Modelling the IEEE 802.11 Wireless MAC Layer under Heterogeneous VoIP Traffic to Evaluate and Dimension QoE

Submitted for the Degree of Doctor of Philosophy

Oliver M. Shepherd

oliver.shepherd@eeecs.qmul.ac.uk

Supervisor : Prof. J. M. Pitts

Networks Research Group
School of Electronic Engineering and Computer Science
Queen Mary, University of London

June 2011
To My Mother & Father and Wife
Tracy, William and Lina
ABSTRACT

As computers become more popular in the home and workplace, sharing resources and Internet access locally is a necessity. The simplest method of choice is by deploying a Wireless Local Area Network; they are inexpensive, easy to configure and require minimal infrastructure. The wireless local area network of choice is the IEEE 802.11 standard; IEEE 802.11, however, is now being implemented on larger scales outside of the original scope of usage. The realistic usage spans from small scale home solutions to commercial ‘hot spots,’ providing access within medium size areas such as cafés, and more recently blanket coverage in metropolitan. Due to increasing Internet availability and faster network access, in both wireless and wired, the concept of using such networks for real-time services such as internet telephony is also becoming popular.

IEEE 802.11 wireless access is shared with many clients on a single channel and there are three non-overlapping channels available. As more stations communicate on a single channel there is increased contention resulting in longer delays due to the backoff overhead of the IEEE 802.11 protocol and hence loss and delay variation; not desirable for time critical traffic.

Simulation of such networks demands super-computing resource, particularly where there are over a dozen clients on a given. Fortunately, the author has access to the UK’s super computers and therefore a clear motivation to develop a state of the art analytical model with the required resources to validate. The goal was to develop an analytical model to deal with realistic IEEE 802.11 deployments and derive results without the need for super computers.

A network analytical model is derived to model the characteristics of the IEEE 802.11 protocol from a given scenario, including the number of clients and the traffic load of each. The model is augmented from an existing published saturated case, where each client is assumed to always have traffic to transmit. The nature of the analytical model is to allow stations to have a variable load, which is achieved by modifying the existing models and then to allow stations to operate with different traffic profiles. The different traffic profiles, for each station, is achieved by using the augmented model state machine per station and distributing the probabilities to each station’s state machine accordingly.

To address the gap between the analytical models medium access delay and standard network metrics which include the effects of buffering traffic, a queueing model is identified and augmented which transforms the medium access delay into standard network metrics; delay, loss and jitter. A Quality of Experience framework, for both computational and analytical results, is investigated to allow the results to be represented as user perception scores and the acceptable voice call carrying capacity found. To find the acceptable call carrying capacity, the ITU-T G.107 E-Model is employed which can be used to give each client a perception rating in terms of user satisfaction.
With the use of a novel framework, benchmarking results show that there is potential to maximise the number of calls carried by the network with an acceptable user perception rating. Dimensioning of the network is undertaken, again compared with simulation from the super computers, to highlight the usefulness of the analytical model and framework and provides recommendations for network configurations, particularly for the latest Wireless Multimedia extensions available in IEEE 802.11.

Dimensioning shows an overall increase of acceptable capacity of 43%; from 7 to 10 bi-directional calls per Access Point by using a tuned transmission opportunity to allow each station to send 4 packets per transmission. It is found that, although the accuracy of the results from the analytical model is not precise, the model achieves a 1 in 13,000 speed up compared to simulation. Results show that the point of maximum calls comes close to simulation with the analytical model and framework and can be used as a guide to configure the network. Alternatively, for specific capacity figures, the model can be used to home-in on the optimal region for further experiments and therefore achievable with standard computational resource, i.e. desktop machines.
ACKNOWLEDGMENTS

I would like to first and foremost thank my supervisor Jonathan Pitts for his continued support and encouragement from prior to commencing the PhD right the way through to professional life. I cannot begin to express my sincere gratitude for the guidance and opportunities.

I would also like to thank Chris Philips and John Schormans for their continued suggestions and encouragement throughout the process.

Finally, I would like to thank everyone in the Department of Electronic Engineering at Queen Mary, University of London for making such a friendly environment. Special thanks to my colleagues; Rupert Ogilvie, Tony Yang, Vindya Wijeratne, Touseef Chaudhery, Keith Jones and Ammar Lilamwala.
# TABLE OF CONTENTS

Abstract ..................................................................................................................... 3  
Acknowledgments .................................................................................................... 5  
Table of Contents...................................................................................................... 6  
List of Figures ........................................................................................................... 9  
List of Tables .......................................................................................................... 13  
List of Symbols ........................................................................................................ 14  
Glossary................................................................................................................... 18  

1 Introduction .................................................................................................... 20  
1.1 Background ................................................................................................................. 20  
1.2 Objectives .................................................................................................................... 20  
1.3 Novelty ........................................................................................................................ 21  
1.4 Thesis Outline ............................................................................................................ 21  
1.4.1 Diagrammatic Overview............................................................................................. 23  

2 Background ..................................................................................................... 25  
2.1 The IEEE 802.11 Protocol ...................................................................................... 25  
2.2 Voice over Internet Protocol ................................................................................... 30  
2.2.1 Overview.................................................................................................................... 30  
2.2.2 Audio related standards ............................................................................................. 31  
2.2.3 Proprietary codecs ....................................................................................................... 32  
2.3 Perceptual Quality...................................................................................................... 32  
2.3.1 Overview.................................................................................................................... 32  
2.3.2 P.800 Mean Opinion Score ........................................................................................... 32  
2.3.3 ITU-T P.861 Perceptual Speech Quality Measure ..................................................... 33  
2.3.4 ITU-T P.862 Perceptual Evaluation of Speech Quality ............................................. 33  
2.3.5 ITU-T G.107 E-Model ............................................................................................... 33  

3 Literature Survey ............................................................................................. 35  
3.1 Modelling the MAC Layer ................................................................................... 35  
3.1.1 Initial MAC Layer Models ....................................................................................... 35  
3.1.2 Un-Saturated MAC Layer Models ........................................................................... 36  
3.1.3 Heterogeneous MAC Layer Models ........................................................................... 38  
3.2 Queue Modelling ...................................................................................................... 39  
3.3 Optimising Network Performance ........................................................................ 40  
3.3.1 Diff-Serv-Like Scheme .............................................................................................. 41  
3.3.2 Distributed Fair Scheduling ..................................................................................... 41  
3.3.3 Virtual MAC ............................................................................................................ 41  
3.3.4 Black Burst ............................................................................................................... 42  
3.3.5 IEEE 802.11e ......................................................................................................... 42  
3.3.6 Adaptive EDCF ....................................................................................................... 43  
3.3.7 Merging and Multicasting Scheme ........................................................................... 44  
3.3.8 Prioritising Voice over Data .................................................................................... 44
3.3.9. Priority Access for the AP ................................................................. 45
3.3.10. Further Methods ........................................................................... 45
3.1 Voice over Wireless LAN ..................................................................... 46
3.1.1. Determining Voice Capacity ............................................................ 46

4 Saturated MAC Model ........................................................................... 47
4.1 Introduction ........................................................................................... 47
4.2 Deriving the Bianchi Model ................................................................. 48
4.3 Comparing with Simulation ................................................................. 55
4.4 Further Explorations ........................................................................... 58

5 Un-Saturated Homogeneous MAC Model .............................................. 62
5.1 Introduction ........................................................................................... 62
5.2 Un-Saturated MAC Model ................................................................. 62
5.3 Convergence to Bianchi Model ............................................................ 67
5.4 Model Limitations ............................................................................... 68
5.5 Comparing with Simulation ................................................................. 69
5.6 Confirming against Large ACK Retransmission Limit ......................... 71

6 Heterogeneous Traffic MAC Model ....................................................... 73
6.1 Introduction ........................................................................................... 73
6.2 Updating the Probabilities ................................................................... 75
   6.2.1. Collision Probability ..................................................................... 75
   6.2.2. Probability of the Channel perceived Busy ................................. 76
   6.2.3. Probability of a Station Transmitting Successfully ...................... 76
   6.2.4. Scrutiny of the Formula for Even Load ....................................... 77
6.3 Updating the Timings .......................................................................... 78
6.4 2 Station Scenario Highlighting the Current Discrepancy ..................... 81
   6.4.1. Solving the State Machine ............................................................ 81
   6.4.2. Full State Investigation ................................................................. 82
6.5 Updating Transmission Probability from Slot Based to Time Based ...... 84
   6.5.1. Identifying the Problem ............................................................... 84
   6.5.2. Updating Transmission Probability Discrepancy ......................... 85
6.6 Calibrating the Model ......................................................................... 86
   6.6.1. Regression Analysis ................................................................... 88
6.7 Validating the Heterogeneous Analytical Model .................................... 90

7 Modelling the Activity Factor of a Cyclic Source ................................... 92
7.1 Introduction ........................................................................................... 92
7.2 Model .................................................................................................. 92

8 Impacts of Buffering Traffic .................................................................. 97
8.1 Introduction ........................................................................................... 97
8.2 Model .................................................................................................. 97
   8.2.1. Queue Size .................................................................................. 98
   8.2.2. Loss ............................................................................................ 101
   8.2.3. Delay .......................................................................................... 101
   8.2.4. Jitter .......................................................................................... 103
9 Exercising the Analytical Model from a QoE Perspective

9.1 Introduction

9.2 Computation Iteration Technique

9.2.1 Connecting the Analytical Models

9.2.2 Building the Heterogeneous Analytical Model

9.2.3 Pre-Execution Optimisation

9.2.4 Iterating to Steady State

9.3 Simulation Technique

9.4 Quality of Experience Measure

9.4.1 R-Factor

9.4.2 Mean Opinion Score

9.5 Benchmarking VoIP QoE

9.5.1 Homogeneous Ad-Hoc Benchmark

9.5.2 Heterogeneous Access Point Benchmark

9.6 Dimensioning VoIP QoE

9.6.1 Packetisation Period

9.6.2 Impacts on the Homogeneous Case

9.6.3 Impact on the Heterogeneous Case

9.6.4 What-If Capabilities

9.7 Analytical Model Usage Conclusions

10 Conclusion & Future work

10.1 Future Direction

10.1.1 Modelling Heterogeneous Applications

10.1.2 Scaling-Up

10.1.3 Fairer Optimisation

Appendix A: Packetisation Optimisation Results

A.1 Heterogeneous Results: Poisson

A.2 Heterogeneous Results: On/Off

A.3 What-If Homogeneous: Poisson

A.3 What-If Homogeneous: On/Off

References

Author’s Publications
LIST OF FIGURES

Figure 1.1 : Research Overview ................................................................. 23
Figure 2.1 : System Diagram : Successful Transmission ................................. 26
Figure 2.2 : System Diagram : Collision and Backoff ..................................... 27
Figure 2.3 : System Diagram : Multiple Collisions ........................................ 27
Figure 2.4 : System Diagram : Pausing Slot Timer for Successful Transmission .. 28
Figure 2.5 : System Diagram : Pausing Slot Timer for Collision ..................... 28
Figure 2.6 : System Diagram : Saturated Condition ....................................... 29
Figure 2.7 : Hidden Node Problem .............................................................. 29
Figure 2.8 : Exposed Node Problem ............................................................. 29
Figure 3.1 : Bianchi Saturated Markov Model with Estimated Mean Backoff Slots 35
Figure 3.2 : Xu’s Un-Saturated Model .......................................................... 37
Figure 4.1 : Bianchi Saturated Markov Model with Expected Mean Backoff Slots 48
Figure 4.2 : Transmission vs. Collision Probability ........................................ 49
Figure 4.3 : Generic Saturated MAC Transition Matrix ................................... 51
Figure 4.4 : Saturated MAC Transition Matrix for m=3 ................................. 51
Figure 4.5 : Bianchi Model Overview .......................................................... 57
Figure 4.6 : Simulations vs. Bianchi Model ................................................... 58
Figure 4.7 : Packet Size vs. Saturated Throughput (Simulation) ..................... 60
Figure 4.8 : Packet Size vs. Saturated Service Time (Simulation) .................... 60
Figure 5.1 : Pitts Un-Saturated Markov Model .............................................. 62
Figure 5.2 : Markov Chain considering ACK Retry Limit .............................. 68
Figure 5.3 : Load vs. Throughput : Analysis and Simulation .......................... 69
Figure 5.4 : Load vs. Service Time : Analysis and Simulation ......................... 70
Figure 5.5 : Load vs. Throughput : Analysis and Simulation (no ACK Retry) ....... 71
Figure 5.6 : Load vs. Service Time : Analysis and Simulation (no ACK Retry) .......... 72
Figure 6.1 : Heterogeneous State Machine per Station Interactions ..................... 73
Figure 6.2 : Heterogeneous State Machine for STA_\text{a} .................. 75
Figure 6.3 : Two Co-existing Stations .................................................. 81
Figure 6.4 : 2 Station Heterogeneous Model showing Incorrect Behaviour ........... 82
Figure 6.5 : 2 Station Heterogeneous Model with Incorrect Behaviour (Top-Down) .... 83
Figure 6.6 : 2 Station Heterogeneous Simulation showing Expected Behaviour .......... 83
Figure 6.7 : 2 Station Heterogeneous showing Expected Behaviour (Top-Down) ....... 83
Figure 6.8 : Highly Loaded Station with a mostly Free Channel ......................... 84
Figure 6.9 : Two Station Validation to Simulation ............................................ 86
Figure 6.10 : Two Station Validation to Simulation (Top-Down View) ............... 86
Figure 6.11 : Calibration Factor ............................................................ 88
Figure 6.12 : Calibration Factor Regression Analysis for N (Linear Scale) .......... 89
Figure 6.13 : Calibration Factor Regression Analysis for N (Log Scale) ............... 89
Figure 6.14 : Calibration Factor Regression Residuals for N ............................. 89
Figure 6.15 : Calibration Factor Regression for Saturated Collision Probability .... 90
Figure 6.16 : Validating Convergence back to the Homogenous Case ............... 91
Figure 7.1 : ON / OFF State Possibilities ................................................ 93
Figure 7.2 : Behaviour of the Proposed ON / OFF Model ............................ 94
Figure 7.3 : ON / OFF Model Analysis .................................................... 95
Figure 8.1 : Queueing Model Delay compared with Simulation ......................... 105
Figure 8.2 : Queueing Model Loss compared with Simulation ......................... 105
Figure 8.3 : Queueing Model Out-of-contract compared with Simulation .......... 106
Figure 9.1 : Analytical Model Connectivity : Homogeneous Poisson ................. 108
Figure 9.2 : Analytical Model Connectivity : Homogeneous On/Off ................. 109
Figure 9.3 : Analytical Model Connectivity : Heterogeneous Poisson ................. 109
Figure 9.4 : Initial $\tau_{\text{iteration} 0}$ vs. Total Iterations Needed ........................................ 110
Figure 9.5 : Maximum $\tau$ at Saturation ........................................................................ 111
Figure 9.6 : Analytical Model Iteration to a Steady State ............................................ 112
Figure 9.7 : Simulation Process .................................................................................. 113
Figure 9.8 : Calculation of Jitter ............................................................................... 115
Figure 9.9 : The R-Factor mapping to Delay and Loss with example Results ............ 117
Figure 9.10 : R to MOS Function [EM00] ................................................................. 118
Figure 9.11 : Homogenous On/Off End-to-End Delay Distribution ......................... 119
Figure 9.12 : Homogenous On/Off End-to-End Delay Percentile ......................... 120
Figure 9.13 : Homogenous On/Off Loss Breakdown ............................................... 121
Figure 9.14 : Homogenous Simulation : On/Off QoE R-Factor ................................. 122
Figure 9.15 : Homogenous Simulation : Poisson QoE R-Factor ................................. 122
Figure 9.16 : Heterogeneous Simulation : On/Off End-to-End Delay Distribution ...... 123
Figure 9.17 : Heterogeneous Simulation : On/Off Effective Loss ............................... 124
Figure 9.18 : Heterogeneous Simulation : On/Off Loss Breakdown .......................... 124
Figure 9.19 : Heterogeneous Simulation : On/Off QoE R-Factor Derivation ............... 125
Figure 9.20 : Heterogeneous Simulation : On/Off QoE R-Factor Derivation ............... 126
Figure 9.21 : Heterogeneous Simulation : Poisson QoE R-Factor Derivation ............... 126
Figure 9.22 : Heterogeneous Simulation : Poisson QoE R-Factor ............................... 127
Figure 9.23 : Analytical System Model Overview ....................................................... 129
Figure 9.24 : Optimising Homogeneous : Analytical Model for Poisson Sources ....... 129
Figure 9.25 : Optimising Homogeneous : Simulation for Poisson Sources ................. 130
Figure 9.26 : Optimising Homogeneous : Analytical Model for On/Off Sources ....... 130
Figure 9.27 : Optimising Homogeneous : Simulation for On/Off Sources ................. 131
Figure 9.28 : Optimising Heterogeneous : Analytical Model for AP with Poisson ...... 132
Figure 9.29 : Optimising Heterogeneous : Simulation for AP with Poisson ............... 133
Figure 9.30: Optimising Heterogeneous: Analytical Model for STAs with Poisson .... 133
Figure 9.31: Optimising Heterogeneous: Simulation for STA with Poisson ............... 134
Figure 9.32: Optimising Heterogeneous: Analytical Model Poisson: Acceptable ....... 135
Figure 9.33: Optimising Heterogeneous: Simulation Poisson: Acceptable ............... 135
Figure 9.34: Optimising Heterogeneous: Analytical Model On/Off: Acceptable ...... 136
Figure 9.35: Optimising Heterogeneous: Simulation On/Off: Acceptable ............... 136
Figure 9.36: Homogeneous Analytical Model Poisson plus 200ms Delay ................. 137
Figure 9.37: Homogeneous Analytical Model On/Off plus 200ms Delay ................. 138
Figure A.1: Optimising Heterogeneous: Analytical Model Pair-Wise with Poisson... 148
Figure A.2: Optimising Heterogeneous: Simulation Pair-Wise with Poisson .......... 148
Figure A.3: Optimising Heterogeneous: Analytical Model for AP with On/Off ....... 149
Figure A.4: Optimising Heterogeneous: Simulation for AP with On/Off ............... 149
Figure A.5: Optimising Heterogeneous: Analytical Model for STAs with On/Off.... 150
Figure A.6: Optimising Heterogeneous: Simulation for STA with On/Off ............. 150
Figure A.7: Optimising Heterogeneous: Analytical Model Pair-Wise with On/Off... 151
Figure A.8: Optimising Heterogeneous: Simulation Pair-Wise with On/Off .......... 151
Figure A.9: Homogeneous Analytical Model Poisson plus 30ms Delay ................. 152
Figure A.10: Homogeneous Analytical Model Poisson plus 60ms Delay ............... 152
Figure A.11: Homogeneous Analytical Model Poisson plus 100ms Delay .............. 153
Figure A.12: Homogeneous Analytical Model Poisson plus 200ms Delay .............. 153
Figure A.13: Homogeneous Analytical Model Poisson plus 30ms Delay .............. 154
Figure A.14: Homogeneous Analytical Model Poisson plus 60ms Delay .............. 154
Figure A.15: Homogeneous Analytical Model Poisson plus 100ms Delay ............. 155
Figure A.16: Homogeneous Analytical Model Poisson plus 200ms Delay ............. 155
**LIST OF TABLES**

Table 4.1 : List of Symbols for the Saturated MAC Analysis ................................................ 50
Table 4.2 : IEEE 802.11 Model Parameters ............................................................................ 57
Table 4.3 : VoIP Model Parameters .......................................................................................... 59
Table 5.1 : Deriving the Idle State Transitions ........................................................................ 63
Table 6.1 : List of Symbols for the Heterogeneous Case ........................................................ 74
Table 8.1 : List of Symbols for Queueing Analysis ................................................................. 98
Table 9.1 : Simulation Parameters .......................................................................................... 114
Table 9.2 : Simulation Statistics ............................................................................................... 115
Table 9.3 : Maximum Number of Acceptable Calls ............................................................. 128
# List of Symbols

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\lambda_{pps}$</td>
<td>Number of Packets arriving within a Service Time</td>
</tr>
<tr>
<td>$\varrho$</td>
<td>Queue Utilisation Probability</td>
</tr>
<tr>
<td>$\varrho'$</td>
<td>Utilisation – the Probability of a Packet Arriving during Transmission</td>
</tr>
<tr>
<td>$\varrho''$</td>
<td>Utilisation – Probability of a Packet being Enqueued during a transmission</td>
</tr>
<tr>
<td>$\sigma$</td>
<td>Minimum Slot Time when Channel is Idle</td>
</tr>
<tr>
<td>$\tau$</td>
<td>Transmission Probability</td>
</tr>
<tr>
<td>$a(k)$</td>
<td>Probability of $k$ Packets arriving within a Service Time</td>
</tr>
<tr>
<td>$b_i$</td>
<td>Probability of being in Backoff state $i$</td>
</tr>
<tr>
<td>$b_{k,n}$</td>
<td>Probability of a Station $n$ being in a given Backoff State, $k$ for a given Station $n$</td>
</tr>
<tr>
<td>$B_{d}(k)$</td>
<td>Delay caused by more than one Packet Arriving together</td>
</tr>
<tr>
<td>$CW$</td>
<td>Number of Slots a Station waits before attempting to Transmit</td>
</tr>
<tr>
<td>$CW_{min}$</td>
<td>The Minimum Range of Slots a Station must Wait to Transmit a Packet</td>
</tr>
<tr>
<td>$CW_{max}$</td>
<td>The Maximum Range of Slots a Station needs to Wait after Successive Collisions</td>
</tr>
<tr>
<td>$E[\text{backoff}]$</td>
<td>Estimated Mean backoff time</td>
</tr>
<tr>
<td>$E[\text{backoff}]_n$</td>
<td>Estimated Mean Slots a Station $n$ needs to wait in order to transmit a packet successfully</td>
</tr>
<tr>
<td>Notation</td>
<td>Description</td>
</tr>
<tr>
<td>----------</td>
<td>-------------</td>
</tr>
<tr>
<td>$E[\text{Sat. Service Time}]$</td>
<td>Estimated Mean Service Time when all Stations are Saturated</td>
</tr>
<tr>
<td>$E[\text{Sat. Service Time}]_{\text{min}}$</td>
<td>Minimum Saturated Service Time when only one Station</td>
</tr>
<tr>
<td>$E[\text{Service Time}]$</td>
<td>Estimated Mean Service Time</td>
</tr>
<tr>
<td>$E[\text{Service Time}]_n$</td>
<td>Estimated Mean Service Time of Station $n$.</td>
</tr>
<tr>
<td>$E[\text{slots}]$</td>
<td>Estimated Mean number of Slots a Station needs to Wait before Transmitting Successfully</td>
</tr>
<tr>
<td>$E[\text{slots}]_n$</td>
<td>Estimated Number of Slots to Wait before Transmitting Successfully for a given Station $n$</td>
</tr>
<tr>
<td>$I_d$</td>
<td>The perceived Delay component QoE degradation of a Voice Call</td>
</tr>
<tr>
<td>$I_e$</td>
<td>The perceived Effective Loss component QoE degradation of a Voice Call</td>
</tr>
<tr>
<td>$j_{\text{lower}}$</td>
<td>The minimum Delay Bound used to determine if a Packet is not Admitted into the De-Jitter Buffer</td>
</tr>
<tr>
<td>$j_{\text{higher}}$</td>
<td>The maximum Delay Bound used to determine if a Packet is not Admitted into the De-Jitter Buffer</td>
</tr>
<tr>
<td>$m$</td>
<td>Number of Times the Range of Slots is Doubled</td>
</tr>
<tr>
<td>MOS</td>
<td>The Mean Opinion Score of the Perceived Experience of a Voice Call</td>
</tr>
<tr>
<td>$p$</td>
<td>Collision Probability</td>
</tr>
<tr>
<td>$p_{i,j}$</td>
<td>Probability of Transitioning from Backoff State $i$ to $j$</td>
</tr>
<tr>
<td>$P{tr}_n^{\text{het}}$</td>
<td>Collision Probability in the Heterogeneous Case for a given Station $n$</td>
</tr>
<tr>
<td>$P{tr}_n^{\text{hom}}$</td>
<td>Collision Probability on the Homogeneous Case for a given Station $n$</td>
</tr>
</tbody>
</table>
\( P(ON) \)  
Probability of a Station being Active and Sending Voice Data

\( P_s \)  
Probability of Channel being Busy due to another Station Transmitting Successfully

\( P_{s_n}^{\text{het}} \)  
Probability of the Channel being Busy due to a Station Transmitting Successfully in the Heterogeneous Case for a given Station \( n \)

\( P_{s_n}^{\text{hom}} \)  
Probability of the Channel being Busy due to a Station Transmitting Successfully in the Homogeneous Case for a given Station \( n \)

\( P_{tr} \)  
Probability of Perceiving the Channel Busy

\( P_{tr_n}^{\text{het}} \)  
Probability of Perceiving the Channel Busy in the Heterogeneous Case for a given Station \( n \)

\( P_{tr_n}^{\text{hom}} \)  
Probability of Perceiving the Channel Busy in the Homogeneous Case for a given Station \( n \)

\( r \)  
Number of Transmission Retries when using a Maximum Number of Re-transmits

\( R \)  
QoE Metric Describing the Perceived Quality of a Voice Call

\( s(k) \)  
Probability of a Queue having \( k \) packets En-Queued

\( S_i \)  
Maximum Number of Slots to Wait in Backoff state \( i \)

\( S_{k_n} \)  
Number of Slots a Station \( n \) waits in each Backoff State, \( k \) for a given Station \( n \)

\( st \)  
Short Hand for the Estimated Mean Service Time

\( T_c \)  
Time of Transmitting a Packet which is not Acknowledged due to a Collision

\( T_{c_{other}} \)  
Time the Channel is Busy due to two Stations Transmitting and Resulting in a Collision
<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T_d(k)$</td>
<td>Delay caused by Packets already in the Queueing System</td>
</tr>
<tr>
<td>$T_{DIFS}$</td>
<td>Time of a DIFS period (used before transmitting and resuming backoff timer)</td>
</tr>
<tr>
<td>$\overline{T}_k$</td>
<td>Mean Slot Time in Each Backoff state $k$</td>
</tr>
<tr>
<td>$T_{s_k}$</td>
<td>Time in Slots a Station $n$ waits in each Backoff State, $k$ for a given Station $n$</td>
</tr>
<tr>
<td>$T_{on}$</td>
<td>Time a Station Remains Active and Sending Voice Data</td>
</tr>
<tr>
<td>$T_{off}$</td>
<td>Time a Station is Idle – No Voice Traffic to Send</td>
</tr>
<tr>
<td>$T_{s}$</td>
<td>Time of a Successful Transmission and Acknowledgment</td>
</tr>
<tr>
<td>$T_{SIFS}$</td>
<td>Time of a SIFS period (used before transmitting acknowledgment)</td>
</tr>
<tr>
<td>$\overline{T}_{slot}$</td>
<td>Mean Slot Time when Backingoff</td>
</tr>
<tr>
<td>$\overline{T}_{slot_n}$</td>
<td>Estimated Mean Slots Time a Station $n$ needs to wait for a given Station $n$</td>
</tr>
<tr>
<td>$u(k)$</td>
<td>An Intermediate Probability used in a Finite Queue case</td>
</tr>
<tr>
<td>$U_d(k)$</td>
<td>Delay caused by the Queue System State</td>
</tr>
<tr>
<td>$W$</td>
<td>The Range of the Minimum Number of Slots a Station needs to Wait – Used to derive $CW_{min}$</td>
</tr>
<tr>
<td>$X_N$</td>
<td>Calibration Factor for a Scenario with a total of $N$ Stations</td>
</tr>
</tbody>
</table>
# Glossary

<table>
<thead>
<tr>
<th>AC</th>
<th>Access Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACK</td>
<td>Acknowledgment</td>
</tr>
<tr>
<td>ADPCM</td>
<td>Adaptive Pulse Code Modulation</td>
</tr>
<tr>
<td>AEDCF</td>
<td>Adaptive Enhanced Distributed Coordination Function</td>
</tr>
<tr>
<td>AIFS</td>
<td>Arbitrary Inter-Frame Space</td>
</tr>
<tr>
<td>AP</td>
<td>Access Point</td>
</tr>
<tr>
<td>BAP</td>
<td>Backhaul Access Point</td>
</tr>
<tr>
<td>CAP</td>
<td>Client Access Point</td>
</tr>
<tr>
<td>C-Pop</td>
<td>City Point of Presence</td>
</tr>
<tr>
<td>CA</td>
<td>Collision Avoidance</td>
</tr>
<tr>
<td>CSMA</td>
<td>Carrier Sense Multiple Access</td>
</tr>
<tr>
<td>CTS</td>
<td>Clear to Send</td>
</tr>
<tr>
<td>CW</td>
<td>Contention Window</td>
</tr>
<tr>
<td>DCF</td>
<td>Distributed Coordination Function</td>
</tr>
<tr>
<td>DHSS</td>
<td>Direct Hopping Spread Spectrum</td>
</tr>
<tr>
<td>DIFS</td>
<td>Distributed Coordination Function Inter-Frame Space</td>
</tr>
<tr>
<td>EDCF</td>
<td>Enhanced Distributed Coordination Function</td>
</tr>
<tr>
<td>FHSS</td>
<td>Frequency Hopping Spread Spectrum</td>
</tr>
<tr>
<td>GAPP</td>
<td>Geometrically Approximated Queueing Model</td>
</tr>
<tr>
<td>HCCA</td>
<td>Hybrid Coordination Function Controlled Channel Access</td>
</tr>
<tr>
<td>HCF</td>
<td>Hybrid Coordination Function</td>
</tr>
<tr>
<td>HECToR</td>
<td>High-End Computing Terascale Resource</td>
</tr>
<tr>
<td>HPC</td>
<td>High Performance Computing</td>
</tr>
<tr>
<td>iLBC</td>
<td>Internet Low Bitrate Codec</td>
</tr>
<tr>
<td>IR</td>
<td>Infrared</td>
</tr>
<tr>
<td>iSAC</td>
<td>Internet Speech Audio Codec</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Full Form</td>
</tr>
<tr>
<td>--------------</td>
<td>-----------</td>
</tr>
<tr>
<td>MAC</td>
<td>Medium Access Control</td>
</tr>
<tr>
<td>MOS</td>
<td>Mean Opinion Score</td>
</tr>
<tr>
<td>NAM</td>
<td>Network Animator</td>
</tr>
<tr>
<td>NS</td>
<td>Network Simulator</td>
</tr>
<tr>
<td>OOC</td>
<td>Out of Contract</td>
</tr>
<tr>
<td>PCF</td>
<td>Point Coordination Function</td>
</tr>
<tr>
<td>PCM</td>
<td>Pulse Code Modulation</td>
</tr>
<tr>
<td>PHY</td>
<td>Physical Layer</td>
</tr>
<tr>
<td>PESQ</td>
<td>Perceptual Evaluation of Speech Quality</td>
</tr>
<tr>
<td>PIFS</td>
<td>Point Coordination Function Inter-Frame Space</td>
</tr>
<tr>
<td>PSTN</td>
<td>Public Switched Telephone Network</td>
</tr>
<tr>
<td>PSQM</td>
<td>Perceptual Speed Quality Measure</td>
</tr>
<tr>
<td>QAP</td>
<td>Quality of Service enabled Access Point</td>
</tr>
<tr>
<td>QoE</td>
<td>Quality of Experience</td>
</tr>
<tr>
<td>QoS</td>
<td>Quality of Service</td>
</tr>
<tr>
<td>RED</td>
<td>Random Early Discard</td>
</tr>
<tr>
<td>RTCP</td>
<td>Real-Time Transport Control Protocol</td>
</tr>
<tr>
<td>RTP</td>
<td>Real-Time Transport Protocol</td>
</tr>
<tr>
<td>RTS</td>
<td>Request to Send</td>
</tr>
<tr>
<td>SIFS</td>
<td>Short Inter-Frame Space</td>
</tr>
<tr>
<td>SLA</td>
<td>Service Level Agreement</td>
</tr>
<tr>
<td>STA</td>
<td>Station</td>
</tr>
<tr>
<td>TxOP</td>
<td>Transmission Opportunity</td>
</tr>
<tr>
<td>VoIP</td>
<td>Voice over Internet Protocol</td>
</tr>
<tr>
<td>VoWLAN</td>
<td>Voice over Wireless Local Area Network</td>
</tr>
<tr>
<td>WLAN</td>
<td>Wireless Local Area Network</td>
</tr>
</tbody>
</table>
1 **INTRODUCTION**

1.1 **Background**

As computers become more common place in the home and office, sharing resources and Internet access locally is a necessity. The simplest method of choice is by setting up a Wireless Local Area Network (WLAN); they are inexpensive, easy to configure and require minimal infrastructure. WLAN was originally designed for mobile devices such as laptops and PDAs but are now the choice for networking solutions for all types of devices including fixed desktop machines.

This increasing trend of Wi-Fi use is also helped by faster Internet access becoming available which allows for real-time applications such as Voice, Audio and Video. An application becoming ever more popular is Internet Telephony; commonly known as ‘VoIP’. Voice over Internet Protocol (VoIP) is a service defined by a set of protocols which allows voice calls to be made via packet switched networks. This technology is becoming more popular because it allows for the integration of voice and data over one network. This arrangement means less infrastructure; the traditional circuit switched networks are sometimes no longer necessary and therefore lower setup and maintenance costs.

Due to the increase of Wi-Fi devices and the use of VoIP, many communication and telephony companies are getting in on the act not just in the home but by providing wireless hot-spots for users to access the Internet away from the home or office. The solutions are all confined to small areas but recently the trend is to create large regions with blanket wireless coverage, for example metropolitan areas such as those seen in Hamburg (Germany), Adelaide (Australia) and Wellington (New Zealand) [VOS07].

Deployment of such networks is a challenge, planning is needed to decide on location and operation of the network. A well thought-out plan of where access points should be placed can greatly increase the number of VoIP calls which can be carried in the network. To evaluate a given plan, simulation must be used which requires a large amounts of computational resource not currently available to most organisations.

To address the network performance issue, an analytical model is needed to quickly and easily find the carrying capacity of the network for a given configuration.

1.2 **Objectives**

The objective of the research is to develop an analytical model to address, from the users’ perspective, the problem of evaluating a wireless.
Current analytical models have assumptions which make them unable to model realistic scenarios. Most WLAN models assume a saturated case scenario in which all stations always have traffic to send; these models are unrealistic as stations usually communicate in a sporadic way, i.e. downloading a web site and not using the network while the users reads, polling email servers every few minutes or making a voice call and not sending data while listening.

The aim here is to increment an existing model [BIA00] to include an un-saturated, heterogeneous traffic case, where different stations have varying amounts of traffic to send, through to a realistic voice traffic model. Furthermore, instead of reporting in medium access delays, the model will report in an acceptable voice call carrying capacity. The carrying capacity relates to Mean Opinion Scores (MOS), allowing the communication of results to telephony companies and industry in general.

1.3 Novelty

- Update the current network analytical model to allow for different traffic loads on each station to be considered.
- Provide an analytical method to model On/Off voice traffic in WLAN.
- Apply queueing and user perception Analytical Models to derive user Quality of Experience; i.e. the users perception of the service.
- Developed a methodology to analyse and compare Quality of Experience and use the methodology to determine carrying capacity.
- Ability to validate the analytical models with simulation, with the use of two national super computing services.
- Optimise voice in WLAN from the perspective of QoE and increase carrying capacity.
- Model large scale WLAN scenarios in the order of 50 Access Points and 200 Stations.

1.4 Thesis Outline

An introduction to the IEEE 802.11 standard is presented in Chapter 2 and details how the channel is shared between many clients and the process by which a client gains access to the channel. Chapter 3 reviews the methods in which the protocol has been modelled and an existing study, which is the most referenced paper in the IEEE 802.11 research area, highlighted. A dimensioning investigation, with the aim of increasing call capacity is also shown. The existing analytical study, shown in Chapter 4, assumes all stations are saturation, that is, all stations have packets to send. The existing analytical model determines medium access delay. It is shown that the model is not sufficient to gain
realistic results due to the saturated assumptions detailed and that an incremental advancement is needed. Chapter 5 details an augmented model which more closely represents a realistic scenario allowing for a range of medium access times to be found from a range of traffic loads. The augmented model highlights that the network can achieve a higher throughput when unsaturated and there is a substantial increase in service time when hitting saturation. Chapter 6 details a further augmentation in which each station can offer a different traffic load and gives different medium access delays for each station. The latter model allows for modelling of an Access Point.

From the combined amendments and augmentations to the original model, which are validated against simulations performed on super computers, a medium access delay per client can now be gained; however, this is not sufficient to quantify call quality. Chapter 7 details the effects of On/Off bursty traffic and a further analytical model is developed. In Chapter 8, a queuing model is used, under certain conditions to give network performance metrics. The analytical model is finally closed, in Chapter 9, with the use of the ITU-T G.107 E-Model which gives users’ perceived Quality of Experience (QoE) measures.

Chapter 9 then shows how the combined analytical model can be used, along with simulation methodology. The new QoE analysis framework is used to benchmark the network and then goes on to show the power of the models in “what-if” evaluations, without the need for super-computers and with a 13,000 to 1 speedup.

The first study “what-if” study explores a method which aims to keep the network out of saturation and hence keep operating in the higher throughput case. The second scenario shows dimensioning of the network in order to investigate effects of adjusting the packetisation interval of a voice codec. In both cases, the value and tradeoffs of each are discussed. Conclusions are drawn in Chapter 10 and thoughts on future work outlined in Section 10.1.
1.4.1. *Diagrammatic Overview*

![Diagram](image)

*Figure 1.1: Research Overview*

- **Saturated Case**
  All stations offering infinite load.

- **Unsaturated Model**  
  [Pitts and Shepherd in PIT08]
  Varying load, showing the transition between unsaturated and saturated model, although equal offered traffic across all stations.

- **Heterogeneous Load**  
  *Results Shown [Shepherd in HER10]*
  Uneven load across stations; each can have a different load and can be used to model an access point and stations.

- **On-Off Traffic**  
  [Shepherd in HER10]
  Considering VoIP traffic sources; stations being in burst state or idle state.

- **Queueing Model**  
  [Shepherd in HER10]
  Taking medium delay (from above models) and finding end-to-end delay, drops from a finite queue and delay variation.
- **ITU G.107 E-Model**
  Taking performance metrics from the Queueing Model to evaluate VoIP call quality from Users Perspective (Quality of Experience).

- **QoE Framework**
  Contour plot of proportion of calls exceeding QoE criterion value for evaluating large numbers of calls.

- **Dimensioning of VoIP (Packetisation Period)**
  Exploring the packetisation period of a voice codec, and hence packet size and packet per second rate, to increase the acceptable call carrying capacity of the network with the use of the QoE Framework.

- **Dimensioning a Network with an of AP (IEEE 802.11e)**
  Highlighting that the Access Point is the bottleneck when servicing bi-directional voice calls, an exploration of contention windows parameters is explored. An increase in capacity is found by giving the AP priority on the network with different contention window parameters. Optimisation is also confirmed with the addition of ‘Best-Effort’ traffic on the network.

  *This was a proof of concept study for a paper and therefore referred to in future work*
2 BACKGROUND

2.1 The IEEE 802.11 Protocol

IEEE 802.11 is a set of standardised protocols for wireless local area networks (WLANs) which specifies a physical (PHY) and medium access control (MAC) layer [P802]. Originally, the PHY layer had three implementations; Infrared (IR), Frequency Hopping Spread Spectrum (FHSS) and Direct Sequence Spread Spectrum (DSSS). DSSS, which supports 1-2Mbps, was adapted in 1999 to produce higher theoretical data rates; IEEE 802.11a at 54Mbps in the 5Ghz range and IEEE 802.11b at 11Mbps in the 2.4Ghz range. It was then the providers’ choice whether to implement a higher data rate or further range solution (lower frequencies travel further). In 2003, the IEEE 802.11g amendment was published which enables 54Mbps in the lower frequency band and is compatible with IEEE 802.11b devices; this combines both of the advantages of IEEE 802.11a and IEEE 802.11b and is therefore the most commonly chosen method for new network implementations.

IEEE 802.11 specify two node types; client Stations (STAs) and Access Points (APs). STAs can be any device with 'Wi-Fi Capability' such as laptops, personal digital assistants, mobile phones or even fixed desktop machines. An AP normally has access to shared resources, such as local network services or the Internet and has additional network management functions to achieve this, such as; broadcasting its presence, assigning unique addresses to each station, access control and encryption. In infrastructure mode, STAs can only communicate with an AP and therefore all inbound and outbound traffic passes through it (whether going to a neighbouring STA or not). In ad-hoc mode however, the STAs can communicate directly which is sometimes more desirable, depending on the final destination of traffic. In ad-hoc mode however, STAs in range of others, on the same frequency band, will be in contention and therefore there may be little or no overall throughput gain with this approach.

The remainder of this chapter details the underlying protocol for communication. As described, an AP is essentially a special STA with an additional network administration role and hence, when considering the wireless communication aspect behaves as if it were an STA; the term STA in this chapter is now used to represent either an AP or STA.

The MAC layer provides methods for control and access of the shared medium. Distributed Coordination Function (DCF) and Point Coordination Function (PCF) were standardised but PCF, an optional protocol which assigns time intervals for contending STAs to access the medium, similar to Time Divisional Multiple Access (TDMA), is not commonly implemented because of its added complexity and need for centralised control.
DCF is on the principle of a carrier sense, multiple access with collision avoidance (CSMA/CA) scheme. When an STA has data to send, it must first check that the channel is idle; this is done by listening to the shared medium and then waiting for a period called the DCF inter-frame space (DIFS). This DIFS time ensures another STA isn’t already in communication, e.g. between sending packets of data. If two stations do transmit at the same time, the data is corrupt and becomes unusable; known as a collision.

To detect collisions, the IEEE 802.11 protocol uses a positive acknowledgment (ACK) mechanism which means each STA must acknowledge a data packet on successful delivery. If no ACK is received, an ACK timer expires and the sending STA will assume a collision. In legacy IEEE 802.11 there will only ever be one ACK waiting to be sent by the last receiving STA; this is because ACKs have a higher priority of accessing the channel. Recall that an STA must only attempt to access the medium after it has been sensed idle for DIFS, plus a backoff time. For an ACK packet, the medium must again be detected idle but for a period called the short inter-frame space (SIFS). As the name implies, SIFS is shorter than DIFS therefore ACK packets are always sent before the next STA obtains the medium or group of STAs contend. Figure 2.1 shows the process for sending a data packet assuming the channel is free.

![Figure 2.1 : System Diagram : Successful Transmission](image)

If the channel is busy, or a collision is detected, the STA must enter into a random exponential backoff mode; this defines how many time slots to wait before accessing the medium and sending data. As all stations choose a slot interval randomly, the probability of two stations starting to transmit at the same time is reduced. The random backoff time is defined as

\[
timeSlots = \text{rand} [0 \rightarrow CW]
\]  

(2.1)

Figure 2.2 shows a scenario where the medium is detected idle by two STAs which then transmit at the same time. After an ACK timer expires, both STAs choose a random number of time slot intervals to wait before re-transmitting.
The random selection of time slots to backoff will decrease the probability that two STAs transmit at the same time; however, it is possible that the STAs choose the same number of time slots, and hence experience another collision. The range of random numbers is therefore increased to further decrease collision probability after subsequent collisions.

To decrease collision probability while keeping backoff times initially short, CW is set to an initial value $CW_{\min}$ and doubled with each collision up to $CW_{\max}$. These parameters are usually defined by $W$, the initial window size and $m$, the maximum number of times the window can be doubled. Therefore,

$$CW_{\min} = W - 1$$  \hspace{1cm} (2.2)

and

$$CW_{\max} = W \cdot 2^m - 1$$  \hspace{1cm} (2.3)

Figure 2.3 shows a scenario in which two STAs experience a collision and then go on to choose a random number of time slots from a bigger range of possible numbers, hence decreasing the probability of subsequent collisions.

The protocol allows for up to 7 retransmissions of a single packet before discarding it; however, the number of times to increase the contention window, $m$, may be less. In the case where $m$ is less than 7 the backoff process will remain at $CW_{\max}$ after $m$ collisions.
During backoff, if the channel is ever perceived busy, the STA must pause its backoff timer until detecting the medium idle for a DIFS period. When the medium is again detected idle, the STA may decrement one time slot and continue. This pausing behaviour means the duration of a single time slot can vary. If the medium is free, a time slot is the standard defined protocol time, usually 50µs, on the other hand, if the medium is busy it could be due to a station transmitting successfully or transmitting resulting in a collision. If a station hears another station transmitting successfully, the slot time includes the time to transmit a packet and the corresponding acknowledgment, as shown in Figure 2.4. If a station hears two other STAs transmitting, results in a collision, the slot time will only be the time to transmit the packet, as shown in Figure 2.5. Figure 2.4 demonstrates pausing the timer of STA 2 while STA 1 transmits successfully and Figure 2.5 where STA 3 is backing off and hears the medium busy caused by STA 1 and STA 2 transmitting and experiencing a collision.

Finally, after a successfully acknowledged packet, if there is another packet immediately ready to send, i.e. waiting in the buffer, the STA must immediately enter into backoff. By entering in backoff immediately, the STA avoids capturing the channel; i.e. constantly transmitting while other stations remain in backoff. This immediate backoff behaviour after successful transmission means that when packets are arriving quicker than the time to service a packet (backoff, successfully send and acknowledged) an STA is always backing off, regardless of the medium state, and hence known as the saturated behaviour of operation. Figure 2.6 shows that, even with only one STA, if a packet arrives while
servicing the previous packet, the STA must backoff even after hearing the medium free for a DIFS period.

![Figure 2.6: System Diagram: Saturated Condition](image)

The methodology described thus far is called *basic mode*, however, there are two issues remaining in the wireless domain known as the 'hidden node problem' and the 'exposed node problem.' The hidden node problem can best be described by imagining a scenario where there is 1 AP and 2 STAs as shown in Figure 2.7.

![Figure 2.7: Hidden Node Problem](image)

Each of the STAs are at opposite edges of the AP’s transmission range and this means that each STA cannot hear the other. Having STA which cannot hear each other can cause problems because the CSMA/CA mechanism requires all stations being able to hear all others. If one of the STAs, the hidden node, sends data the other STA will still perceive the medium as free which could result in it sending and therefore causing a collision. The other issue, the exposed node problem, as shown in Figure 2.8, is where there are two STAs wishing to transmit to two different APs, where the STAs are in range of each other but the APs are out of range of the other and the other transmitting STA. In the exposed node case, the second STA will not transmit according to CSMA/CA but in fact, in this specific scenario, can.

![Figure 2.8: Exposed Node Problem](image)
Both of these issues are addressed in optional handshaking extension of the CSMA/CA called the Request to Send, Clear to Send (RTS/CTS) mechanism. This mechanism essentially allows an STA to request access to send data from the AP. A RTS message is sent out and when the AP decides it is the requesting node’s turn, it will send a CTS message back with parameters which define the duration it has access to the medium. As all stations can hear the AP, they will also receive the CTS destined for the requesting node and know not to access the medium (send an RTS) before the set duration. This solves the hidden node problem because all stations will hear the CTS and, therefore, not access the medium even if it is perceived free. The exposed node issue is also solved because, in the scenario above, each STA can send an RTS to the specific AP and hear a CTS reply but not hear each other’s bound CTS and not be forced to wait.

RTS/CTS is based on a packet payload threshold size, where it is used for larger size packets; it causes a high overhead which is significant at smaller packet sizes. RTS/CTS, therefore is typically not implemented or not used (threshold greater than maximum packet size) because of this overhead and the new issue of contention at the RTS level.

The challenge, highlighted in this dissertation, is capturing the protocol behaviour in an analytical model which can be used to analyse voice traffic in multi-scale networks. The detailed understanding will then be exploited, with the model, to increase voice capacity in multi-scale IEEE 802.11 networks.

## 2.2 Voice over Internet Protocol

### 2.2.1 Overview

VoIP is a service and set of protocols which allow for telephony communications over a packet switched network. Unlike the Public Switched Telephone Network (PSTN) that sets up a circuit through the network, packet switched networks send packets of data independently. The packets are then forwarded by autonomous routers, which decide which of its connections the packet should be sent in order to reach the intended destination. As each packet is independent, packets experience different delays, arrive out of order and some packets may even be lost. The Internet is therefore knows as a ‘best-effort’ network.

VoIP requires standards and protocols to define how the user’s voice has been digitised, which involves sampling, and how the end points communicate. The sampling and digitisation is called encoding and decoding which results in different voice quality and network resource. In this section, some common frameworks, audio related standards and the codecs are investigated. Furthermore, this section explores the methods to determine how the users’ perceptual quality varies with the standards used and the effects of the ‘best-effort’ network.
2.2.2. Audio related standards

The ITU defines frameworks for voice communication in for both analogue circuit switched and digital packet switched communications. ITU-T H.324 and ITU-T H.323 are the two commonly used frameworks for real-time traffic for analogue and digital transmission, respectively. Both frameworks contain standards for signalling, control and multimedia transport; i.e. how the real-time data is encoded and transported over the network. The analogue framework, ITU-T H.324 encodes audio using ITU-T G.723.1 and uses a 33,600 bps modem for transmission.

The ITU-T H.323 framework, for digital communications, is divided into three main components; Data Applications, Media Control and Terminal Control and Management. The Media Control component contains recommendations for encoding and decoding audio and video as well as the transport and control protocol for transmitting the real-time multimedia content. The Terminal Control and Management component contains the overall control and signalling protocols for how the end devices should behave and communicate.

ITU-T H.225.0 defines a signalling protocol within the Terminal Control and Management component of the ITU-T H.323 framework. ITU-T H.225.0 contains methods for the end points to communicate, for example to; setup the call and terminate a call, send each other alerts, progress, notifications and status updates.

Also within the Terminal Control and Management component, ITU-T H.245 provides specific call control once a call has been setup by ITU-T H.225.0. Call Control could include setting up audio and video streams (e.g. beginning a video session during a call) as well as allowing conference calls over unicast or multicast (a network technique to allow a device to send the same data to many devices at once as opposed to one-by-one).

The multimedia content, i.e. the user’s voice, needs to be converted to a digital signal, known as coding, in order to transmit on the network. At the receiver, the encoded digital voice stream needs to be decoded to play back the speech. The coding/encoding technique (codec) can be achieved in many ways which can affect that quality of the received signal, as the encoding technique involves taking samples at a certain rate, and therefore have different bandwidth requirements. One such common codec is ITU-T G.711, known as Pulse Code Modulation (PCM), which encodes at 8000 samples per second, where each sample is 8 bits long. The ITU-T G.711 codec therefore produces 64kb of data per second.

There are also Adaptive Pulse Code Modulation (ADPCM) codecs, such as ITU-T G.723 which can sample at 300 to 3,400 times per second producing 24 or 40 kbps streams. ITU-T G.723 was superseded by ITU-T G.726 which can produce 4 bit rates; 16, 24, 32
and 40 kbps. ITU-T G.723 typically operates at 32 kbps, half the rate of ITU-T G.711 for international communications.

ITU-T G.729 samples speech to allow for 10 milliseconds of voice to be encoded into a single packet. ITU-T G.729 can operate at 8kbps and, with extensions, 11.8kbps or 6.4kbps. ITU-T G.729 is therefore used when bandwidth must be minimised and therefore not recommended for high quality audio.

Within the Media Control component of the ITU-T H.323 framework, a Real-Time Transport Protocol (RTP) defines the packet format that the encoded multimedia data is sent. Real-Time Transport Control Protocol (RTCP) provides control information for RTP and communicates statistics about call quality. The RTCP statistics could be used to change the way packets are transmitted in order to improve the multimedia quality.

2.2.3. Proprietary codecs

As well as the ITU-T recommendations, proprietary codecs exists such as Internet Speech Audio Codec (iSAC) and Internet Low Bitrate Codec (iLBC). iSAC samples at 16,000 times per second and produces a variable bit rate of 10 kbps to 32 kbps. iLBC is a narrowband codec where the bit rate can be 15.2 kbps to 13.33 kbps. Both iSAC and iLBC were created by Global IP Solutions. iLBC requires a licence to use, whereas, iLBC is royalty-free.

2.3 Perceptual Quality

2.3.1. Overview

Perceptual Quality is about determining, or estimating depending on technique, how a user perceives a service. Perception is subjective and can include social and economic issues and is therefore unrepeatable. Deriving Perceptual Quality for a test recording could result in a different score even with the same person. The first approaches used many people in order to derive scores and take an average but later techniques involve more scientific algorithmic techniques. These more scientific techniques are therefore repeatable and therefore easier to benchmark and improve.

2.3.2. P.800 Mean Opinion Score

One method of obtaining perceptual quality scores is by having a number of listeners rate the quality of a test sentence which has been encoded, transmitted and decoded according to the system being tested. ITU-T P.800 defines a Quality Impairment score from 1 to 5, known as; Excellent, Good, Fair, Poor and Bad. Taking the mean of the individual listener’s score gives the Mean Opinion Score (MOS). MOS therefore also
captures the subjectiveness of users; i.e. that the same impairment can be rated differently by a group of individuals for social and economic reasons.

2.3.3. **ITU-T P.861 Perceptual Speech Quality Measure**

The problem with the MOS is that it requires many listeners who score the quality of a voice signal subjectively. Other ethical issues, time and cost also play a factor in aiming to get away from the requirement of a large group of people. The ITU-T began developing and recommending algorithms which allow for simulation based testing; i.e. a proxy to how users’ would rate calls. Once such recommendation is the ITU-T P.861 Perceptual Speed Quality Measure (PSQM). PSQM also allows for reproducible perception scores; important for measuring improvements. PSQM however is based on voice quality alone and not issues such as network delay, loss and jitter; the PSQM scores were higher when compared back to MOS scores under heavy network loads. PSQM was later modified to address the scores when operating in heavily loaded networks. The modification to PSQM, known as PSQM+, was developed while analysing PSQM performance under realistic network load conditions.

2.3.4. **ITU-T P.862 Perceptual Evaluation of Speech Quality**

ITU-T P.862 Perceptual Evaluation of Speech Quality (PESQ) superseded PSQM, which only dealt with narrowband signal at 3.4KHz, to wideband 8kHz. Streaming media however may use the enhancement of PSQM, PSQM+. Another advantage of PESQ is that it can use a Full Reference testing typology which allows comparison to the original reference signal and hence can determine the exact extent of impairment. PESQ has been constantly updated to support new voice codecs. PESQ, like PSQM require a tests setup call. ITU P.563, on the other hand, can be used passively and therefore more representative of the Mean Opinion Score, i.e. not affecting network behaviour where active measurement may.

2.3.5. **ITU-T G.107 E-Model**

MOS, PQSM and PESQ all deal with the received signal and a rating given as to the quality, either unreferenced or compared to the original signal. These perception frameworks however have no way of mapping between network metrics, such as packet delay and loss, to a perceptual rating. The ITU-T G.107 E-Model, works differently to other perception frameworks in that, from a given network performance, a perception score can be derived. The E-Models perception score is known as the R-Factor, which rates a call between 0 and 100, where the higher the number the better the users perceived experience. The network performance is given as parameters to the model which can include; delay, loss and jitter. Delay is the time the signal takes from senders mouth to receivers ear, which includes sampling, packetising, sending, buffering at the receiving end and play back. Loss deals with packets not played back because they are dropped in
the network. Jitter is the delay variance, caused by different packets not taking the same
time to arrive at the receiver. To combat delay variance, the receiving end usually uses a
de-jitter buffer to temporarily store packets, order correctly and play back at a constant
rate. The de-jitter process however introduces more delay and packets may still arrive
too early to be buffered or too late; resulting in additional loss.
3 LITERATURE SURVEY

3.1 Modelling the MAC Layer

3.1.1 Initial MAC Layer Models

In 1996, the same year as the IEEE standardised 802.11, Giuseppe Bianchi published an IEEE letter which details an analytical model [BIA96] in order to derive throughput for a given number of wireless stations. The analysis was based on a Markov model which allowed throughput to be modelled as a function of any number of stations, n, with given contention window parameters (\(CW_{min}, CW_{max}\)). The model was based on collision probability, \(p\), of two stations sending at the same time and therefore advancing through the Markov backoff chain with \(m+1\) states, as shown in Figure 3.1. The model worked by calculating the mean number of slots in each backoff stage, shown above each state in the figure, by multiplying the probability of being in each state with the mean time it takes to decrease the backoff timer by 1.

This model, however, assumes the network is saturated; this is defined as all stations having at least one packet to send, and therefore all always contending. In the saturated case, when a station has successfully transmitted a packet and received an ACK it must return to backoff when the queue is non-empty. The returning to backoff is to avoid a single station capturing the channel and hence attempting to keep the channel access fair among all contenders. In 2000, Bianchi re-published in a Journal; “Performance analysis of the IEEE 802.11 distributed coordination function” [BIA00] and this is now widely used in the wireless. The Bianchi model is therefore the first step in understanding the actual characteristics and then looking into more detailed analysis based on the model.

The Bianchi Model [BIA98, BIA00] specifies a wireless network scenario with N STAs, in which each node has traffic modelled using the Poisson distribution (although at saturation this is irrelevant). In this scenario all STAs use a Tail-Drop queue and are in
saturation, this is defined by saying all stations are always contending for the medium, i.e. all the queues are non-empty. The model proposed in this dissertation is fundamentally based on the Bianchi model and therefore this is explored and derived in more detail in Chapter 4.

### 3.1.2. Un-Saturated MAC Layer Models

In recent years, there have been updates to the Bianchi model to include ‘post-DIFS’ analysis [ZIO02, FOH05]. In the original model, the assumption is that each station is always contending. However, according to the protocol specifications, a station may only begin sending a packet after a DIFS period and, therefore, in these papers, the DIFS period is added after a successful packet transmission. Further updates include more detailed analysis to take account of the fact that, after a collision, only contending stations (the ones that experience the collision) may re-contend [VU06]; this is because all other stations would have stopped their timer. This new analysis is achieved by adding in a channel state function. In 2007, analysis was shown to calculate access delay and standard deviation, it was concluded that the MAC DCF function is prone to long delays and not suited to carry delay sensitive applications [SAK07].

The current review, however, has only included the saturated analysis and an assumption made that this is the maximum throughput achievable. Later an unsaturated Markov chain is formulated to calculate access time delay [ZHA05]; however it is shown versus collision probability and therefore the specific load where the transition between modes cannot be found. A later paper [CHE06], uses a queuing model and, therefore, determines if the network is saturated by the probability of the queue being non-empty; however, this method still uses the backoff analysis from Bianchi and therefore does not consider the fact that a packet arriving when the queue is empty can attempt to transmit immediately after a DIFS period, if the medium is sensed idle and the station is not backing off.

In Xu et al [XU06] this inaccuracy is addressed by an additional state in the Markov model, as shown in Figure 3.2, which represents packets arriving at an empty queue, which is again based on buffer utilisation. This model, however, assumes that the station will always return to an idle state regardless of the queue state and hence breaks the original definition of saturation; finding the queue non empty after a successful transmission requires back-off.
Another line of investigation is pursued by Duffy et al in three successive papers. The authors investigate the saturated model and incrementally provide estimations for the unsaturated case.

The first paper [DUF05a] modifies the saturated model into the unsaturated case by modelling the queue, again with the objective of finding the probability of the queue being non-empty after a successful transmission and hence saturated. The assumption of the model is that the queue can buffer a maximum of one packet and the arrival rate is one packet per state in the Markov model. This utilisation assumption therefore assumes that, after a single collision, a packet has arrived and hence the station will be in saturation. Also, the model assumes that the time for a successful transmission, \( T_s \), is equal to the time of a collision, \( T_c \), which is incorrect due to the considerations of the ACK timeout. The latter assumptions can be easily modified in the model; however, the utilisation based on a queue size of a single packet means the paper cannot claim an accurate model, but a close approximation. The main outcome of the paper, and the rest in the series, is identifying that throughput can be greater in the unsaturated case and therefore opens up a new area of improving performance of using real-time services if saturation can be avoided.

The authors of [DUF05a] modify the model in [DUF05b], still using one packet arrival per state duration, but use feedback from the model to calculate a more accurate utilisation. To achieve this, traffic arrivals are compared to the expected mean state duration time. They find this is still inaccurate due to the final state (a successful transmission) taking the additional time of transmission, \( T_s \), and hence delay results are underestimated.

In the third paper, [MAL07], the authors attempt to remodel the non-empty queue probability more accurately without the buffer limit of one, but determine that due to the
same assumptions in Duffy’s second model, [DUF05b], as the maximum queue size increases so does the variability between the model and simulation.

More recently, Ghaboosi et al [GHA07, GHA08] again focus on the main issue of modelling the unsaturated case; the queue state. In their model, they add a further dimension to the back-off state machine to include the queue state (number of packets buffered). This more complex model, known as Space-Time, is argued to be a better approximation of the behaviour of the queue and hence probability of being in saturation; the service time is derived from calculating utilisation from the number of arrivals per transmission.

A paper which comes closest to MAC behaviour in the unsaturated case is presented by Dao et al [DAO08]. Dao et al again start with the original Bianchi model, add in an extra state to represent the queue being empty to derive utilisation to find the queue state probability. The utilisation is based on service time, as output by the model; this therefore requires iteration to determine a steady state but finds utilisation exactly. In either the saturated or unsaturated case, the model returns to a ‘post-back-off’ phase after a successful transmission, it is after this that the queue state is considered. This phase, however, could be modelled more accurately if the medium idle probability, combined with the queue state probability, was used directly to determine if the station returns to an idle state or immediate back-off.

The key to modelling the MAC layer is to develop a model that allows for stations to be either saturated or unsaturated in the same scenario and derived results from an accurate utilisation based on service time; in other words, different loads per station. Later in this dissertation, results are presented that demonstrate how large the discrepancy between unsaturated and saturated behaviours can become.

3.1.3. Heterogeneous MAC Layer Models

There have been very few papers modelling the IEEE 802.11 MAC layer in a heterogeneous case, perhaps due the lack of an unsaturated case analysis that shows the transition between the unsaturated and saturated case.

In a paper which addresses the heterogeneous case, [BEL05], the authors state that transmission probability, $\tau$, is not dependent on the other stations. The paper highlights the imprecise independent transmission probability, however, report results with this assumption due to the complexity of modelling the dependencies among nodes.

A deviation to the protocol arises when the paper states that the channel is sensed to be idle for a single slot duration before transmission, where in fact the protocol states it is a DIFS time. This slot time assumption means that, as collision probability is conditioned to one slot, the collision probability will be reduced. When modelling, however, it is
necessary to find the mean slot time based on slots where the station perceives the medium idle, medium busy with a successful transmission or medium being busy with a collision. Finding the probabilities of these medium conditions, multiplied by the mean time of each event, the DIFS time can be normalised to slot time; hence finding how many slot times DIFS is equivalent to.

The paper uses a simple adaptive traffic model, which uses the remaining bandwidth available in the system. By using the remaining bandwidth, it is challenge to investigate the fairness of the protocol and provide intuition as to when one station achieves a better service time compared to another.

As explained in Section 3.1.2, Duffy et al in 3 successive papers ([DUF05a], [DUF05b] and [MAL07]) explore the un-saturated characteristic of the MAC layer. In the latter papers, the authors also begin to investigate the heterogeneous traffic case. To achieve heterogeneity, the authors use interacting state diagrams, in which each station now has its own Markov state machine. In this interacting state machine analysis, each station must now have a different collision probability, transmission probability and utilisation.

The results in Duffy’s final paper of the series above, [MAL07], show that the delay estimation is low when compared to simulation due to the mean backoff slot times used. In reality, each backoff slot will be one of three different durations dependent on medium state. If the medium state is idle, the default slot time is used, on the other hand, if the medium is busy, it could be the time for a successful transmission or the time for a collision in the medium.

The results also show that if the buffer capacity is above unity, simulations again deviate from the model due to the simplistic single packet buffer queueing model used.

3.2 Queue Modelling

There have been papers published which highlight the issue of only deriving MAC layer performance. These papers update the model to include impacts of queueing traffic. The interesting part of these papers is that, as has been shown previously, only an accurate saturated case model has been presented in the literature and any queueing theory would need to derive the parameters based on an un-saturated case. Some papers deal with the PCF side of the MAC protocol such as “Queueing Analysis and Delay Mitigation in the Access Point of VoWLAN” [WAN05b] and “A Finite Queueing Model for IEEE 802.11 MAC Protocol” [ZEN04]. PCF however, is a polling mode protocol and therefore the factors which make this an interesting study, namely the service time characteristics as a function of contention, packet size and number of nodes sharing the medium is not explored.
“Queueing Analysis and Delay Mitigation in IEEE 802.11 Random Access MAC based Wireless Networks” [TIC04a] presents a generic traffic and service time model (G/G/1) for evaluating WLAN and determines mean packet delays based on calculating collision probabilities for an arbitrary number of stations. The results presented are similar to simulation; however, considerations such as queue utilisation and whether the queue is full, used to determine drops, are not considered; this infinity model therefore shows delay tending to infinity. The authors then amend the model in “A Queueing Model for Finite Load IEEE 802.11 Random Access MAC” [TIC04b] to take care of this assumption and also provide a closed form solution for queue length. The results from the model compared with simulation show an increased deviation as the number of stations increases. The discrepancy is due to the underlying saturated model used to derive the MAC service time distribution.

Essentially, these queueing analysis papers use either the saturated MAC model or derive an assumed un-saturated case. However, in either case, this inaccuracy affects the results when applying the queueing analysis. One of the directions in this dissertation is, therefore, to combine the un-saturated heterogeneous case model with a new queueing model and determine if this provides more accurate results in terms of delay, delay variance, queue utilisation and packet loss.

3.3 Optimising Network Performance

As stated, as an aside to the analytical model, a higher network throughput was found in the unsaturated case. The aim in this dissertation is to use this fact and show some simple examples of how this problem can be addressed using the analytical model and simulation.

Firstly, a review is carried out to find the feasibility of exploiting the higher throughput before hitting saturation to increase carrying capacity. A number of papers have been published attempting to increase the throughput which can generally be classified into two approach categories; STA-based and queue-based [NI04]. Most of the approaches are designed as a Quality of Service (QoS) approach because, fundamentally, IEEE 802.11 does not support high volumes of traffic, or have any priority mechanism for time critical data. These QoS algorithms and techniques are therefore explored to identify schemes that may assist with congestion avoidance.

Some examples of increasing the carrying capacity in the literature are the Diff-Serv-Like [AAD01], Distributed Fair Scheduling [VAI05], Virtual MAC [VER01] and the Black Burst Scheme [SOB96]. The IEEE subsequently set up a task group (802.11E) with the aim of providing QoS extensions to the legacy IEEE 802.11 framework; this was largely an analysis of the mentioned schemes and deciding which techniques would be feasible in generalised cases. Newer approaches are now based on the later standardised IEEE 802.11e [P802E], such as the Adaptive EDCF Scheme (AEDCF) [ROM03] (Enhanced
Distributed Coordination Function being a new concept from IEEE 802.11e) and others [NI05, BIA05].

The above mentioned schemes are outlined below.

### 3.3.1. Diff-Serv-Like Scheme

The Diff-Serv-Like Scheme [AAD01] specifies three main changes to the MAC layer. The first is the concept of a different backoff function where the new value is a factor of the old; however, the weight used is changed for different traffic priorities. This essentially means each traffic class has its own backoff timer and the higher priority traffic will gain access to the medium before lower priority traffic within each STA. The second method is the idea of setting a DIFS time for each traffic class; this then provides priority across stations so a node with higher priority can access the medium first. The final amendment is providing transmission frame lengths to each STA where a number of packets can be sent within a time interval. The methodology proposed, allows a stations with time critical traffic to have higher priority by using a larger frame length to allow many packets to be sent at once. The methods outlined are essentially the main emphasis of the IEEE 802.11e extension and, therefore, more detailed explanations and analysis can be found below.

### 3.3.2. Distributed Fair Scheduling

The Distributed Fair Scheduling scheme [VAI05] modifies the backoff parameters but, this time, as a function of packet size; smaller packet sizes will be sent more often and so traffic such as VoIP, which usually has smaller packet sizes on average, will be sent more often.

The Distributed Fair Scheduling scheme allows VoIP packets to be sent with higher priority than background traffic only if the background traffic occupied larger packet sizes. An issue, however, that if only VoIP calls are being made the performance will be the same as legacy MAC and at saturation would be equivalent to the varying packet graph shown in Chapter 4, Figure 4.7.

### 3.3.3. Virtual MAC

The Virtual MAC Scheme [VER01] can be viewed as an overlaying MAC layer across all stations with the sole purpose of gathering network performance statistics. By estimating such parameters as delay, jitter and loss the Virtual MAC can then inform higher levels in the network stack of expected performance, which may initiate dynamic modifying of traffic size/weight where implemented. This scheme then goes on to differentiate service levels by stating that
$CW_{\text{min \_ pri}}^\text{high} < CW_{\text{min \_ pri}}^\text{low}, CW_{\text{max \_ pri}}^\text{high} < CW_{\text{max \_ pri}}^\text{low}$  \hspace{1cm} (3.1)

Essentially, each priority has its own contention window function and parameters and can be seen as each physical STA operating as two virtual STAs. The scheme also dynamically changes $CW_{\text{min \_ pri}}^\text{low}$ and $CW_{\text{min \_ pri}}^\text{high}$ dependent on the network statistics, but within the confines of the above rule in Equation 3.1.

### 3.3.4. Black Burst

A different approach to the problem is taken by the Black Burst Scheme [SOB96]; this fundamentally uses channel blocking techniques. Essentially, when the stations arrives in a contention period they all 'fight' for the medium by blocking it, this blocking time interval is a function of the time since the station received the packet. The STA which has waited for the longest time would have the longest pulse and therefore, when it came to the end of blocking, would detect the medium as free and send its data. Each STA, however, is also constrained to access the medium at constant intervals, which the paper claims in some cases, synchronises the nodes and therefore increases channel utilisation. On the other hand this behaviour may cause a degradation in performance when the traffic inbound exceeds the rate at which the STA has access, which becomes more critical as the number of STAs increases.

### 3.3.5. IEEE 802.11e

In the new standardised QoS Enhancements for IEEE 802.11 (802.11e [P802E]), the above schemes were analysed and a combination merged, modified and standardised. IEEE 802.11e APs are known as QoS-enhanced APs (QAP) and similarly STAs are referred to as QSTAs. The basic idea is to mark packets at the sender with a priority level which is then fed into one of a number of virtual queues known as access categories (AC). Background traffic, such as TCP, is given the lowest priority whereas real-time critical packets are placed in a higher priority AC. It is down to changes in the MAC layer as to how these AC are processed; a new Hybrid Coordination Function (HCF) was developed.

Essentially, the HCF is the method used for prioritising the ACs relation to one another within the QSTA and across the medium using Enhanced Distributed Coordination Function (EDCF). In EDCF, access to the medium had been prioritised to stations with higher precedence of data access. This is done by replacing DIFS with the Arbitration Inter-Frame Space (AIFS); this is a variable time, still greater than SIFS, that is set for each AC. The HCF also allows a QSTA to send data for a specific time frame instead of just a single packet by using a Transmission Opportunity (TXOP) time. To complement TXOP, a Block ACK system has been developed that simply allows acknowledgments for all the frames sent in that instance. The TXOP feature, however, requires more
control in the medium to issue these times and this is achieved by the HCF Controlled Channel Access (HCCA). Similar to the PCF in legacy IEEE 802.11, the HCCA essentially means the QAP is responsible for issuing these times and now has a beacon protocol to do so. The new IEEE 802.11e beacon frame, which can assign TXOPs for multiple stations, uses the PCF Inter-Frame Space (PIFS), which is smaller than all of the AIFS in the network. As a result, beacons are sent with priority.

However, each AC also has a TXOP limit and, therefore, the STA may have to share time with other ACs. For additional priority, the backoff window mentioned earlier is now provided for each AC. In total, this means that for four ACs, there are a total of 16 variables to be set up at each station; $CW_{\text{min}}$, $CW_{\text{max}}$, AIFS and TXOP limit.

IEEE 802.11e provides a standard generic set of parameters. However, a simple scenario [NI04] comprising of Voice, Video and Background traffic shows that in fact the standard parameters does not give good performance compared with no QoS. Furthermore, when the number of stations is greater than 10, the voice and background traffic delay shoots up from $<10\text{ms}$ to $500\text{ms}$ and cumulative loss hits 100% in the space of two additional QSTAs.

The paper ([NI04]) does then go on to show a better set of parameters to deal with the higher number of stations; however, the parameters have been developed for the specific traffic models. The scenario the authors are using means that, although traffic can be prioritised, the issue of congestion control is still present. At a certain point, either because of saturation or by hitting the ‘limit’ imposed by the selected parameters, there is an issue with how the overall network performs.

3.3.6. Adaptive EDCF

Due to the static parameters imposed by IEEE 802.11e, the Adaptive EDCF (AEDCF) model [ROM03] essentially replaces the $CW_{\text{max}}$ with a dynamic $CW_{\text{new}}$, for each class of traffic, $i$. This more dynamic approach is based on the old collision window value, $CW_{\text{old}}$, and estimated collision rate at each STA as follows:

$$CW_{\text{new}}[i] = \max(CW_{\text{max}}[i], CW_{\text{old}}[i] \cdot MF[i])$$

where

$$MF[i] = \min((1 + 2i) \cdot f_{\text{avg}}^i, 0.8)$$

$f_{\text{avg}}^i$ is a running average based on $f_{\text{avg}}^{i-1}$ and $f_{\text{curr}}^i$:
\[ f_{\text{avg}}^j = (1 - \alpha) f_{\text{curr}}^j + \alpha f_{\text{avg}}^{j-1} \]  

(3.4)

which is the estimated collision rate at the STA scaled by \( \alpha \); a factor to smooth the estimation.

This adaptive EDCF scheme decreases the collision rate and increases the utilisation of the medium, the higher utilisation allows a higher load in the system and the paper claims the goodput obtained is up to 25% higher in comparison to the original EDCF.

### 3.3.7. Merging and Multicasting Scheme

In “Solutions to Performance Problems in VoIP Over a IEEE 802.11 Wireless LAN” [WAN05a] the authors highlight the issues of bi-directional voice calls, namely that the AP typically serve \( N \) times the load, where \( N \) is the number of STAs. As each station, on average, gets a fair share of the medium, then the AP must handle more packets and thus experience longer delays and drops. To overcome this problem, a multiplexing scheme is proposed, in which the AP can multicast multiple VoIP packets in one transmission. This multiplexing scheme allows the AP to clear substantial amounts of buffered traffic and means that, as the network becomes loaded, the AP’s traffic is not starved until saturation. The multiplexing scheme works due to the fact all stations can hear the AP when it transmits. However, the scheme requires each station to have the relevant demultiplexing algorithm to successfully obtain the data destined for it. It is interesting to note that an acceptable call is defined as a mean delay of no more than 30ms and where no more than 1% of the packets are dropped. These parameters therefore reiterate the issue of how a VoIP call can actually be quantified.

### 3.3.8. Prioritising Voice over Data

In “Novel multiple access protocol for VoIP in WLAN” [HIR02] a method for prioritising VoIP is considered in which the station transmitting a VoIP packet determines if a collision has occurred without waiting for the ACK timeout. The protocol operates as normal however a station which has just transmitted a VoIP packet determines if the medium is idle or busy. By determining the medium state, the VoIP station assumes a collision if the medium is busy as it means that another station is also transmitting. If this occurs, the station sending the VoIP packet assumes a collision and can immediately resend once the medium is idle. By immediately sending, the STA is guaranteed no collisions from other stations as other satiations should wait for a DIFS period. If two VoIP packets collide, or the data packet is smaller than the VoIP packet, then the collision will not be detected by checking the medium after transmission and, therefore, normal operation will occur.
This paper shows a smaller delay time for VoIP packets as, in most cases, VoIP packets will only experience one collision and then be sent without collision. However, this technique can starve low priority traffic and also relies on an assumption that lower priority data has a larger packet size than VoIP packets. If the larger packet size for lower priority traffic assumption is not the case, or only VoIP traffic is using the network, then the operation is as normal.

### 3.3.9. Priority Access for the AP

In “A Quality-Aware VoWLAN Architecture and Its Quantitative Evaluations” [KOG06] the load imbalance on the AP, when using bi-directional VoIP, is again identified. This paper proposes that the AP should have a higher priority in the network to serve the inbound call traffic. To provide this priority, a method is used in which the AP waits for the medium to be idle for a SIFS rather than DIFS period. The results show that a higher, yet fair, throughput can be achieved and is a better solution than using IEEE 802.11e with different prioritised traffic types. The reason for this improvement is that IEEE 802.11e has no method of prioritising the AP over other STAs when they both have the same traffic type to serve. The results are shown in terms of throughput and, therefore, a clear call capacity increase is not shown.

Also in this paper, continuing the increasing throughput theme, wide area wireless networks are also considered. The issue of stations connecting to the AP with the highest received signal strength is highlighted. Stations connecting to an AP with the highest signal strength may not be the best AP to use as there is no consideration of the load of the AP. The scheme presented attempts to distribute stations, with the aim of achieving the highest throughput, by estimating all neighbouring APs load and choosing the one with the lowest. This scheme also would consider the physical rate indirectly as this is one of the parameters affecting throughput. The physical rate aspect of the paper may be used as another performance mechanism in large scale scenarios.

The methods proposed in this paper may increase call capacity; however, the analysis is in terms of throughput and not the components which relate to a call’s quality. To determine call quality and hence call capacity delays and drops need to be calculated and, therefore, the impacts of these schemes in relation to VoIP are complex to determine.

### 3.3.10. Further Methods

Other papers not included in this review have looked into performance enhancements of PCF [WAN06] (though PCF is not widely implemented) and further enhancements specifically on VoWLAN using IEEE 802.11e [QUAN, LIN08]. The latter, “Voice Capacity Analysis of WLANs with Channel Access Prioritizing Mechanism” again identifies the bottleneck in the AP and uses a combination of modifying the contention window, $\text{CW}_{\text{min}}$ and $\text{CW}_{\text{max}}$, and using a smaller medium idle detection time, AIFS, to
prioritise the AP over other stations (using DIFS). Results show that in fact the AIFS mechanism does not increase voice capacity for infrastructure scenarios as hoped. The reason for the lack of capacity increase is due to the protocol behaviour. After any collision, the AP will always decrement its timer due to it only needing to hear the medium idle for a shorter period than the stations. This guaranteed backoff counter decrement means that the AP receives a higher precedence and forces the stations into saturation at a lower load.

3.1 Voice over Wireless LAN

Reviewed so far are the general schemes proposed to enhance the underlying protocol and provide mechanisms to allow real time traffic, such as VoIP, to have higher priority on the medium. The section discusses the efforts made in the field with concentration specifically on Voice over WLAN (VoWLAN).

3.1.1 Determining Voice Capacity

In “Voice Capacity of IEEE 802.11b/a/g Wireless LANs” [MED04], a model is proposed to estimate the number of calls that can be served efficiently in the network. It was found that the capacity was a strong function of channel bit rate, codec packetisation interval, data traffic and packet size. The paper shows that, in a typical IEEE 802.11b network (operating at 11Mbps) using similar parameters as G.711, the total acceptable voice capacity is found to be 21 bi-directional calls in simulation and 22 with their analytical model. These results are based on an assumed criterion of 200ms delay to determine if a call is of acceptable quality; this again highlights the challenge of how to quantify a call’s quality from a users’ perspective.

The paper then goes on to show how the capacity changes with adjustments of the above parameters. The main result of interest is that, if the packetisation interval is doubled from 20ms to 40ms, the call capacity is increased to a 38 calls. This increase is explained as being due to the fact that when the interval is doubled, the number of packets per second is halved and, therefore, the contention overhead introduced by the network (MAC layer) is reduced. This increasing the packetisation interval scheme is explored later in this dissertation to determine the accuracy of the capacity claim and to further explore the opportunity to increase call capacity related to users’ perceived Quality of Experience in Section 9.5 and Section 9.6, respectively.
4 Saturated MAC Model

4.1 Introduction

The contention function of the MAC layer combined with the saturated and unsaturated modes of operation can be a challenge to model. The interaction of STAs on the channel and, therefore, on a given STA, combined with the backoff and counting down variable slots time behaviour, is a complex set of behaviours that need to be considered when developing an analytical model. Furthermore, other factors such as the impacts of different traffic profiles and the buffering of traffic means that there is currently not a fully functioning analytical model developed to investigate the MAC layer performance.

As explained in Section 3.1, the first attempt at a basic model, pre-dating the final protocol release in 1996, was a Markovian based model [HO96] developed to calculate the stationary collision probability based on a two state contention window. This paper, however, does not provide sufficient analysis to model the whole MAC layer and the multiple backoff states. In 1996, Bianchi published an IEEE letter which details a new Markovian model [BIA96] to represent the backoff process a station needs to undertake in order to gain access and successfully send a packet. The model developed, which is based on finding the probability of a station backing off and the time a station needs to wait while backing off, as a function of other stations transmission probability, allows a throughput and service time calculation. The model can be used for any number of STAs and for any contention window parameters ($W$ and $m$ to give $CW_{min}$, $CW_{max}$).

Due to the impact of interacting STAs, Bianchi assumes that each station has the same perception and effect on the channel, hence one state machine is used to represent all STAs in question. The model is based on collision probability, $p$, of two stations transmitting at the same time and therefore experiencing a collision and needing to advance through the developed Markov backoff chain, as shown in Figure 4.1. Each of the $m+1$ states represents a backoff stage, and as an STA progresses through the model, with probability $p$, $CW$ is doubled to represent the exponential distribution of possible backoff slot times. A transition is made back to state 0 on successful transmission, with probability $1-p$. The model also shows, above each state, the mean number of slots an STA will wait in each state. By multiplying through the probability of being in each backoff state by the mean time spent in each state gives an expected mean backoff time.
The Bianchi model, as introduced in Section 4.1, consists of a Markov chain to represent this model assumes a saturated case; this is defined as all stations having at least one packet to send, and therefore all contending. In the saturated case, when a station has successfully transmitted a packet and received an ACK it must return to backoff when the queue is non-empty to avoid capturing the channel. Recalling the operation of the IEEE 802.11 protocol in Chapter 2, when an STA finishes servicing a packet, and there is another packet buffered, the STA must immediately backoff regardless of the channel state to avoid capturing the channel. This is defined in the protocol in an attempt to keep the channel access fair among all contenders. This saturated assumption also means that all STAs have the same impact on one another, i.e. if there are 5 stations they are all affected equally; by 4 stations with equal transmission and therefore collision probabilities.

In 2000, Bianchi re-published in a Journal; “Performance analysis of the IEEE 802.11 distributed coordination function” [BIA00] and this is now widely used in the wireless community as a standard for benchmarking and testing new methods and ideas.

4.2 Deriving the Bianchi Model

The Bianchi model, as introduced in Section 4.1, consists of a Markov chain to represent the transitions though the backoff process. Finding the stationary state probability and combining with the mean time spent in each state, an effective MAC backoff time, \( E[\text{backoff time}] \), can be found; this is the time from when a packet enters the MAC layer until it is successfully acknowledged; this includes the time of the multiple backoffs needed and any collisions experienced.

Although the Bianchi model is derived and solved in the paper [BIA00], it is derived with slightly different terminology here in order to keep the same variables and state based methodology throughout the dissertation.
The Bianchi model first introduces a transmission probability, $\tau$. $\tau$ is the probability, at any given instance, that a station is transmitting. The transmission probability can be used to derive conditional collision probability, $p$, from the perspective of a transmitting station as follows:

$$P(\text{another station transmits } | \text{ transmitting}) = p = 1 - (1 - \tau)^{N-1} \quad (4.1)$$

This is the probability that when a transmitting STA sends a packet it will experience a collision; this is the probability that at least one other station is transmitting.

Figure 4.2 shows the relationship between $\tau$ and $p$ with 5, 10, 20, 50 and 100 stations.

![Figure 4.2: Transmission vs. Collision Probability](image)

To solve the model, state probabilities are derived. Figure 4.3 shows the generic state transition matrix and Figure 4.4 a case where $m=3$.

A reminder of the list of variables used is presented in Table 4.1 with additional symbols introduced in this chapter.
<table>
<thead>
<tr>
<th>Symbol</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>$CW$</td>
<td>Number of Slots a Station waits before attempting to Transmit</td>
</tr>
<tr>
<td>$CW_{\text{min}}$</td>
<td>The Minimum Range of Slots a Station must Wait to Transmit a Packet</td>
</tr>
<tr>
<td>$CW_{\text{max}}$</td>
<td>The Maximum Range of Slots a Station needs to Wait after Successive Collisions</td>
</tr>
<tr>
<td>$W$</td>
<td>The Range of the Minimum Number of Slots a Station needs to Wait – Used to derive $CW_{\text{min}}$.</td>
</tr>
<tr>
<td>$m$</td>
<td>Number of Times the Range of Slots is Doubled</td>
</tr>
<tr>
<td>$p$</td>
<td>Collision Probability</td>
</tr>
<tr>
<td>$\tau$</td>
<td>Transmission Probability</td>
</tr>
<tr>
<td>$Ps$</td>
<td>Probability of Channel being Busy due to another Station Transmitting Successfully</td>
</tr>
<tr>
<td>$Ptr$</td>
<td>Probability of Perceiving the Channel Busy</td>
</tr>
<tr>
<td>$b_i$</td>
<td>Probability of being in Backoff state $i$</td>
</tr>
<tr>
<td>$S_i$</td>
<td>Maximum Number of Slots to Wait in Backoff state $i$</td>
</tr>
<tr>
<td>$\sigma$</td>
<td>Minimum Slot Time when Channel is Idle</td>
</tr>
<tr>
<td>$\overline{T_{\text{slot}}}$</td>
<td>Mean Slot Time when Backing Off</td>
</tr>
<tr>
<td>$\overline{T_k}$</td>
<td>Mean Slot Time in Each Backoff state $k$</td>
</tr>
<tr>
<td>$T_{\text{DIFS}}$</td>
<td>Time of a DIFS period (used before transmitting and resuming backoff timer)</td>
</tr>
<tr>
<td>$T_{\text{SIFS}}$</td>
<td>Time of a SIFS period (used before transmitting acknowledgment)</td>
</tr>
<tr>
<td>$T_s$</td>
<td>Time of a Successful Transmission and Acknowledgment</td>
</tr>
<tr>
<td>$T_c$</td>
<td>Time of Transmitting a Packet which is not Acknowledged due to a Collision</td>
</tr>
<tr>
<td>$E[\text{slots}]$</td>
<td>Estimated Mean Slots a Station needs to Wait before Transmiting Successfully</td>
</tr>
<tr>
<td>$E[\text{backoff}]$</td>
<td>Estimated Mean backoff time</td>
</tr>
<tr>
<td>$E[\text{Saturated Service Time}]$</td>
<td>Estimated Mean Service Time when all Stations are Saturated</td>
</tr>
<tr>
<td>$E[\text{Saturated Service Time}]_{\text{min}}$</td>
<td>Minimum Saturated Service Time when only one Station</td>
</tr>
</tbody>
</table>

*Table 4.1: List of Symbols for the Saturated MAC Analysis*
State probability equations, $b_n$ are derived

$$b_0 = \sum_{k=0}^m b_k (1-p) = 1-p$$ \hspace{1cm} (4.2)

for $k = 1$ to $m-1$

$$b_k = b_{k-1} \cdot p = p^k \cdot (1-p)$$ \hspace{1cm} (4.3)

for $m$

$$b_m = b_m \cdot p + b_{m-1} \cdot p$$ \hspace{1cm} (4.4)

$$b_m = b_m \cdot p + \left[ p^{m-1} \cdot (1-p) \right] \cdot p$$ \hspace{1cm} (4.5)

$$b_m = \frac{p^m \cdot (1-p)}{(1-p)}$$ \hspace{1cm} (4.6)

$$b_m = p^m$$ \hspace{1cm} (4.7)
A simple check shows that $\sum_k b_k = 1$

Next, the mean number of slots to wait in each state, $S_k$ is derived.

$$S_k = \frac{2^k \cdot W}{2} \tag{4.8}$$

The expected mean number of slots a station needs to successfully send a packet can now be found.

$$E[\text{slots}] = 1 + \sum_k b_k \cdot S_k \tag{4.9}$$

The addition of one slot is the slot in which a station transmits successfully.

Transmission probability can now be derived, note the dependence back to collision probability, $p$, hence the state machine requires iteration to find a steady state.

$$\tau = \frac{1}{E[\text{slots}]} \tag{4.10}$$

$$\tau = \left[ 1 + \frac{W}{2} \cdot (1 - p) + \sum_{k=1}^{m-1} \left( 2^{k-1} \cdot W \cdot (1 - p) \cdot p^k \right) + 2^{m-1} \cdot W \cdot p^m \right]^{-1} \tag{4.11}$$

Once a steady state has been found, using numeric methods, the expected mean service time can be calculated by finding the mean time spent in each of the backoff states. The time in each state is comprised of the mean number of slots, multiplied by the mean slot time $T_{\text{slot}}$ as well as the time to transition between states, either due to a collision, $p$ or a successful transmission $(1-p)$ denoted by $T_i$ and $T_j$, respectively.

Where $T_i$ is given as

$$T_i = T_{\text{IFS}} + T_{\text{SIFS}} + 2 \cdot \text{Propagation Delay} + \frac{\text{Data Packet Size} + \text{MAC Header} + \text{ACK Packet Size} + 2 \cdot \text{PHY Header}}{\text{channel bit rate}} \tag{4.12}$$

and $T_j$ given as
\[ T \_\text{DIFS} + \text{Propogation Delay} + \text{ACK Timeout} + \]
\[ T_e = \left( \frac{\text{Data Packet Size} + \text{MAC Header} + \text{PHY Header}}{\text{channel bit rate}} \right) \]  
\[ (4.13) \]

Note these formulas assume that the physical layer header is sent at the same bit rate as the standard channel bit rate. In some implementations, the physical layer header may be sent at a lower bit rate, such as 1Mbps and a data rate of anywhere between 1Mbps and 54Mbps. Essentially, the physical layer header is sent at a lower symbol rate to avoid bit errors; throughout this dissertation, as with the Bianchi analysis, a channel rate of 1Mbps is used in all numerical analysis.

Recall that a slot time can be one of three durations, either the standardised slot parameter, \( \sigma \), when the channel is idle, or if the channel is busy a \( T_i \) or \( T_c \), dependent on another station or stations occupying the channel and sending successfully or experiencing a collision. To find the mean slot time, a probability of hearing the channel busy, \( P_b \), and a probability of hearing another station sending successfully, \( P_s \), or conversely a group of stations transmitting unsuccessfully, \( 1-P_s \), is needed.

The conditional probability of hearing the channel busy, from a backing off STA’s perspective, is simply derived from \( p \), where the N-1 excludes the station in question.

\[ P(\text{channel is perceived busy | backing off}) = P_b = 1 - (1 - \tau)^{N-1} \]  
\[ (4.14) \]

Similarly, the conditional probability another station transmits and that station transmits successfully is given where \( N-2 \) excludes the station in question and the station transmitting; in other words that one and only one station is transmitting.

\[ P(\text{one station is transmitting | backing off}) = P_s = \frac{(N-1) \cdot \tau \cdot (1 - \tau)^{N-2}}{P_b} \]  
\[ (4.15) \]

The mean slot time can now be found.

\[ \overline{T}_{\text{slot}} = (1 - P_b) \sigma + P_b \cdot P_s \cdot T_i + P_b \cdot (1 - P_s) \cdot T_c \]  
\[ (4.16) \]

The time in each state, \( \overline{T}_k \), can now be found.

\[ \overline{T}_k = \overline{T}_{\text{slot}} \cdot S_k + (1 - p) \cdot T_i + p \cdot T_c \]  
\[ (4.17) \]
hence allowing the expected mean backoff time to be found.

\[ E[\text{backoff time}] = \sum_k b_k \cdot T_k \]  
(4.18)

\[ E[\text{backoff time}] = \sum_k \left( b_k \left( \frac{2^k \cdot W}{2} \cdot T_{\text{slot}} + (1 - p) \cdot T_s + p \cdot T_c \right) \right) \]  
(4.19)

\[ \bar{T}_{\text{slot}} \cdot \left[ \sum_k (b_k \cdot S_k) \right] + \]  
\[ E[\text{backoff time}] = T_s \cdot \left[ \sum_k (b_k \cdot (1 - p)) \right] + \]  
\[ T_c \cdot \left[ \sum_k (b_k \cdot p) \right] \]  
(4.20)

\[ \bar{T}_{\text{slot}} \cdot \left[ \sum_k (b_k \cdot S_k) \right] + \]  
\[ E[\text{backoff time}] = T_s \cdot (1 - p) + \]  
\[ T_c \cdot p \]  
(4.21)

Shown with substituted state probabilities

\[ \bar{T}_{\text{slot}} \cdot W \cdot \left[ \frac{(1 - p)}{2} + \sum_{k=2}^{m-1} (2^{k-1} \cdot (1 - p) \cdot p^k) + 2^{m-1} \cdot p^m \right] + \]  
\[ E[\text{backoff time}] = T_s \cdot (1 - p) + \]  
\[ T_c \cdot p \]  
(4.22)

Knowing the proportions of the different times spent backing off, an expected mean saturated service time can be found by normalising by the proportion of time for a successful transmission, hence

\[ E[\text{saturated service time}] = \frac{E[\text{backoff time}]}{(1 - p)} \]  
(4.23)

Finally, the saturated network throughput can be found
Comparing with Simulation

Now a MAC layer model is found, the simulation study can begin. Section 9.2 shows how the analytical model state machine is solved, particularly to arrive at a steady state and is later described in Chapter 6 as the model is augmented. For applying a set of parameters and characteristics to a specified network configuration, simulation provides a good method where most details can be recorded and later analysed. The Network Simulator with Network Animator (NS/NAM) package [NS2]. The output of this type of simulation is, generally, one sequential trace file that can be post-processed using scripts to parse and extract data. Section 9.3 outlines the simulation technique. The benefit of using this specific package is that it is free and open source, which enables lower level protocol manipulation and is widely supported and used in the network research community. This open source structure means the simulation has been validated against many scenarios and in various studies by the research community.

In terms of Wireless Networks, NS2 provides a comprehensive set of protocols that can be called upon a number of different ways to represent the potential real-world uses and multiple traffic models can be used to simulate traffic such as VoIP and other realistic applications.

The Bianchi scenario was setup in NS2 with the IEEE 802.11 parameters defined in the paper, shown in Table 4.2, and was run with varying window parameters $W$ and $m$. Section 9.3 shows the simulation parameters, including time and number of batch repeats. Figure 4.5 shows how the scenario was constructed in the simulator; $N$ stations are created with a simple Poisson traffic source. The model assumes that each STA is saturated and hence always backs off (due to having at least one packet in its queue); this is achieved with an inter-packet time less than the minimum service time achievable.

$$E\left[\text{saturated service time}\right]_{\min} = \frac{W}{2} \cdot \text{Slot Time} + 2 \cdot \text{Propagation Delay} + \text{DIFS} + \text{SIFS} + \left(\text{Data Packet Size} + \text{MAC Header} + \text{PHY Header}\right) + \text{ACK Packet Size} \cdot \text{channel bit rate}$$

(4.25)

$$E\left[\text{saturated service time}\right]_{\min} = 9.782ms$$

(4.26)
Therefore 9.782ms is the best case service time for the saturated service time. The shortest service time is achieved by always hearing the channel idle, while backing off, and hence each slot time is the standard idle slot time defined in the protocol.

Later in the dissertation, different scenarios will be explored including varying number of stations and different packet sizes and, therefore, a best case service time is derived in order to determine a load to ensure saturation. The load which ensures saturation, based on packets per second, for the various scenarios is used to allow for re-use of the simulation setup. Here, the best case is assumed to always have the option of transmitting after a DIFS period and packet transmission time has been discounted.

\[
E[\text{service time}]_{\min} = Propagation \text{ Delay} + DIFS + SIFS + \\
\frac{(MAC \text{ Header} + PHY \text{ Header}) + ACK \text{ Packet Size}}{\text{channel bit rate}}
\]

\[
E[\text{service time}]_{\min} = 0.796\text{ms} \quad (4.28)
\]

To keep a minimum of at least one packet in the queue, the traffic source should inject at least 1,257 packets per second (an inter-packet time of $795.55 \mu s$); however, the simulation is in fact set up with 2,000 packets per second (an additional 37%) to ensure this condition is always met when considering mean-rate-based traffic sources. Furthermore, the simulation scenario is run long enough to ensure the simulation has convergence to a steady state and an investigation of the queue size analysed to ensure accurate simulation results. The simulation scenario is then batch repeated a sufficient number of times (at least 32) to ensure confidence results and to counter any effects of the random number generator; each repeat is seeded with a different number (0 to 31).
Table 4.2: IEEE 802.11 Model Parameters

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Packet Payload</td>
<td>8184 bits</td>
</tr>
<tr>
<td>MAC Header</td>
<td>272 bits</td>
</tr>
<tr>
<td>PHY Header</td>
<td>128 bits</td>
</tr>
<tr>
<td>ACK</td>
<td>112 bits + PHY header</td>
</tr>
<tr>
<td>RTS</td>
<td>160 bits + PHY header</td>
</tr>
<tr>
<td>CTS</td>
<td>112 bits + PHY header</td>
</tr>
<tr>
<td>Channel Bit Rate</td>
<td>1Mbps</td>
</tr>
<tr>
<td>Propagation Delay</td>
<td>1µs</td>
</tr>
<tr>
<td>Slot Time (σ)</td>
<td>50µs</td>
</tr>
<tr>
<td>SIFS (τ_{sifs})</td>
<td>28µs</td>
</tr>
<tr>
<td>DIFS (τ_{difs})</td>
<td>128µs</td>
</tr>
<tr>
<td>ACK_Timeout</td>
<td>300µs</td>
</tr>
<tr>
<td>CTS_Timeout</td>
<td>300µs</td>
</tr>
</tbody>
</table>

The Bianchi scenario was set up with the relevant parameters and was run with varying window parameters \( W \) and \( m \). The Bianchi model also includes NS2 simulation results. This section of the dissertation also shows analytical and simulation results to substantiate NS2’s IEEE 802.11 simulation package and saturated scenario.
Figure 4.6 shows the normalised saturated throughput (throughput divided by the underlying data rate) of 1 to 50 STAs with different backoff parameters. The Bianchi model is shown with lines and simulations (conducted for this dissertation) with points.

![Figure 4.6: Simulations vs. Bianchi Model](image)

As shown, in Figure 4.6, simulation results were within 1% of the model. The results show that overall throughput is a function of the number of STAs sharing access to the channel.

### 4.4 Further Explorations

The Bianchi analysis shows the saturation points of a varying number of stations and it was shown that the channel is in fact underutilised. The underutilisation, at saturation with increasing number of stations, is due to the protocol overhead, namely the constant requirement to back-off.

This dissertation aims to investigate real-time VoIP and choose traffic parameters corresponding to two protocols; G.711 and G.729. In this chapter, an investigation of different packet sizes on the MAC layer is undertaken to see how much the overhead affects the achievable throughput.

G.729 is effectively a constant rate protocol as it has a very small off time of 8ms every 1,004ms, therefore is it transmitting packets for 996ms in a 1,004ms period; just over 0.99 of time. G.711, on the other hand, has a larger off time than on time and has a higher bit rate making it slightly burstier. This burstier traffic therefore models more
accurately the silent listening periods of a voice call and is therefore more reprehensive of traffic patterns seen from a real call between two people having a conversation.

The parameters, shown in Table 4.3, are from [COL01, YAN07]. The aim of the voice parameters is to create the two traffic patterns with On/Off times modelled with a negative exponential distribution. The two voice traffic patterns are configured to produce similar mean data rates to enable comparison of burstier traffic profiles.

<table>
<thead>
<tr>
<th>Packet Size</th>
<th>G. 729</th>
<th>G. 711</th>
</tr>
</thead>
<tbody>
<tr>
<td>bytes</td>
<td>74</td>
<td>214</td>
</tr>
<tr>
<td>Peak Bit Rate</td>
<td>29.6</td>
<td>85.6</td>
</tr>
<tr>
<td>kbps</td>
<td>1.004</td>
<td>0.35</td>
</tr>
<tr>
<td>seconds</td>
<td>0.008</td>
<td>0.67</td>
</tr>
<tr>
<td>seconds</td>
<td>29.366</td>
<td>29.373</td>
</tr>
<tr>
<td>kbps</td>
<td>25</td>
<td>20</td>
</tr>
<tr>
<td>ms</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4.3 : VoIP Model Parameters

\[
\text{^MeanRate} = \frac{\text{PeakBitRate} \cdot \text{MeanOnTime}}{\text{MeanOFFTime} + \text{MeanOnTime}} 
\]

(4.29)

By varying the packet sizes from 200bits up to 8184bits (Bianchi’s original scenario) and ensuring to choose the specified VoIP packet sizes, an estimated maximum number of calls at saturation can be found.
Figure 4.7 and Figure 4.8 show that, with smaller packet sizes, the overall throughput at saturation decreases. Although the service time is lower, the overhead of the back-off means that there is a larger proportion of time waiting to send each packet than the
packet transmission time. In all the scenarios (N = 5 to 100) a maximum of 0.35 normalised throughput was found for packet sizes equal to the G.729 VoIP protocol.

Looking at the service time increase, an estimate of the maximum number of calls that can be served efficiently in legacy IEEE 802.11 can be established. It is common for an end to end maximum criterion of 75ms\(^1\) to be used for VoIP since delay greater than this is seen as poor quality. Currently, only the delay at the MAC layer can be modelled, i.e. excluding queueing delays or other latency in the network. Therefore, a criterion of 30ms\(^1\) is chosen to leave a higher percentage for these other delays. With a 30ms criterion (Red Line) in Figure 4.8, a maximum of only 10 STAs using packet sizes equivalent to VoIP streams are served efficiently before there is a degradation of service. This, in fact, implies that only 5 bi-directional VoIP calls can be made using a channel data rate of 1Mbps. In this scenario, however, each traffic source is mapped to a single station and this therefore can be considered the best case scenario (no bottleneck caused by an Access Point).

There are PHY extensions to IEEE 802.11, namely IEEE 802.11a, IEEE 802.11b and IEEE 802.11g which boast the potential data rates of 54, 11 and 54Mbps respectively. In fact, although it is true packets will get from `a to b` faster, the proportion of time backing off is due to the contention on the channel and backoff time. [MED04] shows that approximately 22 bi-directional calls can be served at 11Mbps; only an increase of 4.4 times capacity for an 11 times increase in channel data rate. The lower capacity increase, compared to the underlying data rate, is due to the overhead introduced by the CSMA/CA scheme.

In this chapter, it has been shown that the Bianchi model analysis accurately captures the behaviour of the IEEE 802.11 MAC layer. The Bianchi model however, assumes all stations are operating in saturated mode. In Chapter 5 a new model is presented which allows a finite load at each station to be modelled.

\[^1\] These assumed criterion values are used simply to give a rough estimate of the number of calls and form the argument for a need to quantify call quality. In Chapter 7, an investigation is performed to highlight how a call can be quantified in terms of user experience without using assumed criterion values.
5 UN-SATURATED HOMOGENEOUS MAC MODEL

5.1 Introduction

In Chapter 4, the Bianchi model was shown which assumes all stations always have at least one packet to send and therefore all operating in a saturated mode. The protocol, as detailed in Chapter 2, describes that when a station is idle and a packet arrives to an empty queue, it can immediately attempt to send if the medium is perceived idle. In this chapter, an augmentation of the model is investigated in order to find a suitable unsaturated analysis. Following the review in Section 3.1, the Pitts Un-Saturated Case Model is explored.

5.2 Un-Saturated MAC Model

Following on from the Bianchi model confirmation, Pitts and Shepherd in [PIT08] revise the model to include the unsaturated case. The unsaturated case achieved by extending the original Bianchi Markov chain, similar to that seen in other published papers [XU06, DAO08], however, not only including the empty queue state, which gives an idle station the opportunity to transmit directly after a DIFS period; moreover, after successful transmission the stations can return to either an idle or non-empty, hence backoff, state. This modification is achieved by introducing utilisation and hence probability of the queue being non-empty, \( \rho' \), and weighting by the probability of a successful delivery which gives \( (1-p)\rho' \), return to backoff state 0, otherwise \( (1-p)(1-\rho') \), return to an empty queue state, state -1, as shown in Figure 5.1 (red text highlights the additional state transitions/probabilities compared to the Bianchi Model).

\[ \text{Figure 5.1: Pitts Un-Saturated Markov Model} \]

\[ \text{† The development of the Un-Saturated MAC Analytical Model was carried out by Pitts as part of a BT Fellowship in 2006 and validated against simulation by Shepherd (Thesis Author).} \]
Essentially, a new idle state, state -1, has been added to the model to represent the case in which an STA finishes transmitting and finds an empty queue. When a packet does arrive, state -1 gives the opportunity of transmitting directly after a DIFS period if the medium is sensed free.

States 0 through to \( m \) represent the backoff process of an STA as in the Bianchi model; the difference here is that each of these states has two transitions. The two transitions are based on the probability of a successfully transmission, \( 1 - p \), but weighted by \( \rho' \). The weighting is used to determine if the station has a non-empty queue, in which case return to -1, or at least one packet buffered, hence return to state 0.

The additional state transitions from state -1 are derived:

<table>
<thead>
<tr>
<th>Condition</th>
<th>Transition from State -1 to</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Medium is busy, with probability ( P_{in} ), begin backoff</td>
<td>State 0</td>
<td>( P_{tr} )</td>
</tr>
<tr>
<td>Medium is free, with probability ( 1 - P_{in} ) so the station transmits which results in a collision with probability ( p )</td>
<td>State 0</td>
<td>( (1 - P_{tr}) \cdot p )</td>
</tr>
<tr>
<td>Medium is free, with probability ( 1 - P_{in} ) so the station transmits which is successful with probability, ( 1 - p ), queue found empty, ( 1 - \rho' )</td>
<td>State -1 (packet send directly after DIFS next packet has the opportunity to send after DIFS)</td>
<td>( (1 - P_{tr}) \cdot (1 - p) \cdot (1 - \rho') )</td>
</tr>
<tr>
<td>Medium is free with probability ( 1 - P_{in} ) so the station transmits which is successful, ( 1 - p ), queue found non-empty, ( \rho' )</td>
<td>State 0 (packet send directly after DIFS, but the next packet needs to backoff)</td>
<td>( (1 - P_{tr}) \cdot (1 - p) \cdot \rho' )</td>
</tr>
</tbody>
</table>

Table 5.1 : Deriving the Idle State Transitions

Therefore, the state transition from state -1 to state -1 is given

\[
p_{-1,-1} = (1 - P_{tr}) \cdot (1 - p) \cdot (1 - \rho')
\]

and state -1 to state 0 is given

\[
p_{-1,0} = (1 - P_{tr}) \cdot (1 - p) \cdot \rho'
\]
\[ p_{-1,0} = P_n + (1 - P_n) \cdot p + (1 - P_n) \cdot (1 - p) \cdot \rho' \]
\[ = P_n + (1 - P_n) \cdot (p + (1 - p) \cdot \rho') \]  \hspace{1cm} (5.2)

New state probabilities are now derived (red shows the differences between the unsaturated and Bianchi Model):

\[ b_{-1} = \sum_{i=0}^{m} b_i \cdot (1 - p)(1 - \rho') + b_{-1} \cdot (1 - P_n) \cdot (1 - p) \cdot (1 - \rho') \]  \hspace{1cm} (5.3)

\[ b_0 = b_{-1} \left[ P_n + (1 - P_n) \cdot (p + (1 - p) \cdot \rho') \right] + \sum_{i=0}^{m} b_i \cdot (1 - p) \cdot \rho' \]  \hspace{1cm} (5.4)

for \( k = 1 \) to \( m - 1 \)

\[ b_k = b_{k-1} \cdot p = b_0 \cdot p^k \]  \hspace{1cm} (5.5)

and for \( k = m \)

\[ b_m = b_m \cdot p + b_{m-1} \cdot p \]  \hspace{1cm} (5.6)

\[ b_m = b_m \cdot p + \left[ p^{m-1} \cdot b_0 \right] \cdot p \]  \hspace{1cm} (5.7)

\[ b_m = \frac{p^m \cdot b_0}{1 - p} \]  \hspace{1cm} (5.8)

The collision probability, \( p \), the conditional medium busy probability from the perspective of a backing off station, \( P_t \), and the probability of a station transmitting successfully from the perspective of a backing off station, \( P_s \), remain the same as the Bianchi Model.

\[ P(\text{another station transmits} \mid \text{transmitting}) = p = 1 - (1 - \tau)^{N-1} \]  \hspace{1cm} (5.9)

\[ P(\text{channel is perceived busy} \mid \text{backing off}) = P_n = 1 - (1 - \tau)^{N-1} \]  \hspace{1cm} (5.10)
\[ P(\text{one station is transmitting | backing off}) = P_s = (N - 1) \cdot \tau \cdot (1 - \tau)^{N-2} \quad (5.11) \]

The mean number of slots in each state, which was given as, \( S_k \) is updated to include the -1 state. The number of slots is found as the proportion of idle slot times, \( \sigma \), within a DIFS period, \( T_{difs} \).

\[ S_{-1} = \frac{T_{difs}}{\sigma} \quad (5.12) \]

The remaining mean slots within each state \( k \) remains the same

\[ S_k = \frac{2^k \cdot W}{2} \quad (5.13) \]

\( \overline{T_{slot}} \), the mean slot time and \( E[\text{slots}] \) also remain the same but where \( k \) now ranges from -1 to \( m \)

\[ \overline{T_{slot}} = (1 - P_r) \cdot \sigma + P_r \cdot P_s \cdot T_s + P_r \cdot (1 - P_s) \cdot T_c \quad (5.14) \]

\[ E[\text{slots}] = 1 + \sum_k b_k \cdot S_k \quad (5.15) \]

An additional mean time in each state, \( \overline{T}_k \), is extended to include the -1 state

\[ \overline{T}_{-1} = T_{difs} + (1 - P_r) \cdot (1 - p) \cdot T_s + (P_r + (1 - P_r) \cdot p) \cdot T_c \quad (5.16) \]

Where the remaining times in each state remain the same as the Bianchi Model

\[ \overline{T}_k = \overline{T}_{slot} \cdot S_k + (1 - p) \cdot T_s + p \cdot T_c \quad (5.17) \]

The expected mean backoff time can now be found

\[ E[\text{backoff time}] = \sum_k b_k \cdot \overline{T}_k \quad (5.18) \]
\[ E[\text{backoff time}] = b_{-1} \left( T_{\text{diff}} + (1 - P_{tr}) \cdot (1 - p) \cdot T_s + (P_{tr} + (1 - P_{tr}) \cdot p) \cdot T_c + \sum_{k=0}^{m-1} b_k \cdot \frac{2^k \cdot W}{2} \cdot T_{\text{slot}} + (1 - p) \cdot T_s + p \cdot T_c \right) \] (5.19)

\[ E[\text{backoff time}] = T_s \left( \sum_{k=0}^{m} b_k \cdot (1 - p) + b_{-1} \cdot (1 - P_{tr}) \cdot (1 - p) \right) + T_c \left( \sum_{k=0}^{m} b_k \cdot p + b_{-1} \cdot P_{tr} \cdot (1 - P_{tr}) \cdot p \right) + T_{\text{diff}} \cdot b_{-1} \] (5.20)

And by normalising by the proportion of time, \( T_s \),

\[ E[\text{service time}] = \frac{E[\text{backoff time}]}{\sum_{k=0}^{m} b_k \cdot (1 - p) + b_{-1} \cdot (1 - P_{tr}) \cdot (1 - p)} \] (5.21)

The utilisation, \( \rho' \), needed to determine queue state after a successful transmission can be found

\[ \rho' = \min \left( 1, \frac{\lambda_{\text{sta}} \cdot E[\text{service time}]}{1} \right) \] (5.22)

Here the probability of a queue being in a non-empty state is derived by finding utilisation as a function of service time calculated with the model. This method requires iteration to a steady state to derive the estimated mean service time. This method of deriving utilisation and hence queue being non-empty had not been presented in the literature.

Finally, the throughput formula is updated

\[ \text{throughput} = \frac{N \cdot \text{Payload}}{\text{Channel Bit Rate}} \cdot \min \left( \lambda_{\text{sta}} \cdot \frac{1}{E[\text{service time}]} \right) \] (5.23)
5.3 Convergence to Bianchi Model

When $\rho=1$, the stations are operating in the saturated mode of operation, therefore, this should be able to be found mathematically; the model should simplify down to the Bianchi Model. Bianchi Model state probabilities are shown in Chapter 4.

for $k = -1$

$$b_{-1} = \sum_{i=0}^{m} b_i \cdot (1-p) \cdot (1-\rho') + b_{-1} \cdot (1-P_r) \cdot (1-p) \cdot (1-\rho')$$  \hspace{1cm} (5.24)

$$b_{-1} = \sum_{i=0}^{m} b_i \cdot (1-p) \cdot (1-1) + b_{-1} \cdot (1-P_r) \cdot (1-p) \cdot (1-1)$$  \hspace{1cm} (5.25)

$$b_{-1} = 0$$ \hspace{1cm} (5.26)

for $k = 0$

$$b_0 = b_{-1} \left[ P_r + (1-P_r) \cdot \left( p \cdot (1-p) \cdot \rho' \right) \right] + \sum_{i=0}^{m} b_i \cdot (1-p) \cdot \rho'$$  \hspace{1cm} (5.27)

$$b_0 = 0 \cdot \left[ P_r + (1-P_r) \cdot \left( p \cdot (1-p) \cdot 0 \right) \right] + \sum_{i=0}^{m} b_i \cdot (1-p) \cdot 1$$  \hspace{1cm} (5.28)

$$b_0 = 1-p$$ \hspace{1cm} (5.29)

for $k = 1$ to $m-1$

$$b_k = b_0 \cdot p^k = (1-p) \cdot p^k \cdot$$  \hspace{1cm} (5.30)

$$b_k = p^k \cdot (1-p)$$ \hspace{1cm} (5.31)

for $k = m$

$$b_m = \frac{p^m \cdot b_0}{1-p}$$  \hspace{1cm} (5.32)
\[ b_m = p^m \cdot \frac{(1 - p)}{1 - p} = p^m \]  

(5.33)

\[ b_m = p^m \quad \checkmark \]  

(5.34)

The remaining probabilities and timings match the Bianchi model as only the addition of \( k = -1 \) was added, which is now multiplied by \( b_1 (\neq 0) \).

### 5.4 Model Limitations

As stated, there is an assumption that the un-saturated case model proposed allows for an infinite number of re-transmissions if there are successive collisions. To explain this assumption, the state diagram would have to be modified to have a limit of \( r \) backoff states, where \( r \) is the maximum number of retransmission attempts. Furthermore, if in a backoff state, where \( r \geq \text{state} > m \), the mean number of slots to backoff should remain at \( 2^{m \times \text{W}} \) in other words, no longer doubling. Figure 5.2 shows how the updated state model may look; however, it should be noted that when in state \( r \), if a packet experiences a collision, the time for this packet should be ignored as it will be dropped and not reach its intended destination. The updated limited ACK retransmission model may also be used to find the probability of dropping a packet due to the ACK retry limit. Note that in the current un-saturated model, once a packet reaches the MAC layer it will never be dropped.

![Figure 5.2: Markov Chain considering ACK Retry Limit](image-url)
5.5 Comparing with Simulation

Using the un-saturated analysis, the expected mean service time and throughput are investigated at different loads and compared with simulation. Simulation setup is shown in Section 9.3. Each STA is configured with a negative exponential traffic source where the mean inter-arrival time is decreased from 100 seconds to 0.01 seconds; equivalent to 0.01 to 100 packets per second. As load increases up to and past saturation, $\rho' \geq 1$, the results are expected to converge to the Bianchi model results as mathematically shown in Chapter 4.

As shown, in Figure 5.3, there is evidence to support the idea that analysis represents the behaviour at the MAC layer behaviour found by simulation. The analysis compared to simulation is accurate in the un-saturated and saturated states for all $N$ values tested (5 to 100). There is however a deviation of the results at the point where the mode changes between unsaturated and saturation. The deviation is due to the model’s assumption that packets which experience a collision are allowed to reattempt transmission, after a further backoff increment, an infinite number of times. The assumption is detailed in Section 5.4 in which the protocol and therefore simulation has a limit on the number of re-transmissions, resulting in a smaller delay as packets that experience multiple collisions are dropped.
Clearly, the results in Figure 5.4 show that there is a sharp increase in effective service time just before hitting saturation. This increase is more intense as the number of STAs increase and is a critical find for any application running on IEEE 802.11; however, more critical is the impact on delay dependent traffic.

The throughput results, shown in Figure 5.3, correspond to the increase in effective service time. As service time increase, throughput drops and vice versa. At large numbers of STAs (>20) there is the substantial drop in throughput. Results clearly show that the maximum achievable throughput and lower service time is not at saturation but at a slightly lower load. The throughput drop at the transition between saturated and unsaturated is dependent on the number of stations and is more severe as more stations contend.

The effective service time includes waiting to access the medium, defined by the backoff timer, and packet transmission time, defined by the underlying data rate. At a data rate of 1Mbps the packet transmission time would be 8.184ms and at 11Mbps (IEEE 802.11b) would be 0.744ms. From the results, the saturated service time for 5 to 100 stations is in the range 50ms to 1000ms, that’s over 10 times the transmission time. Therefore, even at increased data rates, the onset of saturation would still be severely affected network performance.
This un-saturated model shows the transition between the un-saturated and saturated modes of operation. It is clearly shown that as the network reaches saturation, throughput drops and service times can increase substantially. It is therefore evident that some form of congestion avoidance could be beneficial to keep the network out of saturation which may be achieved at the MAC layer or by adapting the queue mechanism from the simple Tail-Drop; the objective would be to maintain a lower offered load to the MAC layer by decreasing the probability of the buffer always being non-empty, \( \rho' \). This congestion avoidance / increased capacity analysis is used as an aside to demonstrate the usefulness of the novel analytical models proposed after further augmentations in Chapter 6.

5.6 Confirming against Large ACK Retransmission Limit

The difference between simulation and the un-saturated model is due to the model not considering an ACK retry limit (set to 7 in simulation); often IEEE 802.11 is defined so that after a packet experiences a fixed number of collisions it is dropped; this has the effect of decreasing effective service time.

Figure 5.5 and Figure 5.6 show simulation results with an ACK retransmission limit of 1,000,000. The retransmission limit used here is a closer approximation to the unlimited retransmission assumption used in the un-saturated model. The results show that the analytical model is now much closer to simulation compared with Figure 5.4. The assumed behaviour results in an increased effective service time as packets are not dropped at the MAC layer due to the standardised, lower retransmission limit.

![Figure 5.5: Load vs. Throughput: Analysis and Simulation (no ACK Retry)](image-url)
In this chapter, an un-saturated model was shown which gives accurate throughput and service time values for all number of stations tested ($N=5$ to 100) with a uniform load, hence homogeneous, across all stations. The model, however, does not include the ability to accurately model a case in which stations are operating with different traffic profiles, i.e. a heterogeneous case; this is explored in Chapter 6.
6 HETEROGENEOUS TRAFFIC MAC MODEL

6.1 Introduction

The model thus far only considers a scenario where all stations have the same load; known as the homogeneous case. This homogeneous case cannot be used for realistic scenarios where all stations have different traffic profiles, such as in the case in which there is an Access Point. In this chapter, a heterogeneous load case is developed. In order to develop this heterogeneous model, each station will now have its own state machine, transmission probability, \( \tau \), and therefore collision probability \( p \). To calculate an uneven load case, the probability of another station transmitting, \( P_{tr} \), and another station transmitting successfully \( Ps \) must also be recalculated to reflect the different states of neighbouring stations.

Figure 6.1: Heterogeneous State Machine per Station Interactions

To represent the different probability and timing values now present for each station, the symbols are updated for the per station case as shown in Table 6.1.
<table>
<thead>
<tr>
<th>Symbol</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\rho'_n$</td>
<td>Utilisation – Probability of a Packet being En-queued during a transmission</td>
</tr>
<tr>
<td>$p_{n}^{\text{hom}}$</td>
<td>Collision Probability on the Homogeneous Case for a given Station $n$</td>
</tr>
<tr>
<td>$p_{n}^{\text{het}}$</td>
<td>Collision Probability in the Heterogeneous Case for a given Station $n$</td>
</tr>
<tr>
<td>$P_{tr_{n}}^{\text{hom}}$</td>
<td>Probability of Perceiving the Channel Busy in the Homogeneous Case for a given Station $n$</td>
</tr>
<tr>
<td>$P_{tr_{n}}^{\text{het}}$</td>
<td>Probability of Perceiving the Channel Busy in the Heterogeneous Case for a given Station $n$</td>
</tr>
<tr>
<td>$p_{s_{n}}^{\text{hom}}$</td>
<td>Probability of the Channel being Busy due to a Station Transmitting Successfully in the Homogeneous Case for a given Station $n$</td>
</tr>
<tr>
<td>$p_{s_{n}}^{\text{het}}$</td>
<td>Probability of the Channel being Busy due to a Station Transmitting Successfully in the Heterogeneous Case for a given Station $n$</td>
</tr>
<tr>
<td>$b_{n,k}$</td>
<td>Probability of a Station $n$ being in a given Backoff State, $k$ for a given Station $n$</td>
</tr>
<tr>
<td>$S_{k_{n}}$</td>
<td>Number of Slots a Station $n$ waits in each Backoff State, $k$ for a given Station $n$</td>
</tr>
<tr>
<td>$T_{k_{n}}$</td>
<td>Time in Slots a Station $n$ waits in each Backoff State, $k$ for a given Station $n$</td>
</tr>
<tr>
<td>$\bar{T}<em>{\text{slot}}</em>{n}$</td>
<td>Estimated Mean Slots Time a Station $n$ needs to wait for a given Station $n$</td>
</tr>
<tr>
<td>$T_{e_{\text{other}}}$</td>
<td>Time the Channel is Busy due to two Stations Transmitting and Resulting in a Collision</td>
</tr>
<tr>
<td>$E_{{\text{slots}}}_{n}$</td>
<td>Estimated Number of Slots to Wait before Transmitting Successfully for a given Station $n$</td>
</tr>
<tr>
<td>$E_{{\text{backoff}}}_{n}$</td>
<td>Estimated Mean Slots a Station $n$ needs to wait in order to transmit a packet successfully</td>
</tr>
<tr>
<td>$E_{{\text{Service Time}}}_{n}$</td>
<td>Estimated Mean Service Time of Station $n$.</td>
</tr>
</tbody>
</table>

Table 6.1: List of Symbols for the Heterogeneous Case
This chapter highlights the need for more accurate timings in the underlying model and that a move from a slot based to time based analysis is needed in order to capture the behaviour of the heterogeneous case. It is also found, from a study of the simplest scenario of 2 interacting stations, that calibration of the model is needed.

6.2 Updating the Probabilities

In the heterogeneous traffic model, as each station is offering a different load and they are represented by separate state machines, each station will impact on and perceive the channel conditions differently. To consider this heterogeneous behaviour, the underlying probabilities that are affected by other stations, including collision probability, the probability of hearing the medium busy and the probability of another station sending successfully, need to be reformulated. Figure 6.2 highlights the updated state machine with STA specific probabilities.

![Figure 6.2 : Heterogeneous State Machine for STA](image)

6.2.1. Collision Probability

In the uniform, homogeneous traffic case, the conditional collision probability, \( p \), is given as:

\[
P(\text{another station transmits} \mid \text{transmitting}) = p^\text{hom} = 1 - (1 - \tau)^{N-1}
\]

(6.1)

This is derived from the complementary probability of all stations not transmitting, except for the station in question. In the heterogeneous case, where each of the \( N \) stations has its own transmission probabilities, \( \tau_1, \tau_2, \ldots, \tau_N \), this would become the complementary probability of \( (1 - \tau_1)(1 - \tau_2) \ldots (1 - \tau_N) \). The final division cancels out the station in question from the numerator. The heterogeneous collision probability is then given as
6.2.2. Probability of the Channel perceived Busy

The probability of at least one other station transmitting, from the perspective of a backing off station, $P_{tr}$, was shown to be equivalent to the conditional collision probability, $p$, in the homogenous model. The same holds, for deriving the probability from the perspective of station $n$, in the heterogeneous case.

$$P_{n}(\text{channel is perceived busy} \mid \text{station } n \text{ backing off}) = P_{tr}^{\text{het}} \equiv p_{n}^{\text{het}}$$ (6.3)

6.2.3. Probability of a Station Transmitting Successfully

The probability of another station transmitting, and that transmission not resulting in a collision, $P_{s}$, is given below in the homogeneous case (Equation 5.11).

$$P_{s}^{\text{hom}} = (N-1)(\tau)(1-\tau)^{N-2}$$ (6.4)

So, from the perspective of one station, one of the stations is transmitting; $\tau_1$ or $\tau_2$ or ... or $\tau_{N-1}$ or $\tau_N$ but not $\tau_n$, the station in question, which gives $\tau_1 + \tau_2 + ... + \tau_{N-1} + \tau_N - \tau_n$. Then, the other N-2 stations, not including the station being considered, $n$, or the station transmitting, $tx$, are all not transmitting; $(1 - \tau_1)(1 - \tau_2) ... (1 - \tau_{N-1})(1 - \tau_N) / (1 - \tau_n)(1 - \tau)$. Again, division is used to cancel out the two stations not considered. This then gives:

$$P_{s}^{\text{het}} = \left[ \sum_{n=1}^{N} \left( (\tau_n) \cdot \frac{\prod_{i=1}^{N}(1-\tau_i)}{(1-\tau_n)(1-\tau_n)} \right) \right] / P_{tr}^{\text{het}}$$ (6.5)

This leads to a summation where $\tau_{tx} = \tau_1$ or $\tau_2$ or ... or $\tau_{N-1}$ or $\tau_N$. However, the case where $\tau_{tx} = \tau_n$, must be subtracted, finally giving:

$$P_{s}^{\text{het}} = \left[ \sum_{n=1}^{N} \left( (\tau_n) \cdot \frac{\prod_{i=1}^{N}(1-\tau_i)}{(1-\tau_n)(1-\tau_n)} \right) - \left( (\tau_n) \cdot \frac{\prod_{i=1}^{N}(1-\tau_i)}{(1-\tau_n)^2} \right) \right] / P_{tr}^{\text{het}}$$ (6.6)
6.2.4. Scrutiny of the Formula for Even Load

Here, the new heterogeneous probability formulas are shown where all $\tau$'s are set equal; this is the case where all stations operate with the same load and therefore should simplify to the homogeneous case. Collision probability, $p_n$, and the probability that another station is sending successfully, $P_s$, are shown where $\forall \tau_n^{\text{het}} = \tau^{\text{hom}}$; the output of both correspond to the homogenous case as shown.

\[
p_h = 1 - \left[ \prod_{i=1}^{N} \left( 1 - \tau_i \right) \right] \left( 1 - \tau_n \right)
\]  
\[6.7\]

\[
p_n^{\text{het}} = 1 - \left[ \frac{(1 - \tau)^N}{(1 - \tau)} \right]
\]  
\[6.8\]

\[
p_n^{\text{het}} = p_n^{\text{hom}} = 1 - (1 - \tau)^{N-1} \quad \text{(where $\tau$'s are the same across all STAs)} \]
\[6.9\]

$P_s$ is given:

\[
P_s^{\text{het}} = \left[ \sum_{i=1}^{N} \left( \tau_i \cdot \frac{\prod_{j=1}^{N} (1 - \tau_j)}{(1 - \tau_n)(1 - \tau_n)} \right) - \left( \tau_n \cdot \frac{\prod_{j=1}^{N} (1 - \tau_j)}{(1 - \tau_n)^2} \right) \right] / \text{Ptr}^{\text{het}}
\]  
\[6.10\]

\[
P_s^{\text{het}} = \left[ \sum_{i=1}^{N} \left( \tau_i \cdot \frac{(1 - \tau)^N}{(1 - \tau)^N} \right) - \left( \tau_n \cdot \frac{(1 - \tau)^N}{(1 - \tau)^N} \right) \right] / \text{Ptr}^{\text{het}}
\]  
\[6.11\]

\[
P_s^{\text{het}} = \left[ N \cdot \tau \cdot (1 - \tau)^{N-2} - \tau \cdot (1 - \tau)^{N-2} \right] / \text{Ptr}^{\text{het}}
\]  
\[6.12\]

\[
P_s^{\text{het}} = \left[ (N-1) \cdot \tau \cdot (1 - \tau)^{N-2} \right] / \text{Ptr}^{\text{het}}
\]  
\[6.13\]

\[
P_s^{\text{het}} = P_s^{\text{hom}} = \frac{(N-1)(1 - \tau)^{N-2}}{\text{Ptr}} \quad \text{(where $\tau$'s are the same across all STAs)}
\]  
\[6.14\]
6.3 Updating the Timings

The original homogeneous state machine will be used, and therefore the same state probabilities.

\[
b_{-1} = \sum_{i=0}^{m} b_i \cdot (1 - p_n) \left(1 - \rho_i'\right) + b_{-1} \cdot (1 - P_{i,\text{tr}}) \cdot (1 - p_n) \cdot (1 - \rho_i') \tag{6.15}
\]

\[
b_{0,\text{tr}} = b_{-1,\text{tr}} \left[ P_{i,\text{tr}} + (1 - P_{i,\text{tr}}) \cdot (p_n + (1 - p_n) \cdot \rho_i') \right] + \sum_{i=0}^{m} b_{i,\text{tr}} (1 - p_n) \rho_i' \tag{6.16}
\]

\[
b_{k,\text{tr}} = b_{0,\text{tr}} \cdot p_n^k \tag{6.17}
\]

\[
b_{m,\text{tr}} = \frac{p_n^m \cdot b_{0,\text{tr}}}{1 - p_n} \tag{6.18}
\]

Next, for each state, the mean number of slots a station needs to backoff is found. Note that state -1 is amended to represent a more realistic timing; when a station perceives the medium busy it, will not wait the entire DIFS period. The remainder of the slots per state remain are the same.

\[
S_{-1} = (1 - P_{i,\text{tr}}) \cdot \frac{T_{\text{difs}}}{\sigma} \tag{6.19}
\]

\[
S_k = \frac{2^k \cdot W}{2} \quad 0 \leq k \leq m \tag{6.20}
\]

The mean number of slots a station backs off for remains the same.

\[
E[\text{slots}]_n = 1 + \sum_k b_{k,\text{tr}} \cdot \overline{S_k} \tag{6.21}
\]

Similarly, the mean backoff time is found.

\[
E[\text{backoff time}]_n = \sum_k b_{k,\text{tr}} \cdot \overline{T_k} \tag{6.22}
\]
Where $T_k$ is defined as the number of slots in each state, $S_k$, multiplied by the mean slot time, $T_{\text{slot}}$, added to the transition time out of a given state multiplied by its probability.

$T_k$ is derived in the homogeneous model.

\[
\overline{T}_{-1} = T_{\text{diff}} + \left(1 - P_{tr}ight) \cdot (1 - p_n) \cdot T_s + \left(P_{tr} + (1 - P_{tr}) \cdot p_n \right) T_c
\]  

(6.23)

However, if the medium is busy with probability $P_{tr}$, the station does not experience a $T_c$ time as shown above, but experiences the time another station transmits for, $T_c^{\text{other}}$, and it may be that the transmission is already occurring when in state -1.

\[
\overline{T}_{-1} = T_{\text{diff}} + \left(1 - P_{tr}ight) \cdot (1 - p_n) \cdot T_s + \left(1 - P_{tr} \right) \cdot p_n \cdot T_c + P_{tr} \left( \frac{T_c^{\text{other}}}{2} \right)
\]  

(6.24)

for $T_k$ where $k = 1$ to $k = m$

\[
\overline{T}_{-1} = S_k \cdot \overline{T}_{\text{slot}} + T_s \left(1 - p_n \right) + T_c \cdot p_n
\]  

(6.25)

Where the mean slot time, $T_{\text{slot}}$, is made up of weighting three possible times; σ when the medium is free, $T_s$ when the medium is busy due to one station transmitting successfully or $T_c$ if more than one station is transmitting.

\[
\overline{T}_{\text{slot}} = \left(1 - P_{tr} \right) \sigma + P_{tr} \cdot P_{S_n} \cdot T_s + P_{tr} \cdot (1 - P_{S_n}) \cdot T_c
\]  

(6.26)

However, $T_{\text{slot}}$ is amended so that when another station transmits and experiences a collision, originally shown as $T_s$, in fact this will just be the proportion of time to transmit, now represented as $T_c^{\text{other}}$ ($T_c^{\text{other}} = T_c$ without the SIFS time and ACK timeout). Furthermore, a DIFS time is added to each time case to determine the medium is again free, hence a station would be pausing its timer for this as well.

\[
\overline{T}_{\text{slot}} = \left(1 - P_{tr} \right) \sigma + P_{tr} \cdot P_{S_n} \cdot \left(T_s + T_{\text{diff}} \right) + P_{tr} \cdot (1 - P_{S_n}) \cdot \left(T_c^{\text{other}} + T_{\text{diff}} \right)
\]  

(6.27)

Note, as the model assumes all stations have the same packet size, $T_s$, $T_c$ and $T_c^{\text{other}}$ are the same for all stations.
Finally, the normalisation between the mean service time, \( E[\text{backoff time}]_n \), and the estimated service time, \( E[\text{service time}]_n \), is found by normalising the probability of a successful transmission, \( p(T_s) \).

\[
E[\text{backoff time}]_n = \sum_k b_{k n} \cdot T_{k n} \tag{6.28}
\]

\[
E[\text{backoff time}]_n = T_{s n} \cdot \left[ \sum_{k=0}^m b_{k n} \cdot \frac{2^k \cdot W}{2} \right] + T_{v n} \cdot \left[ \sum_{k=0}^m b_{k n} \cdot (1 - p_n) + b_{-1 n} (1 - p_n) \cdot (1 - p_n) \right] + T_{d n} \cdot b_{-1 n} \tag{6.29}
\]

\[
E[\text{service time}]_n = \frac{E[\text{backoff time}]_n}{p(T_s)_n} \tag{6.30}
\]

Where \( p(T_s) \) is given

\[
p(T_s)_n = \sum_{k=0}^m b_{k n} \cdot (1 - p_n) + b_{-1 n} (1 - p_n) \cdot (1 - p_n) \tag{6.31}
\]

Other fundamental probabilities, such as the probability of finding a packet waiting in the buffer after a successful transmission, \( \rho'_n \), remain the same, with the exception of \( \tau \), the probability of a station transmitting.

\[
\rho'_n = \min \left( 1, \lambda_{m n} \cdot E[\text{service time}]_n \right) \tag{6.32}
\]

The transmission probability of the original model is given and later amended in Section 6.5.2 after more detailed investigation and justification in Section 6.4 and Section 6.5.

\[
\tau_n = \Pr\{\text{transmit} | \text{packet to send}\}_n \cdot \Pr\{\text{packet to send}\}_n = \frac{1}{E[\text{slots}]_n} \cdot \rho'_n \tag{6.33}
\]
6.4 2 Station Scenario Highlighting the Current Discrepancy

To demonstrate the heterogeneous traffic model, the simplest case is taken where two stations operate with different loads.

![Two Co-existing Stations](image)

Figure 6.3 : Two Co-existing Stations

6.4.1. Solving the State Machine

When there are two co-existing stations, there are some fundamental probability relationships.

Collision probability, $p_a$, is from the perspective of a transmitting station is simply the transmission probability, $\tau_a$, of the other (recalling that collision probability is conditioned on the station in question transmitting).

\[ p_a = \tau_b \]  
\[ p_b = \tau_a \]  

(6.34)  
(6.35)

Perceiving the medium busy, $P_{tr_a}$; from the perspective of a station backing off, the medium will be busy only when the other station is transmitting

\[ P_{tr_a} = \tau_b \]  
\[ P_{tr_b} = \tau_a \]  

(6.36)  
(6.37)

The probability of another station sending successfully, $P_{s_a}$; from the perspective of a station backing off, in the two station case will always be 1 given the assumption of no other interference on the channel.

\[ P_{s_a} = 1 \]  
\[ P_{s_b} = 1 \]  

(6.38)  
(6.39)
Hence the mean slot time duration becomes

\[
\overline{T_{\text{slot}_a}} = (1 - P_{\text{ta}}) \cdot \sigma + P_{\text{ta}} \cdot (T_s + T_{\text{difs}})
\]  

(6.40)

\[
\overline{T_{\text{slot}_b}} = (1 - P_{\text{tb}}) \cdot \sigma + P_{\text{tb}} \cdot (T_s + T_{\text{difs}})
\]  

(6.41)

6.4.2. Full State Investigation

The heterogeneous traffic case results, for two stations, are shown in Figure 6.4, where the x and y axis are the load for each station in packets per second and the z axis is the effective mean service time. To the right of Figure 6.4, a 3D view is shown to determine which station perceives a higher effective mean service time, where STA \( a \) is coloured and STA \( b \) is white; hence in the coloured areas on the left top down view, STA \( a \) has a longer effective service time.

Figure 6.6 shows, from simulations, the expected mean service times for two stations. As shown, simulation suggests that the station with a higher load always experiences a lower service time, hence the current heterogeneous model does not match the behaviour of the protocol.

![Figure 6.4: 2 Station Heterogeneous Model showing Incorrect Behaviour](image)
Figure 6.5: 2 Station Heterogeneous Model with Incorrect Behaviour (Top-Down)

Figure 6.6: 2 Station Heterogeneous Simulation showing Expected Behaviour

Figure 6.7: 2 Station Heterogeneous showing Expected Behaviour (Top-Down)
6.5 Updating Transmission Probability from Slot Based to Time Based

6.5.1. Identifying the Problem

Previous models are based on a slotted behaviour, i.e. the probability of an event in each discrete slot unit. Transmission probability is based on a station transmitting for a single estimated mean time slot duration, \( T_{\text{slot}} \), out of the total slot times needed to access the medium and send successfully. In the homogeneous case, this discrete behaviour is acceptable as the mean slot time is equal across all stations; however, in a heterogeneous case, slot times across stations vary.

Take a station that has an extremely high load operating with another station with an insignificant load. The higher load station will almost always perceive the medium free and rarely experience collisions. The higher load station may always need to backoff due to always having at least one packet to send; however, the higher load station’s slot times will be insignificant (equal to the standardised slot time, \( \sigma \)) when perceiving an almost free medium.

![Figure 6.8: Highly Loaded Station with a mostly Free Channel](image)

Figure 6.8 shows the system diagram for the higher load station; previously, transmission probability was calculated by one busy slot, i.e. the one that represents the sending of a packet successfully, divided by the number of slots needed to access the medium, scaled by the probability having a packet to send.

\[
\tau_n = \Pr\{\text{transmit} \mid \text{packet to send}\}_n \cdot \Pr\{\text{packet to send}\}_n = \frac{1}{E[\text{slots}]_n} \cdot \rho' \quad (6.42)
\]

In this case, the transition probability would be

\[
\tau_n^{\text{slot based}} = \frac{1}{E[\text{slots}]_n} \cdot \rho' = \frac{1}{17} \cdot 1 = 0.059 \quad (6.43)
\]

by inspection, however, the proportion of time the station is transmitting is more closely represented as:
The 16 represents the mean number of slots the station will backoff; this is due being in saturation. The 50µs represents the mean slot time, which is the standardised slot time, σ, used to decrement of the backoff counter when the medium is idle. The standardised slot time is used because the channel is almost always free (P_{tr} ≈ 0) when shared with a single, low loaded station.

6.5.2. Updating Transmission Probability Discrepancy

In order to update the transmission probability to show a more realistic case, a timing approach is used. An augmentation from the homogenous case is the additional consideration of a station transmitting resulting in a collision. Essentially, the probability of transmitting successfully or transmitting resulting in a collision is found from the $E[\text{backoff time}]$ formulation.

$$p(T_s)_n = \left[ b_{-1} \cdot (1 - P_{tr}) \cdot (1 - P_n) \right] + \left[ (1 - b_{-1}) \cdot (1 - P_n) \right]$$  \hspace{1cm} (6.45)

$$p(T_c)_n = \left[ b_{-1} \cdot (1 - P_{tr}) \cdot P_n \right] + \left[ (1 - b_{-1}) \cdot P_n \right]$$  \hspace{1cm} (6.46)

The transmission time of each of the above events is found and divided by the total backoff time, giving a proportion of time a station transmits.

$$\tau^\text{time proportion}_n = \frac{\text{packet size}}{\text{data rate}}\left(\frac{16 \cdot 50\mu s + \text{packet size}}{\text{data rate}} + T_{difs}\right) \approx \frac{9\text{ms}}{800\mu s + 9\text{ms}} = 0.918$$  \hspace{1cm} (6.44)

This is then weighted by the utilisation, $\rho'$:

$$\tau_n = \rho' \cdot \frac{\left[ T_s \cdot p(T_s)_n + T_c^\text{other} \cdot p(T_c)_n \right]}{E[\text{backoff time}]_n}$$  \hspace{1cm} (6.48)
6.6 Calibrating the Model

Using the new time based transmission probability, a steady state service time is found for two interacting STAs. Results show that the station with a higher load achieves a lower service time, as does simulation. A problem, however, is that the service time is now not accurate for the saturated or homogeneous cases.

![Figure 6.9: Two Station Validation to Simulation](image)

![Figure 6.10: Two Station Validation to Simulation (Top-Down View)](image)

It is found that, with the new time based transmission probability; the correct behaviour of the higher loaded station having a lower service time is achieved. The new time based probability, however, now causes the state machines to iterate to an incorrect result,
which is known from comparison to the homogenous case. The iteration to the
incorrect results therefore gives erroneous expected mean service times in the case where
all stations’ transmission probabilities are equal.

To take into consideration the incorrect transmission time it is found that the model
needs to be calibrated to the saturated condition for a specific number of stations, $N$.
The calibration factor is represented as $X_N$, for which the value is fixed regardless of load
for a specific number of stations.

Recall the transmission probability for the saturated case

$$
\tau_{saturated} = \frac{1}{E[\text{slots}]} \quad (6.49)
$$

The new transmission probability should be equivalent in the saturated case ($\rho' = 1$)

$$
\tau_{time-based} = \rho'_n \cdot \left[ \frac{T_s \cdot p(T_s) + T_{other} \cdot p(T_{other})}{E[\text{backoff time}]} \right] = \left[ \frac{T_s \cdot p(T_s) + T_{other} \cdot p(T_{other})}{E[\text{backoff time}]} \right] \quad (6.50)
$$

therefore

$$
\left[ \frac{T_s \cdot p(T_s) + T_{other} \cdot p(T_{other})}{E[\text{backoff time}]} \right] \cdot X_N = \frac{1}{E[\text{slots}]} \quad (6.51)
$$

$$
X_N = \frac{E[\text{backoff time}]}{E[\text{slots}] \cdot \left[ T_s \cdot p(T_s) + T_{other} \cdot p(T_{other}) \right]} \quad (6.52)
$$

$$
X_N = \frac{T_{slot} \cdot \sum_k b_k \cdot S_k + T_s \cdot (1 - p) + T_{other} \cdot p}{1 \cdot \sum_k b_k \cdot S_k \cdot \left[ T_s \cdot (1 - p) + T_{other} \cdot p \right]} \quad (6.53)
$$

$$
X_N = \frac{T_s \cdot (1 - p) + T_{other} \cdot p}{1 + W / (1 - p) + \sum_{k=1}^{m-1} \left( 2^{k-1} \cdot W \cdot (1 - p) \cdot p^k \right) + 2^{m-1} \cdot W \cdot p^m} \quad (6.54)
$$
where $p$ is the collision probability for the saturated case of $N$ co-existing stations.

The Calibration Factor, $X_N$ is shown from $N=2$ to $N=100$ in Figure 6.11.

![Figure 6.11: Calibration Factor](image)

### 6.6.1. Regression Analysis

After the Calibration Factor was found, regression analysis was undertaken to determine a generic equation that could be used. Firstly, the calibration factor as a function of the number of stations being modelled, $N$, was found. The regression equations below show a cubic and 4\textsuperscript{th} degree order (usually the higher degree, the more accurate the function will be) and plotted in Figure 6.12 and Figure 6.13. The residuals of the regression analysis are shown in Figure 6.14, which effectively shows the discrepancy between the original calibration factor and the formula derived from regression analysis. Figure 6.14 shows that fairly high orders of $N$ are needed and residuals show that even then, the function is inaccurate, particularly for low numbers of stations.

\[
X_N \quad \text{(cubic)} = 1.15 \cdot 10^{-6} \cdot N^3 - 2.44 \cdot 10^{-4} \cdot N^2 + 0.019 \cdot N + 0.14 \quad (6.55)
\]

\[
X_N \quad \text{(4th)} = -2.30 \cdot 10^{-8} \cdot N^4 + 5.88 \cdot 10^{-6} \cdot N^3 - 5.60 \cdot 10^{-4} \cdot N^2 + 0.03 \cdot N + 0.1 \quad (6.56)
\]
Figure 6.12: Calibration Factor Regression Analysis for N (Linear Scale)

Figure 6.13: Calibration Factor Regression Analysis for N (Log Scale)

Figure 6.14: Calibration Factor Regression Residuals for N
Similarly, regression analysis was conducted to derive a calibration factor function from the saturated collision probability. Again two degrees of analysis are shown, in this case linear and quadratic, it is shown that the calibration factor is approximately equal to the saturated collision probability for a given number of stations.

\[ X_N (\text{linear}) = 0.98 \cdot p_N + 0.04 \]  
\[ X_N (\text{quadratic}) = 0.12 \cdot p_N^2 + 0.87 \cdot p_N + 0.06 \]

Figure 6.15: Calibration Factor Regression for Saturated Collision Probability

Figure 6.15 shows the comparison between the calibration factor and saturated collision probability and that \( X_N \approx p_{n_{\text{sat}}} \). It is concluded that \( p_{n_{\text{sat}}} \) may be used for the calibration factor, although the fully derived \( X_N \) should be used for more accurate analysis and as such will be used throughout this dissertation.

6.7 Validating the Heterogeneous Analytical Model

Shown in Section 6.2 is the derivation of \( p_{\alpha}, \text{Ptr}_{\alpha}, \) and \( P_{s_{\alpha}} \) for the uneven load case where each station has a different transmission probability, \( \tau_{\alpha} \). A correspondence to the even load case is shown to be equivalent. Here, the interacting state machine model and optimisation technique for deriving \( \tau_{\alpha} \) and \( \varphi'_{\alpha} \) are shown. Results are close to the homogeneous model.
A validation is shown in Figure 6.16 between the homogenous (with unlimited acknowledgment timeouts as shown in Section 5.6) and heterogeneous model. The black lines are the service times of the original homogenous model and coloured lines show the heterogeneous model with calibration for the corresponding number of stations.
7 MODELLING THE ACTIVITY FACTOR OF A CYCLIC SOURCE

7.1 Introduction

Chapters 4 to 6 show the development of an analytical model from a saturated wireless network, through to an unsaturated case, which is constrained to each station offering the network an equivalent load, through to a heterogeneous case in which stations can offer different traffic loads. The latter model allows modelling of scenarios where one station, namely an access point, can serve a much higher rate of traffic and therefore the model can now be used for more realistic scenarios. Take for example a voice call; a common scenario may be clients communicating between themselves, such as in an office environment (known as homogenous in this dissertation’s terminology) or conversely all calls may be destined for clients outside of the wireless domain. The latter scenario, assuming bi-directional calls on an infrastructure setup, means that one particular station, the access point, is effectively serving $n$ times the traffic, where $n$ is the number of stations (known as heterogeneous in this dissertation’s terminology).

Although different loads can be modelled, voice traffic has a significantly different traffic profile in comparison to Poisson. The traffic profile of voice is typically offering high data rates for part of the time and no traffic for another time interval – i.e. the talking and listening periods. The more sporadic nature of voice traffic, where a station offers a block of packets into the network and then remains idle for a time is known as an ON-OFF traffic source.

7.2 Model

To consider the ON-OFF nature of voice and other bursty traffic sources, a further augmentation of the model needs to be constructed. It is known that the ON periods can be described using the Negative Exponential distribution; i.e. Poisson like. The simplistic intuitive method is to find out how often, on average, a station is in each of the ON or OFF states. By finding the probability of a station being in either an ON or OFF state, the original Poisson model can be used for each of the permutations of stations being in either an ON or OFF case.

The probability a station is transmitting (i.e. in the ON State) can be found. The ON and OFF periods are the talk-spurt and silence properties of voice conversations.

$$p(ON) = \frac{T_{on}}{T_{on} + T_{off}}$$  (7.1)
Now take for example \( N = 3 \) stations (STA\(_a\), STA\(_b\), and STA\(_c\)), each station could be either in an ON or OFF period hence the combinations could be found (a total of \( 2^3 = 8 \)).

![Figure 7.1: ON / OFF State Possibilities](image)

The probability of \( k \) out of \( N \) stations being active, i.e. being in the ON period, can be found.

\[
p(k \text{ stations ON}) = \binom{N}{k} \cdot (p(ON))^k \cdot (1 - p(ON))^{N-k}
\]

(7.2)

Now, using the MAC layer analytical model, the estimated mean service, \( E[ST]_N \), time can be found where \( k \) stations are active with a Poisson Source of the ON data rate. By weighting each of the service times derived from the MAC layer models explored, with the probability of each occurrence of ON and OFF stations, an estimated mean service time can be derived.

\[
E[ST]_N = \sum_{k=0}^{N} p(k \text{ stations on}) \cdot ST(k)
\]

(7.3)

The presented solution works well in the lower load cases (\( \rho' < 0.9 \)). However, a problem occurs when a station can no longer serve all of its packets in the ON period. After a given load, where the time to service the burst of traffic is longer than the ON period itself, packets will remain buffered and the station will continue to need access to the network in the OFF period. The maximum service time before needing to buffer packets can be determined as the time of the ON period, \( T_{on} \), divided by the number of packets in the burst. Essentially, after maximum service time limit, the ON time appears longer from the perspective of other stations on the network. If the service time multiplied by the burst size in packets per second is greater than \( T_{on} \), a new \( T_{on} \) is needed. This new ON time, which now considers that some packets will be buffered and then transmit after the application has stopped bursting packets, is calculated. The updated time is derived by finding the number of packets remaining in the queue after the original \( T_{on} \), capped to the queue size, and multiplying through by the known service time.
\[ packets_{Unserved} = \max \left( 0, burstSizePPS - \frac{T_{ON}}{E[ST]} \right) \]  \hspace{1cm} (7.4)\\

\[ packets_{Buffered} = \min \left( queueSize, packets_{Unserved} \right) \]  \hspace{1cm} (7.5)\\

\[ T_{ON}^{NEW} = T_{ON}^{OLD} + E[ST] \cdot packets_{Buffered} \]  \hspace{1cm} (7.6)\\

The new ON time expresses the time it takes a station to clear all of its packets and hence be offering the network data. As ON time increases OFF time decreases until the station is saturated. With increasing ON time, the probability of offering the network traffic also increases, hence Equation 7.1 is re-evaluated. The model therefore iterates to a steady state – if the new ON time is greater than the one used in Equation 7.1 then the model is not at steady state.

Essentially, this means that when a station is unable to serve all of its packets within the standard ON time, the ON time and hence ON probability increases. The increase in ON probability means the estimated mean service time will also increase and hence the iteration will always cause the ON probability to iterate to 1. The model therefore has two modes of operation, which reflect that of the MAC layer models. Either the packets can be served in the application ON time or the station hits saturation; hence the sharp increase in service time at this point.

![Figure 7.2: Behaviour of the Proposed ON / OFF Model](image)

Figure 7.2 shows how the two modes of operation affect the buffering of traffic; either unsaturated and a near empty queue or saturated, in this case between 15 and 16 calls. The two model behaviour is similar to that of the MAC layer; saturated or unsaturated.
Figure 7.3 : ON / OFF Model Analysis

Figure 7.3 investigates the feasibility of the model. The number of stations is increased from 1 to 20 stations and the estimated mean service time is found with the model proposed and simulation. The solid green line represents the peak data rate of the traffic source; this is the load when a station is in the ON period. The peak data rate is the load used to calculate the estimated mean service time of an ON station. The solid blue line shows simulation runs (one for each number of stations) to be compared with the new proposed model, shown in red. For the feasibility study of this model, an alternative approach of calculating a mean data rate for each station is also shown; this mean data rate is calculated by taking the ON time data rate and scaling by the $T_{ON}$ (the dotted green line).

Initial results show that there is a good correlation when operating between 1 and 13 stations and it is also shown that the model will converge with simulation at high numbers of stations because this will be the saturated region. The area of concern is at the transition of modes between unsaturated and saturated; the two state ON / OFF model behaviour is reflected in Figure 7.2 with the sharp increase in service time at 16 stations. Note that the simulation does not show a sharp increase of service time as it flips between the two modes which is due to the queueing behaviour; the probability of a non-empty queue does not immediately change from 0 to 1 as seen when increasing load but has a more graceful degradation.
It is clear that the queue is affecting the performance of the network, as shown in the MAC layer models and in this Chapter and an investigation is needed. Furthermore, with the aim, set out at the beginning of the dissertation, of modelling a voice from the users’ perceived Quality of Experience, the impacts of buffering traffic is clearly needed.
8 IMPACTS OF BUFFERING TRAFFIC

8.1 Introduction

An unsaturated, heterogeneous model has been developed, which can model both Poisson and On/Off traffic profiles, giving the estimated mean delay for each STA or AP. Thus far, the MAC layer has only been considered in terms of delays; the analysis does not include any queueing delay, dropped traffic or delay variance. To incorporate these aspects, in order to quantify realistic traffic models and network usage, a queueing model must be introduced.

To develop the model to include the VoIP application level, the model must include end-to-end delay, variation and packet loss. The delay variance will then enable jitter to be measured, this is used to find the ‘out of contract’ packets which are those arriving outside a range where they can be used. This ‘out of contract’ window represents the de-jitter buffer on the receiver which has a limited capacity and is calculated as the mean delay plus or minus a de-jitter buffer.

The MAC layer model is not sufficient at providing the parameters needed as it does not consider the impacts of buffering traffic (at the sender and receiver). A combined MAC and queueing model is required.

8.2 Model

The challenge in modelling the total delays in the system, at the MAC layer and queue, is that the service time of the MAC is non-deterministic and cannot be estimated based on a standard distribution. This constraint means that traditional M/M/1 and M/D/1 models cannot be used. However, a formula can be derived for an individual discrete load / service time.
In this Chapter, several symbols are introduced and derived as pre-shown in Table 8.1.

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>$st$</td>
<td>Short Hand for the Estimated Mean Service Time</td>
</tr>
<tr>
<td>$a(k)$</td>
<td>Probability of $k$ Packets arriving within a Service Time</td>
</tr>
<tr>
<td>$\rho$</td>
<td>Queue Utilisation Probability</td>
</tr>
<tr>
<td>$\lambda_{PPS}$</td>
<td>Number of Packets arriving within a Service Time</td>
</tr>
<tr>
<td>$s(k)$</td>
<td>Probability of a Queue having $k$ packets Buffered</td>
</tr>
<tr>
<td>$u(k)$</td>
<td>An Intermediate Probability used in a Finite Queue case</td>
</tr>
<tr>
<td>$Td(k)$</td>
<td>Delay caused by Packets already in the Queueing System</td>
</tr>
<tr>
<td>$Ud(k)$</td>
<td>Delay caused by the Queue System State</td>
</tr>
<tr>
<td>$Bd(k)$</td>
<td>Delay caused by more than one Packet Arriving together</td>
</tr>
<tr>
<td>$j_{lower}$</td>
<td>The minimum Delay Bound used to determine if a Packet is not Admitted into the De-Jitter Buffer</td>
</tr>
<tr>
<td>$j_{higher}$</td>
<td>The maximum Delay Bound used to determine if a Packet is not Admitted into the De-Jitter Buffer</td>
</tr>
</tbody>
</table>

Table 8.1: List of Symbols for Queueing Analysis

8.2.1 Queue Size

By starting with the updated MAC model, proposed in Chapter 6, a service time, $st$, can be determined for a given load and number of stations, $N$. Load, $\lambda_{PPS}$, is expressed in packets per second, which is the reciprocal of the inter-packet time; i.e. the time between each packet arriving.

\[
st = unsaturatedMAC(N, \lambda_{PPS})
\]

\[
st_i = heterogeneousMAC(n = i, \lambda_{PPS_i}, \lambda_{PPS_i}, \ldots, \lambda_{PPS_{i-n-1}}, \lambda_{PPS_n})
\]

Packet arrival probability of each station is based on the Poisson distribution

\[
a(k) = Pr\{k \text{ arrivals in a service time}\}
\]
By normalising $s(t)$ to 1, an expected mean arrival rate, $E[a]$ and hence utilisation, $\rho$, can be found:

$$\rho = E[a] = \frac{1}{\lambda_{pps}}$$  \hspace{1cm} (8.4)

Now by deriving balancing equations, each queue state probability can be derived from the previous state equation, for example $s(1)$ from $s(0)$:

$$s(0) = s(0) \cdot a(0) + s(1) \cdot a(0)$$  \hspace{1cm} (8.5)

The empty queue state 0 can be reached from state 0 when there are no arrivals in a service time interval or state 1 (1 packet buffered) with no arrivals (the buffered packet is served).

$$s(1) \cdot a(0) = s(0) - s(0) \cdot a(0)$$  \hspace{1cm} (8.6)

$$s(1) = s(0) \cdot \frac{1 - a(0)}{a(0)}$$  \hspace{1cm} (8.7)

Then continuing this process gives a general equation:

$$s(k - 1) = s(0) \cdot a(k - 1) + s(1) \cdot a(k - 1) + s(2) \cdot a(k - 2) + \ldots$$

$$s(k - 1) = s(0) - s(0) \cdot a(0) + s(k - 1) \cdot a(1) + s(k) \cdot a(0)$$  \hspace{1cm} (8.8)

Which rearranges to:

$$s(k) = \frac{s(k - 1) - s(0) \cdot a(k - 1) - \sum_{i=1}^{k-1} s(i) \cdot a(k - i)}{a(0)}$$  \hspace{1cm} (8.9)

Hence, now all states can be found from $s(0)$, which in an infinite buffer case is simply the probability the system is empty:

$$s(0) = 1 - E[a]$$  \hspace{1cm} (8.10)
The model, however, needs to consider a finite buffer case to derive lost traffic and the above $s(0)$ representation does not consider this. Also the $s(X)$, where $X$ is the queue size, must be calculated. Due to one packet being serviced in each normalised service time slot the only way to arrive in state $s(X)$ is by the queue being empty, e.g. $X$ packets arriving as no packets are served. If the queue is non-empty then, as the packets arrive, one is serviced and therefore not arriving at $s(X)$. Hence the state equation is simply:

$$s(X) = s(0) \cdot A(X) \quad (8.11)$$

Where

$$A(X) = 1 - a(0) - a(1) - \ldots - a(k - 1) \quad (8.12)$$

However, $s(0)$ is still unknown, [PIT01] goes on to introduce a new variable $u(k)$, which is similar to the $s(k)$ formulation above, and a new $s(0)$ calculation is created to consider the finite buffer analysis:

$$u(k) = \frac{s(k)}{s(0)} \quad (8.13)$$

so

$$u(0) = 1 \quad (8.14)$$

Then

$$u(1) = \frac{1 - a(0)}{a(0)} \quad (8.15)$$

so

$$u(k) = \frac{u(k - 1) - a(k - 1) - \sum_{i=1}^{k-1} u(i) \cdot a(k - i)}{a(0)} \quad (8.16)$$

$$u(X) = A(X) \quad (8.17)$$
Therefore $u(k)$ can now be evaluated based on the state probabilities summing to 1:

$$\sum_{i=0}^{X} s(i) = 1 \quad (8.18)$$

Then

$$\sum_{i=0}^{X} \frac{s(i)}{s(0)} = \frac{1}{s(0)} = \sum_{i=0}^{X} u(i) \quad (8.19)$$

So

$$s(0) = \frac{1}{\sum_{i=0}^{X} u(i)} \quad (8.20)$$

Finally all queue states can be found and therefore queue size probability and the estimated mean queue size can be found.

$$s(k) = s(0) \cdot u(k) \quad (8.21)$$

$$Q_{\text{size}}(\rho, X) = E[Q_{\text{size}}] = \sum_{k=0}^{X} k \cdot s(k) \quad (8.22)$$

### 8.2.2. Loss

Loss Probability (LP) can also be calculated, which is the ratio between lost (offered - carried) and carried, hence:

$$LP(\rho, st, X) = \frac{\text{Offered} - \text{Carried}}{\text{Offered}} \quad (8.23)$$

$$LP(\rho, st, X) = \frac{E[a] - \rho}{E[a]} = \frac{E[a] - (1 - s(0))}{E[a]} \quad (8.24)$$

### 8.2.3. Delay

Now the delay may be found at the queue, again for the single load case. The total time $T_d$ is made up of the delay caused by other packets in the system (i.e. buffered), $U_d$ and
packets arriving at the same time, \( B_d \) as a given packet. \( U_d \) is based on the state of the system \( s(k) \).

\[
T_d = U_d + B_d \tag{8.25}
\]

\[
\Pr\{U_d = k\} = U_d(k) = s(k) \tag{8.26}
\]

\[
B_d(k) = \frac{1 - \sum_{i=0}^{k} a(i)}{E[a]} \tag{8.27}
\]

The total mean delay is calculated by convolving \( U_d \) with \( B_d \).

\[
T_d(k) = \Pr\{U_d = 1 \text{ and } B_d = k - 1\} + \Pr\{U_d = 2 \text{ and } B_d = k - 2\} + \ldots \tag{8.28}
\]

\[
T_d(k) = \sum_{j=1}^{k} U_d(j) \cdot B_d(k - j) \tag{8.29}
\]

The estimated mean queuing delay can therefore be found as a multiple of normalised service time intervals:

\[
delay(p, X) = E[Q\text{delay}] = \sum_{k=0}^{X} k \cdot T_d(k) \tag{8.30}
\]

And finally for a non-deterministic service time, such as that seen in WLAN, the mean delay of the system (queuing delay plus service time) can be found for a given load:

\[
Total\text{Delay}(N, \lambda_{pps}, X) = st(N, \lambda_{pps}) \cdot \text{delay} \left( \frac{\lambda_{pps}}{st(N, \lambda_{pps})}, X \right) \tag{8.31}
\]

With the constrain that if \( \rho \geq 1 \) then \( Total\text{Delay}(N, \lambda_{pps}, X) = st(N, \lambda_{pps}) \cdot X \)

The assumptions of this model are that \( E[a] = \text{Lambda} \), for example a Poisson traffic distribution, the packets are of fixed size and the queue is tail drop with a finite size, \( X \).
8.2.4. Jitter

The aim of this model is to derive values that can be used to evaluate a voice call. To fully evaluate a voice call, the jitter must be calculated. Jitter affects the number of packets arriving outside of a de-jitter buffer, which can be calculated as the mean delay +/- half of a de-jitter buffer. Packets arriving outside this range are not usable, even though they arrived at the receiver and considered out-of-contract (OOC). To derive out-of-contract, the probability of packet delay outside of the de-jitter buffer range needs to be calculated.

When finding \( T_d(k) \), a probability delay distribution is derived. By then calculating the number of service times the de-jitter buffer equates to, the probability of delay being outside this window can be found.

The de-jitter boundary is found, in integer multiples of service times. Half of the de-jitter buffer, expressed in multiples of service times, is then added and subtracted from the estimated mean delay to give a range of packet delays admitted into the de-jitter buffer.

\[
\begin{align*}
    j_{\text{lower}} &= \text{int} \left[ E[T_d(k)] - \left( \frac{st}{0.5 \cdot j} \right) \right] \quad \text{where } j \geq 0 \\
    j_{\text{higher}} &= \text{int} \left[ E[T_d(k)] + \left( \frac{st}{0.5 \cdot j} \right) \right] \quad \text{where } j \leq X
\end{align*}
\]

(8.32) \hspace{1cm} (8.33)

\[
OOC(j_{\text{lower}}, j_{\text{higher}}) = 1 - \sum_{k=j_{\text{lower}}}^{j_{\text{higher}}} T_d(k)
\]

(8.34)

It should be noted that due to the rounding process needed for the discrete model, results may vary from simulation. The rounding error is introduced because the delay variation is expressed in integer service time units; the jitter results should be tolerable at low service times; however as service time increases, the rounding will be more extreme therefore causing variation between the analysis and simulation.

8.2.5. Geometrically Approximated

The above method requires substantial calculations to arrive at a steady state solution. Another method which could be used is by approximating the offered traffic as a geometric distribution; in this method only ‘excess’ arrivals are considered, i.e. packets that need to be buffered, which simplifies the derivation [PIT01]. Such as Excess
Geometric Approximated Poisson Process (GAPP) model, the formula derived in the above reference for queue state and loss are shown directly as functions of $\lambda$.

$$GAPP_{-s}(k) = \left(1 - \frac{\lambda \cdot e^\lambda - e^\lambda - \lambda^2 + \lambda + e^{-k}}{\lambda - 1 + e^{-\lambda}}\right) \left(\frac{\lambda \cdot e^\lambda - e^\lambda - \lambda^2 + \lambda + e^{-k}}{\lambda - 1 + e^{-\lambda}}\right)^k$$  

(8.35)

$$GAPP_{-LP}(X) = 1 - \sum_{k=0}^{X} GAPP_{-s}(k)$$  

(8.36)

$$GAPP_{-LP}(X) = \left(\frac{\lambda \cdot e^\lambda - e^\lambda - \lambda^2 + \lambda + e^{-k}}{\lambda - 1 + e^{-\lambda}}\right)^{X+1}$$  

(8.37)

### 8.2.6. Results

The new non-deterministic service time queueing model was used to find the total system delay, drops and jitter and a comparison to simulation results. It is intuitive that the delay will be similar to the service time at low network utilisation and when the network is saturated, the delay will be approximately equivalent to the queue size multiplied by the service time.

The Bianchi scenario parameters are used for comparison to simulation. The Bianchi parameters are; a Poisson (equivalent to negative exponential in simulation) traffic source, a fixed packet size of 8184 bits and a tail-drop buffer of length 50 packets. Simulation samples were taken over different fixed load values, however, a continuous line is used in the graphs for ease of comparison.
Figure 8.1: Queueing Model Delay compared with Simulation

Figure 8.2: Queueing Model Loss compared with Simulation
Figure 8.3: Queueing Model Out-of-contract compared with Simulation

The results illustrate, in Figure 8.1, Figure 8.2 and Figure 8.3, that the delay, loss and out-of-contract are close to simulation in the M/G/1 model proposed. The discrepancy between the model and simulation are due to (a) the queueing model assuming a deterministic service time for each load, whereas the simulation results are obtained from a mean service time of all packets, (b) the discrete nature of the model compared to the time-based simulation and (c) the de-jitter buffer being an integer multiple of service time in the model.

Queueing models normally cannot be used for a utilisation of greater than 1, which holds true for the GAPP model. In this study, it is found that close approximations can be found from the model where $\rho > 1$; this is therefore the chosen model.

This new model then gives the total system delay, including the queue and non-deterministic service time, loss and out-of-contract not currently shown in the literature. Due to the comparison between a discrete queueing model and continuous simulation, there are discrepancies; however, it is concluded this is accurate enough to capture the behaviour of the buffering packets.
9 Exercising the Analytical Model from a QoE Perspective

9.1 Introduction

In this Chapter, we begin by showing the methodology used in order to use the analytical model and perform simulation. The analytical model has been augmented in many stages including the final heterogeneous model, the VoIP model to allow for bursty traffic and a queueing model. These individual components can be connected together depending on the scenario in question. The simulation technique is then shown, which was ran on the UK’s supercomputing services HPCx and HECTOR.

This chapter then goes on to introduce the method of deriving QoE using the ITU-G.107 E-Model and the framework built around it to allow for state space explorations to be presented on a single QoE contour plot.

Using the QoE methodology, a IEEE 802.11 scenario, with an AP and an increasing number of stations, is explored in order to benchmark the number of calls the network can serve where all the users perceive an acceptable QoE.

Using the QoE contour plot methodology, that allows for state space exploration, a dimensioning study is undertaken to demonstrate the analytical model at scale and increase the number of acceptable calls the can be served. The method used to increase call capacity is to send more data in one transmission and hence decrease the contention time overhead at the MAC layer. This dimensioning study is compared to the simulation ran on the super computers.

Finally, this chapter demonstrates the ‘what-if’ capabilities of the analytical model by showing how the dimensioning studies, and point at which the most acceptable calls can be served, with additional delay added to represent traffic destined for somewhere outside of the wireless domain being modelled.

9.2 Computation Iteration Technique

This dissertation proposes many augmentations to the Bianchi MAC layer model to allow results to be quantifies as user perceived Quality of Experience. In order to achieve this, the analytical models have been augmented in discrete sections throughout the dissertation, firstly identifying the issues and proposing novel MAC layer models and then using these new models to produce results more commonly seen by network administrators. Augmentations include the modelling of On/Off traffic, such as VoIP traffic with the Activity Factor model, a queueing model to show performance on an end-to-end bases with the additional performance degradation due to buffering traffic and finally the ITU-T G.107 E-Model to derive results for Quality of Experience. The
final model was then used in a methodology which allowed multiple exploration state spaces to be analysed in a single, yet very powerful contour plot. In this section the models computational techniques are explored, specifically how the MAC layer models iterate to a steady state and the models themselves, which have been presented in this dissertation in Sections, are explored in a form which shows how they can be linked together to provide different solutions for wireless modelling.

9.2.1. Connecting the Analytical Models

There have been several MAC layer analytical models which have been explored and derived in this dissertation, the first being the Bianchi Saturated case, this was then augmented to derive results for stations which are not operating at saturation and finally how a group of stations, potentially with very different traffic profiles on each was derived, specifically that stations could operate an On/Off traffic source such as VoIP. The dissertation then went on to show performance degradation when buffering packets and analysed results by deriving the human perceived quality of experience.

Figure 9.1 shows how the analytical models link together in a scenario where all stations are operating with a homogeneous traffic profile of the same data rate. As shown, the homogeneous analytical model is used in the first block and the bold inputs show what parameters are required to set-up the scenario; in this case the mean number of packets per second, number of station and in the queueing model, the number of packets which can be buffered. In this simple example, the output of the homogeneous model is the estimated mean service time and utilization of the stations; this is the same for all stations. The queueing model takes these inputs and determines the queue capacity and hence delay, the delay variance and proportion of packets that will be lost. The final Quality of Experience methodology block, comprised in part by the ITU-T G.107 E-Model then outputs a single R-Factor (as all stations are contributing and perceiving the same traffic in the homogeneous case).

![Analytical Model Connectivity: Homogeneous Poisson](image)

*Figure 9.1: Analytical Model Connectivity: Homogeneous Poisson*

In Figure 9.2, the homogeneous model is again shown but where traffic is using an On/Off pattern, in this case, the maximum number of packets per second is input and
results are presented in the form of the number of stations being on; a mean service time when 1, 2, ... N stations are transmitting. The activity factor then takes in the results where n stations are transmitting and weights that against the probability of each number of stations transmitting at the same time, defined by the on and off time input and results in an estimated mean service time and utilization. Again the queueing model takes in the estimated mean service time and utilization a above and through the ITU-T G.107 E-Model outputs an R-Factor.

![Analytical Model Connectivity: Homogeneous On/Off](image1)

**Figure 9.2: Analytical Model Connectivity: Homogeneous On/Off**

In the scenario where stations are operating with different traffic profiles, say where there is an AP transmitting more data the STAs, an estimated mean service time for each station is output along with a utilisation for each station. Figure 9.3 shows this heterogeneous Poisson case with the queueing and ITU-T G.107 E-Model; the queueing and ITU-T G.107 E-Model are ran for each station and the QoE methodology proposed can be used to show an R-Factor per station.

![Analytical Model Connectivity: Heterogeneous Poisson](image2)

**Figure 9.3: Analytical Model Connectivity: Heterogeneous Poisson**

### 9.2.2. Building the Heterogeneous Analytical Model

The analytical model was implemented in National Instruments LabVIEW v8.6 (the author is a LabVIEW Certified Associate Developer), a graphical programming language. LabVIEW allows for a state machine programming paradigm; this is used to represent each station and initial guess-time tariffs, $r_{\text{iteration } 0}$ and $\rho'_{\text{iteration } 0}$, based on a pre-execution study carried out in Section 9.2.3, for $r$ and $\rho'$, for all $n$. Iteration of each $N$ state machines
gives a new value of $\tau_n$ and $\rho'_n$ for each station and an optimisation process is used to adjust the parameters and run all of the stations’ state machines again. The iteration process is continued until the absolute difference between the inputs and outputs is less than $10^{-6}$; this is the criteria for identifying steady state.

The optimisation simply involved taking the mid-point between the input and output for all $\tau_n$ and $\rho'_n$. The probabilities for $p_n$, $P_{tr}$ and $P_s$ are recalculated before entering the next block of N iterations.

All state machines converge at a steady state approximately between 20 and 50 iterations for all attempted saturated, homogeneous and heterogeneous load cases throughout the dissertation. The system used to automate the solving of the analytical mode had the relevant process to report the number of iterations needed and report an error on more than 200 iterations.

### 9.2.3. Pre-Execution Optimisation

Due to the iteration process, and that initial guess-imates ($\tau_n^{\text{iteration 0}}$ and $\rho'_n^{\text{iteration 0}}$) are given to the program, an exploration of the initial input is shown.

In order to decrease the number of iterations, the initial input for $\tau_n$ is shown against the number of iterations to reach steady state. The scenario being used is the 5 station homogeneous load case with a number of packets per second of 14; this is close to saturation. The expected $\tau$ is 0.029174 and $\rho'$ is 0.293411, $\rho'_n^{\text{iteration 0}}$ is set to 0.3 and $\tau_n^{\text{iteration 0}}$ ranges between 0.01 to 0.9.

![Figure 9.4: Initial $\tau_n^{\text{iteration 0}}$ vs. Total Iterations Needed](image_url)
Figure 9.4 shows that the number of iterations is lower when the estimate is nearer the output value, as expected. In most scenarios, however, this is not known.

![Figure 9.4: Maximum \( \tau \) at Saturation](image)

In Figure 9.5, it is shown that the range of \( \tau^{\text{iteration 0}} \) at saturation, is between 0.0556 and 0.0137 for \( N=2 \) to \( N=100 \). A value is of the initial \( \tau^{\text{iteration 0}} \) is chosen to be 0.03 as a fairly close initial guess for low \( N \) values (<30), the most common scenarios, but still suitable for the extended range. Utilisation can vary between 0 and 1 and therefore any initial mid-range guess-estimate would be valid; the value of 0.3 remains in use.

### 9.2.4. Iterating to Steady State

Now the initial inputs to the model have been derived, the model needs to iterate to find a steady state; this is where the outputs of the models match the inputs; in this case the \( \tau \)'s and \( \rho \)'s. In order to derive the inputs for the next iteration, when the model has not settled to steady state, the median of the current input and output is obtained. Figure 9.6 shows the process of iteration.
All the MAC layer models in this dissertation require initially seeding with input parameters $\tau$, transmission probability and $\rho$, utilisation. The initial seeding is also required when there are multiple stations; each state machine for each station needs seeding. In order to iterate to a steady state, the median is taken from the output $\tau$ and $\rho$ and the current input $\tau$ and $\rho$. It is found that, throughout all the models tested, the model never showed signs of iterating way from a steady state, i.e. example iterating continuously and never resolving. The iterating away from steady state was checked by monitoring the number of iteration each state machine made before reaching steady state. To check that a state machine didn’t more than 100 times, a system was designed to flag
a warning to the user was implemented. The system was never triggered while producing results for the dissertation.

9.3 Simulation Technique

In this dissertation, the simulations are used to validate the analytical models as they are derived and for comparisons in the later ‘what-if’ scenarios which explore large parameter state spaces, for example different number of stations and different packet sizes.

To derive results, simulations have written using NS-2 and have been run on UK national supercomputing services HPCx and HECTOR. The NS-2 simulations, with modification to the underlying reporting package, output a file per traffic source which holds every packet’s sent and received time. Once a single simulation has been run, a post processing step is run, which is made up of several Perl scripts which read all of the files produced by NS-2 and use the raw packet sent and receive times to derive a mean end-to-end delay, proportion of packets lost (those packets which do not have a receive time logged) and throughout.

The ITU-T G.107 E-Model is also written as a Perl script and is a sub-section of the overall Quality of Experience methodology presented in this dissertation. The final output is typically several CSV files (a file holding a table of values) which show the summaries of delay, loss, jitter, R-Factor, contour plots for traffic sources, stations and overall. The entire process can be seen in Figure 9.7. Figures which show plots are created by using MATLAB 2009 to read in the files and, with a little manual processing of axis labels, scales and colour scheme several FIG files.

Table 9.1 shows the simulation parameters used for each single simulation including the overall time which is comprised of a Start-up time, Capture Time and Cool Down Time. The Start-up time is used to ensure that the network has reached a steady state, that is where the mean delay, loss and jitter is approximately constant; at the beginning of the simulations queues build and results from the first minute can be variable and not represent the true performance over a long run. At the beginning of the Start-up time,
each of the traffic sources is started over a 10 seconds period. The start-up process involved randomly choosing a time to start each traffic source within the 10 second period ensures that the traffic sources do not become synchronised, for example all stations sending and waiting at the same time.

The next phase of simulation is the capture time; this is where the simulation begins logging packets send and received times in order to derive the mean end to end delay, loss, jitter and R-Factor. The final phase is the cool down period where all the traffic sources are switched off, yet we still capture packet activity which in this case will be packets being received; the cool down period is to ensure we don’t artificially report losses due to not waiting for packets to be received after the traffic has been terminated.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Time (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simulation Time</td>
<td>2500</td>
</tr>
<tr>
<td>Start-up Time</td>
<td>1000</td>
</tr>
<tr>
<td>Capture Time</td>
<td>1000</td>
</tr>
<tr>
<td>Cool Down Time</td>
<td>500</td>
</tr>
<tr>
<td>Batch Repeats</td>
<td>32</td>
</tr>
</tbody>
</table>

*Table 9.1: Simulation Parameters*

In order to ensure results are as accurate as possible, all simulations are repeated 32 times with a different random number generator seed; this is to ensure that the results are not a function of a specific seeding. To derive results for a single data point we then average the mean delay, loss, jitter and R-Factor to produce final results. Note that another approach is to take a mean of all packets end-to-end time and apply a weight of the number of packets; however, due to the long simulation time we find there is no difference.

Simulations for a single point, over the batch repeats, typically consist of 17,000 packets per station as shown in Table 9.2. The contour plots produced in the homogeneous and heterogeneous case and with both Poisson and On/Off traffic models and required over 6 billion packets to be simulated. The large number of packets simulated indicates (a) the motivation for using the UK’s national super computing services and (b) that the analytical model can derive these results far more computationally cheaper.
Interestingly, all the papers dealing with increasing Voice over WLAN (VoWLAN) capacity or using Quality of Service (QoS) techniques to provide better service, choose almost arbitrary criterion values such as a maximum delay and loss tolerance such as in [WAN05a]. Although choosing such values is considered acceptable, having a cut off of a certain delay or dropped amount does not take into consideration other important factors. VoIP for example can tolerate a slightly higher delay if there are minimal drops and conversely an extremely low delay may counter more drops. Another major factor is the issue of jitter, the packet delay variation from the mean; this measure allows for calculation of packets which may arrive but outside a window where they can be played back (even with a large client side buffer). In this dissertation, jitter is calculated by finding the number of packets that arrive outside of the mean delay +/- half of a defined de-jitter buffer (150ms throughout). Packets arriving outside of this range are effectively lost as they cannot be held in the finite de-jitter buffer and are known as out-of-contract. The de-jitter process is shown in Figure 9.8.

![Figure 9.8: Calculation of Jitter](image)

To quantify a VoIP call in terms of acceptable or unacceptable quality and the deviation from these, some papers have explored the use of the ITU-T G.107 E-Model [EM00] such as “Voice over IP Performance Monitoring” [COL01] and “Assessment of VoIP Network Using an R-Value Embedded Simulation Model” [TAN04] which combine packet delay and loss\dropped results into a single value which rates the user's perceived

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Time (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean Number of Packets Captured per Station</td>
<td>17,000</td>
</tr>
<tr>
<td>Packets Simulated for a single point</td>
<td>544,000</td>
</tr>
<tr>
<td>Packets Simulated for a Contour Plot (Homogenous)</td>
<td>1,018,368,000</td>
</tr>
<tr>
<td>(2 to 25 Stations = 234, Packet Sizes = 8)</td>
<td></td>
</tr>
<tr>
<td>Packets Simulated for a Contour Plot (Heterogeneous)</td>
<td>2,036,736,000</td>
</tr>
<tr>
<td>(Add in AP transmitting equivalent number of packets)</td>
<td></td>
</tr>
</tbody>
</table>

Table 9.2: Simulation Statistics
quality. This call rating, known as the R-Value, or R-Factor, is the one chosen here to report user experience and find acceptable call network carrying capacity.

### 9.4.1. R-Factor

Highlighted earlier was the need to quantify a call as being of either acceptable or unacceptable quality without choosing arbitrary criterion for delays and drops etc. To provide a QoE measure for VoIP, the E-Model, defined in ITU-T G.107 [EM00], is used. The ITU-T G.107 E-Model encapsulates the voice performance measurements, such as delay, loss and jitter, to produce a quality rating between 0 and 100 called the R-Factor. The R-Factor summarises the users’ perceived experience; [EM00] states that a value below 60 can generally be said to be unacceptable and over 70 is said to be of Public Switched Telephone Network (PSTN) quality. The standardised R-Factor is given below.

\[
R = R_o + I_s - I_d - I_e + A
\]  

(9.1)

Where, \( R_o \) represents the signal to noise ratio between the sender and receiver, including the impacts of noise in the room. \( I_s \) represents the impairments of the voice signal, such as that cause by the microphone quality. \( I_d \) is delay factor and \( I_e \) is the equipment impairment factor caused by encoding or caused during transmission.

In this dissertation, the impacts of delay and loss will be investigated, hence components of \( I_d \) and \( I_e \) respectively. The loss is comprised of packets dropped in the network or which arrive either too late or too early to be served, known as out of contract. The ITU-T G.107 E-Model goes on specify default values for those components which make up \( R_o, I_s \) and \( A \) to give

\[
R = 94.2 - I_d - I_e
\]  

(9.2)

which was later modified, in the year 2000, to

\[
R = 93.2 - I_d - I_e
\]  

(9.3)

Derived from regression analysis [COL01] and validated [TAN04], the G.711 specific R-Factor formula is found, including the delay component, \( I_o \) and loss, \( I_s \) component formulas for the packet switch protocol. Note the papers use the 94.2 base R-Factor value.

\[
R(G.711) = 94.2 - I_d - I_e
\]  

(9.4)
\[ I_d(G.711) = 0.024d + 0.11(d - 177.3)H(d - 177.3) \]  \hspace{1cm} (9.5)

Where \( d \) is delay in milliseconds and \( H(x) \) is a Heaviside function

\[
H(x) = \begin{cases} 
0, & \text{for } x < 0 \\
1, & \text{otherwise}
\end{cases}
\]  \hspace{1cm} (9.6)

\( I_e \), again specified on a per codec basis, is shown below:

\[ I_e(G.711) = 30\ln(1 + 15l) \]  \hspace{1cm} (9.7)

Where \( l \) is the proportion of packets lost or arriving at the receiver but being ‘out of contract; essentially, those packets arriving too late or too early to be useful. The proportion of packets out of contract is calculated as those packets with a delay outside of a de-jitter range. The de-jitter range is the mean packet delay +/- half the de-jitter buffer.

Figure 9.9, shows the mapping between overall loss, which includes the out-of contract packets (those effective loss from jitter) and delay. The contours show the values of the R-Factor and a knee at 177.3ms shows where the Heaviside function takes effect in the delay component.
9.4.2. Mean Opinion Score

In the telephony community, a parameter commonly used is the Mean Opinion Score (MOS). MOS scores were originally obtained by having volunteers listening to a phone call and rating it from 1 to 5, where 1 is very poor quality and 5 is excellent. Hence, it is important to be able to show results in this format for industry. The R-Factor can be converted to MOS with the formula below [EM00].

\[
R \leq 6.5 \quad MOS = 1
\]  \hspace{1cm} (9.8)

\[
6.5 \leq R \leq 100 \quad MOS = 1 - \frac{7}{10^3} R + \frac{7}{6.25 \cdot 10^3} R^2 - \frac{7}{10^6} R^3
\]  \hspace{1cm} (9.9)

\[
R \geq 100 \quad MOS = 4.5
\]  \hspace{1cm} (9.10)

The conversion is visualised in Figure 9.10 with generalised descriptions.

Figure 9.10 : R to MOS Function [EM00]

9.5 Benchmarking VoIP QoE

In this chapter, an investigation into the performance of IEEE 802.11 with the full analytical model and simulation is undertaken. The aim is to determine the maximum number of acceptable VoIP calls, where acceptability is based on Quality of Experience (QoE). Unlike the majority of papers shown in Section 3.3 which use fixed delay or loss criterion, the employment of the ITU-T G.107 E-Model is used to give the first insights
of (a) the actual number of calls which can be served determined by QoE and (b) the QoE when an AP exists.

As stated in the literature review, some papers [WAN05a] explore voice performance based on inflexible delay and loss thresholds to determine if a call is acceptable or not. This method means that there is no flexibility in quantifying a call, for example, a call that experiences higher loss and lower delay may still be acceptable. The more flexible acceptability criterion of, for example higher loss and lower delay, is shown in the contour plots produced in Chapter 9, Figure 9.9 when analysing the ITU-T G.107 E-Model.

With the use of R-Factors, a more realistic number of calls can be found from the various components affecting a user’s call experience. Firstly, a benchmarking of QoE is found in the homogeneous case, i.e. where there are only stations communicating on an ad-hoc basis and secondly, where a single AP provides bi-directional call capability. Due to the AP having a higher load \((N \times STA \text{ load})\) increased delay and loss are expected and therefore the two directions of the bi-directional call QoE (to and from the AP) are plotted separately.

### 9.5.1. Homogeneous Ad-Hoc Benchmark

In this section, 1 to 25 STAs have a single G.711 voice source. As the number of STAs increase, end-to-end delay, effective loss and R-Factor are measured.

![Figure 9.11: Homogenous On/Off End-QtoQEnd Delay Distribution](image)

```
Figure 9.11 : Homogenous On/Off End-to-End Delay Distribution
```
Figure 9.11 shows the mean end-to-end delay across all STAs as the number of contending STAs increase. The figure also shows the 10\textsuperscript{th} through 90\textsuperscript{th} percentiles in order to show the spread of delay results. Figure 9.12 then goes on to show the percentile values verses delay, for each number of STAs (in colour groups of 5 STAs). To produce this figure, the delay values are sorted in ascending order and the values of delay are referenced at the 10\textsuperscript{th} through 90\textsuperscript{th} percentile as well as 99\textsuperscript{th}, 99.9\textsuperscript{th} and 99.99\%. Note that the first group of 5 stations show a vertical drop due to not having enough data.
Figure 9.13 shows the mean proportion of packets dropped from the STA's buffer, the proportion of packets arriving but either too late or too early, i.e. Out of Contract (OOC) and the resulting total proportion of packets effectively lost. OOC is calculated by finding those packets arriving outside of a range of the mean delay, in this case 75ms away from the mean delay; representing a 150ms de-jitter buffer in the receiver. Note the sudden increase in loss between 14 and 15 STAs, due to the increase of jitter and hence OOC packets, is caused by the high variability of the number of packets held in the queue. In the lower number of STA case ($N < 10$), most packets are served immediately and all packets are served before the next burst of packets, hence lower values for dropped and OOC. On the other hand, at large numbers of STAs, the queue is almost always full and hence an increase in mean delay, but decrease of delay variability. It is important to note that when the queue is almost always full, although delay increases, the delay variance decreases for those packets admitted into the buffer and hence the drop in OOC.

Finally, Figure 9.14 shows the resulting R-Factor for the On/Off Ad-Hoc case. Taking an acceptability R-Factor of 60, it is shown that 13 streams can be served; however, if a voice call is made up of two streams to form a bi-directional call, the network can only serve 6 calls where all users perceive an acceptable QoE.

Figure 9.15 shows the same scenario but where the traffic is modelled using the Poisson distribution, but having the same mean data rate as the On/Off source. In this figure, the sharp drop shows the more pronounced flip from the unsaturated to saturated mode.
of operation, whereas the On/Off has a slightly more graceful degradation, except it starts earlier. The later but sudden drop means 17 streams can be served (8 bi-directional). However, network operators would want to avoid this due to the extent of the ‘cliff.’

Figure 9.14: Homogenous Simulation: On/Off QoE R-Factor

Figure 9.15: Homogenous Simulation: Poisson QoE R-Factor
9.5.2. **Heterogeneous Access Point Benchmark**

One of the main novelties of this dissertation is the derivation of a heterogeneous model. With the heterogeneous model, different nodes (an AP or STA) can offer the network different traffic loads and therefore allowing realistic real-world scenarios. In this chapter, the same scenario is taken as in Section 9.5.1 but where an AP has a voice stream back to each STA and therefore all bi-directional voice calls. In infrastructure mode, this may represent either a voice call with someone outside of the local wi-fi network, or even a STA to STA call, recalling that in infrastructure mode STAs cannot communicate directly between themselves.

![Figure 9.16: Heterogeneous Simulation: On/Off End-to-End Delay Distribution](image-url)
Figure 9.16 shows the mean end-to-end delay of the stations as well as the mean delay of all traffic being served by the AP. Again percentiles are shown to highlight the spread of delay results. It is shown that the AP’s delay is higher due to the additional number of packets it needs to serve in order to provide \( n \) bi-directional traffic sources. Figure 9.17 shows the effective loss, which comprises of dropped packets and those deemed as out of contact.

Figure 9.18 shows the two components of loss, dropped and OOC, have been broken down per direction. From the results, it is shown that the delay and loss are higher from the
AP, due to the higher utilisation, and therefore more packets on average are buffed in the queue. Note that the fluctuations of the dropped results for the STAs, between 1 and 10 stations, is caused by single packet losses being exaggerated on the log scale.

In this scenario, again it is interesting to note that when the queue is full and there are packets being dropped, the proportion of out-of-contract packets is reduced. The reduction is due to the queue being full therefore any admitted packets will have a similar delay (MAC delay x Queue Size), therefore reducing the delay variance; i.e. packet delays become closer to the mean and are admitted into the de-jitter buffer at the receiver.

Figure 9.19 : Heterogeneous Simulation : On/Off QoE R-Factor Derivation
Figure 9.20: Heterogeneous Simulation: On/Off QoE R-Factor

Figure 9.21: Heterogeneous Simulation: Poisson QoE R-Factor Derivation
Figure 9.19 illustrates the conversion from the delay and effective loss metrics into R-Factor using the ITU-T G.107 E-Model contour plot investigated earlier. Although the contour plot is not convenient to read values for specific number of stations, it can be used to show why the R-Factor is degrading. The results show that the AP’s R-Factor is degrading due to increasing number of packets lost while the STA’s performance degradation is due to both increasing delay and then, at saturation, increasing loss.

In Figure 9.20, again the two directions of traffic flow are plotted. From this graph, using a criterion of R≥60 as described earlier, an acceptable number of calls is shown to be 6, note that the lower R-Factor must be taken to ensure acceptable quality in both directions.

Figure 9.21 again shows the R-Factor on the contour plot for the Poisson case; it is shown that in the Poisson case, both the AP and STAs degradation is caused by a similar increase in delay and loss. The similarities are apparent because of the more pronounced flip from saturated to unsaturated modes. Figure 9.22 shows the sharper drop of R-Factor for the STAs but occurring at a slightly higher number of calls than the On/Off source. The Poisson case can also serve a maximum of 7 acceptable calls.

In this chapter, benchmarking analysis has been conducted to determine the maximum number of acceptable calls which can be simultaneously made in IEEE 802.11. Table 9.3 summarises the maximum number of bi-directional calls which can be made, where each traffic flow direction (from or to the AP) must achieve an R-Factor greater or equal to 60.
9.6 Dimensioning VoIP QoE

9.6.1. Packetisation Period

As discussed in Chapter 9.5, the protocol behaviour means there can be a large time overhead to consider when sending a packet. In this study, also shown in Hershey, Pitts and Shepherd [HER10], the aim is to explore network behaviour (dimension) for voice calls by increasing the packetisation period in the voice codec. By increasing the packetisation period, packets which contain more data can be sent less frequently, therefore decreasing the time overhead per bit ratio. However, there is a trade-off as there will be an additional delay introduced by the longer packetisation intervals.

This dimensioning study presents results from the analytical models and simulation to further investigate the accuracy of each the models under different scenarios. Simulation is achieved using NS-2 running on HECToR, a UK national supercomputing service, 2 to 25 voice calls are simulated in the wireless network. Each STA has a single unidirectional G.711 call operating at a mean data rate 29.37kbps with varying packetisation intervals from 5ms to 200ms; equivalent to packet sizes from 486b to 17,120b, respectively. Traffic is modelled as negative exponential (Poisson equivalent) and as an On/Off source with the On/Off times shown in Table 1. The MAC data rate is set to 1Mbps and a finite buffer size of 25 packets is used. Each simulation ran until steady state and batch repeated 32 times; more details are shown in Section 9.3. The results shown (for the two traffic profiles in both homogenous and heterogeneous) required 52,224 simulations with approximately 6,500,000,000 packets simulated. Without the use of HECToR, a standard desktop PC would have taken two years at an average of 20mins per simulation and hence the need for an analytical model is clearly crucial. The augmented analytical model takes 20 minutes for the full set of results, hence a 13,000 times faster compared with simulating on a PC.

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Maximum AP can Support</th>
<th>Maximum Stations</th>
<th>Maximum Bi-directional</th>
</tr>
</thead>
<tbody>
<tr>
<td>Poisson without AP</td>
<td>N/A</td>
<td>17</td>
<td>8</td>
</tr>
<tr>
<td>ON / OFF without AP</td>
<td>N/A</td>
<td>13</td>
<td>6</td>
</tr>
<tr>
<td>Poisson with AP</td>
<td>7</td>
<td>15</td>
<td>7</td>
</tr>
<tr>
<td>ON / OFF with AP</td>
<td>6</td>
<td>11</td>
<td>5</td>
</tr>
</tbody>
</table>

Table 9.3: Maximum Number of Acceptable Calls
9.6.2. **Impacts on the Homogeneous Case**

The homogeneous case is investigated first in which each STA has a single unidirectional G.711 voice call. The scenario represents an ad-hoc usage of the network where there are no APs being used. The service time is found from the homogeneous MAC layer analytical model and converted to the users perceived Quality of Experience with the Activity Factor Model, the Buffering analysis and ITU-T G.107 E-Model. A system model is shown in Figure 9.23 to highlight the inputs and outputs of each model and how the analytical models are linked together. The simulation return end-to-end delay of each packet and proportion of packets not received. The effective loss is calculated from the per packet delay results and the QoE is also found using the ITU-T G.107 E-Model. Both an On/Off voice traffic profile is used as well as a Poisson equivalent in which the mean number of packets sent in the cyclic source is used in an always-on traffic profile.

![Analytical System Model Overview](image)

The mean QoE from the analytical model and simulation are plotted as contour plots for both traffic profiles, in which the acceptable quality of 60 is highlighted with a bold yellow line.

![Optimising Homogeneous: Analytical Model for Poisson Sources](image)
Figure 9.25: Optimising Homogeneous: Simulation for Poisson Sources

Figure 9.26: Optimising Homogeneous: Analytical Model for On/Off Sources
Figure 9.24 and Figure 9.25 show the analytical model QoE and simulation of the Poisson source, respectively and Figure 9.26 and Figure 9.27 the analytical and simulation results of the On/Off source. The 60 R-Factor acceptable user perception score is highlighted for clarity. The contour plots allow a huge amount of data to be displayed in a single graph including the extent of degradation as the number of stations increase; i.e. the spacing between the contours.

All figures show that with the small packet sizes (<500bits) only a few calls can be served (<7); this is due to the high time overhead caused by the contention at the MAC layer.

Conversely, when the packet sizes are large (>9000bits) and packetisation period is long, each packet already has a long delay before it enters service, hence a lower QoE. The results show a similar optimal point of acceptable call carrying capacity at approximately 8,000bits.

The point at which most stations are served is found where the packetisation period (shown on the right hand side of the figures) is approximately to be 100ms. The 100ms packetisation period is equivalent to sending 3.5 packets per second to achieve the same mean data rate (mean data rate divided by packet size of 100ms of encoded speech; 29.37kbps / 8560bits). At the 100ms point, over 20 calls can be served compared to the 12 calls at the standard 20ms level; a 66% increase.

In the Poisson case, the drop between acceptable and extremely poor perceived quality occurs within a few calls and clearly shows that in fact the network should not be
operating at a 60 R-Factor due to the risk of a sharp decrease if utilisation increases only slightly.

The On/Off case, on the other hand, shows a more graceful degradation as shown by the contour spacing difference between Figure 9.25 and Figure 9.27 as the off period allows the network to catch up with itself and, at low loads, return to an idle queue empty state before the subsequent burst. When the load increases and subsequent bursts arrive at non-empty buffers, the performance begins to degrade.

### 9.6.3. Impact on the Heterogeneous Case

In this chapter, the same methodology for generating the results is used as in Section 9.6.2 except that in these scenarios, an AP is used and modelled with the Heterogeneous Case Model. All STA traffic is destined for the AP and the AP has a voice source to each STA, therefore bi-directional voice calls. The ad-hoc case above showed an increase of 12 to 20 calls; however this was based on 1 call per STA. If STAs were communicating in pairs (therefore bi-directional calls) this would be equivalent to 6 and 10 calls respectively.

![Figure 9.28: Optimising Heterogeneous: Analytical Model for AP with Poisson](image)
Figure 9.29: Optimising Heterogeneous: Simulation for AP with Poisson

Figure 9.30: Optimising Heterogeneous: Analytical Model for STAs with Poisson
Figure 9.31: Optimising Heterogeneous : Simulation for STA with Poisson

Figure 9.28 and Figure 9.29 show the results for the calls that originate at the AP and Figure 9.30 and Figure 9.31 the results for the STAs using the analytical model and simulation, respectively. It is shown that the AP clearly only achieves a good perception rating with a low number of calls; due to the AP having to serve as many calls as there are STAs. The AP has the same queue size limit as the STAs and therefore, as the AP serves more traffic, delay and loss increases as the number of STAs increases. On the other hand, STAs have a good quality rating far past the APs having no recordable rating; this behaviour means that someone making a call originating in the wireless network to someone outside would be heard very well but not hear any reply due to the AP backlog.

There is again a difference between the analytical model and simulation, particularly the Access Point. However, the optimal point of operation appears around the same level; approximately 100ms.
The mean QoE is found for each bi-directional call and the AP, STA and mean QoE, known as the Flow, are plotted for both the analytical model and simulation for the Poisson case in Figure 9.32 and Figure 9.33, respectively. The results show that there is a discrepancy between the analytical model and simulation and therefore the model should be taken as a safe estimate of the optimal point and number of calls that can be served. The simulation results show that there is a potential increase in capacity from 7 bi-directional, as shown in the Benchmarking Table 9.3 and at the 20ms packetisation period in Figure 9.33, to 11 bi-directional calls, as shown with the maximum stations server by AP in Figure 9.33, using the dimensioning method presented.
Detailed results for the On/Off case are shown in Appendix A for the AP, STAs and Flow for both the analytical model and simulation and the acceptable quality level shown in Figure 9.34 and Figure 9.35 for the model and simulation, respectively. Results again show that the AP is underestimated; however the STA perception is slightly over estimated. Again the optimal point, which is at a packetisation period of 80ms, shows an increase from 7 to 11 bi-directional calls. The original 7 calls is shown in Table 9.3 and at the 20ms packetisation period plot (20 being the standardised packetisation period) in Figure 9.35.
9.6.4. What-If Capabilities

The augmented analytical model allows for what-if analysis without the need for simulation. In this section, additional delay is introduced which represents the case where the sender and receiver are not in the same wireless domain, for example there may be an additional network to consider. Again the perceived QoE is plotted as contour plots.

![Contour plots showing perceived QoE with additional delay](image)

*Figure 9.36: Homogeneous Analytical Model Poisson plus 200ms Delay*
Figure 9.36 and Figure 9.37 show an additional delay of 200ms on the Poisson and On/Off analytical results, respectively. Appendix A also shows both traffic cases with 30, 60 and 100ms additional delay. The what-if analysis shows that the behaviour of the network is similar but the optimal point of operation is shown around the 60ms packetisation period.

The main differences are that as delay increases, R-Factor decrease across the board and the maximum number of calls which can be served decreases (although only from 21 to 19 with the additional 200ms delay. In the 100ms scenario, there will never be a call perceiving a QoE R-Factor above 90, regardless of packetisation period or number of calls on the network and similarly none above an R-Factor of 80 in the 200ms case.

The optimal point which would satisfy most of the experiments thus far appears to be a packetisation period of between 80ms and 100ms, which is equivalent to 6848 bit to 8560 bit packets. This could be implemented either at the application layer, by multiplexing 4 or 5 packets together at the MAC layer or setting the TxOP time accordingly to send 4 or 5 normal size G.711 packets. The latter approach would need care as TxOP is expressed as time as opposed to number of packets, hence additional dynamics would need to be considered such as the mean service time, which in turn is affected by the data-rate which may change due to the dynamic physical layer data rate used to combat interference. [SHA04].
9.7 Analytical Model Usage Conclusions

The MAC layer analytical model is accurate for the saturated and non-saturated cases with the physical layer and retransmission limit assumptions. The physical layer assumption is that there are no other devices transmitting or noise on the channel other than the stations being modelled. The retransmission assumption assumes a station can contend for the channel, for each packet, an unlimited number of times. The unlimited retransmission behaviour means that no packets will be lost at the MAC layer but does can cause higher delay for the packet in question and those buffered. With the addition of the On/Off and queueing model, variations between the analytical model and simulation are due to the discrete calculation of the de-jitter buffer.

The analytical model and simulation results for Poisson and On/Off traffic sources show similar behaviour patterns in the contour plots. The contour plots, which represent QoE, show the same overall shape and degradation extent (spaces between the contour lines).

In the Poisson case, the maximum number of stations at an acceptable quality threshold (a PQ of 60) can be considered an approximation as the degradation occurs before simulation.

In the On/Off case, the model appears to be showing a higher number of acceptable calls. The higher number of calls shown in the analytical model, compared to simulation, should be well noted. The On/Off results do show a more graceful degradation compared to the Poisson case.

Generally, in the extreme cases there may be discrepancy, however, the model is extremely useful in giving a behavioural overview of the perceived Quality of Experience and to limit the scope of investigation, hence negate the need for extremely large scale supercomputing resource.

The method of dimensioning, with the objective being to maximise QoE, is extremely useful. The QoE framework has a relationship to revenue as the more calls which can be severed, while the user perceives an acceptable QoE, the more revenue the company can generate. The analytical models are therefore extremely useful for the telecommunications industry.

An augmented analytical model has been developed to cope with On/Off traffic in a Wi-Fi environment. The model has been validated against extensive super-computer simulations with over 3 billion packets being simulated. It is recognised that the model in the On/Off case is an overestimation of QoE, however the model gives a 1 in 13,000 speed up compared to simulation. The model may be used for extremely quick
approximate results and, if more detailed granularity of capacity need, used to identify an area to focus in on, such as the optimum packet payload size when using VoIP.

The dimensioning framework and methodology give the opportunity to increase call capacity using QoE and the value of the analytical model has been shown with ability to perform ‘what-if’ analysis on network delay which provides validity of conclusions.
10 Conclusion & Future Work

In this dissertation, an investigation was undertaken to identify some challenges of using time critical services over the more widely used WLAN and addressed these with novel analytical models, large scale simulations and dimensioning to increase acceptable call capacity.

This dissertation outlines the increasing trend using IEEE 802.11 with highly time critical services such as Internet Telephony. The introduction explains in detail how IEEE 802.11 operates and the issues that arise from using such networks which rely on a shared resource. These contention based protocols, and the algorithms used, try to fairly distribute resources; however the attempt to be fair can in fact lead to poor resource usage and unfair distribution of the resource. This dissertation also states that to simulate such networks, where there are numerous events for the simulator to handle, high performance supercomputing resource is a necessity, particularly when needing multiple repeats of the same experiment to derive statistically confident results. Simulation in this dissertation typical consists of 32 batch repeats, seeded differently to ensure the averaging of results is not a function of the random number generator. Each simulation ran for at least 60 minutes where the first 50% of time is devoted to setup and reaching steady state. Finally, this dissertation states an additional challenge of how to equate the end-to-end network performance and the unfair behaviour of the protocols to the users perceived perception of the network.

To address the issues; IEEE 802.11 performance, need for super computers and deriving the users perceived quality of experience, a combined analytical and simulation study was undertaken. The motivation for the analytical model was to utilise the compute power available and the desire to capitalise on this, i.e. having the ability to compare analytical models to simulation. The dissertation was then tied together with dimensioning of the network in order to increase call carrying capacity from the users’ perceived experience perspective.

In order to address the modelling and maximising call challenges, a literature review was undertaken which highlighted several opportunities to build on existing analytical models and pull together numerous dimensioning attempts. The main analytical model which stood out was by G. Bianchi; the Bianchi model is the seminal paper in the field as it finds, with close correlation to simulation, the service time and throughput of an IEEE 802.11 network with a given number of stations. The Bianchi model however assumes all stations are continuously offering traffic to the network.

It was shown in the Bianchi case, stations always needed to backoff to avoid a capturing the channel as packets were arriving faster than they could be served. On the other hand, a station can immediately transmit when data arrives to the station which is idle.
The change of behaviour and the traffic assumptions were termed the unsaturated and saturated modes of operation.

The Pitts model aimed to model both saturated, as assumed in the Bianchi model, and unsaturated traffic. The model was validated by the dissertation author using large scale simulations on super computers. The model showed that, as the mode of operation switches from unsaturated to saturated, a significant increase in service time accompanied by a substantial drop in throughput for 20 stations or more. These results again underpinned (a) the need for supercomputers to validate the model, particularly when repeated several times, and (b) the need for techniques to avoid switching from unsaturated to saturated.

Current models however do not tell us the performance of the network where there is an access point present (and normally needing to send more data than the stations), the effect of buffering traffic, the effects of on-off sources such as voice or the users’ perceived quality of the network. This dissertation addresses these four points and demonstrates the value with dimensioning to increase call capacity.

Firstly, an analytical model is constructed to model performance where different stations have different amounts of traffic to send over the network; this is termed the heterogeneous case and can therefore model an access point which has more traffic to send than the stations. This inclusion of an access point is absolutely critical when modelling typical usage of IEEE 802.11 and allowing the modelling of bi-directional voice calls which may not have a destination within the given wireless network. To construct the analytical model, the previous state machine based model which modelled a whole network, was now used to model each station/access point individually. In order to achieve the one state machine per station, the probabilities involved had to be re-derived due to the effect stations have on each other. Initial results showed a discrepancy in the service time for two stations compared to simulation. Simulation showed that, when there are two stations, the highly loaded station should achieve a lower service time and vice-versa. It was also found that there had to be a modification from slot based derivation to time based as each station now had different timings for a slot (previously all stations have the same duration for each slot). The modification from a slot based to a time based approach involved calibrating the slot times of the stations to the saturated service time results for a given number of stations. Results showed that service times were approximately equal to simulation. There were three major outcomes; this was the first notable model to derive service times for an access point and an arbitrary number of stations, it was evident now from the analytical model that the access point does not have fair access to the channel compared to traffic load and critically, there was a 1 to 13,000 speedup shown from simulation to modelling compared with a typical desktop machine.
To address the impact of buffering, a queueing model was introduced. In itself, a standard queueing model, but this model had to be applied in such a way that it related back to the state machines for each station. Normally, one calculation of results can be used for a range of loads, however it was found that the model needed recalculating for each load as the service time distribution didn’t fit any standard distributions.

The on-off traffic sources, typically seen with voice calls, were addressed by calculating the probability of stations being active (on) or inactive (off). The model allowed each possible state (a given number on and a given number off) to be found and the resulting service times were weighted by the probability of each case.

Finally, a framework was derived in order to show users’ perceived quality of experience using the ITU G.107 E-Model. In this dissertation, the R-Factor was inspected compared to delay, loss and jitter and contour plots were used in order to show how user perceived quality of experience degrades with respect to different network configurations and load.

To demonstrate the usage of the model and address the previous results, which showed degradation when moving from the unsaturated mode of operation to the saturated mode of operation, dimensioning was undertaken. It was clear that the aim was to decrease the number of packets each station had to deal with and a number of schemes came to light. The main scheme involved sending multiple packets in one transmission and, with the use of the analytical model, it was shown that a higher number of calls can be served. The sending multiple packets in one transmission approach meant that stations in back-off had to wait for the channel for a longer period of time, but this far outweighed the advantages of keeping out of saturation. Results showed that sending multiple packets was feasible and three methods were constructed to allow this to be achieved; (a) change the protocol to allow three packets to be transmitted at once or (b) change parameters in the application to send longer packets, i.e. record the sender’s voice for a longer period before packetising the data and sending and (c), the most elegant solution which didn’t involve redefining the protocol or re-writing applications, was to use the Wireless Multimedia Extensions to IEEE 802.11. In IEEE 802.11e, the QoS and multimedia extensions, there is an additional parameter which allows multiple packets to be sent in one transmission opportunity, this parameter is advertised to the stations from an access point. Using the contour plot framework with both simulation and analytical models, results were shown across an entire dimensioning state space and the optimum found. The optimum depended on keeping out of saturation for higher loads while considering the trade-off that additional delay was introduced by holding on for the packets until the next transmission opportunity. Again, results showed that there was a discrepancy between the analytical model and super computer simulations of between 1 and 5 stations, however, the optimal capacity occurred at the same point which allowed for very small scale simulation to be used to determine a definitive number of stations. The results also showed a number of important contributions, such as (a) the effects on
QoE with on-off traffic sources compared to Poisson sources, (b) the effects on QoE with and without an access point, (c) a 1 to 13,000 increase in speed with the analytical model compared to simulation on a desktop machine, (d) the QoE framework allows for vast dimensioning state spaces to be explored, (e) there is scope to increase capacity, in some cases by 43% and (f) a transmission opportunity parameter that can be directly used.

10.1 Future Direction

There are several directions for further academic and industrial research to build on and follow from this dissertation. Firstly, augmenting the analytical model to deal with scenarios where stations have different packet sizes to allow the modelling of different services across stations; this would aid in QoS provisioning where each station has two applications using the network. Secondly, augmenting to allow larger scale scenarios to be modelled where partial overlap between stations and access points exist, particularly when different channel and power strategies are employed; this would allow modelling of the growing popularity of wireless blanket metropolitan area networks. Larger scale investigation may also require an update to the QoE Methodology to consider the number of clients that can access a network under different physical layer conditions. Thirdly, alternative strategies could be employed to improve access point QoE and hence allow the access point a fairer usage of the channel compared with its load. Finally, existing published methods could be re-evaluated with the QoE framework presented to allow for easy comparisons of published strategies which currently use strict criterion to measure acceptable performance.

10.1.1. Modelling Heterogeneous Applications

The analytical model presented in this dissertation has the capability of: evaluating a heterogeneous case where each station may have different traffic loads; evaluating infrastructure mode (with an AP) rather than limited to ad-hoc mode; deriving service times for sporadic traffic sources (with a known On/Off distribution); deriving end-to-end delays, loss and jitter caused by buffering traffic and deriving the users experienced QoE of various different scenarios to give a methodology to allow dimensioning of the network in order to increase acceptable call carrying capacity.

The model however does have assumptions/limitations of use imposed by the calibration of slot to time based transmission probability. The calibration factor effectively finds the time based proportion of channel usage in the saturated case and normalised the interacting state machine model accordingly. This calibration methodology however assumed that all stations are operating with a single packet size across all stations. This packet size assumption means that different traffic types on each station cannot be calculated; take for example two stations with Voice over IP at 214 byte packets and a station browsing the web with 1,000 byte packets. This packet size assumption
effectively means all stations should be using at least the same application and hence traffic profile (with exception of bit rate). In order to allow the model to be further augmenting to allow each station to be operating with different packet sizes, an update to the saturated case model would be needed beforehand to derive a new calibration factor, however it is likely even the saturated model would need to become time based rather than slot based.

### 10.1.2. Scaling-Up

Clearly, the analytical model, which can now approximate service times for any number of stations offering a different load, can be expanded for use in much larger scale scenarios where there are several access points and potentially hundreds of stations. Metropolitan Area Networks typically have in the order of 50 access points in a two kilometre area [BT10] providing blanket coverage for potentially hundreds of subscribers. The analytical model presented thus far assumes that all stations can hear each other, however in a metropolitan area, stations and access points can only hear close-by stations and access points.

To augment the model, each station/access point would need their neighbour’s (other stations/access points it can hear) transmission probabilities. The complexity increases as each station would need a list of transmission probabilities of its neighbours, which in turn would be a function of its neighbours’ neighbours etc. To solve this potential interacting station scenario, each station’s Markov chain would have to iterate once, then its neighbours and then its neighbours’ neighbours etc. continuously in order to find steady state. The question arises in the order of which stations iterate and could take huge numbers of iterations to find the steady state and hence service times and QoE scores.

The first question expected from an internet service provider may address how does the model deal with different channel and power settings. As mentioned in the background chapter, IEEE 802.11 can operate on three separate channels and this could be easily mapped on to the analytical model by effectively solving three different network scenarios (1 for each channel) as there is no cross channel interference. The complexity increases when a service provider uses all channels and a range of power settings to improve network performance. Another assumption of the analytical model presented in the dissertation is that there are no physical layer bit errors; the model assumes an error free channel with no interference other than the stations being modelled.

If a station experiences bit errors due to stations on close channels or the packet not being received due to power settings being too low, the result would be no acknowledgment and the station would proceed up its backoff phase, hence go to the next Markov state. One way of modelling this is to add on an additional probability, which maps from channel condition to bit error, to the existing collision probability, in
effect this could be a range of probabilities from interference for different channels; the existing collision probability being of the channel the station is operating. Power could be similarly mapped so to derive a probability of a packet not being received even when there is no collision.

From a QoE point of view, the augmentations in the analytical model would mean that there may also need to be an augmentation of the user perception methodology. Take power settings as an example, a decrease in power could cause less interference between stations, however the trade-off is the number of stations that can access the network will also increase. Stations which are out of service range, in an advertised blanket metropolitan area effectively have a QoE of 0.

In order to incorporate the loss of service for some users’ results an option would be to normalise that QoE results from the users which can get connected by the proportion of area covered. This however can get complex due to the underlying population distribution of targeted users in a certain area.

10.1.3. Fairer Optimisation

Optimisation studies in this dissertation have been based around keeping stations out of saturation. The analytical model investigation highlighted clearly how this should be achieved; namely of keeping a station’s buffer empty by allowing more data to be sent in one transmission either by multiplexing packets at the application layer or using the transmission opportunity in IEEE 802.11e. The dimensioning was undertaken from the perspective of users’ Quality of Experience; the objective being to maximise the number of ‘acceptable’ calls in the network, i.e. the maximum number of bi-directional calls which could be served where all users experience acceptable quality in both up and downstream directions.

Optimisation results presented solutions which could improve voice capacity; however it was shown that the next step in increasing call capacity is to prioritise the access point on the network. Optimisation schemes which prioritise the access point have been shown in the literature review (some pre and post-dating the draft IEEE 802.11e standard) but most required modifying the underlying IEEE 802.11 protocol, unlike the TxOP scheme proposed which at most required installing or configuring IEEE 802.11e, the standardised IEEE 802.11e QoS extension to the underlying protocol.

To prioritise the access point, while keeping the same additional requirements, of only configuring IEEE 802.11e, a method which comes to light is giving the access point lower contention window parameters allowing it priority accessing the channel. IEEE 802.11e provides methods to advertise different contention window parameters and as such, the access point could advertise standard parameters to IEEE 802.11 capable stations with itself using lower $CW_{\min}$ and $m$ values. The justification behind giving the
access point lower contention window parameters rather than advertising higher values for the stations is that stations incompatible with IEEE 802.11e (i.e. those with legacy IEEE 802.11) can also operate in the improved network environment.

Further investigations could be to consider differentiated services for best effort TCP and prioritised VoIP. Again, the requirements mentioned, namely configuring IEEE 802.11e, can easily handle differentiated service by virtual queues in stations with parameters advertised by the access point; again the dimensioning should consider stations which may not be IEEE 802.11e compatible.

In addition to the further dimensioning and increasing call direction, there have been numerous strategies published; the notable contributions being shown in the literature review. These other published schemes have reported success with metrics such as delay, loss and throughput; sometimes using fixed criteria such as no more than 1% loss. In order to evaluate the best approaches, a future direction could be the employment of the QoE methodology framework to re-evaluate these proposed methods from a user’s perceived QoE perspective. The impetus here, as in this dissertation, is to link back to the industry understanding of QoE, that relates to industry Mean Opinion Scores, and more importantly that QoE proportion of acceptable calls has a direct relationship to revenue.
APPENDIX A: PACKETISATION OPTIMISATION RESULTS

A.1 Heterogeneous Results: Poisson

Figure A.1: Optimising Heterogeneous: Analytical Model Pair-Wise with Poisson

Figure A.2: Optimising Heterogeneous: Simulation Pair-Wise with Poisson
A.2 Heterogeneous Results : On/Off

Figure A.3 : Optimising Heterogeneous : Analytical Model for AP with On/Off

Figure A.4 : Optimising Heterogeneous : Simulation for AP with On/Off
Figure A.5: Optimising Heterogeneous: Analytical Model for STAs with On/Off

Figure A.6: Optimising Heterogeneous: Simulation for STA with On/Off
Figure A.7: Optimising Heterogeneous: Analytical Model Pair-Wise with On/Off

Figure A.8: Optimising Heterogeneous: Simulation Pair-Wise with On/Off
A.3 What-If Homogeneous : Poisson

Figure A.9 : Homogeneous Analytical Model Poisson plus 30ms Delay

Figure A.10 : Homogeneous Analytical Model Poisson plus 60ms Delay
Figure A.11: Homogeneous Analytical Model Poisson plus 100ms Delay

Figure A.12: Homogeneous Analytical Model Poisson plus 200ms Delay
A.3 What-If Homogeneous : On/Off

Figure A.13: Homogeneous Analytical Model Poisson plus 30ms Delay

Figure A.14: Homogeneous Analytical Model Poisson plus 60ms Delay
Figure A.15: Homogeneous Analytical Model Poisson plus 100ms Delay

Figure A.16: Homogeneous Analytical Model Poisson plus 200ms Delay
REFERENCES


Personal Communications


Research Grants


Super Computing Services


AUTHOR’S PUBLICATIONS


