special Issue on Advanced Mobility Management and QoS Protocols for Next-Generation Wireless Internet, May 2004.

[Xiao and Ni, 1999] Xipeng Xiao, and Lionel M. Ni, "Internet QoS: a big picture". [Journal Paper] IEEE Network, vol.13, no.2, March-April 1999, pp.8-18. Publisher: IEEE, USA.

[Xu, et.al 2001] S. Xu, et.al, Does the IEEE 802.11 MAC Protocol Work Well in MultiHop Wireless Ad Hoc Networks, IEEE Comm. Mag. , June 2001.

[Yavatkar et al. 1999] R. Yavatkar, D. Hoffman, Y. Bernet, F. Baker, "SBM (Subnet Bandwidth Manager): A Protocol for RSVP-based Admission Control over IEEE 802-style networks", May 1999.

[Ye, et. al 2003] F. Ye, et. al, Improving Spatial Reuse of IEEE 802.11 Based Ad Hoc Networks, Globecom 2003.

[Yeh and Chen 2002] Jing-Yuan Yeh and Chienhua Chen, "Support of multimedia services with the IEEE 802-11 MAC protocol", Communications, 2002, ICC 2002, IEEE International Conference on, Volume: 1, 28 April-2 May 2002 Page(s): 600 –604.

[Yoon et al. 2000] Suk-Un Yoon; Ji-Hoon Lee; Ki-Sun Lee; Chul-Hee Kang, "QoS support in mobile/wireless IP networks using differentiated services and fast handoff method", Wireless Communications and Networking Conference, 2000. WCNC. 2000 IEEE, Volume: 1, 2000, P266 -270 vol.1

[Zhang et al. 1993] Lixia Zhang, Steve Deering, Deborah Estrin, Scott Shenker, and Daniel Zappala, "RSVP: A New Resource ReSerVation Protocol," IEEE Network Magazine, 7(5):8–18, September 1993.

[Zhao et al. 2000] Weibin Zhao, David Olshefski and Henning Schulzrinne, "Internet Quality of Service: an Overview", CUCS-003-00 Columbia Technical Report, Columbia University, 2000.

[Zhu et al. 2004] J. Zhu, Xingang Guo ; Yang, L.L. ; and Conner, W.S., "Leveraging Spatial Reuse in 802.11 Mesh Networks with Enhanced Physical Carrier Sensing', ICC 2004.

Appendix A Acronyms and Abbreviations

2G	Second Generation
3G	Third Generation
AC	Access Category
ACK	Acknowledgement
AGC	Automatic Gain Control
AIFS	Arbitration Inter-Frame Space
AIFSN	AIFS Number
AP	Access Point
ATM	Asynchronous Transfer Mode
BER	Bit Error Rate
BSS	Basic Service Set
CCK	Complementary Code Keying
CFP	Contention Free Period
CSMA/CA	Carrier Sense Multiple Access with Collision Avoidance
CTS	Clear To Send
CW	Contention Window
DCF	Distributed Coordination Function
DIFS	DCF Inter-Frame Space
EDCF	Enhanced Distributed Coordination Function
ERP	Extended Rate Physical Layer
FCS	Frame Check Sequence
FIFO	First In First Out
Gbps	Giga bits per second
HCF	Hybrid Coordination Function
IBSS	Independent Basic Service Set
IEEE	Institute of Electrical and Electronic Engineers
IFS	Inter-Frame Space
ISM	Industrial, Scientific and Medical
LAN	Local Area Network
MAC	Medium Access Control
MIB	Management Entity
NAV	Network Allocation Vector
OFDM	Orthogonal Frequency Division Multiplexing
PC	Point Coordinator
PCF	Point Coordination Function
PHY	Physical Layer Specification
PLCP	Physical Layer Convergence Protocol
QoS	Quality of Service
RTS	Request To Send
SAP	Service Access Point
SIFS	Short Inter-Frame Space
SIR	Signal to Interference Ratio
SME	Station Management Entity
TCP	Transmission Control Protocol
TXOP	Transmission Opportunity
UDP	User Datagram Protocol

Appendix B Control & Management Mechanisms for OBSS Coordination

B.1 Overlapping Messages Collection

The overlapping messages are crucial information needed for overlapping resource management and decision-making in an OBSS environment, which will be discussed in the next section. This section describes the processes and procedures for collecting this overlapping information for all BSSs. We firstly depict specifications of information/messages required by the RM to work out the resource allocation scheme; and then give detailed protocols for exchange of information among the network entities required to build the initial topology (overlapping relationship among BSSs) "database" at the Resource Manager and to subsequently maintain/update that information. Finally, the detailed specifications of processes running within each agent are further described in a form of the state transition diagrams.

B.1.1 Specifications of Overlapping Information

The overlapping messages describe the current overlapping situations (topology) among the BSSs in the OBSS environment and are required at the RM in order to make decision on resource allocation scheme to suit the current overlapping situation. The RM needs to know the overlapping relationships among the BSSs with the identifications of the involved BSSs, and the client number of the BSS in the respective overlapping area to perform computation for a proper resource allocation scheme for OBSS coordination. The information are stored and recorded in each entity including RM, AP and clients.

Overlapping Records in RM and BSS

The information record in RM and BSS (AP) consists of elements as below. Each BSS is identified by its BSSID and SSID. If an IBSS or another ESS were detected overlapped, the RM would recognize by reading a foreign SSID. Each overlapping relationship can be expressed by a pair of BSSs (one reporting BSS and one detected BSS), which overlap each other. And all pairs reported in the RM form a topology database (overlapping relationships), which reflects the current overlapping situation in the OBSS environment. The number of clients overlapped in the reporting BSS indicates the overlapping severity in this BSS.

- BSS identification (BSSID) of two BSSs (the reporting BSS and the detected BSS), which overlap each other. This field of Basic Service Set Identification uniquely identifies each BSS. The value of this field, in an infrastructure BSS, is the MAC address currently in use by the client in the AP of the BSS.
- ESS identification (SSID) of two BSSs, which overlap each other. The Service Set Identifier indicates the identity of the ESS.
- Number of clients overlapped in the reporting BSS: These clients should firstly be associated with the reporting BSS, and secondly hear the foreign AP, which associates with the detected BSS. If number reaches 0, it indicates that such overlapping is currently inactive.

The overlapping situation between a pair of BSSs can be detected and reported by either two parties or one party. For example, two BSS A and B overlap and also detect each other; RM would have two reports separately from BSS A and BSS B. These two records are called counterpart record to each other.

Overlapping Records in Clients

The client record provides the raw overlapping information on each client, while RM and BSS (AP) records describe the overlapping information on the BSS level and are summary of client records. Therefore, client record is slightly different from the BSS and RM records, as discussed below. The clients would not know the number of clients overlapped, but it is necessary to keep the client overlapping status updated, using a status timer, in order to maintain RM and BSS records updated as well.

- BSS identification (BSSID) of two BSSs (the reporting BSS and the detected BSS), which overlap each other.
- ESS identification (SSID) of two BSSs, which overlap each other.
- Timer of the record keeps the track of the current status of the overlapping existence.

The details of the record data structures are described below.

Message Name	Field Order	Field Name	Type	Valid Range	Description
Client Record	1	BSSID	Octet String	1-32 octets	The BSSID of the detected BSS
	2	SSID	Octet String	1-32 octets	The SSID of the detected BSS
	3	Timer	Real	0 - lifetime	The current lifetime of the record before it expires.
BSS Record	1	BSSID	Octet String	1-32 octets	The BSSID of the detected BSS
	2	SSID	Octet String	1-32 octets	The SSID of the detected BSS
	3	Client number	Integer	>= 0	The current number of clients, which belong to the reporting BSS and are overlapped from the detected BSS.
RM Record	1	BSSID	Octet String	1-32 octets	The BSSID of the reporting BSS
	2	SSID	Octet String	1-32 octets	The SSID of the reporting BSS
	3	BSSID	Octet String	1-32 octets	The BSSID of the detected BSS
	4	SSID	Octet String	1-32 octets	The SSID of the detected BSS
	5	Client number	Integer	>= 0	The current number of clients, which belong to the reporting BSS and are overlapped from the detected BSS.

Table 10 Overlapping Records Structures

B.1.2 Overlapping Information Collection Protocols

Collection Hierarchy

The overlapping information collection follows a hierarchy structure from client stations, the AP to the RM, as illustrated in Figure 73. Each station and AP has responsibility to participate in collecting this information. The overlapping information are widely collected in the level of clients and reported to AP. AP plays as a gateway for message/report transfer between RM and clients. It would filter the messages received from all its clients and only the necessary summary of the information would be forwarded to RM. In such way, it provides necessary information for management in the AP level and reduces amount of information received in RM, comparing to the method, where all messages are directly transferred from clients to RM.

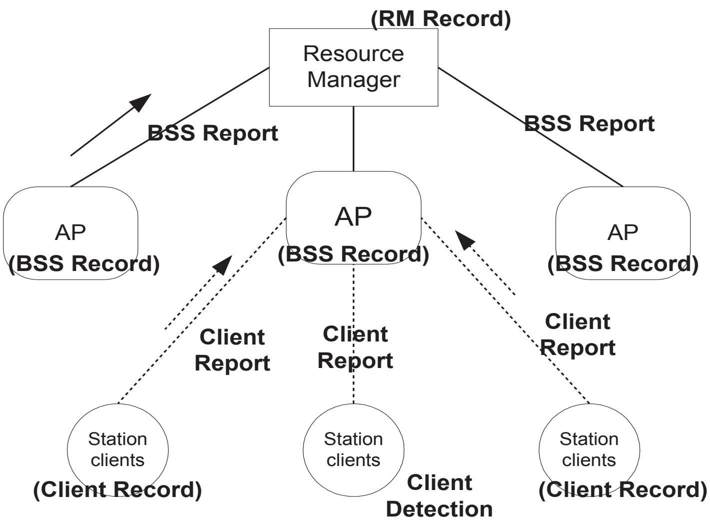


Figure 73 Overlapping Message Collection Hierarchies

Overlapping Report Structure

Record Name	*Field Order*	*Field Name*	*Type*	*Valid Range*	*Description*
Client Report	1	*BSSID*	*Octet String*	*1-32 octets*	*The BSSID of the detected BSS*
	2	*SSID*	*Octet String*	*1-32 octets*	*The SSID of the detected BSS*
	3	*Status Code*	*Boolean*	*True, false*	*The current status of overlapping existence from the detected BSS.*
BSS Report	1	*BSSID*	*Octet String*	*1-32 octets*	*The BSSID of the detected BSS*
	2	*SSID*	*Octet String*	*1-32 octets*	*The SSID of the detected BSS*
	3	*Client number*	*Integer*	*>= 0*	*The current number of clients, which belong to the reporting BSS and are overlapped from the detected BSS.*

Table 11 Overlapping Report Messages Structures

The report message is different from the records, which have been discussed. However, it is based on the relevant record at the same level. There are two types of report messages being transfer in the hierarch above, the client report and BSS report message. Client report message includes the BSSID & SSID of the detected BSS and the status code of the overlapping with this detected BSS. The status code indicates whether such client still can detect/hear the overlapping from this detected BSS and is used to calculate the number of its clients overlapped in the receiving AP. BSS reports message is the original copy of the RM record, including a pair of BSSID & SSID, and the number of its clients overlapped in the reporting BSS.

Collection Interactions and Procedures

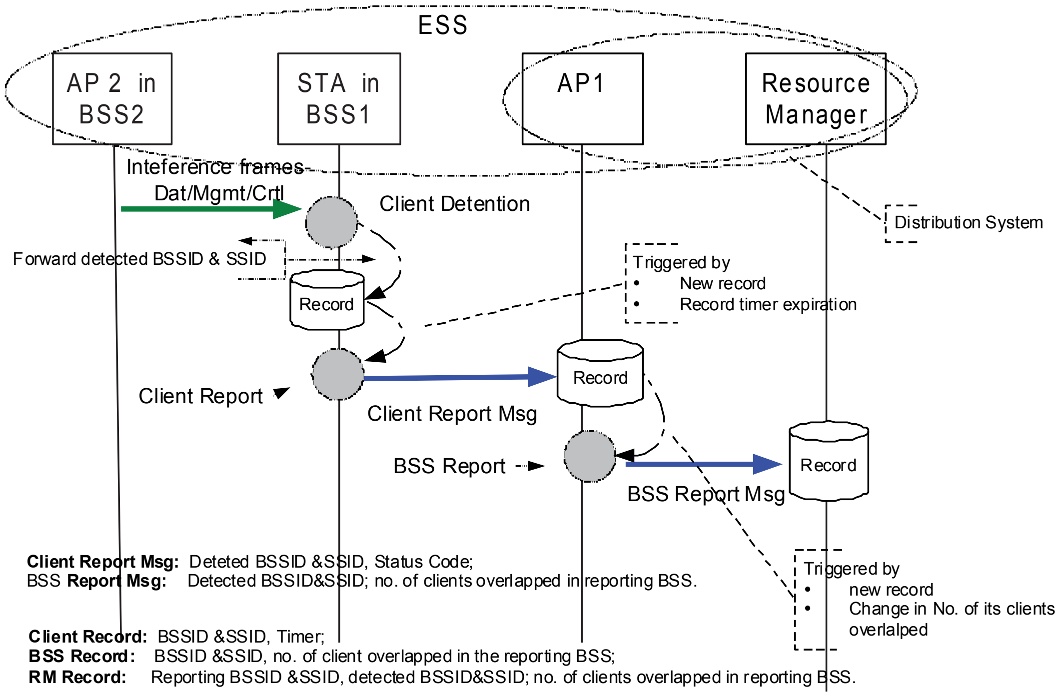


Figure 74 Time Sequence Diagram for Message Collection

In order to keep RM updated with the current overlapping situation, the information collected should be 'soft' status, which means that they are regularly updated. Therefore, either client, RM or AP should maintain their own record list; keep the record of the information they collect and update the record in certain ways. The update mechanism in client level is a timer. The update period (record lifetime) should be efficient enough to capture the overlapping movement timely. The update mechanism in AP and RM is based on their received reports. Particularly, if the number of its clients overlapped in the BSS report or RM record is 0, it indicates the removal of the overlapping.

The collection procedures cover across two sub-network domains, the individual BSS and the supporting distribution system. These messages are forwarded around the data path from the stations to AP via the BSS air media and from AP to the RM via the supporting distribution system. The combined procedure is described in the time sequence diagram to illustrate time sequence of protocols operations, as in Figure 74.

The report actions at levels of both AP and clients follow the event-triggered mechanism. The client would only report, when any of its record timers expires or a record is newly established. The AP would only report when there is change/update in its record list (either the record removal, record update in number of client overlapped, or receipt of a new record). These procedures are detailed in the sub-protocols of client-detection, client-report and BSS-report.

Client Detection in Individual BSS

Every client station, including AP station, all participates in the client detection procedure. The low-level detection enables most possible overlapping to be discovered in the OBSS environment. The overlapping situation may be detected by one station or simultaneously by more than one station, depending on the severity of the overlapping. However, all stations, would detect the overlapping situation, and report the finding to their AP individually.

The client detection procedure is used to detect the current OBSS information in each client. The purpose of this procedure is for client to be aware of the current OBSS interference and find out the identity of the potential interference source. The key information to be found is the BSSID and SSID of the overlapping BSS and its current overlapping status. The detection procedure occurs at the MAC layer making use of the existing standard protocols and services, but does not require any additional MAC layer management and control services in the client. However, additional protocols for client detection are needed in the MAC layer.

Before depicting the procedure, we firstly further clarify the overlapping situations, which the RM concerns about. As discussed in the previous sections, when two BSSs are located nearby, depending to their distance, they have two types of overlapping degree. If distance is close, at least one of APs would be heard by the stations on the other BSS, so called AP-overlapped. If the distance is far away, only some stations of each BSS can hear from each other, so called client-overlapped. The RM only needs to know the situation of AP-overlapped in order to work out the resource allocation scheme. The following procedure describes how such situation can be detected in the individual BSS.

We start with the introduction of address specification of the MAC frame format in [IEEE 802.11 1999]. There are four address fields in the MAC frame (data, control or management frame). These fields are used to indicate the BSSID, source address, destination address, transmitting station address, and receiving station address. The BSSID field is a 48-bit field and the same format as an IEEE 802 MAC address. This field uniquely identifies each BSS. The value of this field, in an infrastructure BSS, is the MAC address currently in use by the AP station of the BSS. Therefore, if a frame is sent from an AP, the BSSID is the same as the transmitting station address.

Upon the receipt of any frame from other BSS, either clients or AP should confirm if such frame is sent from another nearby foreign AP, by checking if the receiving BSSID is not its own BSSID and if the receiving BSSID is equal to the transmitting station address in the received frame. If the received frame is confirmed from a foreign AP, either the client or AP can hear other nearby AP, belonging to other BSS, then the AP-overlapped situation exists. Such situation needs to be detected; and the detected BSSID would be recorded and forwarded to client report procedure. The simple detection procedure repeats when a frame is received.

Client Reporting in Individual BSS

The client reporting occurs within the domain of each individual BSS. The procedure is used to deliver the detected OBSS information from a client to its AP and only occur there is either new client record established or existing client record expired in the client. The purpose of this procedure is for client to forward the current interference source information upstream. The information to be reported from the client to the AP includes the overlapping pair and its overlapping status.

Once the client detects the existence of an AP-overlapped situation and records the identities of the foreign BSS, it firstly confirms if it is a new discovery by comparing the detected identities with the existing records (identities and timer), which have been stored and kept updated in the clients. The new discovery would then trigger the client to report the new overlapping BSS to its AP within their own BSS. It also triggers the establishment of the client record for the new OBSS discovery and initialises an expiration timer for the new record in the client. If the finding already exists, it would only reset the expiration timer of the existing record. Once the timer expires, the corresponding record becomes inactive and would be removed. It also triggers such client to report the record removal with the report element of an inactive status code. The design of the client expiration timer should be efficient enough to capture the overlapping movement timely. It is one of the performance tuning parameters

The information to be reported from stations to AP is the overlapping pair (the Identification of the nearby Overlapping BSS) and its status code. The report messages can be transferred either on the application layer as data content, or within MAC layer as a new management frame. The later sections give details of the preferred procedure, the new management service in the context of IEEE 802.11.

BSS Reporting from AP to RM

The BSS reporting occurs within the domain of supporting distribution system. The procedure is used to deliver the detected OBSS information from the AP to the RM and only occur there is a new BSS record established, an existing record removal or update of an existing BSS record in the AP. The purpose of this procedure is for the AP to forward the current interference source information upstream. The information to be reported from the AP to RM includes the overlapping pair and its number of clients overlapped.

The BSS record update or establishment is triggered by the receipt of the client report. As discussed, each BSS record is identified by the detected/foreign BSSID and indicates the current overlapping situation (number of clients overlapped in the BSS). If the client report is with an active status code, AP should firstly identify if the foreign BSS is new to AP by checking the foreign BSSID in the client report. If the detected/foreign BSSID in the client report is matched with one of foreign BSSID in the existing BSS records, it means that some of its other clients have already detected and reported this overlapping foreign BSS; AP should only update the existing BSS record by increasing its number of its client overlapped by 1. If the foreign BSSID in the client report is new to the AP, AP should then establish a new BSS record by recording this foreign BSSID and assigning value of 1 as its number of its client overlapped. If the client report is with an inactive status code, it indicates that the reporting client previously detected and reported the foreign BSS, and no longer hears it any more. AP should then update the record by decreasing its number of client overlapped by 1. AP would also further examine if the updated number of client overlapped reaches zero. In the case of non-zero number, it means that other clients detects and reports such foreign BSS, and the overlapping from this foreign BSS is currently active. In the case of a zero number, this record becomes inactive and would be removed.

Once the BSS report is received in RM, RM is then informed of the BSS record update, removal or establishment as discussed above. The reporting message is simply a copy of the BSS record; any BSS record update, removal or establishment would trigger the BSS reporting. The information reported from AP to the RM includes the overlapping BSS pairs, number of the client overlapped in the reporting BSS. The message can be transferred either on the application layer as data content, or within MAC layer with a new management frame in the supporting distribution system. The later sections give details of the preferred procedure, this new management service in the context of IEEE 802.11.

B.1.3 Entities Processes for Record Collections

The record collection protocols are supported by the management entities of client, AP and RM. The entire process consists of three sub-processes, client process, BSS process and RM process. A particular overlapping record collection agent would be implemented in each entity to execute their respective processes/process as described below in Figure 75.

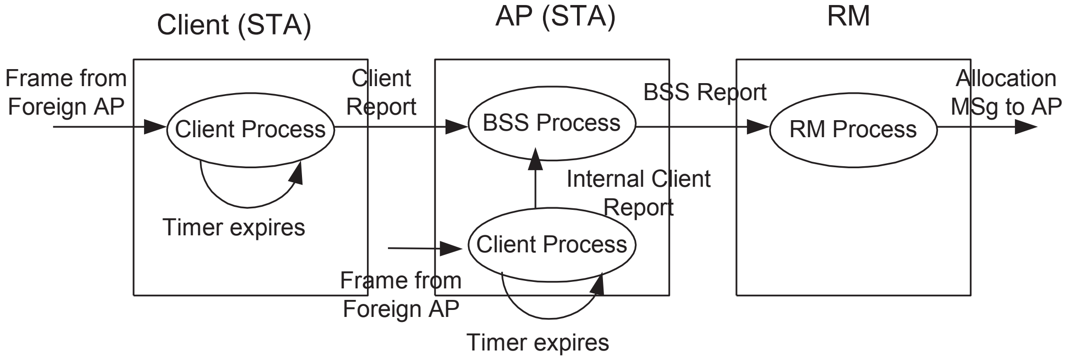


Figure 75 Processes Relationship

Process at Client

The client process executes the client detection and client report protocols. As stated in the Figure 75, the process is initialised by either receipt of foreign frame or expiration of any record timers. The foreign frame, which initialises the process, must meet the criteria of:

- BSSID in the received frame = Transmitting STA address in the received frame.
- BSSID in the received frame = BSSID value of the receiver STA.

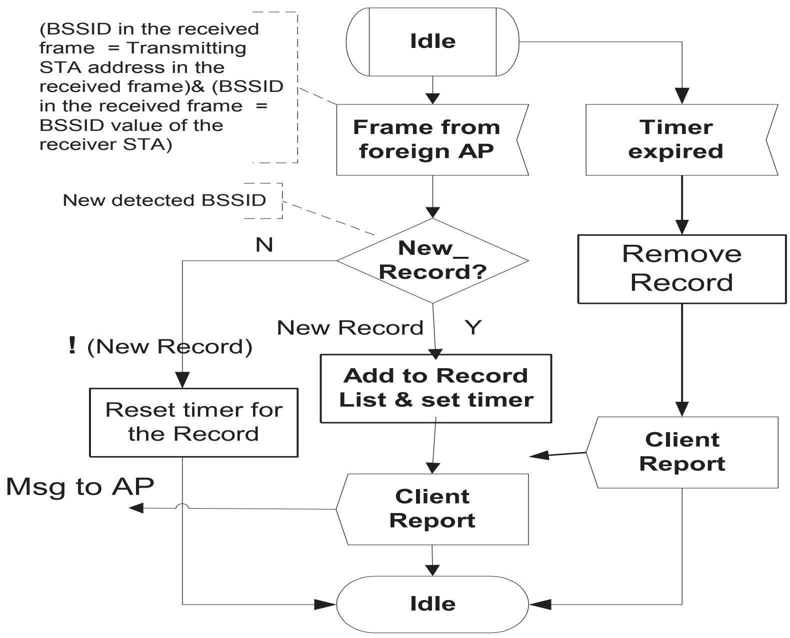


Figure 76 State Transition Diagram of a Client Process

If there is no record matching to this detected/foreign BSSID, the overlapping from this foreign BSS is new to the client. The client would create a new record, containing the detected BSSID and initialises the timer for the new record. If the detected BSSID is not new to the client, the corresponding record would be updated by resetting its existing timer. The record would be only removed if its timer expires. The process output is the client report message, which would be transmitted from the client to the AP. There is a record removal; the client would send a report with an inactive status code. On the contrary; the client would send a report with an active status code, if there were a new record.

Processes at AP

AP itself is firstly a client station, and it also carries particular function of MAC coordination and overlapping resource management as AP. Therefore, AP entity would perform a normal client process as a client and a BSS process as a point coordinator.

The details of the client process are the same as discussed above, except that a client report forward does not need any transmission on the air medium. Instead, it is a simple message transfer between two sub management entities (client sub-entity and point coordinator sub-entity) within the same AP station.

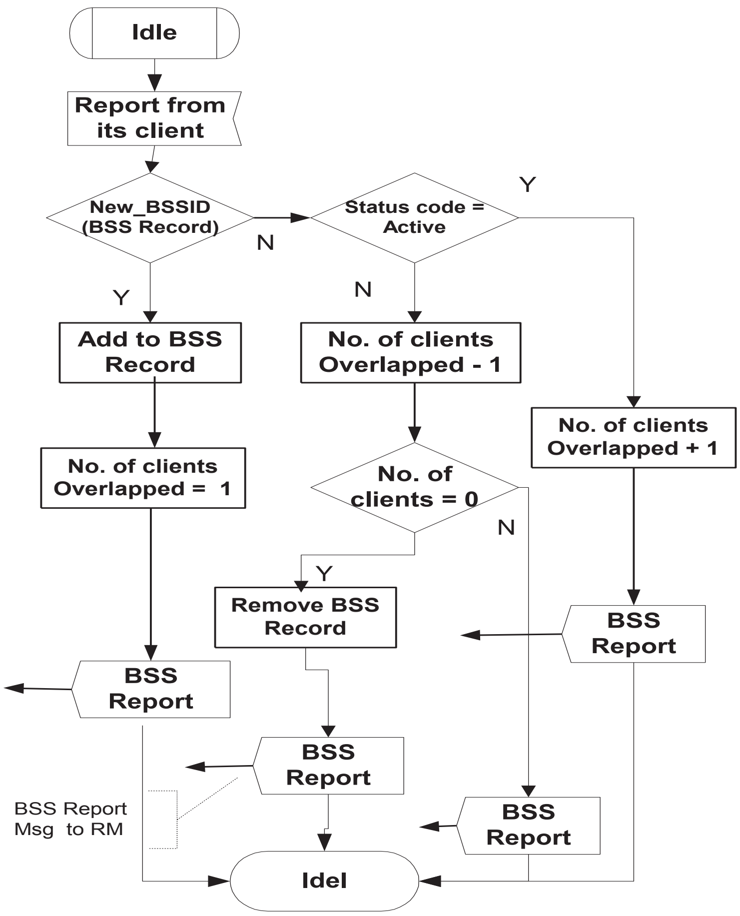


Figure 77 State Transition Diagram of an AP

If there is no BSS record matching to the detected foreign BSSID in the client report, the overlapping from this foreign BSS is new to the BSS. The AP would create a new record, containing the detected BSSID and assign 1 to the number of clients overlapped for this record. If the detected BSSID is not new to the AP, the corresponding record would be updated by re-adjusting the value of the no. of clients overlapped. The record would be only removed if the no. of clients overlapped reaches zero. The output of process is the BSS report message, which would be transmitted from the AP to the RM. The report of the record removal would have 0 value of number of clients overlapped.

Process at the RM

As stated in Figure 77, the process is initialised by receipt of BSS report from any of its BSSs. The RM would justify if the report were new finding of overlapping (new overlapping pair) by the criteria below:

- Reporting BSSID in the report = any of the reporting BSSID in the RM records.
- & Detected BSSID in the report = the detected BSSID in the record, which has the matching reporting BSSID.

If there is no RM record matching to the reported overlapping pair, the overlapping from the detected BSS to the reporting BSS is new to the RM. The RM would then create a new record, containing the detected BSSID and reporting BSSID with its existing value of no. of clients overlapped. If the pair is not new to the client, the corresponding record would be updated by re-adjusting the value of the no. of clients overlapped. The record would be only removed if the no. of clients overlapped is zero.

The outputs of the RM process are the resources allocation computation and the allocation sub-processes. The resources allocation computation sub-process performs necessary calculations to work out the allocation schemes for each BSS based on the existing RM records and current traffic condition reports. The allocation sub-process is responsible for transmitting the allocation schemes from the RM to the respective APs. These details of these sub-processes would be discussed in the following Sections.

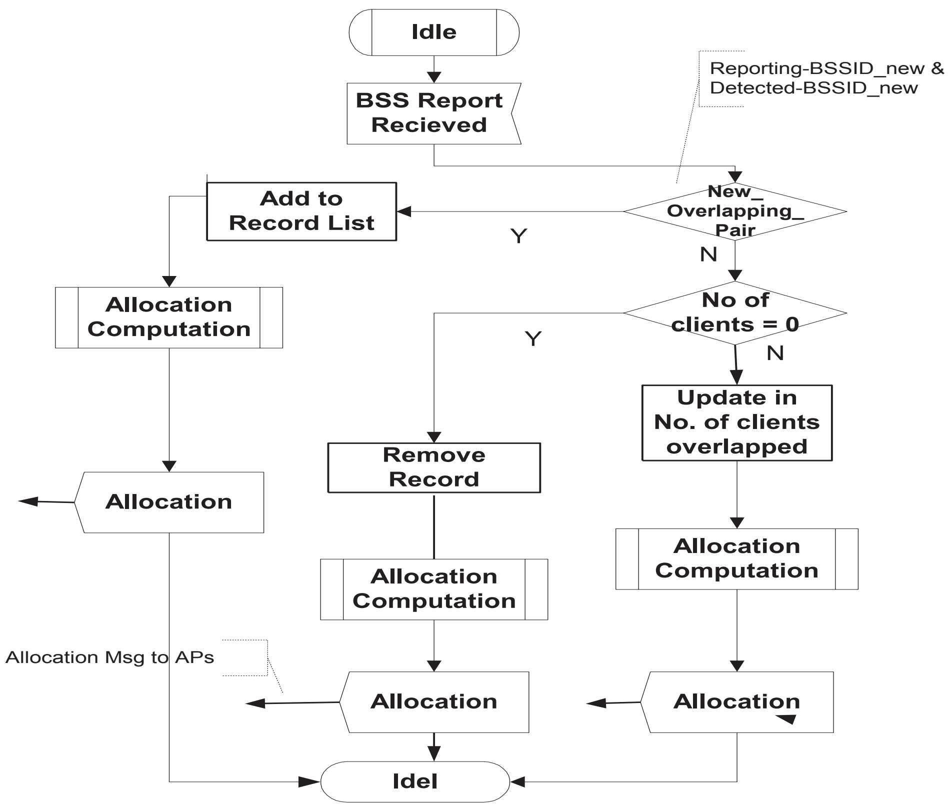


Figure 78 State Transition Diagram of a RM

B.2 Allocation Message Transfer

The allocation message describes the resources allocation scheme, which is determined in RM. The BSS learns of the resources allocation scheme by receiving and reading the allocation message from RM.

Allocation Message Specification

Allocation message to a BSS should be able to explain the allocation scheme to both AP and its clients. They are instructions to formulate the OBSS unique superframe. The allocation message is specific to each BSS, and it should include information of:

- Downlink duration.
- Superframe duration.
- The duration and starting time of the time span, which the receiving BSS is assigned.
- Beacon interval.
- Synchronization information.

As discussed, resource allocation execution in clients and AP are parts of operation of MAC coordination functions in the BSS. In the standard IEEE 802.11, beacon is used for transports of medium access guidelines for implementation of MAC coordination functions, including polling and contention parameters for all clients. Therefore, the allocation message, which carries the information for resource allocation execution/ MAC coordination, can be interpreted as relevant parts of standard beacon information elements.

The relevant parts are Timestamp, beacon interval and CF parameter set. These items describe the complete implementation of a superframe and provide synchronization for the AP and clients. The definitions of these elements are detailed in [IEEE P802.11 1999]. The values of these elements are determined in RM and specific to each reuse group. Table 12 shows how these standard IEEE P802.11 information elements interpret the necessary contents/information carried in the allocation message.

Order	***Information Element***	***Interpretation***
1.	*Timestamp*	*Entire OBSS synchronization - TSF (timing synchronization function)*
2.	*Beacon Interval*	*Time unit between two neighbouring Beacon*
8.	*CP parameter set*	• *CFPCount indicates how many DTIMs appear before the next CF time span commences for its group (Starting time of its assigned time span). A CFPCount of 0 indicates that the current DTIM marks the start of its active span.* • *CFPPeriod indicates the number of DTIM intervals between the start of CFPs (superframe duration).* • *CFPMaxDuration is defined here as duration of a time span.* • *CFPDurRemaining indicates the length of downlink period.*

Table 12 Allocation Message Information Elements in A Beacon

Allocation Message Transfer Process

The allocation messages transfer occurs within the domain of the supporting distribution system and the BSS. The procedure is used to deliver the resource allocation message from the RM to the AP and then from AP to its clients. The purpose of this procedure is for the RM to update the allocation scheme for the addressed BSS in a timely method.

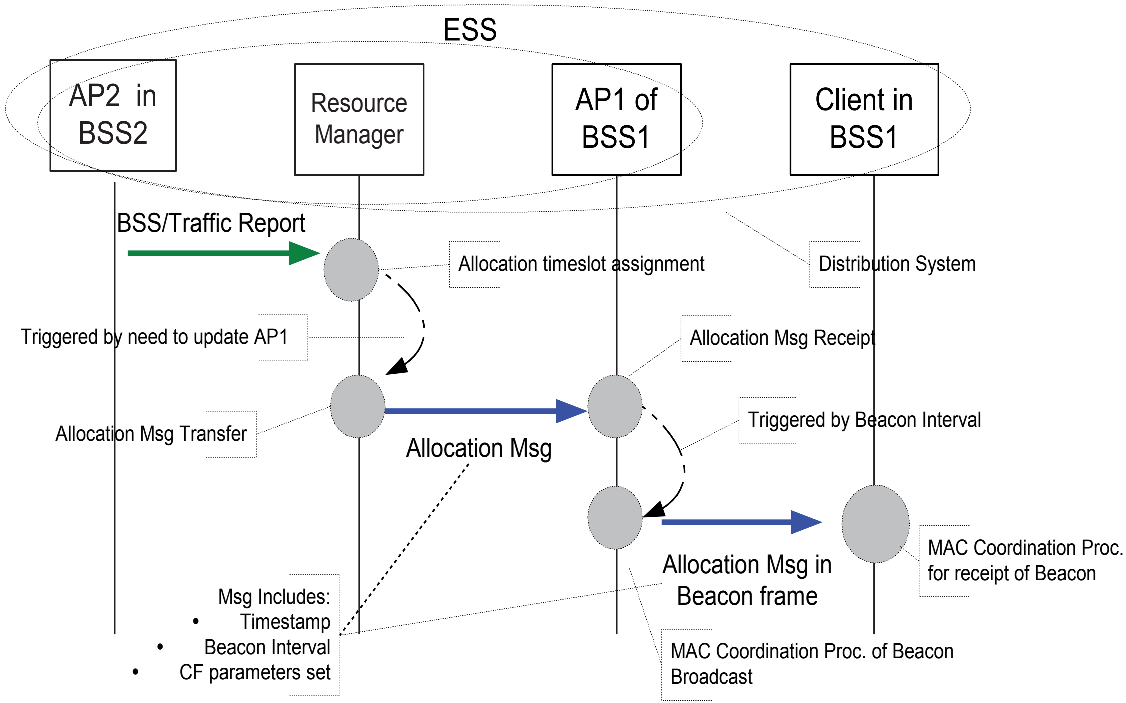


Figure 79 Allocation Message Transfer

The message destinations are clients (via AP) and it is not necessary to be transferred directly from RM to clients, because AP has control over clients for resource allocation and procedure is defined in the IEEE 802.11 standard. The message transfer between AP and client would be the standard procedure of beacon transfer within the BSS as defined in [IEEE P802.11 1999]. As discussed above, the allocation message can be interpreted as beacon information elements and is carried in the beacon frame when it is transferred from AP its clients. Therefore, we only concern how the message is transferred from the RM to AP as below.

The message transfer from the RM to AP can occur in two ways: either based on requests or in a repetition period (periodic update). The periodic update for message transfer operates in a way similar to a beacon transmission in [IEEE P802.11 1999]. The updates are scheduled every per-defined repetition period for transmission. The repetition period has to be defined a relative short time period, such as a beacon interval size; otherwise, the message may become out-dated when they are arrived in the AP. However, if the update were transferred so frequently, it would create relatively large amount of overhead in the network. More importantly, in case of message losses within the transport path, the BSS would not be aware of the losses.

Therefore, it comes to question how often the RM has an allocation updates to the BSS. As discussed, an allocation update to a BSS only occurs when such conditions below are met:

- The RM has recognized changes on traffic demands and overlapping topology based on the reports received from any of its BSSs.

- And then the RM performs the resource allocation re-arrangement and it results in needs for updates to a list of the addressed BSSs.
- The BSS is on such list.

Not all BSSs need to be involved in resource allocation update. For example, one BSS experiences high downlink traffic demand, while there is no any topology change, the allocation re-arrangement would only result in assigning the second group to the BSS, while keeping the superframe structure unchanged. In such case, only this BSS would receive an allocation message to allow this BSS access the additional second time span for downlink traffic.

Also, as discussed above, only overlapping message reports or traffic demand reports trigger the allocation re-arrangement/update. And changes on overlapping topology and traffic demands do not occur frequently and periodically; therefore the irregular changes do not lead to a periodical need for allocation updates to BSSs.

Based on the rationale above, it is considered that the allocation message transfer is more effective with the request mechanism as compared to the periodic mechanism. And we consider topology and traffic condition reports as the requests for allocation update. In such way, the reports collected in the RM fully describe the current environment timely; the RM then is able to respond in time and transfers allocation message to the addressed BSS in a highly effective way.

The process details are shown in Figure 79. The implementation of the allocation message transfer from the RM to AP can be achieved in two ways. They are either forwarded as MAC management frames or as data over IP. However, it is preferable to be implemented as MAC management frames as bridge operation within 802 sub-network groups brings simplicity. The Appendix section gives implementation details of this new management service in the context of IEEE 802.11.

B.3 Slot Duration Re-adjustment

As discussed in Chapter 6, the design of resource allocation scheme is based on traffic conditions and the RM overlapping topology. In the previous section 6.5.2, we have mainly described how the scheme is dynamic to the current overlapping topology. In this section, we further focus on how RM re-adjusts the designed timeslot duration based on the traffic condition.

Re-adjustment objectives

The allocation scheme needs to be re-adjusted according to the varieties of the traffic patterns from time to time. The system divides the uplink and downlink traffic to be transferred separately in the contention period and polling period during a superframe period. Uplink traffic demands are not always the same as the downlink traffic demands. Often, downlink traffic is more dominant nowadays and uplink/downlink proportion varies from time to time. Also, each BSS may experience different traffic pattern (uplink/downlink demands) at the same time. The system needs to make leverage of pattern demands from all BSSs from time to time. Therefore, the design of uplink/downlink slot duration should be dynamical to the current traffic conditions in all BSSs.

Furthermore, the admitted Traffic Streams (TS, delay-sensitive applications) always need to be paid attention to, as they have strict requirements on the delay performance. Certain guarantee needs to be made to ensure the enough resources allocated to these applications as promised, when re-adjusting the slot duration.

Traffic Report information needed for Re-adjustment

According to the objectives above, the traffic report information needed in RM for slot duration re-adjustment should include:

- Current uplink traffic demands: being overloaded or not
- Current downlink traffic demands: being overloaded or not
- Current proportions of admitted TS in uplink
- Current proportions of admitted TS in downlink

Report. Name	*Field Order*	*Field Name*	*Type*	*Valid Range*	*Description*
Traffic Report	*1*	*Status Code*	*Boolean*	*True, false (2 octets)*	*Uplink Demand*
	2	*Status Code*	*Boolean*	*True, false (2 octets)*	*Downlink Demand*
	3	*TS Uplink*	*Real*	*>=0 (2 octets)*	*The current TS proportions in uplink*
	4	*TS Downlink*	*Real*	*>=0 (2 octets)*	*The current TS proportions in downlink*

Table 13 Traffic Information Structure

These traffic information needed would be available in AP, as each AP is the gateway for all traffic. During the polling (downlink) period, AP plays as a point coordinator for polling medium access, it is aware if the downlink sub-system can satisfy the transmission demands. During the contention period, AP listens to channel and monitors transmission attempts from all its clients and then is aware of the uplink traffic conditions status. AP would also have information on proportion of all admitted Traffic Streams from the admitted TSPEC elements, which are recorded in AP when these Traffic Streams are registered and admitted.

Traffic Report Transfer Process

Traffic information items are collected in AP, and they need to be transferred from AP to the Resource Manager. We recall that in Section 6.5, the BSS report protocol (for the purpose of overlapping message collection) is defined to transfer BSS report from AP to the RM. These two types of information, the BSS report and traffic report, are handled in the same entities around the same transfer path; we consider that the traffic report may share the transport mechanism with the BSS report process.

The traffic report occurs at a periodic method. The traffic conditions vary from time to time, even though the changes may not be necessarily large. AP should report the traffic condition timely, and let RM estimate the traffic condition changes and justify the need for update. Therefore, traffic information should be updated in RM with the current traffic conditions via the periodical report from AP. The repetition period should be one of the performance tuning parameters.

Therefore, the traffic report process in AP has two possible triggers, the traffic report timer and the awareness of need for BSS report as shown in Figure 80 above. If a traffic report is triggered by a BSS report, the timer resets, so no duplicate copy of traffic report would be transfer until the next repetition period is reached. The output of the process is the traffic report message forwarded from AP to RM.

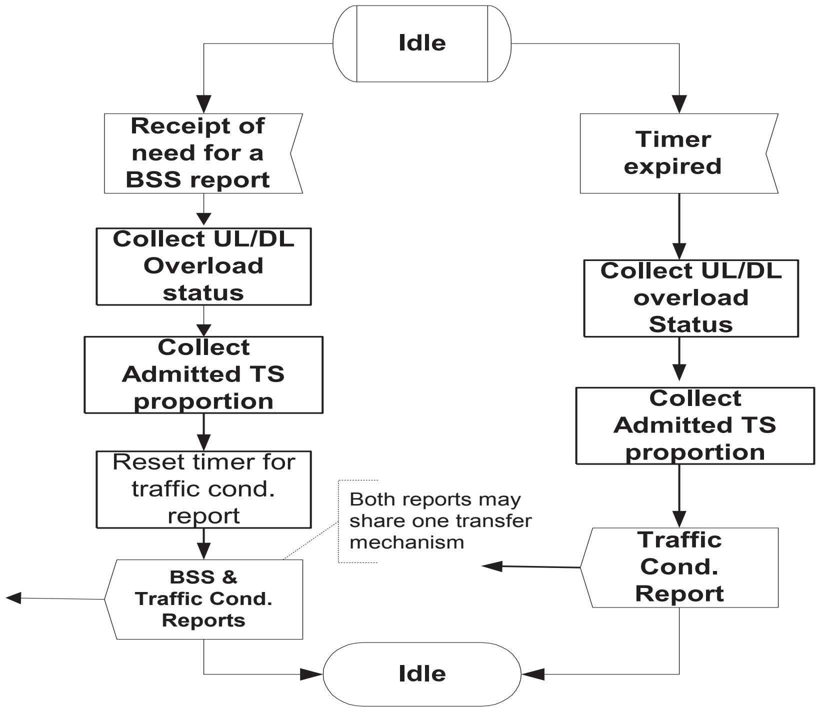


Figure 80 Traffic Report Collection Process in AP

Slot Duration Re-adjustment Process

The slot duration re-adjustment process is part of resource allocation protocol shown in Figure 81. It consists of two parts: traffic condition justification and timeslot increment/decrement. Further details will be given in following Figure.

The traffic condition justification process may result in timeslot re-assignment (increment/decrement), or result with no change on the allocation scheme. Upon receipt of reports from its BSSs, the RM firstly figures out the average uplink and downlink traffic status and leverages the demands from all BSSs. The justification of being overloaded or underloaded in an OBSS environment is as below:

- If the number of BSSs with overloaded status is X% larger than the number of BSSs with underloaded status, we consider the system overloaded in average.
- If the number of BSSs with underloaded status is X% larger than the number of BSSs with overloaded status, we consider the system underloaded in average.

X should be one of the performance tuning parameters. If both uplink and downlink are overloaded, there is no possibility for uplink/downlink re-adjustment. Instead, the admission control should be re-adjusted in RM. If both uplink and downlink are underloaded, there is no need for re-adjustment as the system satisfies all demands.

If the justification process results in need for timeslot re-assignment, the timeslot increment/decrement would perform. The timeslot increment/decrement would not make any effect on the existing results of group assignment in RM. The RM would only try to re-adjust the uplink period by a re-adjustment increment or decrement Y. In order to provide guarantee on all admitted TS on all BSS, the re-adjusted scheme must still satisfy the current sessions for all admitted TS. In details, it is:

- When decreasing the uplink slot, the adjusted uplink should be less than any of uplink admitted TS proportions among all BSSs;
- When decreasing the downlink slot, the adjusted downlink time spans should not be less than any of downlink admitted TS proportions among all BSSs.

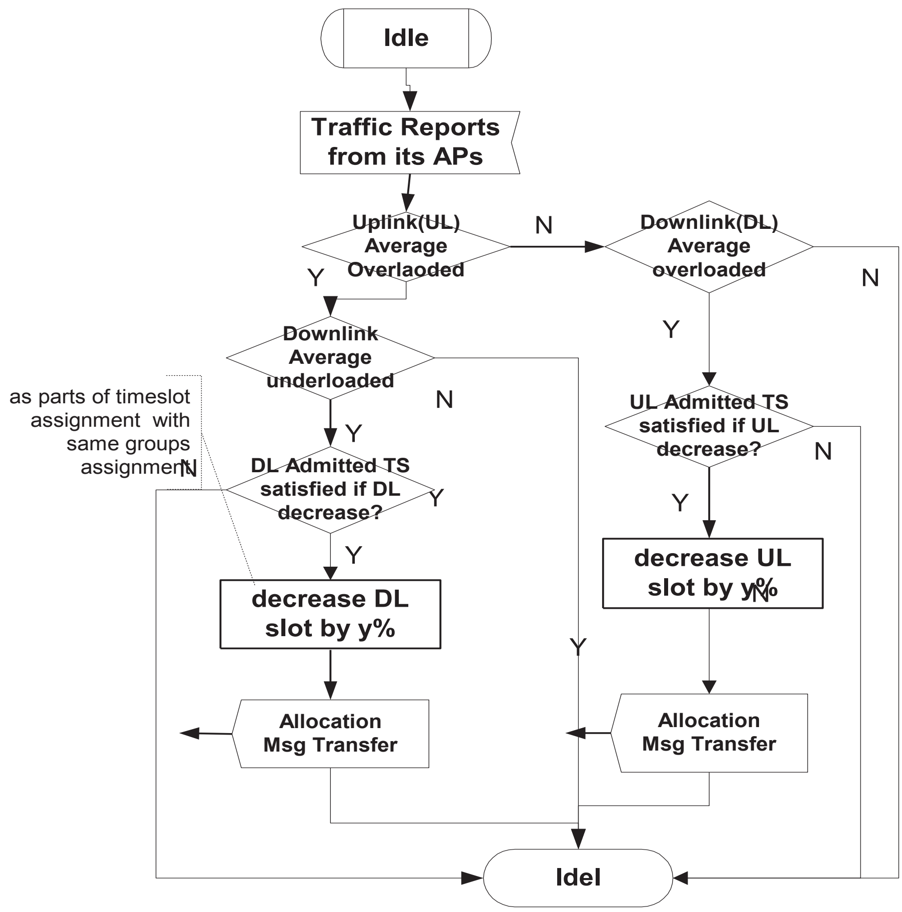


Figure 81 Slot Re-adjustment Protocol

The slot re-adjustment increment/decrement Y should be one of the performance tuning parameters.

The process completes with output of allocation message transfer, which carries the new allocation scheme for the BSSs. The re-adjustment would then make effect on the traffic transmission and the results would be again reported to RM as feedbacks via the periodical traffic report to justify the need for further re-adjustment. In such way, the uplink and downlink allocation would be balanced in an appropriate point.

B.4 Mutual Synchronization across OBSS

Maintaining synchronization

In accordance with the resources coordination principle, all APs and clients within the OBSS environment shall be synchronized to a common clock using the similar timing synchronization function (TSF) kept in all BSSs. The RM shall be the timing master and shall initialize the TSF timer. Each AP should also maintain its local TSF timer as defined in [IEEE P802.11 1999].

The RM would regularly transmit Allocation message that contains a copy of its TSF timer to synchronize the APs in all participating BSSs. A receiving AP shall always accept the timing information in the Allocation message sent from the RM servicing the OBSS environment /ESS. If the AP's TSF timer is different from the timestamp in the received Allocation message, the receiving AP shall set its local timer to the received timestamp value.

The RM shall generate the transmission need for allocation message transfer to a BSS when a resource allocation update for such BSS occurs. The RM shall define the timing for the entire OBSS environment by transmitting allocation message regularly. As the transmission is over a fixed DS network, the actual transmission of an allocation message would have more reliable performance compared to the beacon transmission over the wireless medium.

Timer Accuracy

Upon receiving an allocation message, the AP shall update its TSF timer according to the following algorithm: The received timestamp value shall be adjusted by adding an amount equal to the receiving AP's delay through its local PHY components plus the time since the first bit of the timestamp was received at the MAC/PHY interface. The AP's TSF timer shall then be set to the adjusted value of the timestamp.

Appendix C an Implementation of Messages Transfer for OBSS Coordination in the context of IEEE 802.11

As seen in Chapter 6, in the entire allocation scheme, there are four types of messages transferred with the OBSS environment. They are client report (overlapping information collection), BSS report (overlapping information collection), traffic report, and the allocation messages.

The messages transfers are preferable to be implemented as MAC management frames, as bridge operation within 802 sub-network groups brings simplicity and unique for system management; and the existing of IEEE 802 MAC service also provides ease of implementation. Additionally, the BSS report and traffic report are able to share the transport facility between AP and RM, as they have same transport path. The following sections give implementation details of these new management services in the context of IEEE 802.

- Client Report

The client report occurs only between AP and client in the BSS domain. Therefore, it only requires making use of reserved management function in IEEE 802.11 for the client report service as below.

- The Protocol

For this purpose of client report, a new MAC sublayer management entity (MLME) service, the “Client-Report” service can be used. This service is requested by the station management entity (SME) residing within each client via a management primitive MLME-CLIENT-REPORT.request, in order to report existing overlapping BSSs from the station to AP. The primitive MLME-CLIENT-REPORT.confirm returns the report results to the SME.

Upon a need for client report, SME of this station issues the MLME-CLIENT-REPORT.request. The client would then send a CLIENT-REPORT management frame, to its AP, with the new detected BSS information. This information includes the BSSID, SSID of the detected BSS and status code of overlapping. An MLME- CLIENT-REPORT.confirm is issued by the MLME indicating all of the OBSS information received by the AP, upon the receipt of the ACK from the AP.

- The Frame Format

The subtype value of the frame control field is defined 1101 (currently reserved for management function.). The frame body of a management frame with subtype OBSS Report request and confirm contains the information shown below.

Client Report Requests (Information éléments)

Order	*Information*	*Note*
1	*BSSID of detected BSS*	
2	*SSID of detected BSS*	
3	*Status code of overlapping*	

Client Report Confirm

Order	*Information*	*Note*
1	*Status Code*	

Table 14 Management Frame Format of Message Transfer in IEEE 802.11

Layer Management (SAP Interfaces)

As seen, there are additional management entities required for coordination management; therefore, additional SAP (Service Access Point) interfaces are defined here for these various entities to interact. These interfaces are all only with the MLME domain.

<u>MLME-CLIENT-REPORT.request</u>

The primitive reports a change in the overlapping situation.

The primitive parameters are as follows:

MLME- CLIENT-REPORT.request (
Client Report Information Elements
)

This primitive is generated by the SME of client to report the information of newly detected overlapping BSS to its AP. The MLME subsequently issues a MLME- CLIENT - Report.confirm that reflects the results.

<u>MLME- CLIENT -REPORT.confirm</u>

This primitive confirms the client report procedure.

The primitive parameters are as follows:

MLME- CLIENT-REPORT.confirm (
ResultCode)

This primitive is generated by the MLME as a result of an MLME-CLIENT Report.request to the AP. The SME is notified the result of the client-report procedure.

BSS Report and Traffic Report

We recall the assumption of the supporting distribution system; All MAC frames of 802.11 WLAN are supported in bridge operation (Refer to ANSI/IEEE Std 802.1D, "Media Access Control (MAC) Bridges"). Based on these rationales, a 802 MAC frame can be transferred via simple bridging within the distribution system. Therefore, we define BSS-TRAFFIC-REPORT and ALLOCATION management services to support the message transfer from AP to the RM within the supporting distribution system, using the bridge operation.

- The Frame Format

Order	*Information*	*Note*
1	*BSSID of detected BSS*	
2	*SSID of detected BSS*	
3	*No. of clients overlapped in the reporting BSS*	
4	*Current uplink traffic demands: being overloaded or not*	
5	*Current downlink traffic demands: being overloaded or not*	
6	*Current proportions of admitted TS in uplink*	
7	*Current proportions of admitted TS in downlink*	

Table 15 BSS Traffic Report request (Information elements) in IEEE 802.11

- The Protocol

This BSS-TRAFFIC-REPORT service is requested by the station management entity (SME) residing within each AP via a management primitive MLME-BSS-TRAFFIC-REPORT.request, in order to report existing overlapping BSS information and traffic information. Upon a need for traffic or BSS report, the SME of AP issues the MLME-BSS-TRAFFIC-REPORT.request. The AP shall then transmit a traffic report or BSS report to a Resource Manager. As the report occurs in a repetition period, there is no need for a receipt confirmation for the report from the RM

The subtype value of the frame control field is defined 1110 (currently reserved for management function.). The frame body of a management frame with subtype OBSS Associate request and confirm contain the information shown below. The BSSID of the sender BSS is written in the field of 'address 3' in the MAC header.

- Layer Management

This mechanism supports the process of report of a BSS to a Resource Manager.

MLME- BSS-TRAFFIC-REPORT.request

The function of this primitive is to report to a Resource Manager.

The primitive parameters are as follows:

MLME- BSS-TRAFFIC-REPORT.request (
MAC address of the AP initiating the report,
MAC address of the RM with which the initiating AP will report to,
BSSID and SSID of the sender BSS,
BSS Traffic Report Information elements,
ListenInterval
)

This primitive is generated by the SME of an AP for its BSS to make a report to a RM when such BSS detect a need for traffic report or traffic report. It initiates a report procedure.

Allocation

As the rationales above for traffic or BSS report, we define OBSS-allocation management service to support the message transfer from the RM to AP within the supporting distribution system, using the bridge operation. This service is requested by the management entity residing within RM via a management primitive MLME-OBSS-Allocation.reqest, in order to transfer the updated allocation message. The primitive MLME-OBSS-allocation.confirm returns the allocation results to the management entity.

- The Protocol

Upon completion of allocation computation in the Resource Manager, it firstly find out, which BSS needs to be allocation-updated. After the addressed BSS is confirmed with its BSSID, RM should prepare the particular allocation content for this addressed BSS and its management entity issues the MLME-OBSS- Allocation.Request. The RM shall then transmit an Allocation Request frame to the AP of the addressed BSS. If an Allocation confirm frame is received with a status value of "successful" in the Resource Manager, the addressed BSS is then allocation-updated and the MLME shall issue an MLME-OBSS-Allocation.confirm indicating the successful completion of the operation. The Allocation Beacon frame carries the BSSID of the addressed BSS, Timestamp, Beacon Interval, and its OBSS CP parameter set.

Whenever an Allocation request frame is received at an AP and its BSS is allocation-updated, the AP shall transmit an Allocation confirm frame with a status code. The Allocation Confirm frame is only directed to the RM to inform the receipt of allocation. A timeout procedure is carried out if the allocation is a failure operation.

- The Frame Format

The subtype value of the frame control field is defined 1110 (currently reserved for management function.). The frame body of a management frame with subtype OBSS Allocation Request and confirm contain the information shown below. The BSSID of the sender BSS is written in the field of 'address 3' in the MAC header.

OBSS Allocation Request

Order	*Information*	*Note*
1	*Timestamp*	
2	*Beacon Interval*	
3	*OBSS CP parameter set*	

OBSS Allocation Confirm

Order	*Information*	*Note*
1	*Status Code*	

Table 16 OBSS Allocation Frame Format in IEEE 802.11

- Layer Management

This mechanism supports the process of allocation message transfer from a RM to an addressed BSS.

MLME-OBSS-Allocation.Request

The function of this primitive is to convey message from a RM to an AP.

The primitive parameters are as follows:

MLME-OBSS-Allocation.Request (
MAC address of the AP receiving the allocation message,
Timestamp,
Beacon Interval,
OBSS CP parameter set,
ListenInterval
)

This primitive is generated by the management entity of a RM to update an addressed BSS with resource allocation scheme when such RM confirms the need for update for this BSS. It initiates a resource allocation procedure. The MLME subsequently issues a MLME-OBSS-allocation.confirm that reflects the results.

MLME- OBSS-Allocation.confirm

This primitive reports the results of a resource allocation attempt with a specified peer MAC entity that is acting as an AP.

The primitive parameters are as follows:

MLME-OBSS-Allocation.confirm (
ResultCode
)

This primitive is generated by the MLME as a result of an MLME-Allocation.Beacon to allocate resource to an AP when the RM receives one of OBSS-Allocation.confirm frames from the AP. The MLME is informed the results of allocation to the Resource Manager.