An Empirical Study of Computational Optimisation Techniques for Microstrip Antennas

Robert Woodhouse

Submitted to the University of York,
for the Degree of Doctor of Philosophy

Department of Electronics, University of York

2010

Contents

Abstract **xiii**

1 Introduction to The Thesis **1**

2 Introduction to Optimisation **3**
2.1 Overview . . . 3
2.2 Global and Local Maxima . . . 4
2.3 Constrained Optimisation . . . 4
2.4 Fitness Landscape . . . 4
2.5 Computational Optimisation Algorithms . . . 6
2.6 Multi-Objective Optimisation . . . 7
2.6.1 Pareto Optimality and Dominance . . . 7
2.6.2 Fitness Assessment in Multi-Objective Optimisation . . . 8
2.7 Representation . . . 9
2.8 Redundancy . . . 10

3 Optimisation Techniques **12**
3.1 Overview . . . 12
3.2 Gradient Methods . . . 12
3.3 Particle Swarm Optimisation (PSO) . . . 15
3.4 Simulated Annealing (SA) . . . 15
3.5 Genetic Algorithms (GA) . . . 16
3.5.1 Example: Evolution of Wire Antennas . . . 18
3.5.2 Selection in GAs . . . 19
3.5.3 Elitism . . . 20
3.5.4 Mutation Schemes . . . 20
3.5.5 Steady-State and Generational GAs . . . 22
3.5.6 GA Analysis . . . 22
3.5.7 GA Control Parameters . . . 24
3.6 Genetic Programming (GP) . . . 25
3.6.1 Tree Based GP and Program Bloat . . . 26
3.7 Cartesian Genetic Programming (CGP) . . . 26
3.7.1 Example: Representing Combinational Logic Circuits . . . 27
3.7.2 CGP, Redundancy and Neutral Search . . . 28

4 Introduction to Antennas **32**
4.1 Definition . . . 32
4.2 Reciprocity . . . 32
4.3 The Near Field and Far Field Regions . . . 33
4.4 Isotropic Antenna . . . 34
4.5 Directionality . . . 35
4.5.1 Antenna Coordinate System . . . 35
4.5.2 Solid Angle . . . 36
4.5.3 Radiation Intensity . . . 36
4.5.4 Directivity . . . 37
4.5.5 Radiation Patterns . . . 38
4.5.6 Units . . . 39
4.5.7 Half Power Beam Beam Width (HPBW) . . . 39
4.6 Bandwidth . . . 40
4.6.1 Input Impedance . . . 40
4.6.2 Radiation Efficiency . . . 41
4.6.3 Bandwidth Plots . . . 42
4.6.4 Quality (Q) Factor . . . 43
4.7 Polarisation . . . 44
4.7.1 Elliptical Polarisation . . . 45
4.7.2 Linear Polarisation . . . 46
4.7.3 Circular Polarisation . . . 46
4.7.4 Visualising Polarisation . . . 47

5 Introduction to Microstrip Antennas **48**
5.1 MSA Patch Shapes . . . 49
5.2 MSA Feed Types . . . 49
5.2.1 Microstrip Line Feed . . . 49
5.2.2 Probe Feed . . . 50
5.2.3 Aperture Coupled Feed . . . 51
5.2.4 Proximity Feed . . . 51
5.3 RMSA Operation . . . 52
5.3.1 Radiation Mechanism . . . 52
5.3.2 Loss Mechanism . . . 53
5.3.3 Current, Voltage and Input Impedance . . . 54
5.4 Main Characteristics of MSAs . . . 55
5.5 Parametric Analysis of RMSAs . . . 58
5.5.1 Effect of Relative Permittivity (ϵ_r) . . . 58
5.5.2 Effect of Patch Width (W) . . . 59
5.5.3 Effect of Substrate Thickness (h) . . . 61

6 MSA Analytical Models **62**
6.1 The Transmission Line Model . . . 62
6.1.1 Patch Dimensions . . . 63
6.1.2 Radiating Slots . . . 64
6.1.3 Input Impedance . . . 65
6.1.4 Radiation Pattern . . . 66
6.2 The Cavity Model . . . 69
6.2.1 Overview . . . 69
6.2.2 Fields Inside The Cavity . . . 71
6.2.3 Input Impedance . . . 76
6.2.4 Radiated Fields . . . 77
6.2.5 Other Patch Shapes . . . 79
6.3 Conclusion . . . 79

7 Compacting and Multi-Banding MSAs **80**
7.1 Introduction . . . 80
7.2 Conventional Methods for Compacting and Multi-Banding MSAs . . . 81
7.2.1 Changing the Properties of the Dielectric . . . 81
7.2.2 Including Shorting Posts and Strips . . . 81
7.2.3 Slits in the Patch . . . 82
7.2.4 Parasitic Patches . . . 84
7.3 Computational Optimisation Of MSAs . . . 85
7.4 Boolean Grid Representation . . . 86
7.5 Applicability of Computational Techniques to Boolean Grid Optimisation . . . 87
7.5.1 Inapplicable Techniques . . . 87
7.6 Genetic Algorithms (GA) . . . 87
7.7 Genetic Programming (GP) . . . 88
7.7.1 Example i . . . 88
7.7.2 Example ii . . . 89
7.8 Cartesian Genetic Programming (CGP) . . . 90
7.9 Previously Computationally Optimised MSAs . . . 91
7.9.1 Dual Band MSA Optimised by GAs using Parallel Computation . . . 91
7.9.2 Dual Band MSA Evolved with Increased Efficiency MoM . . . 95
7.9.3 Circularly Polarised MSA . . . 98

8 MSAs for Mobile Communication: Size and Environment **100**
8.1 MSA Size and Bandwidth . . . 100
8.1.1 General Relationship Between Antenna Size and Bandwidth . . . 100
8.1.2 Relationship Between MSA Size and Bandwidth . . . 101
8.1.3 Size of Grid Based MSA and Bandwidth . . . 103
8.1.4 Determination of Grid Size Example . . . 106
8.2 Antenna Size and Directionality . . . 107
8.2.1 General Relationship Between Antenna Size and Directionality . . . 107

8.2.2 RMSA Size and Directionality . . . 113
8.2.3 Grid based MSA Size and Directionality . . . 116
8.2.4 Conclusion of Antenna Size and Directionality Analysis . . . 117
8.3 Propagation in Mobile Communications: Depolarisation . . . 118
8.3.1 Signal Strength Variation: Path Loss . . . 118
8.3.2 Localised Urban Propagation Effects . . . 121
8.3.3 Sources of Depolarisation . . . 123
8.3.4 Conclusion of Depolarisation Analysis . . . 124
8.4 Grid Resolution . . . 125
8.5 Conclusion . . . 125

9 Optimisation of Fitness Evaluation 126
9.1 Analysis of FDTD Modeling Parameters . . . 127
9.1.1 Microstrip Transmission Lines . . . 127
9.1.2 RMSAs . . . 129
9.2 Complete Structure FDTD Modeling . . . 131
9.2.1 Microstrip Transmission Line Base Structures . . . 131
9.2.2 RMSAs . . . 133
9.2.3 Random Grid MSAs . . . 135
9.3 Conclusion . . . 137

10 Empirical Study 138
10.1 Justification of the Empirical Study . . . 138
10.2 Computational Techniques Used . . . 139
10.3 Optimisation Specification . . . 139
10.4 Practical Considerations . . . 140
10.4.1 Electromagnetic Simulation Software . . . 140
10.4.2 Substrate Parameters . . . 142
10.4.3 Grid Size and Resolution . . . 143
10.4.4 Number of Runs and Evaluations per Run . . . 143
10.5 Optimisation Technique Control Parameters . . . 144
10.5.1 Fitness Function . . . 145
10.5.2 Implementation of Neutral search . . . 147
10.6 Instruction Sets . . . 149
10.6.1 Instruction Set Types and List/Graph Length . . . 150

11 Empirical Study Statistical Results 152
11.1 Results Overview . . . 152
11.2 Results Description . . . 155
11.3 Results Analysis . . . 157
11.4 Results Conclusion . . . 158

12 Evolved Antennas from the Empirical Study **159**
12.1 Antenna 1 . . . 159
12.1.1 Bandwidth . . . 159
12.1.2 Radiation Pattern . . . 160
12.1.3 Surface Current Distribution . . . 163
12.2 Antenna 2 . . . 167
12.2.1 Bandwidth . . . 167
12.2.2 Radiation Pattern . . . 168
12.2.3 Surface Current Distribution . . . 170
12.3 Conclusion . . . 172

13 Conclusion **173**

14 Further Work **175**
14.1 MSAs with Higher Bandwidth Feeds . . . 175
14.2 Printed Monopoles . . . 175
14.3 Wire Antennas . . . 177
14.4 Planar Microwave Structures . . . 178

References **179**

List of Figures

2.1 One dimensional fitness landscape. 4
2.2 Example of a 2 dimensional fitness landscape with no dominant global maximum. [1] 5
2.3 Example of a 2 dimensional fitness landscape with a clear global maximum. [1] 5
2.4 Generic computational optimisation algorithm. 6
2.5 Pareto front for a dual objective minimisation. 8
2.6 Two dimensional wire antenna optimisation: (a) polar representation and (b) Cartesian representation. 9
2.7 Effect of redundancy in a one dimensional fitness landscape. 10
2.8 6 variable list. 11
2.9 9 variable list with 3 redundant variables. 11

3.1 Gradient ascent of a simple hill. [2] 13
3.2 Gradient ascent with different starting points. [3] 14
3.3 Generic genetic algorithm. 17
3.4 Two dimensional wire antennas and their chromosomes. 18
3.5 Mathematical expression represented with a tree. [4] 25
3.6 Circuit encoded by chromosome. 27
3.7 $1 + \lambda$ algorithm. 28
3.8 The best fitness values from 2 x 100 evolutionary runs with allowed (diamonds) and forbidden (crosses) neutral mutations. [5] 29
3.9 Minimum computational effort for all mutation probabilities versus genotype length (in nodes) for two-bit multiplier with gate set AND, OR, NAND, NOR. [6] 30
3.10 Average phenotype length at end of an evolutionary run for all mutation probabilities versus genotype length (in nodes) for two-bit multiplier with gate set AND, OR, NAND, NOR. [6] 31
3.11 Average proportion of active nodes in genotype at conclusion of evolutionary run for all mutation probabilities versus genotype length (in nodes) for two-bit multiplier with gate set AND, OR, NAND, NOR. [6] 31

4.1 Linearly polarised electromagnetic wave. [7] 33
4.2 Isotropic antenna. 35
4.3 Spherical coordinates. 35

4.4 Graphical representation of 1 steradian. [8] . . . 36
4.5 Elevation plane radiation pattern of a dipole antenna. . . . 38
4.6 Half power beam width. [9] . . . 39
4.7 Power flow of an antenna in transmission. . . . 40
4.8 S_{11} of Rectangular Microstrip Antenna (RMSA). . . . 42
4.9 Frequency response of an electronic filter with variable Q. [10] . . . 43
4.10 Polarisation: (a) elliptical, (b) linear, (c) circular. [11] . . . 47

5.1 Geometry of RMSA. [12] . . . 48
5.2 Common MSA patch shapes. [12] . . . 49
5.3 Microstrip line fed MSA. [13] . . . 50
5.4 Probe fed MSA. [13] . . . 50
5.5 Aperture coupled fed MSA. [14] . . . 51
5.6 Proximity fed MSA. [15] . . . 51
5.7 RMSA and its radiating slots. [16] . . . 52
5.8 Top view of RMSA and its radiating slots. . . . 54
5.9 Side view of RMSA. [17] . . . 54
5.10 Top view of surface current distribution of RMSA. . . . 54
5.11 Current (I), voltage (V) and input impedance ($|Z|$) along the length of an RMSA. [18] . . . 55
5.12 RMSA radiation patterns: (a) E plane, (b) H plane. [19] . . . 56
5.13 Circularly polarised square and circular MSAs with thin diagonal centre slots. [20] . . . 57
5.14 Nearly square patch circularly polarised MSA. [17] . . . 57
5.15 Effect of ϵ_r on patch length and bandwidth for fixed frequency RMSA (centre frequency is 1.8 GHz), data from [21] . . . 58
5.16 Effect of ϵ_r on directivity for fixed frequency RMSA (centre frequency is 1.8 GHz), data from [21] . . . 59
5.17 Effect of W on the H plane radiation pattern of RMSAs. . . . 60
5.18 RMSA directivity against patch width. Directivity is in dB. [20] . . . 60
5.19 RMSA bandwidth against substrate thickness. [20] . . . 61

6.1 Microstrip transmission line: (a) geometry, (b) electric field lines & (c) effective dielectric constant. [17] . . . 62
6.2 Physical and effective lengths of an RMSA. [17] . . . 64
6.3 Transmission line model of an RMSA. . . . 65
6.4 Microstrip line fed RMSA and its radiating slots. [16] . . . 67
6.5 RMSA radiation pattern. [22] . . . 68
6.6 Charge distribution and current density on microstrip patch [12]. . . . 69
6.7 Top surface of patch from above. . . . 70
6.8 Cavity geometry [17]. . . . 71
6.9 Electric field configurations (modes) inside the cavity [17]. . . . 75
6.10 Equivalent circuit model of an RMSA. . . . 77

6.11 Radiating slots of RMSA for the dominant mode [17]. 78
6.12 Non-radiating slots of RMSA for the dominant mode [17]. 78

7.1 Top view of probe fed (a) half wavelength RMSA and (b) quarter wavelength shorted RMSA. 82
7.2 Surface current distributions for RMSAs with (a) long narrow slits and (b) triangular notches. [23] 82
7.3 Geometry of RMSA with slits. [24] 83
7.4 Measured S11 of RMSA with slits. [24] 83
7.5 Geometry of MSA with directly coupled and parasitic patches. [25] 84
7.6 Measured S11 of MSA with directly coupled and parasitic patches. [25] . . . 85
7.7 4 x 4 grid and its corresponding bit string. 86
7.8 Resulting pattern of GP list on 4 x 4 grid. 89
7.9 Resulting pattern of GP list with redundant instructions on 4 x 4 grid. 90
7.10 6 vertex graph. 90
7.11 Geometry of MSA ($\alpha = 0.7$). [26] 93
7.12 Normalised Co-polar component of radiated electric field ($\alpha = 0.7$). [26] . . . 94
7.13 Normalised Cross-polar component of radiated electric field ($\alpha = 0.7$). [26] . 94
7.14 S_{11} of optimised MSA ($\alpha = 0.7$). [26] 95
7.15 Mother structure MSA. [27] 95
7.16 Creation of mother structure Z matrix and substructure Z matrix. [27] . . . 96
7.17 Geometry of evolved dual band MSA. [27] 97
7.18 S_{11} of evolved dual band MSA. [27] 97
7.19 Regions of MSA able to be optimised. [28] 98
7.20 Trimmed square circularly polarised MSA. [17] 98
7.21 Geometry of patch of circularly polarised MSA. [28] 99
7.22 Measured axial ratio of circularly polarised MSA. [28] 99

8.1 Radiansphere of an antenna. 100
8.2 Side view of an MSA and the wasted space in its radiansphere. 102
8.3 Meander line folded monopole. [29] 102
8.4 View from above of a square grid based MSA and its radiansphere. 105
8.5 Path difference geometry of two source interference pattern. 107
8.6 Interference patterns from two coherent sources for increasing source separations. [30] 108
8.7 Geometry of ith source point and jth observation point. 109
8.8 Geometry of ith source point and jth observation point. 110
8.9 Path from ith source point to jth observation point. 110
8.10 Average 3 dB beamwidth of main lobe against antenna (source block) dimension. 111
8.11 RMSA radiation pattern. 113
8.12 RMSA directivity against ϵ_e. 114
8.13 Equivalent RMSA sources. 116

8.14 Beamwidth of grid based MSA against grid size. . . . 117
8.15 Typical ray geometry in urban streets. [31] . . . 118
8.16 Path loss example geometry. . . . 119
8.17 Measured path loss and least squares approximation versus range for sample data set. [32] . . . 120
8.18 Distribution of measured fading error compared with normal distribution. [32] 120
8.19 View from the receiver . [33] . . . 121
8.20 Top view of measurement site location. [33] . . . 122
8.21 Azimuth-elevation plane of received signal. [33] . . . 122
8.22 Azimuth-delay plane of received signal. [33] . . . 123

9.1 FDTD simulation run time against number of FDTD cells * number of time steps (simulations were performed on a windows desktop PC). . . . 127
9.2 Input impedance of 50 Ohm transmission line with matched load against boundary size. . . . 128
9.3 Input impedance of 50 Ohm transmission line with matched load against time steps. . . . 129
9.4 Input impedance of 50 Ohm probe fed RMSA against boundary size. . . . 130
9.5 S11 of 50 Ohm probe fed RMSA against boundary size. . . . 130
9.6 Input impedance of 50 Ohm probe fed RMSA against time steps. . . . 131
9.7 Input impedance of 50 Ohm transmission line with 100 Ohm load. . . . 132
9.8 Input impedance of 50 Ohm transmission line matching stub and 100 Ohm load. . . . 133
9.9 S11 of 50 Ohm 2GHz probe fed RMSA. . . . 134
9.10 S11 of 50 Ohm 3GHz probe fed RMSA. . . . 134
9.11 Random grid based MSA. . . . 135
9.12 S11 of Random grid based MSA (Fig. 9.11). . . . 136
9.13 S11 of another Random grid based MSA. . . . 136
9.14 Illustration of antenna grid cell overlap: a) no overlap and b) 1 FDTD cell overlap. . . . 137

10.1 Yee cell. [34] . . . 141
10.2 Example of a 'set 2 cells in vicinity' instruction. . . . 149

12.1 Top view of antenna 1. . . . 159
12.2 Return loss of antenna 1. . . . 160
12.3 View from above of a square grid based MSA and its ground plane. . . . 161
12.4 Radiation pattern of antenna 1 at 3.5 GHz (plane 1). . . . 162
12.5 Radiation pattern of antenna 1 at 3.5 GHz (plane 2). . . . 162
12.6 Surface current distribution of antenna 1 at 3.5 GHz. . . . 165
12.7 Surface current distribution of antenna 1 at 4 GHz. . . . 166
12.8 Top view of antenna 2 (scale is in cm). . . . 167
12.9 Return loss of antenna 2. . . . 167

12.10Radiation pattern of antenna 2 at 3.5 GHz (plane 1). 168
12.11Radiation pattern of antenna 2 at 3.5 GHz (plane 2). 169
12.12Surface current distribution of antenna 2 at 3.5 GHz. 170
12.13Surface current distribution of antenna 2 at 4 GHz. 171

14.1 Printed monopoles: (a) straight & (b) bent. [35] 176
14.2 (a) inverted-f antenna & (b) meander line antenna. [36] 176
14.3 Evolved antenna for NASA's ST5 satellite. [37] 177
14.4 Edge coupled microstrip Tx line band pass filter. [38] 178

List of Abbreviations

Microstrip Antenna	MSA
Rectangular Microstrip Antenna	RMSA
Genetic Algorithm	GA
Genetic Programming	GP
Cartesian Genetic Programming	CGP

Abstract

There are many computational optimisation techniques, several of which have been applied to real world problems, such as wire antennas, building structures and turbine blade profiles. Some of these techniques are relatively well known within the scientific and engineering communities, such as genetic algorithms. Microstrip antennas (MSAs) are widely used, especially for mobile communications applications, due to their low profile and low cost. An empirical study was performed to ascertain which computational optimisation technique is the most efficient when optimising MSAs. In this context, the most efficient technique refers to the one that has the highest probability of finding a solution that meets the required specification when all techniques have the same computational time allocated to them. It was found that genetic algorithms, the simplest technique used, is the most efficient of those that were tried. The main reason for this was concluded to be due to the relatively low number of fitness evaluations performed per run. Other, more complex, techniques are likely to to be more efficient when more fitness evaluations (run time) are available.

Chapter 1

Introduction to The Thesis

The most accurate way to compare the effectiveness of several different stochastic techniques is empirically. The most significant disadvantage of this approach is that it can take a particularly long time. Several different computational optimisation techniques have been previously applied to the optimisation of antennas, and indeed to MSAs, many of which have generated impressive results. Computationally optimising antennas can be markedly time consuming due to the relatively long fitness evaluation time. Determining the most efficient technique is a worthwhile pursuit because it should enable significant time savings.

An empirical study of computational optimisation techniques for the optimisation of any type of antenna does not appear in the literature. This is the main reason why this study is novel. The most likely reason why no empirical studies have been performed is that they are particularly time consuming. This study itself may have been markedly time consuming, but the results is has yielded could potentially save much more time in the future.

Another source of novelty for this study is that it includes a comprehensive and complete analysis of the phenomena that affect the performance of grid based MSAs. This analysis includes, for instance, investigations into the relationship between the size of grid based MSAs and bandwidth and directionality. Currently, in the literature, there are in-depth analytical analyses of simple geometry MSAs. There are also comparisons of measured against simulated results of specific geometries of MSAs for characteristics such as bandwidth. However, there is no complete single, yet multi-faceted, investigation into all of the areas that are relevant to the computational optimisation of MSAs.

Additionally, this study includes a thorough investigation into which computational optimisation techniques can be applied to grid based MSAs. A similar investigation can not be found in the literature. Another relevant point to be made is that this study was not just a pure academic exercise. Some of the resulting antenna designs were built and tested and then investigated further using computational electromagnetic modeling software.

This thesis begins with a description of the key principles and phenomena common to all computational optimisation techniques in chapter 2. In chapter 3, the main computational optimisation techniques currently used in the scientific and engineering communities, and how they work, are described. The fundamental aspects of antenna characterisation are explained in chapter 4. As this study is involving MSAs, a comprehensive investigation of their operation, characteristics and modeling was conducted. The findings of this can be seen in chapters 5 & 6. Furthermore, as size is such an important feature in many MSA applications, a comprehensive study of MSA performance with respect to antenna size was conducted, which is described in chapter 8. As well as exploring the relationship between size and both bandwidth and directionality, an extensive investigation into depolarisation in urban/sub-urban environments was also performed. This can also be seen in chapter 8.

An insight into conventional techniques for broad-banding and multi-banding MSAs was necessary for this study, which is shown in chapter 7. Also in this chapter is an analysis of how computational optimisation techniques can be practically applied to the optimisation of MSAs. Finally in this chapter are some particularly notable examples of previously computationally optimised MSAs.

As fitness testing, i.e., the accuracy of the computer modeling, of the antennas is such a vital component of their successful optimisation, the extensive investigation that was conducted is documented in chapter 9.

In chapter 10 is a description of all the other practical considerations and parameters necessary to fully implement the study. To summarise, the key features of the study, GA, GP & CGP were the techniques used. Several different parameter sets were tried for each technique. Each parameter set was run 10 times and 1000 fitness evaluations were given for each run.

The statistical results generated by the study, and what they reveal are discussed in chapter 11. Several of the antenna designs that were generated during the course of the study were built and tested. Two of them are shown in chapter 12.

Chapter 2

Introduction to Optimisation

2.1 Overview

The goal of optimisation is to find the best solution within a given problem space. More specifically, computational optimisation techniques typically aim to either maximise or minimise a real valued function.

This can be expressed formally [39]:

$$\begin{aligned} &\text{Given:} && f: X \longrightarrow \Re \\ &\text{where:} && \boldsymbol{x} \in X \text{ and } \boldsymbol{x} = x_1, x_2, \ldots x_n \\ &\text{The goal (maximisation) is to find } \boldsymbol{x}_0 \text{ such that:} && f(\boldsymbol{x}_0) \geq f(\boldsymbol{x}) \, \forall \, \boldsymbol{x} \end{aligned}$$

Alternatively, the goal may be to minimise the function f, i.e.:

$$\text{The goal (minimisation) is to find } \boldsymbol{x}_0 \text{ such that:} \quad f(\boldsymbol{x}_0) \leq f(\boldsymbol{x}) \, \forall \, \boldsymbol{x}$$

The function f is known as the *fitness function* or the *objective function*. The 'surface' of the fitness function is referred to as the *fitness landscape*. It can be thought of as having 'peaks' that correspond to high fitness function values and 'valleys' that correspond to low values. The set X of all possible input argument vectors $\boldsymbol{x}$ is known as the *search space*. n is the number of dimensions of the search space. The variables x (within $\boldsymbol{x}$) are also known as *decision variables* or *parameter variables*.

In the case of maximisation, the goal is to reach the global maximun of the fitness function (i.e. the highest 'peak' in the fitness landscape) or a local maximum that meets the required specification. In minimisation the opposite is true, i.e., the global minimum or a sufficiently low local minima is the goal.

2.2 Global and Local Maxima

The global maxima occurs at the point where the fitness function has its maximum value in the whole of the problem space. Local maxima occur where the fitness function reaches a maximum only in the 'local' vicinity.

The *global* maximum occurs at a point $\boldsymbol{x}_0$ if:

$$f(\boldsymbol{x}_0) \geq f(\boldsymbol{x}) \, \forall \, \boldsymbol{x}$$

A *local* maximum occurs at a point $\boldsymbol{x}_0$ if there exists $\epsilon > 0$ such that:

$$f(\boldsymbol{x}_0) \geq f(\boldsymbol{x}) \text{ when } |\boldsymbol{x}_0 - \boldsymbol{x}| \leq \epsilon$$

2.3 Constrained Optimisation

In *unconstrained optimisation*, the whole of n dimensional space may be searched, i.e.:

$$X = \Re^n \qquad \Longrightarrow \qquad -\infty \leq x_i \leq \infty$$

Most optimisations are however, in practice, *constrained* , i.e.:

$$X \subset \Re^n \qquad \Longrightarrow \qquad x_i^{min} \leq x_i \leq x_i^{max}$$

2.4 Fitness Landscape

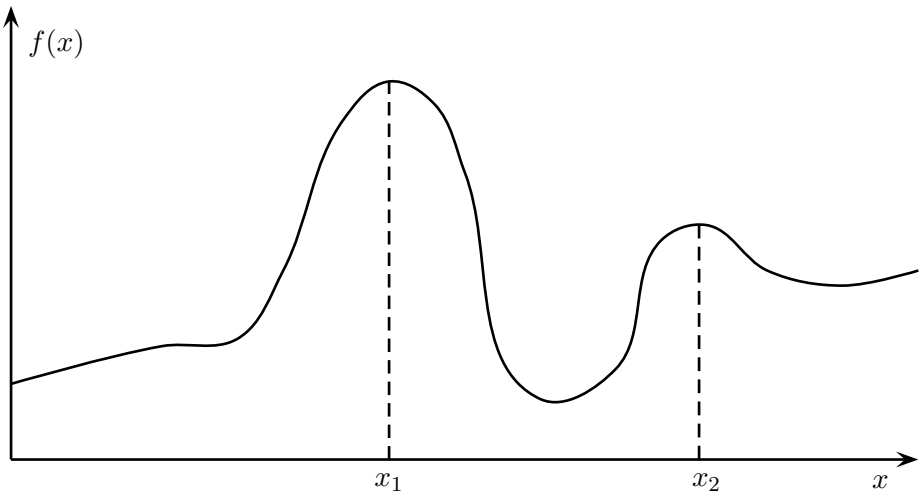


Figure 2.1: One dimensional fitness landscape.

The idea of the fitness landscape is useful in visualising the concepts of local and global maxima. Fig. 2.1 is an example of a one dimensional fitness landscape, so $\boldsymbol{x} = x$. The global maximum occurs at x_1, and a local maximum occurs at x_2.

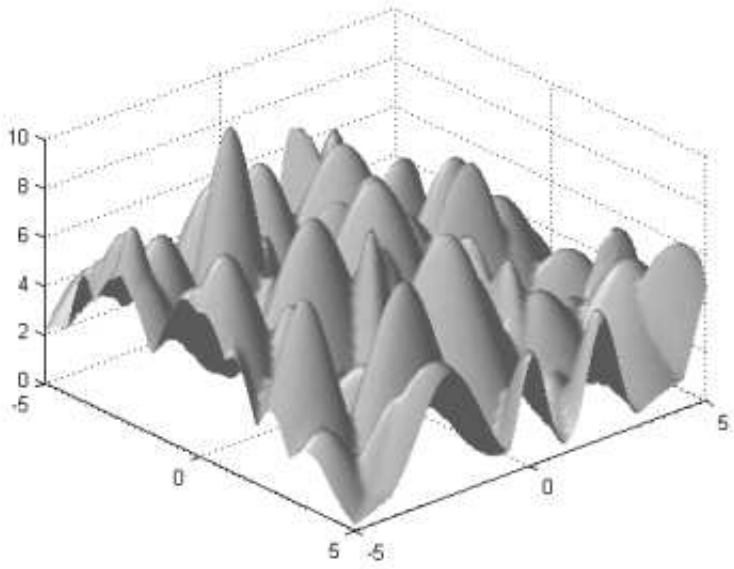


Figure 2.2: Example of a 2 dimensional fitness landscape with no dominant global maximum. [1]

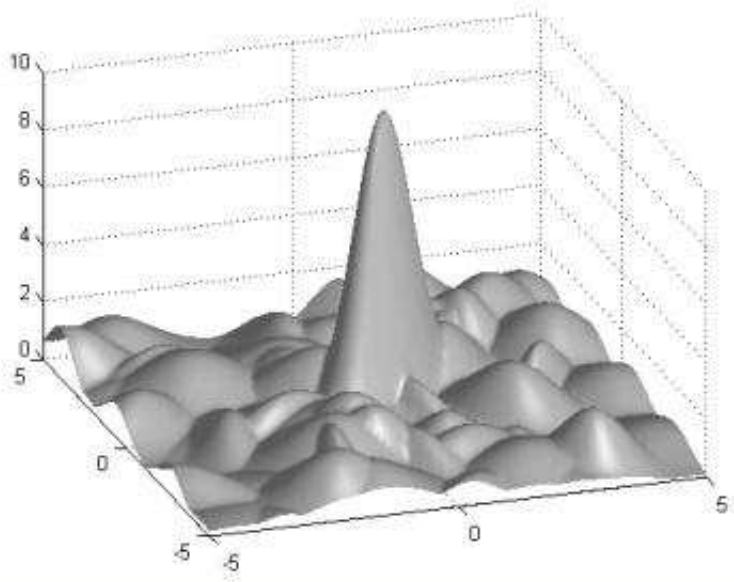


Figure 2.3: Example of a 2 dimensional fitness landscape with a clear global maximum. [1]

2.5 Computational Optimisation Algorithms

Computational optimisation algorithms can generally be divided into one of two categories: *deterministic* and *stochastic*. Deterministic algorithms use fixed rules at all stages of their operation. For example, parameters controlling how far in the search space that solutions can move, are dependent on fixed functions. Stochastic algorithms, conversely, incorporate probabilistic (random) features into their operation. Often, parameters controlling how far in the search space that solutions can move have a degree of randomness.

All computational optimisation algorithms are naturally iterative. They involve a series of discrete time steps, often referred to as *generations*. The algorithm operates on a collection of solutions, known as the *population*. The population may contain only one member (*individual*) or may have hundreds of thousands of individuals. The population size may be allowed to change as the algorithm progresses or it could be fixed.

The general basic concept behind the vast majority of computational optimisation algorithms is the same. The algorithm terminates when either a pre-determined number of generations have been completed, or when a required fitness has been achieved.

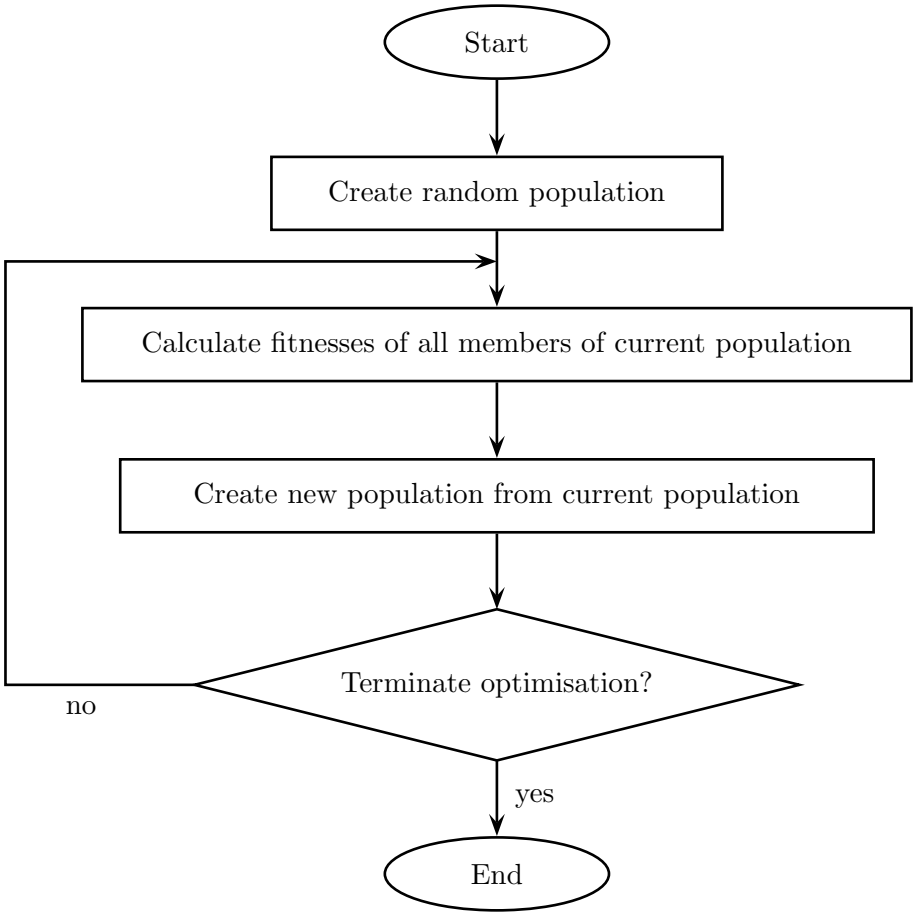


Figure 2.4: Generic computational optimisation algorithm.

2.6 Multi-Objective Optimisation

Many real world design tasks have several, possibly conflicting, objectives. In manufacturing, for example, cost and reliability are both important but competing requirements. Typically, a balance has to be found between the various competing objectives. In terms of optimisation, multiple objectives equals multiple fitness functions. For an optimisation involving n dimensional problem space with m objectives:

$$\begin{aligned} \text{decision variable vector:} \quad & \boldsymbol{x} = (x_1, x_2, \ldots x_n) \\ \text{fitness function vector:} \quad & \boldsymbol{f} = (f_1(\boldsymbol{x}), \ldots f_m(\boldsymbol{x})) \end{aligned}$$

2.6.1 Pareto Optimality and Dominance

In the case of single objective optimisation it is straightforward to determine which is the best solution in the population. This is because each individual has a single scalar fitness function value, which can be easily compared with that of the other individuals.

In multi objective optimisation, there are several fitness functions and so assessing the quality of individual solutions is non-trivial. Consider two solutions (A and B) in a dual objective optimisation. A is better than B for the first objective, but B is better than A for the second objective, but which is the best overall?

The concepts of *Pareto optimality* and *dominance* are useful when dealing with multi-objective optimisation.

In the case of maximisation, a decision vector $\boldsymbol{a}$ is said to dominate a decision vector $\boldsymbol{b}$ (also written as $\boldsymbol{a} \succ \boldsymbol{b}$) if and only if [**?**]:

$$\begin{aligned} & \forall\, i \in \{1, \ldots, m\} : \quad f_i(\boldsymbol{a}) \geq f_i(\boldsymbol{b}) \\ \wedge \quad & \exists\, j \in \{1, \ldots, m\} : \quad f_j(\boldsymbol{a}) > f_j(\boldsymbol{b}) \end{aligned} \tag{2.1}$$

If a solution is not dominated by any other individual in the population then it is said to be Pareto optimal or non-dominated. Depending on the population size, there can be several individuals which are Pareto optimal. This group of individuals is known as the *Pareto optimal set*. When plotted graphically, if possible, the Pareto optimal set forms a boundary in the fitness function space, which is known as the *Pareto front* , as in Fig. 2.5.

In Fig. 2.5, solution A is better than solution B for f_1 but it is the other way round for f_2. Neither can be regarded as being better than the other despite the fact that they are significantly different solutions. Since they are Pareto optimal, they lie on the Pareto front. Even though solution C is better than both A and B in one of the objectives, overall it is inferior to both of them, i.e., it is dominated by them. As a multi-objective optimisation algorithm proceeds, the Pareto front can be thought of as gradually moving closer to the optimum solution (the origin in Fig. 2.5). In most practical cases, it is impossible to reach the theoretically optimum solution due to fundamental physical laws etc.

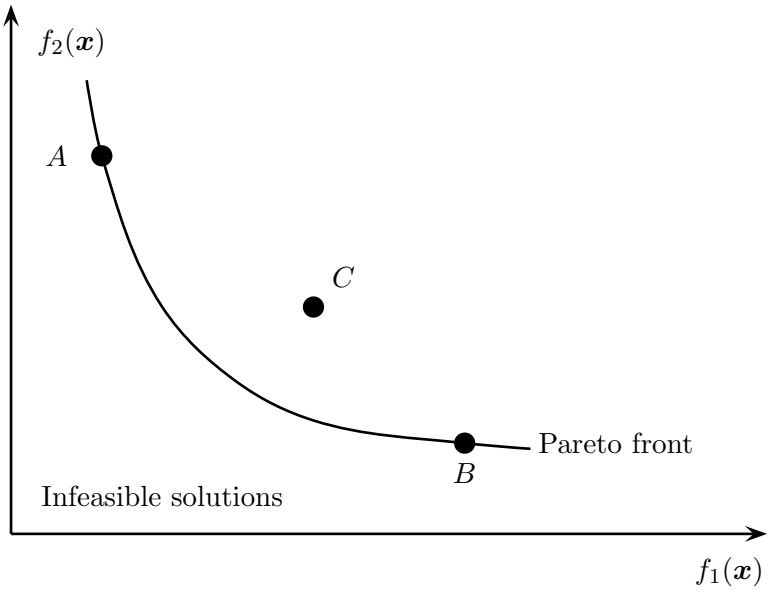


Figure 2.5: Pareto front for a dual objective minimisation.

2.6.2 Fitness Assessment in Multi-Objective Optimisation

There are several approaches to the problem of how to assess fitness in multi-objective optimisation [**?**]. The simplest approach is to generate a single overall scalar fitness function using a weighted sum of the fitness functions:

$$f(\boldsymbol{x}) = \sum_{i=1}^{m} w_i \, f_i(\boldsymbol{x}) \qquad\qquad 0 \leq w_i \leq 1 \tag{2.2}$$

Other techniques include: evaluating the population for each objective individually and then combining selected individuals based on their individual strengths. Another technique involves ranking each individual according to how many other individuals dominates it.

2.7 Representation

Representation refers to how solutions are encoded in the optimisation algorithm. There are usually several different ways of representing the object that is being optimised.

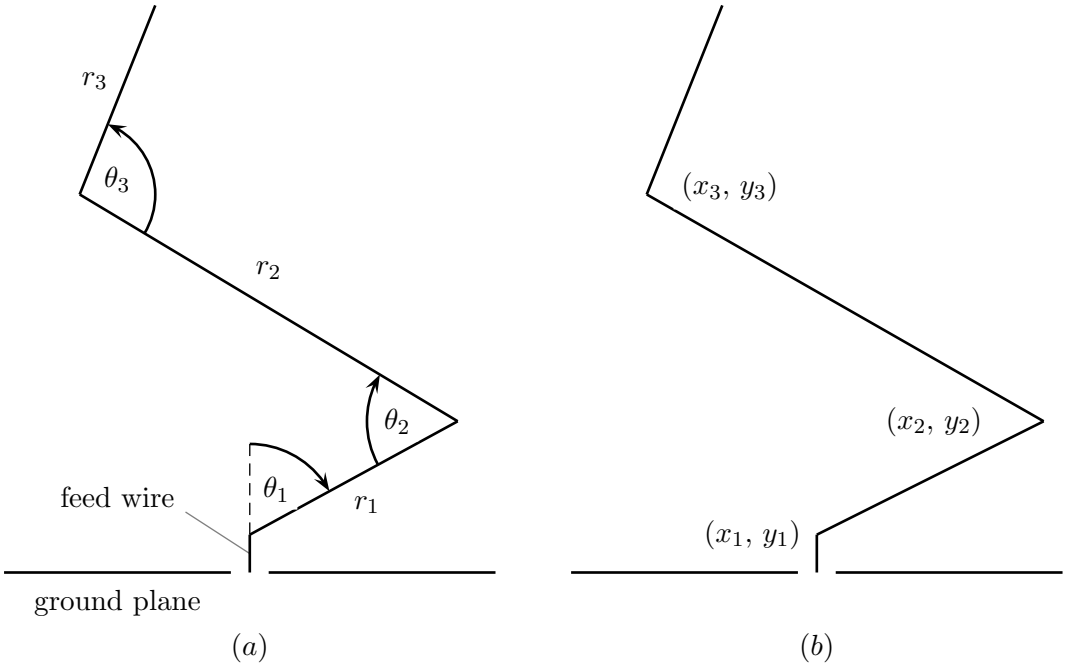


Figure 2.6: Two dimensional wire antenna optimisation: (a) polar representation and (b) Cartesian representation.

Fig. 2.6 shows two examples of how a two dimensional wire monopole antenna could be represented. The antenna consists of a ground plane, a short feed wire and three lengths of wire. The three lengths of wire can vary in length and orientation. During the course of the optimisation, the antennas will be modified such that their fitnesses improve, i.e., they will conform to the required specification more and more closely.

The antennas would actually be represented as lists of six floating point variables:

$$\begin{aligned} \text{polar:} \quad & r_1,\, \theta_1,\, r_2,\, \theta_2,\, r_3,\, \theta_3, \\ \text{Cartesian:} \quad & x_1,\, y_1,\, x_2,\, y_2,\, x_3,\, y_3, \end{aligned}$$

It is these variables that would be manipulated by the optimisation algorithm. Due to various constraints and limitations, many possible representations may have to be ruled out. Some representations have flaws which are not obvious but which become clear at run time.

2.8 Redundancy

Redundancy, also known as *neutrality*, is potentially a very useful characteristic of a given representation. Not all representations have redundancy. Redundancy refers to the fact that some parts of the representation may be inactive, i.e., they are not used in the encoding of the object. The key feature of redundancy is that currently inactive parts of the individual may become re-activated at a later time, i.e., in a later generation, and vice versa.

When currently inactive parts of the individual are modified, this will have no effect on the individuals fitness. However, when these parts are re-activated, large changes in the individual's characteristics, and thus fitness, are possible. This gives the optimisation the extremely advantageous ability of being able to avoid becoming 'trapped' at local maxima (maximisation).

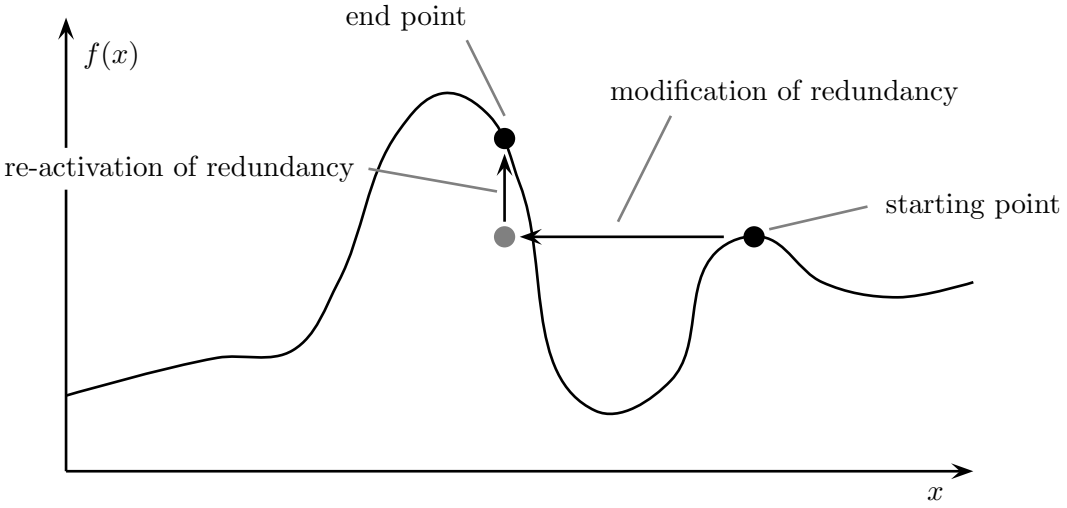


Figure 2.7: Effect of redundancy in a one dimensional fitness landscape.

In all optimisations, all of the solutions lie on the surface of the fitness landscape. Modifying inactive parts of an individual will not change its fitness ($f(\boldsymbol{x})$), but it will change its decision vector ($\boldsymbol{x}$). As a result, the individual can be viewed as having moved to a new location that has the same fitness value. This new location may or may not be on the surface of the fitness landscape, because it it not a real location. This is why the use of redundancy is akin to 'tunneling' in the fitness landscape. When the previously modified inactive parts of an individual are re-activated, dramatic changes in the individual's fitness can result. In some instances these changes could be significant improvements. The amount of change that is possible depends upon how much of the representation is potentially redundant.

The six variable list of Fig. 2.8 has the same representation as the Cartesian format described in section 2.7. The start and end markers signify partly how the algorithm would read the each individual. It would read the variables in pairs and construct the antenna accordingly.

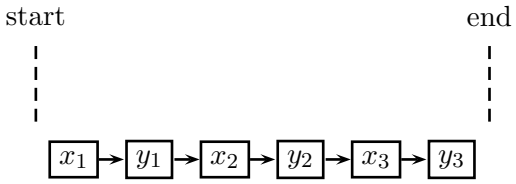


Figure 2.8: 6 variable list.

The nine variable list of Fig. 2.8 could also be used for the Cartesian representation of the three wire monopole antenna described in section 2.7.

In the nine variable list, three of the variables would always be redundant. The variables themselves (a_i) would be able to be modified by the optimisation. However, in addition, the start position, and thus the end position, would also be able to be moved. The end position would always be six places after the start position.

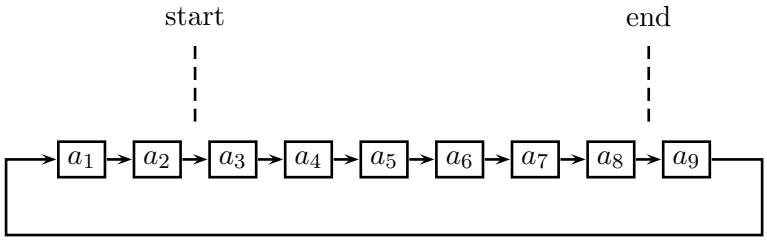


Figure 2.9: 9 variable list with 3 redundant variables.

In Fig. 2.9 the list would be read as follows:

$x_1 = a_3$ $y_1 = a_4$ $x_2 = a_5$ $y_2 = a_6$ $x_3 = a_7$ $y_3 = a_8$

The variables: a_1, a_2 and a_9 are currently redundant.

Chapter 3

Optimisation Techniques

3.1 Overview

There are many different optimisation techniques. Some arise from a mathematical analysis of the optimisation task, whilst others are inspired by biological or physical processes. Various optimisation techniques have been applied in a wide variety of fields including: scheduling, robotics, structural engineering, wing shapes, electronic circuits and antennas.

3.2 Gradient Methods

The term *Gradient methods* includes any method that involves calculating the first, or higher, derivative of the fitness function ($f(\boldsymbol{x})$). The simplest methods, and thus easiest to implement, involve the calculation of only the first derivative.

$$\begin{aligned} \text{decision variable vector:} \quad & \boldsymbol{x} = (x_1, x_2, \ldots x_n) \\ \text{fitness function:} \quad & f(\boldsymbol{x}) \\ \text{gradient:} \quad & \nabla f(\boldsymbol{x}) = \left(\frac{\partial f(\boldsymbol{x})}{\partial x_1}, \cdots, \frac{\partial f(\boldsymbol{x})}{\partial x_n} \right) \end{aligned} \tag{3.1}$$

The gradient has the important property that at any point in search space ($\boldsymbol{x}$), it always points into the direction of the maximal increase of the objective function. As such, the gradient is always perpendicular to the *contour* ($f(\boldsymbol{x}) = \text{constant}$) at that point.

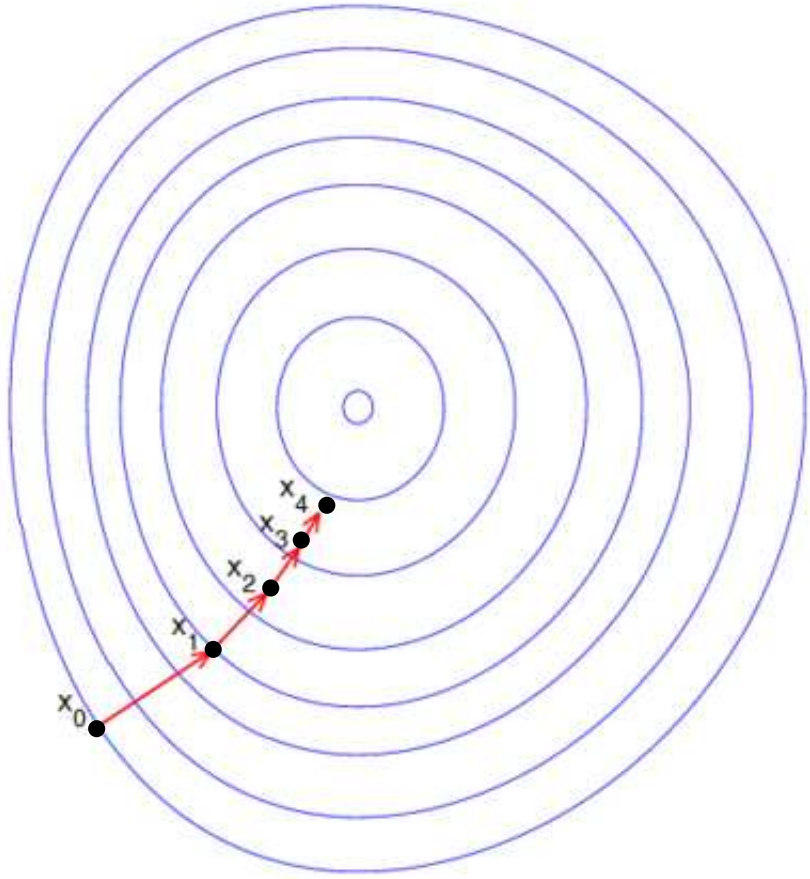


Figure 3.1: Gradient ascent of a simple hill. [2]

In Fig. 3.1, the blue lines are the contours, i.e., the lines of constant fitness value. If the contours are increasing towards the centre, then Fig. 3.1 describes the ascent of a 'hill'. At the points where the gradient (red arrows) is calculated, it always points in the direction of maximum increase. The gradient ascent algorithm is an iterative algorithm, in which the next point, x^{t+1}, is found by moving a distance, dependent on the gradient, $\nabla f(\boldsymbol{x}^t)$, from the current point, x^t, [39]:

$$\boldsymbol{x}^{t+1} = \boldsymbol{x}^t + \alpha_t \nabla f(\boldsymbol{x}^t) \tag{3.2}$$

For gradient descent, the $+$ is simply swapped for a $-$ sign. α_t is a positive scalar called the *step size*. The magnitude of the step size greatly affects the performance of the optimisation. A very small step size will result in many iterations which will be inefficient. A very large step size will lead to many overshoots, i.e., overshooting the top of the hill and having to come back again.

The *method of steepest ascent* is the same as that described in eqn. 3.2, but where the step size is chosen to deliver the maximum amount of increase of the fitness function at each individual step [39]:

$$\alpha_t = \arg\min\ f(\boldsymbol{x}^t + \alpha \nabla f(\boldsymbol{x}^t)) \qquad \alpha \geq 0 \tag{3.3}$$

In other words, at each step, starting from the point $\boldsymbol{x}^t$, a line search is conducted in the direction $\nabla f(\boldsymbol{x}^t))$ until the highest fitness function value $f(\boldsymbol{x}^{t+1})$ is found. The distance moved is then α_t.

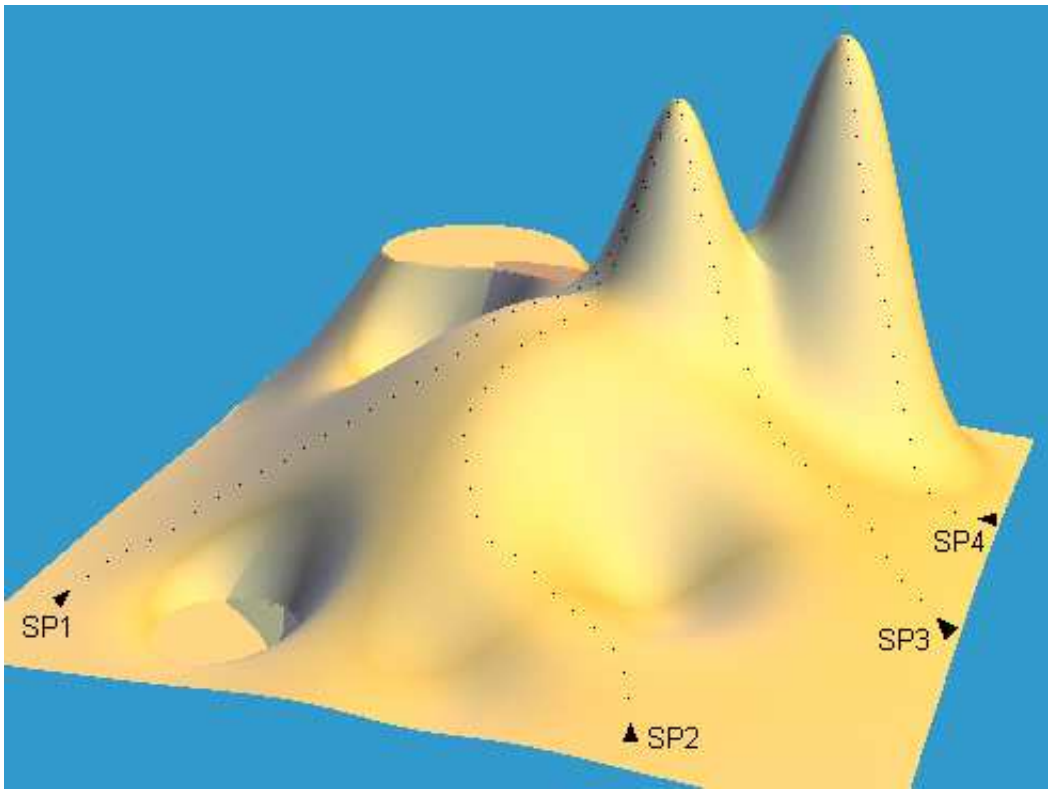


Figure 3.2: Gradient ascent with different starting points. [3]

Despite being relatively simple to implement and effective, gradient methods suffer from several significant drawbacks. Fig. 3.2 highlights one of them, which is that gradient methods do not necessarily find the global optimum. As can be seen, whether or not the global optimum is found depends on the starting point. This problem can be addressed by detecting when the gradient is zero or virtually zero, and then re-starting the optimisation at another random starting point. The more times this cycle is repeated, the more likely it is that the global optimum, rather than just a local optimum, will be found. This gradient ascent with random re-start algorithm clearly has a significantly much longer run time than just a single gradient ascent run.

Another important disadvantage of gradient methods is that they require continuous, differentiable variables. Many real world optimisation problems involve integer or Boolean variables and as such gradient methods can not be applied to them.

As well as having to be continuous and differentiable, all of the decision variables must be orthogonal to each other. If this is not the case then the calculation of the gradient would be practically, if not actually, impossible.

3.3 Particle Swarm Optimisation (PSO)

Particle swarm optimisation emulates a swarm of insects looking for food. Integral to its operation are the key features of memory and communication. PSO has been used in many optimisation applications and has been modified extensively [40].

The basic algorithm [41] starts by randomly positioning a population of pre-determined size within the search space. Each individual has two key variables associated with them. These are their positions (decision vector $\boldsymbol{x}$) and their velocities, $\boldsymbol{v}$. Every iteration, each particle updates, if necessary, the location of its own personal best fitness value, p. Additionally, the current global best g, i.e., the best location from all members of the population, is recorded. Below is given the update equations of the basic PSO algorithm for the *ith* individual, where t is the current iteration:

$$\boldsymbol{v}_i^{t+1} = c_0\,\boldsymbol{v}_i^t \;+\; c_1\,r_1\,(g^t \;-\; \boldsymbol{x}_i^t) \;+\; c_2\,r_2\,(p_i^t \;-\; \boldsymbol{x}_i^t) \tag{3.4}$$

$$\boldsymbol{x}_i^{t+1} = \boldsymbol{x}_i^t \;+\; \boldsymbol{v}_i^{t+1} \tag{3.5}$$

Where c_0, c_1 and c_2 are constants used for weighting the various individual terms. In order to inject some randomness into the swarm, r_1 and r_2 are uniformly distributed random numbers in the range 0 to 1. As the optimisation proceeds, the swarm incrementally converges on the optimum solution. As with gradient methods, PSO requires continuous search variables. It is therefore not applicable to search spaces that have integer or Boolean search variables.

3.4 Simulated Annealing (SA)

Simulated annealing [42] is inspired by the process of annealing in metals. Annealing metals involves heating the metal and letting it cool slowly. This has the effect of producing a well-ordered crystalline structure in the metal. Due to the low incidence of defects, the metal is both strong and malleable.

The goal of simulated annealing is that the system should reach a state that has a sufficiently low energy, i.e., high stability. The chance that the system moves from the current state s to a candidate new state s', depends on the acceptance probability P. The acceptance probability depends upon the energy of the current state $E(s)$, the energy of the candidate new state $E(s')$ and and on a global time-varying parameter T called the temperature.

When $E(s')$ is lower, i.e., better, than $E(s)$, then the system will always move into s'. However, when $E(s')$ is higher, i.e., worse, than $E(s)$, there is a chance that the system will still move into s'. This feature is included in SA in order to avoid the system becoming trapped at local optima.

When T is high, i.e., near the start of the optimisation, then the chance of accepting higher energy states is high. As T goes to zero, the chance of accepting higher energy states also goes to zero. As such, when T is sufficiently low, the optimisation will only accept lower energy states. SA behaves in such a way that at the start of the optimisation, i.e., when T is high, large fluctuations in the current system state and its energy are possible. As the system cools, the fluctuations reduce in magnitude and the algorithm becomes more conservative.

The acceptance probability (P) is defined as follows:

$$P = \begin{cases} 1 & E(s') \leq E(s) \\ \exp\left(\frac{E(s) - E(s')}{T}\right) & E(s') > E(s) \end{cases} \tag{3.6}$$

The main control parameters in SA apply to the *cooling schedule* or, as it is also known the *annealing schedule*. The parameters involved are the starting temperature T_0 and a parameter controlling how much T is decreased each iteration. The two most basic cooling schedules are (t = current iteration):

$$\text{exponential:} \quad T_{t+1} = \alpha\, T_t \qquad 0 < \alpha < 1 \tag{3.7}$$

$$\text{linear:} \quad T_{t+1} = T_t - \delta\, t \qquad 0 < \delta\, t \tag{3.8}$$

SA can be used to optimise search spaces that include integer, Boolean or real valued variables.

3.5 Genetic Algorithms (GA)

Genetic algorithms are based on biological evolution. The goal of GAs is to transform an initial population of into a set of acceptable solutions. In each *generation* (iteration) of a GA, a new population is created from the current one. This process involves reproductive operators that are based on biological processes. The basic terminology of GAs is based on that described in section 2.5.

The basic principle behind all GAs is the same. This is that fitter individuals in the population have a higher chance of reproducing, i.e., they have a higher chance of passing their characteristics onto the next generation. As such, advantageous characteristics become more and more widespread in the population as time passes. Incrementally the solutions will converge on either local optima or the global optimum.

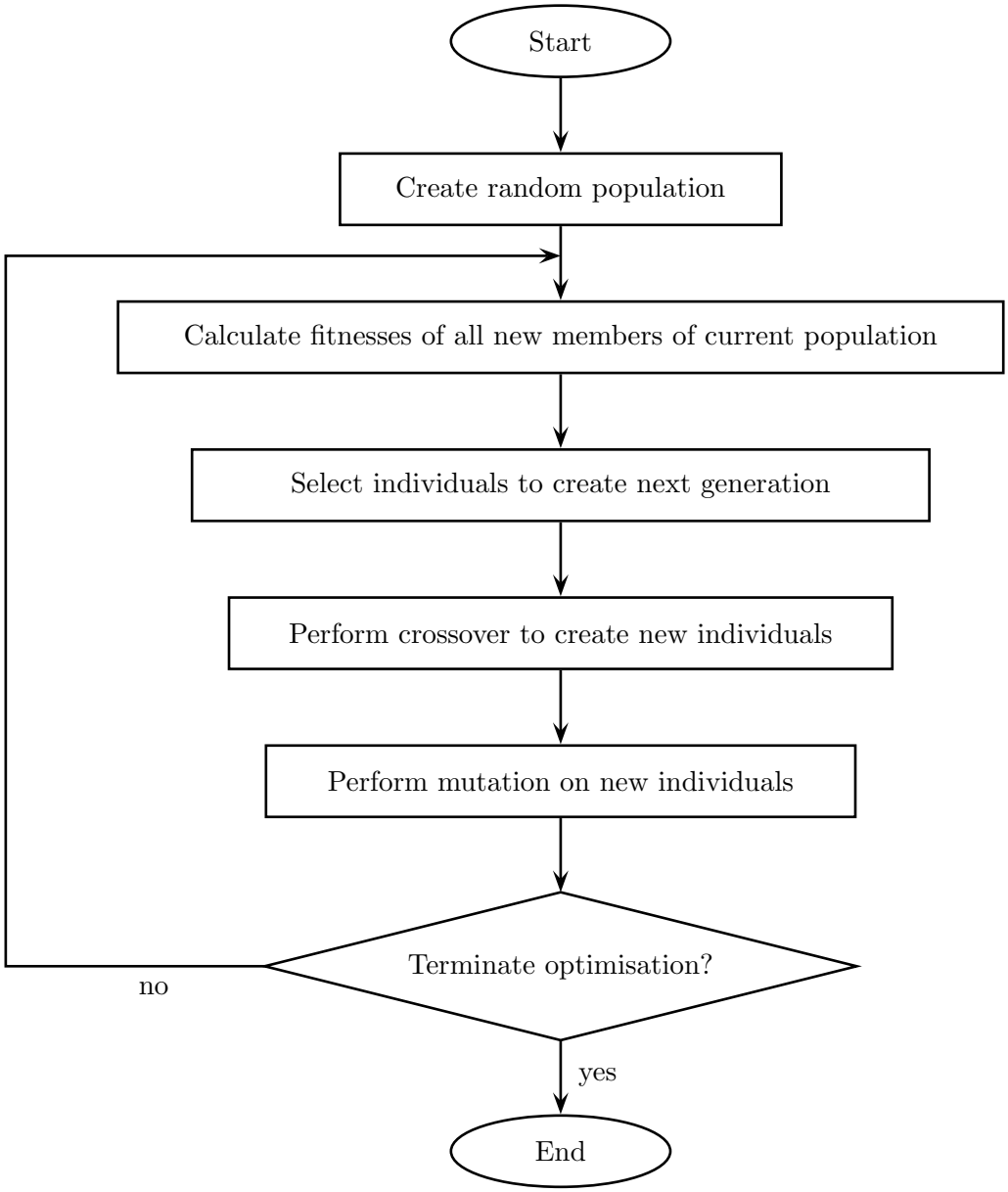


Figure 3.3: Generic genetic algorithm.

GAs were not the first biologically inspired optimisation techniques but they have become a widely used and successful technique. John Holland developed the basic ideas behind GAs in the late 1960s and early 1970s [43].

The structure of all GAs is the similar to that illustrated in Fig. 3.3. The basic principle is that all the members of the current population have their fitnesses evaluated and then certain individuals are selected based on their fitnesses. A new population is then created from these individuals.

The algorithm is terminated when either a pre-determined fitness value has been reached by at least one individual, or when a pre-determined number of generations have been completed.

In GAs, each individual has a corresponding *chromosome*. Individual and chromosome are effectively interchangeable terms. Chromosome refers more to the actual encoding (representation) of a particular individual. A chromosome is made up of one or more *genes*. The chromosome length indicates how many genes make up each chromosome. Genes encode for the actual functionality of the chromosome and how that functionality is dispersed and connected.

The new population is created from the current one using two main operators: *crossover* and *mutation*.

Crossover is a process which closely resembles sexual reproduction in the biological world. It involves 2 chromosomes (parents) swapping an equal size section of themselves, so that the 2 new chromosomes produced (offspring) both inherit characteristics from both parents.

Mutation involves picking genes within a chromosome at random, according to a given probability, and then altering those genes by a random amount. This probability is known as the *mutation rate*. There are various schemes for determining the mutation rate, of which the main ones are discussed later.

3.5.1 Example: Evolution of Wire Antennas

In this example, two dimensional, non-branching wire antennas are being evolved. The antennas are represented by alternate angles and lengths of wire, starting from a feed wire. The chromosome length is 6.

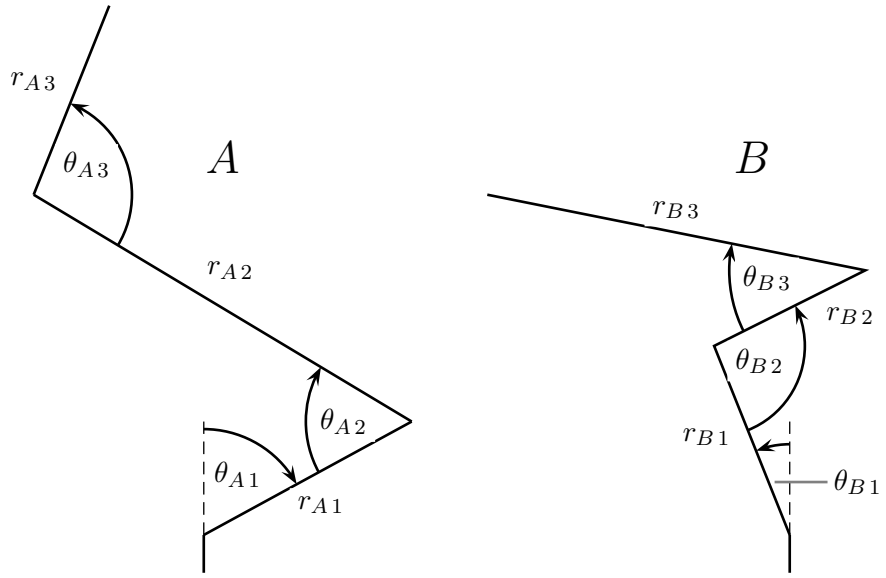


Figure 3.4: Two dimensional wire antennas and their chromosomes.

The chromosomes of the two antennas in Fig. 3.4 are:

$$\begin{array}{lcccccc} A: & \theta_{A1} & r_{A1} & \theta_{A2} & r_{A2} & \theta_{A3} & r_{A3} \\ B: & \theta_{B1} & r_{B1} & \theta_{B2} & r_{B2} & \theta_{B3} & r_{B3} \end{array}$$

The effect of two point crossover is illustrated below:

$$\begin{array}{cccccc} \theta_{A1} & r_{A1} & \theta_{A2} & r_{A2} & \theta_{A3} & r_{A3} \\ \theta_{B1} & r_{B1} & \theta_{B2} & r_{B2} & \theta_{B3} & r_{B3} \end{array} \quad \Longrightarrow \quad \begin{array}{cccccc} \theta_{A1} & r_{A1} & \theta_{B2} & r_{B2} & \theta_{B3} & r_{A3} \\ \theta_{B1} & r_{B1} & \theta_{A2} & r_{A2} & \theta_{A3} & r_{B3} \end{array}$$

The effect of mutation (after the crossover stage) can be seen below:

$$\begin{array}{cccccc} \theta_{A1} & r_{A1} & \theta_{B2} & r_{B2} & \theta_{B3} & r_{A3} \\ \theta_{B1} & r_{B1} & \theta_{A2} & r_{A2} & \theta_{A3} & r_{B3} \end{array} \quad \Longrightarrow \quad \begin{array}{cccccc} \theta_{A1} & r_{A1} & \theta'_{B2} & r_{B2} & \theta_{B3} & r_{A3} \\ \theta_{B1} & r'_{B1} & \theta_{A2} & r_{A2} & \theta'_{A3} & r_{B3} \end{array}$$

The primed values above (θ'_{B2} etc.) indicate those that have been modified by a random amount.

3.5.2 Selection in GAs

The purpose of the selection stage is to choose the individuals in the current population that will produce the next generation. The most important feature of selection is that individuals with a better fitness have a higher chance of being selected than worse fitness individuals. Selection strategies are generally *deterministic* or *stochastic*. There are several selection strategies [44], the main ones being described below.

The simplest approach is known as *population decimation*. This technique involves automatically discarding the individuals with lower fitness. The population is ranked according to fitness, with the best at the top. The minimum number of individuals required to produce the next generation is taken from the top of the list, or, all those individuals below a minimum required fitness value are discarded. Population decimation is therefore deterministic.

In *proportionate selection*, sometimes called *roulette wheel selection*, the probability p that an individual i will be selected from a population of size m is:

$$p_i = \frac{f_i(\boldsymbol{x})}{\sum_{i=1}^{m} f_i(\boldsymbol{x})} \tag{3.9}$$

Individuals with relatively better fitnesses will tend to be selected more often than those individuals with worse fitnesses. However, the better individuals are not guaranteed to be selected, and, relatively unfit individuals still have some of being selected Proportionate selection is therefore a stochastic strategy.

Tournament selection involves randomly picking two individuals and choosing the one with the best fitness with a probability of p_{sel}. Typically, $0.5 < p_{sel} < 1$, and so the strategy is stochastic because the fitter individual is not guaranteed to be selected. However, when $p_{sel} = 1$, the fitter individual will always be selected and so it becomes deterministic.

The potential advantage of stochastic strategies is that they give the less fit individuals a chance to take part in the creation of the next generation. Despite having relatively low fitnesses, these individuals may have some beneficial characteristics. With stochastic selection strategies there is a chance that these beneficial characteristics may survive to the next generation and become part of individuals with better fitnesses.

3.5.3 Elitism

Elitism is a selection strategy that is used in conjunction with other strategies, such as those mentioned above. It involves copying a given number of the fittest individuals in the current generation straight over into the next generation.

By using elitism, it can be guaranteed that the current best individuals, and thus their characteristics, will not be lost. The key advantage of elitism is that it has no cost. Due to the fact that the individuals are copied straight over, they do not need to have their fitnesses re-evaluated.

The inclusion of elitism in a GA in not necessarily always beneficial. In particular, when concerns exist about premature convergence, i.e., becoming trapped at the first local optima reached, then it may be best to avoid elitism [44] [45].

3.5.4 Mutation Schemes

The *mutation scheme* refers to how the correct mutation rate p_m for a particular point in the operation of GA is determined ($0 < p_m < 1$). When the search variables contain integer or real values, other mutation parameters may also be required. Various mutation schemes are considered first for the case of Boolean variables (binary-coded) because their implementation is simpler.

When a chromosome undergoes mutation in a binary-coded GA, each gene is selected in turn and a random number p is generated in the range of 0 to 1. If $p \leq p_m$ then the gene is mutated by having its value inverted, i.e., $0 \rightarrow 1$ or $1 \rightarrow 0$. If $p > p_m$ then the gene remains unchanged.

The simplest mutation scheme for a binary-coded GA is to have a *fixed* mutation rate, i.e., p_m never changes. A *deterministic* mutation rate is one which the mutation rate changes according to a fixed schedule:

$$p_m = p_{m0} \left(1 - \frac{g}{G}\right) \tag{3.10}$$

where:

$$\begin{aligned} p_{m0} &= \text{starting mutation rate} \\ g &= \text{current generation} \\ G &= \text{total number of generations} \end{aligned}$$

When using the deterministic scheme, the GA starts with a high mutation rate which decreases as the GA progresses. The purpose of this is to enable large mutations at the start of the GA when the fitnesses are poor. It also limits the amount of mutation nearer the end of the GA when the fitnesses should be better and thus only small improvements in fitness are possible. The potential drawback of this scheme is that it does not take into account the actual fitnesses of the population. Different runs of the same GA with the same parameters will not necessarily converge at the same rate.

Adaptive mutation schemes, on the other hand, do take fitness into account. A simple example of an adaptive scheme is:

$$p_m = p_{m0} \left(1 - \frac{f}{f_{max}}\right) \tag{3.11}$$

where:

$$\begin{aligned} p_{m0} &= \text{starting mutation rate} \\ f &= \text{current best fitness} \qquad 0 \leq f \leq f_{max} \\ f_{max} &= \text{maximum possible fitness} \end{aligned}$$

In real-coded GAs, fixed, deterministic and adaptive mutation schemes can still be used. In other words, the mutation rate can be controlled in exactly the same way as described above. However, another variable (*mutation range*) is also required [44]. The mutation range is usually the width of a given symmetric distribution with zero mean. Most often, the chosen distribution is either Gaussian or uniform. For a Gaussian distribution the mutation range will be the standard deviation and for the uniform distribution it will be the width. As with the mutation rate, the mutation amount can be either fixed, deterministic or adaptive.

3.5.5 Steady-State and Generational GAs

GAs generally fall into one of two categories: *steady-state* and *generational*. In generational GAs, a large proportion, if not all, of the individuals in each new generation is completely new. Not many individuals survive from one generation to the next unchanged.

In steady-state GAs, only a relatively small proportion of the current population is replaced by new individuals. Many individuals survive completely unchanged from one generation to the next. In many applications, steady state GAs have been shown to exhibit faster convergence the generational GAs [44].

3.5.6 GA Analysis

Schema theory was the first significant theoretical analysis of GAs [43] [46]. A Schema , also known as a similarity template, is used to describe all the possible chromosomes which have identical gene values (*alleles*) at several different gene locations. Schema theory was developed for binary-coded GAs. For example consider two 5-bit chromosomes:

$$A = 01101$$
$$B = 11000$$

The two strings both a 1 in the second position and a 0 in the fourth position. As such they are both represented by the following schema H_1:

$$H_1 = *1*0*$$

The symbol * means 'don't care', i.e., it can be either a 0 or a 1. Due to the fact that there are 3 don't cares in H_1, there are in total 8 (2^3) strings that H_1 represents:

$$H_1 = *1*0* =$$

$$01000$$
$$01001$$
$$01100$$
$$01101$$
$$11000$$
$$11001$$
$$11100$$
$$11101$$

Two important attributes of schemata are their order $o()$ and their defining length $\delta()$. The order is equal to the number of defined positions:

$$o(*1*0*) = 2 \qquad o(1****) = 1 \qquad o(10**0) = 3$$

The defining length is the distance between a schema's outermost defining positions:

$$\delta(*1*0*) = 2 \qquad \delta(1****) = 0 \qquad \delta(10**0) = 4$$

For chromosomes of length L, a schema of order o will have 2^{L-o} possible chromosomes. The *fitness* of a particular schema in a given population is the average fitness of all its representatives in that population.

The main finding of schema theory is that short (low defining length), low order, above average schema will grow exponentially [44] [39]. This leads on to *building block theory* which states that solutions comprised of schemata with the above mentioned characteristics (building blocks) will grow in number and dominate the population as the optimisation progresses. It has therefore been conjectured that GAs will operate most successfully when the most important contributions to the fitness come from short, low order schemata.

Schema theory also attempts to provide an explanation as to why GAs are such effective optimisers. In a population of size n and chromosome length l there could be $n2^l$ different schemata. In reality there will be less than this due to similar chromosomes, particularly as the population converges. The key concept, known as *implicit parallelism*, is that each time a single chromosome has its fitness evaluated, simultaneously many schemata will be evaluated as well. Holland argued that if n chromosomes are evaluated every generation then around n^3 schemata are automatically evaluated as well.

Another interesting finding of schema theory is that of *deception*. Deception describes the situation when the optimal chromosome bears little resemblance to near optimal chromosomes. A deceptive fitness function is one which is said to have this property.

Table 3.1: Non-deceptive function.

$\boldsymbol{x}$	000	001	010	011	100	101	110	111
$f(\boldsymbol{x})$	0	5	5	10	5	10	10	15

Table 3.2: Deceptive function.

$\boldsymbol{x}$	000	001	010	011	100	101	110	111
$f(\boldsymbol{x})$	15	14	12	3	13	2	1	20

In tables 3.1 and 3.2, $\boldsymbol{x}$ is the decision vector and $f(\boldsymbol{x})$ is the fitness. For both functions, the chromosome 111 is the fittest. This optimal solution is most likely to be created, via crossover and mutation, from the chromosomes that are the most similar to it (011, 101 and 110).

In the non-deceptive function, the fitness is simply proportional to the number of 1s in the chromosome. As such, the near optimal chromosomes (011, 101 and 110) all have high fitnesses. As the GA progresses, the number of 1s in the chromosomes of the population will increase. The optimum solution will easily be reached because there is a clear general progression towards it.

On the other hand, it is far less likely that the optimum solution will be reached in the case of the deceptive function. This is because all the chromosomes that are similar to it (011, 101 and 110) all have low fitnesses and will thus become diminished in the population.

The major criticism of schema theory is that it reveals very little about what is actually going on inside a GA. Consequently, it does not yield any particularly useful information such as how to determine the optimum choices of various control parameter or schemes (mutation rate and selection strategy etc.). An empirical study remains the most reliable way to determine the optimum control parameters.

3.5.7 GA Control Parameters

The functioning of GAs can be regarded as combining both an exploration of new regions in the search space and the exploitation already characterised regions. This balance between exploration and exploitation is controlled by the GAs control parameters, i.e., the crossover and mutation rates and the population size.

A large population has increased diversity and thus at the start of the GA, results in a wide exploration of the search space. This reduces the chance of becoming trapped at a local optimum. However, the disadvantage is that this will tend to lead to an increased convergence time. This is because at the start, the GA will resemble a random search. The population can be thought of as containing a large amount of 'noise', i.e., many poor individuals. As such it will take the GA several generations to remove this noise from the population.

On the other hand, a population that is too small, will not enable a sufficiently wide exploration of the search space. Consequently, the GA will tend to converge on a local optimum.

De Jong (1975) performed an early systematic study of how varying control parameters affected the performance of GAs [47]. The optimisation problems were a, now widely used, range of binary coded test functions. De Jong's results indicated that an optimum population size was between 50 and 100 individuals. Furthermore, De Jong concluded that the optimum rate for single point crossover was approximately 0.6 and that the optimum mutation rate was 0.001.

Other studies, such as Grefenstette (1986) and Schaffer et al (1989) [47], have also systematically tested a wide range of control parameter combinations. Additionally, a review of GAs in electromagnetic applications yielded broadly similar findings [48], see table 3.3.

Table 3.3: Optimum control parameters.

Study	Optimum population size	Optimum crossover rate	Optimum mutation rate
De Jong (1975)	50 - 100	0.6	0.001
Grefenstette (1986)	30	0.95	0.01
Schaffer et al (1989)	20 - 30	0.75 - 0.95	0.005 - 0.01
Johnson & Rahmat-Samii (1997)	30 - 100	0.6 - 0.9	0.01 - 0.1

Analytical approaches to finding the optimal population size have also been considered, such as using schema theory, e.g. Goldberg (1989). The expression Goldberg derived for the optimum population size for binary coded problem of chromosome length L, is approximately [49]:

$$\text{population size} = 1.65 * 2^{0.21\,L} \tag{3.12}$$

This results in much larger optimum population sizes than those determined by empirical methods. As such, analytically obtained optimum control parameter values are rarely used in practical applications of GAs.

3.6 Genetic Programming (GP)

In GAs, individuals represent objects, i.e., their representation directly encodes the objects parameters. In the example of section 3.5.1 each individual represents a wire antenna. In GP, individuals represent a program, i.e., a sequence of instructions, rather than an object. Other than this important difference, GP and GAs are effectively the same, i.e., the population is evolved using a GA. For instance, in GP crossover and mutation are used to create the next generation and the same selection methods can also be employed.

John Koza is largely credited with the discovery of GP in 1992 [50]. He mainly uses tree structures to form the chromosomes, because this is the natural form of most LISP programs, which is the language that he used to develop GP. Fig 3.5., below, shows how a mathematical expression can be encoded in a tree structure.

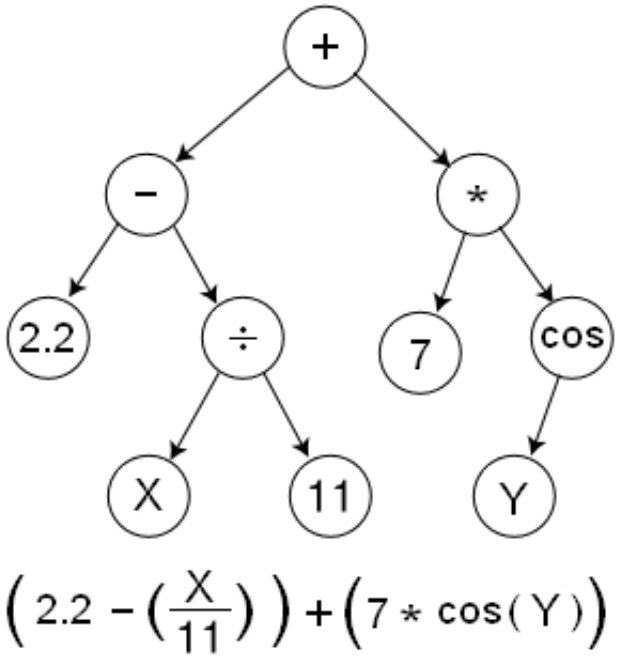


Figure 3.5: Mathematical expression represented with a tree. [4]

Due to the fact that GP is essentially the same as GAs, apart from the representation of course, it may be unclear as to why GP can potentially be a significantly more powerful optimisation technique. The power of GP comes from the power of the instructions. This can be clearly seen when considering the optimisation of the *ones max* problem.

The ones max problem is a simple testing function for comparing the efficiencies of different optimisation techniques. It is a binary coded problem in which the goal is to achieve one or more individuals whose chromosomes contain only 1s. For example, when optimising 8 bit chromosomes, the fittest possible chromosome would be 11111111, and the worst would be 00000000.

The performance of GAs for the ones max problem is that they typically achieve a fairly good fitness, i.e., mostly 1s, relatively quickly, but that the last few 1s become progressively harder to obtain. On the other hand, if a GP system has an instruction which sets all of the bits to 1, then it is possible that the optimum solution could well be reached in a single generation.

3.6.1 Tree Based GP and Program Bloat

GP systems can use structures other than trees, but trees constitute convenient structures for representing programs, as can be seen in Fig. 3.5. It is essential in virtually all tree based GP systems that the trees can vary in size as the optimisation progresses, so as to enable variable length programs. However, this gives rise to a commonly observed problem in tree based GP called *program bloat*. This refers to the rapid growth of the tree size of the population. The size of the programs often grows excessively beyond the typical program size that the optimum solution requires. Program bloat is usually due to trees containing large sections of code that have very little effect on the overall program. In other words, these sections do not contribute significantly to the fitness of the individual. At the end of the optimisation, the resulting solutions are often far larger than they realistically need to be and are thus impractical.

3.7 Cartesian Genetic Programming (CGP)

Cartesian Genetic Programming was invented by Julian Miller [51]. It is the same as GP except that whereas GP uses a tree to hold the program statements, CGP uses a graph. The motivation behind this is the fact that graphs are far more flexible and versatile data structures than trees. In fact a tree is a special case of a graph (acyclic).

The directed graph encoding is achieved by giving the chromosome a 2 dimensional grid format. The chromosomes are actually just strings of integers but they represent a 2 dimensional grid of functional blocks. Each functional block has a fixed maximum number of inputs and one output. The actual function that a particular block can perform depends on the function set that is available. If the chromosomes, for example, represent combinational digital logic circuits then the functions might include AND, OR, NAND etc.

One of the main advantages of CGP over GP is that redundancy can be easily implemented. Another advantage is that, due to the fact that the chromosome length is fixed, CGP produces solutions that are free from program bloat. One of the areas in which CGP has been most successfully exploited is in the evolution of combinational logic circuits [52].

3.7.1 Example: Representing Combinational Logic Circuits

In this example, a 2 input, 1 output circuit is represented. The 2 inputs (A and B) are assigned to input numbers 0 and 1 respectively. The first 2 integers of each block are the inputs to that block, and the last integer is the logic function of that block. Block 3, for example, has a function code of '1' so is therefore an AND gate. Its first input comes from block 2, and the other input is from the circuit input B. The output of the circuit (F) is the output of block 5. The output of block 4 does not propagate through to the output, therefore it has no effect on the output. Block 4 is consequently redundant.

Function Set:

1 = AND
2 = OR
3 = NAND
4 = NOR

Chromosome:

block 2	block 3	block 4	block 5	output
003	211	024	231	5

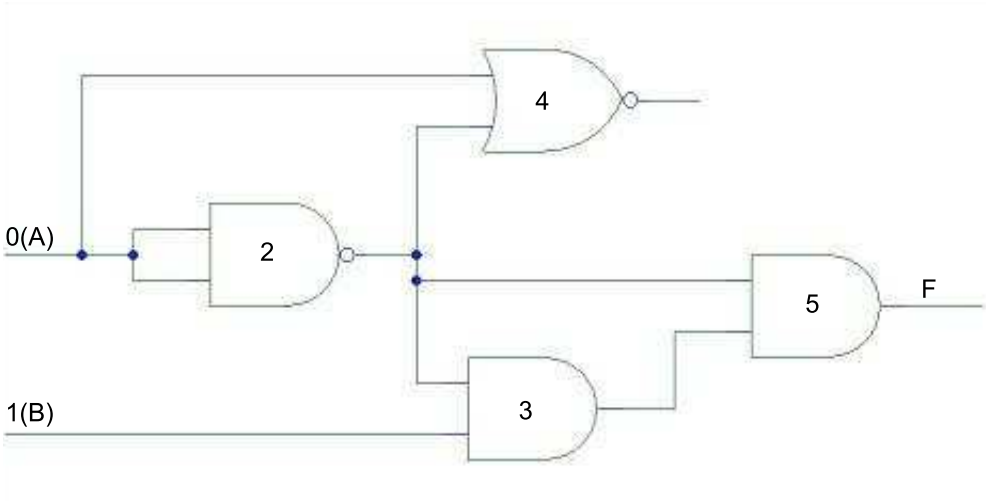


Figure 3.6: Circuit encoded by chromosome.

In this example, each individual in the population has a chromosome consisting of 4 blocks and a single output block number. The length of each chromosome is therefore 13 genes long. Reflecting the biological inspiration behind GAs and thus GP and CGP, the contents of a chromosome is referred to as its *genotype*. The active part of the genotype is known as the *phenotype*. Due to the fact that block 4 is redundant, the phenotype of the individual in this example is: 003 211 231 5. As there is a direct mapping between the phenotype and the circuit which it describes, both the active part of the genotype and the circuit itself can be regarded as being the phenotype. Due to redundancy, the length of the phenotype is always less than or equal to the length of the genotype. In this example, the genotype length is 13 whilst the phenotype length is 10.

3.7.2 CGP, Redundancy and Neutral Search

The impressive efficiency of CGP in evolving combinational logic circuits is due to its use of redundancy. When evolving logic circuits, Miller uses the $1 + \lambda$ algorithm, where λ is a positive integer. The reason for this is that the $1 + \lambda$ algorithm is one of the simplest GAs and is thus simple to implement.

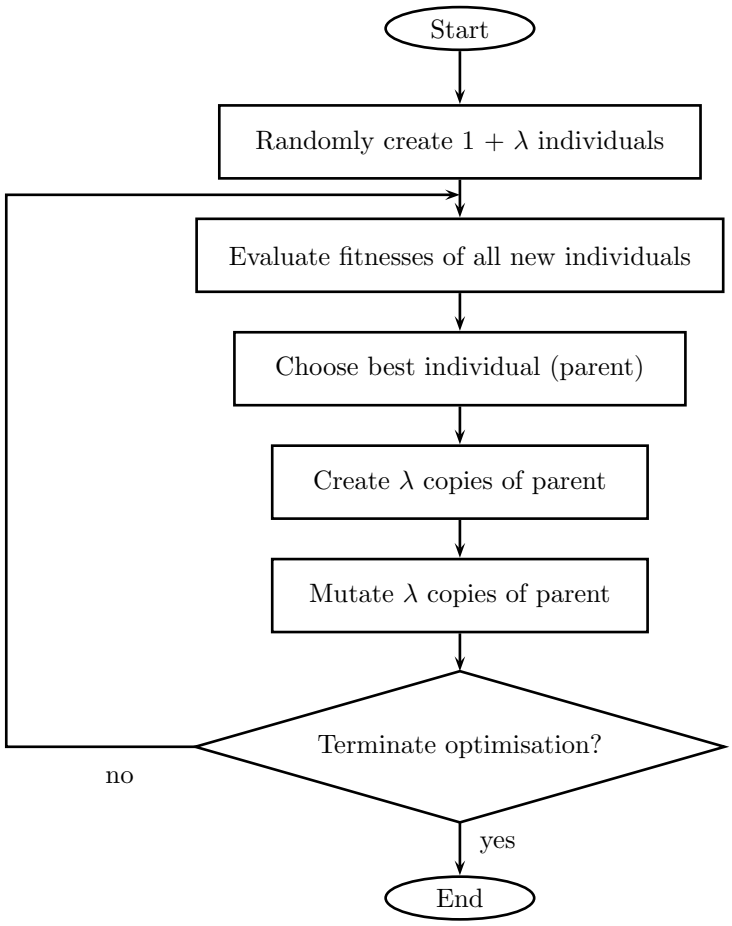


Figure 3.7: $1 + \lambda$ algorithm.

In the $1 + \lambda$ algorithm, each generation the fittest individual (parent) is selected and all the others are discarded. λ copies are then made of the parent, and these copies are then mutated. The fitnesses of these new individuals is then evaluated and the process is repeated. The $1 + \lambda$ algorithm is a highly generational GA because all individuals except the fittest are discarded every generation. The $1 + \lambda$ algorithm is further simplified by not using crossover in the creation of the next generation.

As illustrated in section 3.7.1, the representation implicitly enables redundancy. However, the beneficial effects of redundancy can be significantly increased by using what is known as *neutral search*. Miller et al implements neutral search in $1 + \lambda$ CGP [5] by enabling the algorithm to choose a new parent even if the new fittest members of the population have fitness values that are equal to the fitness of the previous parent. The advantage that neutral search can have is illustrated in Fig. 3.8. The results of 100 runs, each of 10 million generations, in the evolution of the three bit binary multiplier can be seen. A three bit multiplier simply multiplies two 3 bit numbers together to produce a 6 bit result. Multiplier functions have often been used to assess the performance of optimisation algorithms for Boolean problems because they are hard to evolve.

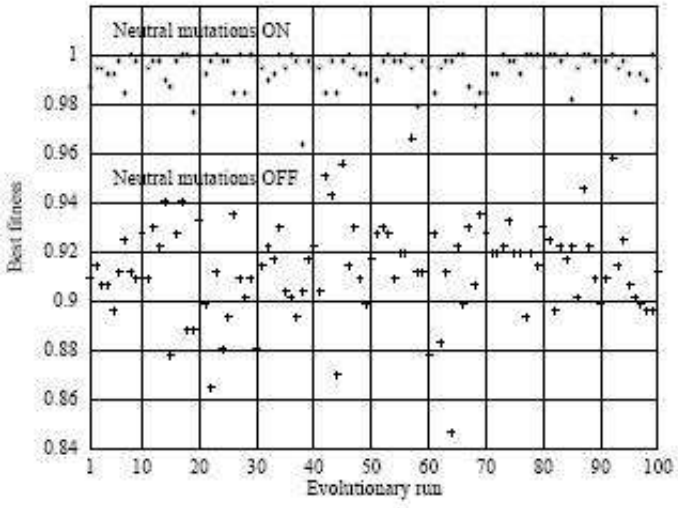


Figure 3.8: The best fitness values from 2 x 100 evolutionary runs with allowed (diamonds) and forbidden (crosses) neutral mutations. [5]

When neutral search was forbidden, a new parent was chosen only when a fitter individual than the current parent was generated. In Fig. 3.8, a fitness of 1 is the best possible, i.e., a perfect solution. When neutral mutations were allowed 27 of the 100 runs generated perfect solutions. These perfect solutions contained between 21 and 24 logic gates, which is competitive with conventional design methods. When neutral mutations were forbidden, the attained fitnesses were generally high but there were no perfect solutions. This is a clear demonstration of how the use of redundancy can aid evolutionary search by preventing the optimisation becoming trapped at local optima.

As well as helping search techniques to find solutions that it otherwise could not reach, the use of redundancy can also greatly improve the efficiency of the search. Miller and Smith [6] show that by using CGP with much greater genotype lengths than the solution actually requires, significant improvements in the efficiency of the search can be achieved. In order to measure efficiency Miller and Smith used the widely accepted *minimum computational effort*, which was first proposed by John Koza. This is the minimum number of genotype evaluations (fitness evaluations) required to give a 0.99 probability of success in an evolutionary run. It does not take into account the length of genomes.

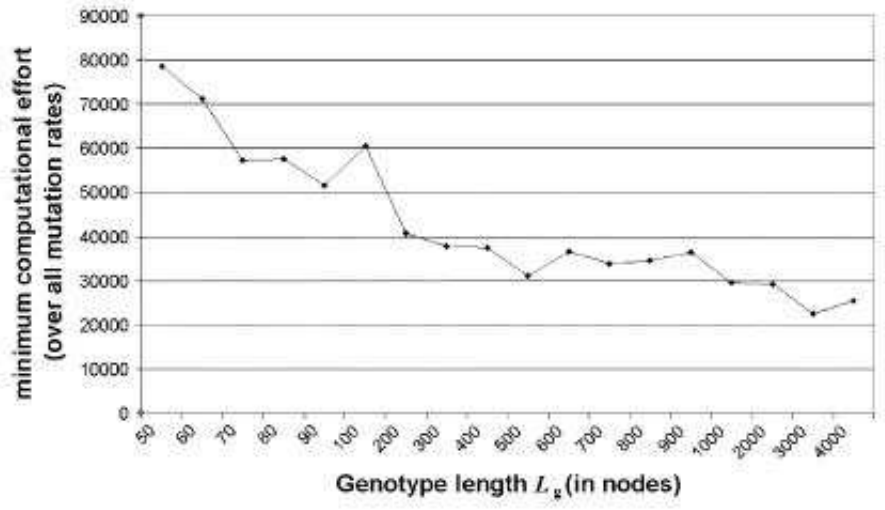


Figure 3.9: Minimum computational effort for all mutation probabilities versus genotype length (in nodes) for two-bit multiplier with gate set AND, OR, NAND, NOR. [6]

In [6], the genotype length and mutation rate were varied when attempting to evolve an even three parity function and a two bit multiplier. As can be seen in Fig. 3.9 there is a consistent drop in computational effort as the genotype length increases. In other words, for a given optimisation goal, fewer fitness evaluations are required the longer the genotype, i.e., chromosome length, becomes.

As the genotype length increases, so does that of the phenotype (active part of the genotype). This would appear perfectly natural but an interesting observation is that the phenotype length increases much less than linearly with genotype length, see Fig. 3.10. When the data was fitted to a polynomial curve it was:

$$L_p \approx 2.24\, L_g^{0.53} \tag{3.13}$$

Another interesting observation is that as the genotype length increases the average proportion of inactive (redundant) nodes (logic gates) increases rapidly, as can be seen in Fig. 3.11. With relatively long genotype lengths (~ 4000), when the optimisation is most efficient, only 5% of the nodes of the genotype are active, and so 95 % are redundant.

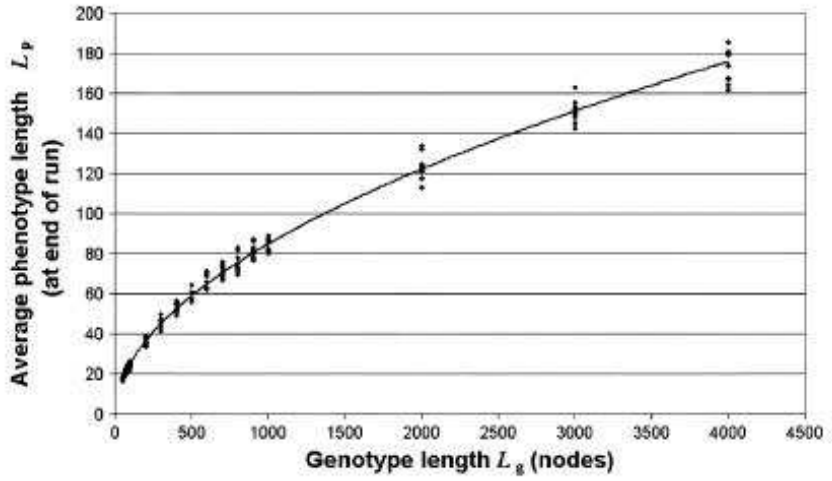


Figure 3.10: Average phenotype length at end of an evolutionary run for all mutation probabilities versus genotype length (in nodes) for two-bit multiplier with gate set AND, OR, NAND, NOR. [6]

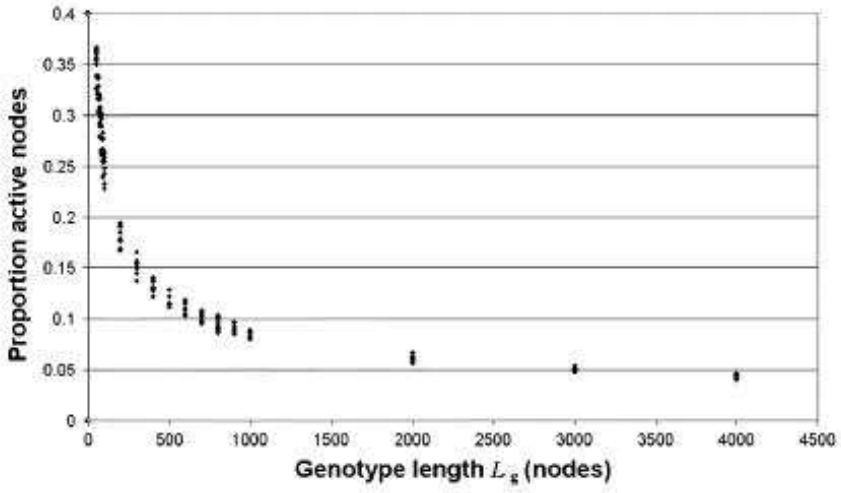


Figure 3.11: Average proportion of active nodes in genotype at conclusion of evolutionary run for all mutation probabilities versus genotype length (in nodes) for two-bit multiplier with gate set AND, OR, NAND, NOR. [6]

Miller et al have shown that by using a representation that implicitly has redundancy (graph) and by further exploiting this redundancy by using neutral search that efficient evolution of complicated logic circuits is possible. Furthermore, CGP uses one of the most basic selection algorithms $(1 + \lambda)$. Furthermore, it does require crossover, although crossover could be incorporated into it, or elaborate selection methods, because its effectiveness comes from redundancy and neutral search.

Chapter 4

Introduction to Antennas

4.1 Definition

According to [53]:

An antenna is a device which converts some of a guided electromagnetic wave propagating along a waveguide or transmission line into an unguided electromagnetic wave propagating through free space. An antenna is therefore a transducer. Generally, an antenna will also intercept some of the energy of an incident electromagnetic wave propagating through free space, and transduce it into an electromagnetic wave propagating along the waveguide or transmission line. 2 other mechanisms are also present:

(a) Loss

In either the transmission or reception case, some energy will be dissipated as ohmic losses within the antenna.

(b) Scattering

Some of the incident energy will be scattered from the antenna, i.e., in the case of transmission, it will be reflected back along the waveguide or transmission line towards the source. Scattering is described in terms of either the antenna input S parameter (reflection coefficient) or in terms of the energy reflected from the antenna structure.

4.2 Reciprocity

The underlying physics that determine how antennas work is linear. Maxwells equations and Ohms law are the main laws when describing the electromagnetics of antenna operation. Probably the most significant and useful consequence of the linearity of antennas is the principle of reciprocity. This states that an antenna in reception has a receive power pattern identical in shape to its radiation pattern in transmission. In other words an antenna will behave the same when it is transmitting or receiving. This is true for all antennas.

It is often more convenient, especially when considering the directive properties of antennas, to consider antennas in transmission. For this reason, the analysis of bandwidth and directionality is considered in the transmission case.

4.3 The Near Field and Far Field Regions

The purpose of communications antennas, when in transmission, is to radiate electromagnetic waves into free space. An electromagnetic wave, also known as a radiation field, can be defined as consisting of an electric ($\boldsymbol{E}$) and a magnetic ($\boldsymbol{H}$) field, which are in phase with each other but are orthogonal to each other and to the direction of propagation.

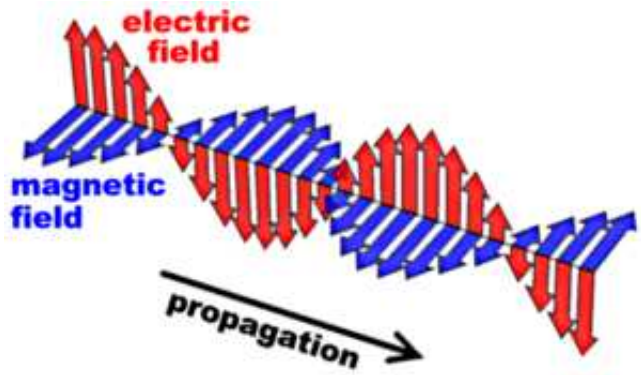


Figure 4.1: Linearly polarised electromagnetic wave. [7]

Spherical coordinates are commonly used when analysing antennas, and so the direction of propagation is radially away from the antenna, i.e., in the r direction. Additionally, in an electromagnetic wave traveling through a constant medium, the ratio of the electric to magnetic field magnitudes is a constant which depends on the medium.

When in transmission, as well as generating electromagnetic waves, all antennas also generate induction fields in their near vicinity. This region, close to the antenna, in which the induction fields are significant, is known as the near field region. In this region, the magnitudes of the induction fields may well be larger than that of the field components of electromagnetic waves.

For all real antennas, the induction fields of the near field zone, decay away faster with distance than radiation fields. This gives rise to the fact that at relatively large distances away from the antenna, only radiation fields will be significant. This region is known as the far field region.

All the fields generated by an antenna in transmission, whether they are induction fields or are components of electromagnetic waves, decay with distance (R) according to a $1/R^n$ function. For induction fields $n \geq 2$, but for radiation fields $n = 1$. This difference results in the faster decay of the radial fields.

There is no exact boundary between the near and far field regions because they merge into one another. A comprehensive analysis of the near-far field boundary of antennas can be found in [54]. For an infinitesimally small (Hertzian) dipole the distance from the dipole at which the induction fields and radiation fields are equal in magnitude is:

$$R_{boundary} = \frac{\lambda}{2\,\pi} \qquad (m) \tag{4.1}$$

However, for practical antenna measurements, the approximate distance, $R_{boundary}$, that is commonly used is:

$$R_{boundary} \approx \frac{2\,D^2}{\lambda} \qquad (m) \tag{4.2}$$

where:

$$D = \text{largest dimension of the antenna} \qquad (m)$$
$$\lambda = \text{free space wavelength} \qquad (m)$$

When expressed in spherical coordinates, an electric field at a particular point has the following form:

$$\boldsymbol{E} = \hat{\boldsymbol{r}} E_r + \hat{\boldsymbol{\theta}} E_\theta + \hat{\boldsymbol{\phi}} E_\phi \tag{4.3}$$

The characteristic impedance to electromagnetic waves (η_0) of free space is:

$$\eta_0 = \frac{|\boldsymbol{E}|}{|\boldsymbol{H}|} = \sqrt{\frac{\mu_0}{\epsilon_0}} \approx 377 \qquad (\Omega) \tag{4.4}$$

The differences between the near and far field regions can be summarised as follows:

Near field zone:

$$E_r \neq 0 \qquad \frac{|\boldsymbol{E}|}{|\boldsymbol{H}|} \neq \eta_0$$

Far field zone:

$$E_r \approx 0 \qquad \frac{|\boldsymbol{E}|}{|\boldsymbol{H}|} = \eta_0 \qquad \boldsymbol{E} = -\eta_0\, \hat{\boldsymbol{r}} \times \boldsymbol{H} \qquad \boldsymbol{H} = \frac{1}{\eta_0}\, \hat{\boldsymbol{r}} \times \boldsymbol{E}$$

4.4 Isotropic Antenna

An isotropic antenna is a hypothetical antenna that radiates or receives equally in all directions. Isotropic antennas do not exist physically but represent convenient reference antennas for expressing directional properties of physical antennas.

In Fig. 4.2 the isotropic antenna is radiating the input power (Pin) equally in all directions. The power density S at radius R from the antenna is then:

$$S(R) = \frac{\text{input power}}{\text{area of sphere}} = \frac{P_{in}}{4\pi R^2} \qquad (Wm^{-2}) \tag{4.5}$$

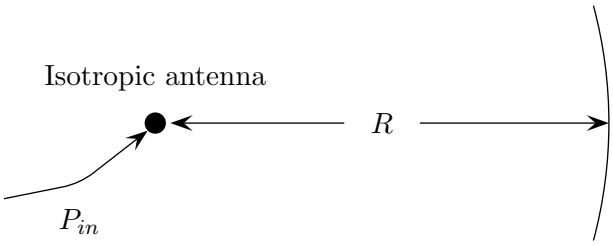


Figure 4.2: Isotropic antenna.

4.5 Directionality

This section introduces the various concepts associated with the directional properties of antennas. The term directionality has been used as the heading of this section because the term directivity has a more specific meaning which is given below.

4.5.1 Antenna Coordinate System

The directional properties of an antenna depend on its orientation in 3 dimensional space. A coordinate system is therefore required to accurately determine the direction in which an antenna is pointing. Since energy radiates outwards from an antenna when it is radiating, spherical coordinates are the most natural coordinate system.

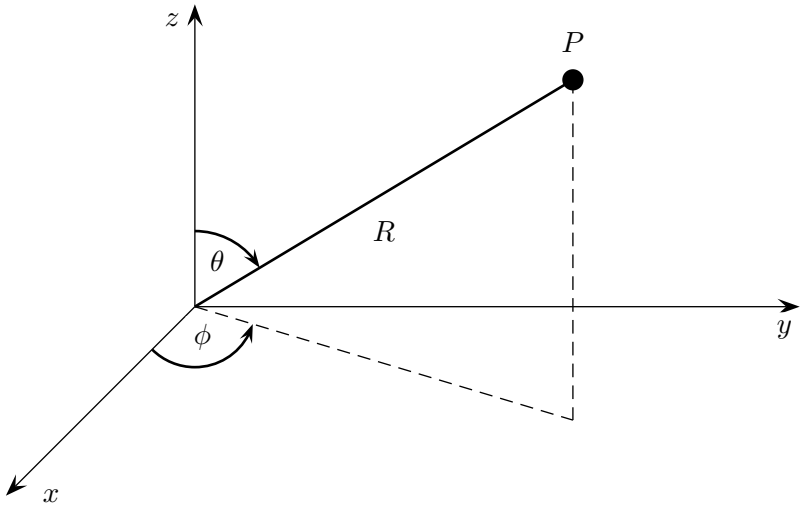


Figure 4.3: Spherical coordinates.

All real antennas do not radiate equally in all directions. More power is radiated in certain directions than in others. This leads to the various parameters that describe the directional properties of antennas being functions of θ and ϕ.

4.5.2 Solid Angle

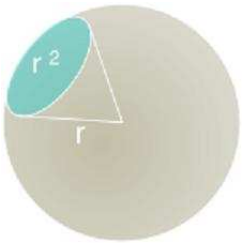


Figure 4.4: Graphical representation of 1 steradian. [8]

Plane angles, which are usually measured in degrees or radians, exist in two dimensional space (planes). Solid angles are the three dimensional equivalent of plane angles. The most common unit of measuring solid angles is the steradian.

One steradian is defined as [12, page 34]: the solid angle with its vertex at the centre of a sphere of radius r that is subtended by a spherical surface area equal to that of a square with each side of length r. Since the area of a sphere of radius r is $4\pi r^2$ there are 4π radians in a closed sphere.

Thus the area subtended by the differential solid angle element $d\Omega$, with a radius of r, is:

$$dA = r^2 d\Omega = r^2 \sin\theta \, d\theta \, d\phi \tag{4.6}$$

4.5.3 Radiation Intensity

Radiation intensity, U, in a given direction is defined as [12, page 38]: the power radiated from an antenna per unit solid angle. Radiation intensity is a far field parameter and is related to the radiated power density S as follows:

$$U = U(\theta, \phi) = r^2 \, S \tag{4.7}$$

where:

$$U = \text{radiation intensity } (W/steradian)$$
$$S = \text{radiation density } (W/m^2)$$

The total radiated power can be found by integrating the radiation intensity over the entire solid angle that surrounds the antenna:

$$P_{rad} = \int_0^{4\pi} U \, d\Omega = \int_0^{2\pi} \int_0^{\pi} U \sin\theta \, d\theta \, d\phi \tag{4.8}$$

Due to the fact that isotropic antennas radiate equally in all directions, their radiation intensity U_0 is constant with respect to direction:

$$U_0 = \frac{P_{rad}}{4\pi} \tag{4.9}$$

4.5.4 Directivity

The directivity, D, of an antenna is defined as [55]: the ratio of maximum radiation intensity to the average (isotropic) radiation intensity.

$$D = \frac{U_{max}}{U_0} \tag{4.10}$$

The proportion of input power that is actually radiated by the antenna depends upon the antenna's radiation efficiency ξ. Radiation efficiency is discussed in more detail in section 4.6.

The gain of an antenna, G, is a function of both its directivity and its radiation efficiency:

$$G = \xi\, D$$

$$P_{rad} = \xi\, P_{in} \tag{4.11}$$

where:

$$\begin{aligned} P_{in} &= \text{input power to antenna } (W) \\ P_{rad} &= \text{radiated power } (W) \\ \xi &= \text{radition efficiency} \quad (0 \leq \xi \leq 1) \end{aligned}$$

Eqn. 4.11 can probably best be explained with reference to Fig. 4.7. and section 4.6.1. P_{in} is the the power going into the antenna and P_{rad} is the power being radiated from the antenna when in transmission. If the antenna was constructed of perfectly conducting materials, then no energy would be lost in the antenna itself and all of P_{in} would be radiated. This would correspond to a radiation efficiency (ξ) of 1 (100 %). In reality, all practical antennas have some energy lost within them and so have radiation efficiencies of less than 1. The radiation intensity observed in a given direction is thus related to the radiated and input power as follows:

$$U(\theta, \phi) = \frac{P_{rad}}{4\pi}\, D(\theta, \phi) = \frac{\xi P_{in}}{4\pi}\, D(\theta, \phi) \tag{4.12}$$

4.5.5 Radiation Patterns

All real antennas radiate more power in certain directions than in other directions. A radiation pattern plot is a graphical representation of how the magnitude of an antenna's radiated power varies with direction from the antenna. The directions in which the antenna radiates most of its power are known as lobes. The directions in which it radiates the least power are known as nulls. Lobes ans nulls can be seen in Figs. 4.5 & 4.6.

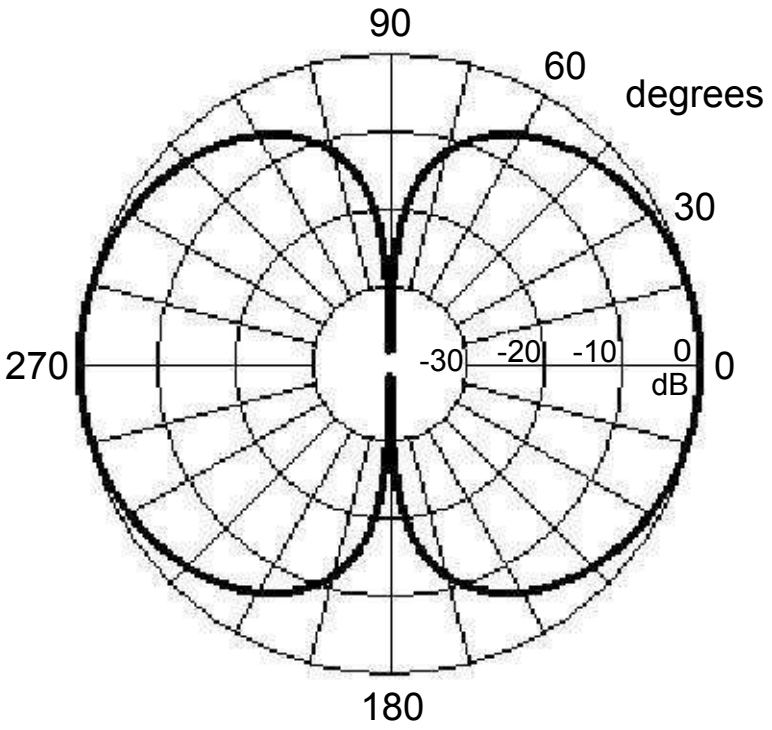


Figure 4.5: Elevation plane radiation pattern of a dipole antenna.

The radiation pattern describes the relative strength of the radiated power in various directions from the antenna, at a constant distance. Due to the reciprocity, the radiation pattern is a reception pattern as well. The radiation pattern is three-dimensional, but usually the measured radiation patterns are two dimensional slices of the three-dimensional pattern, typically in horizontal or vertical planes. Radiation patterns are presented in either rectangular or polar format.

Conventionally, the power density is plotted relative to the maximum power level, i.e., relative to the power density that occurs at the direction in which the antenna is radiating most power. It is also usually expressed in dB. Alternatively, the power density can be expressed in dBi. This is dB relative to an isotropic antenna that is radiating the same total amount of power as the antenna in question. For example, if in a particular direction an antenna has a directivity of +3dBi, then the antenna is radiating twice as much power in this direction as an isotropic antenna that is radiating the same total amount of power.

4.5.6 Units

The gain of an antenna is often quoted with reference to an isotropic antenna. This means that the power from the antenna is measured relative to the power from an isotropic antenna. This ratio, when expressed in dBs, is in dBi. Since isotropic antennas are not practical antennas, often antenna gain is expressed relative to a simple dipole. This ratio, when expressed in dBs, is in dBd.

$$0 \text{ dBd} = 2.15 \text{ dBi}$$

4.5.7 Half Power Beam Beam Width (HPBW)

The half power, or 3 dB, beam width is the angular width over which the main (highest magnitude) lobe radiation intensity is greater than or equal to half (-3 dB) of the maximum radiation intensity.

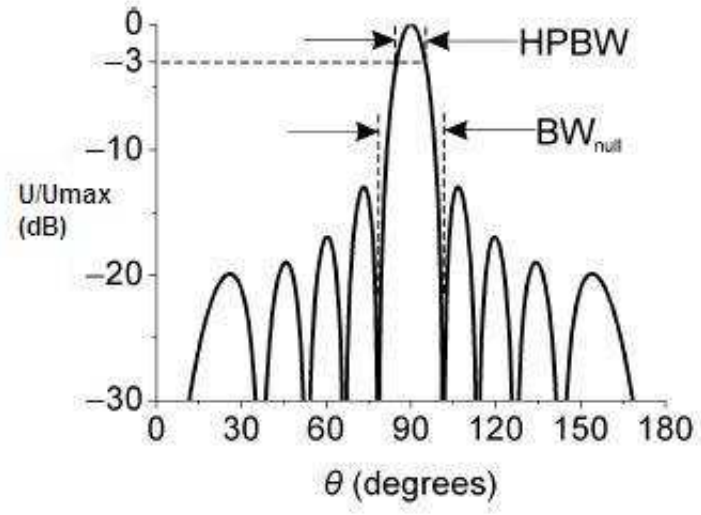


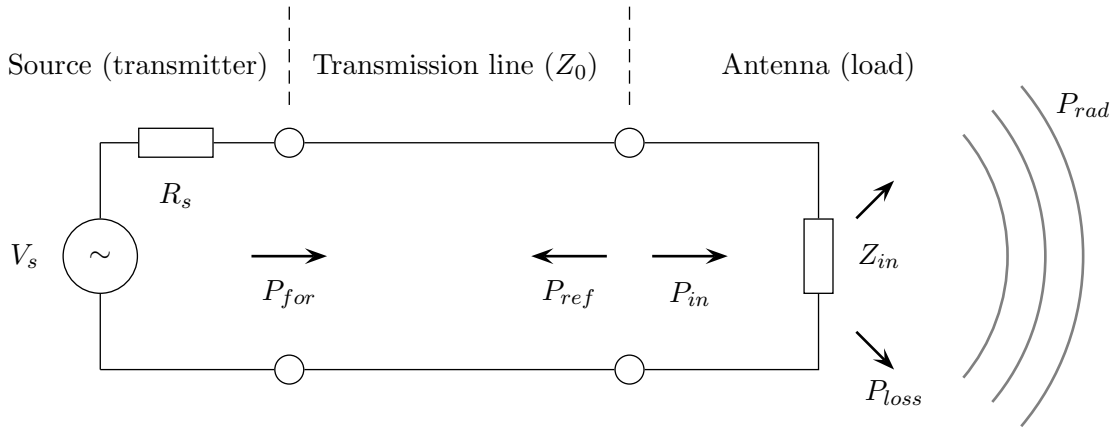
Figure 4.6: Half power beam width. [9]

4.6 Bandwidth

This section introduces the various concepts associated with the bandwidth of antennas.

The bandwidth of an antenna refers to the range of frequencies over which the antenna can operate satisfactorily. It is usually defined by impedance mismatch (impedance bandwidth). Antennas in transmission are modeled simply as a complex load impedance. This is because antennas both dissipate energy, through radiation and various losses, and store energy, in near fields. The dissipation of energy gives rise to a resistive part of the input impedance and the storage of energy gives rise to a reactive part. When the antennas's input impedance is poorly matched to the transmitter's and transmission line's impedance, a large proportion of the incident, i.e., power from the transmitter, will be reflected back from the antenna's input back towards the transmitter. This will reduce the amount of radiated power and could also damage the transmitter.

4.6.1 Input Impedance



P_{for} = incident (forward) power from transmitter
P_{ref} = reflected power from antenna's input
P_{in} = input power to antenna
P_{loss} = power lost in antenna
P_{rad} = radiated power
Z_{in} = antenna's input impedance

Figure 4.7: Power flow of an antenna in transmission.

In Fig. 4.7 the transmission line has a characteristic impedance of Z_0 ohms and is assumed to be lossless. The antenna's input impedance is complex:

$$Z_{in} = R_{in} + j\,X_{in} \tag{4.13}$$

How well an antenna's input impedance is matched to that of the characteristic impedance of the system that is driving it (typically 50 Ω) is commonly expressed using the voltage reflection coefficient, ρ.

$$\rho = \frac{Z_{in} - Z_0}{Z_{in} + Z_0} \tag{4.14}$$

The proportion of power reflected from the antenna's input is equal to ρ^2. The relationship between the forward, reflected and input powers and ρ are as follows:

$$P_{for} = P_{ref} + P_{in} \qquad \Longleftrightarrow \qquad P_{in} = P_{for} - P_{ref}$$
$$\text{where:} \quad P_{in} = (1 - \rho^2)\, P_{for}$$

An antenna is a one port network and so only has one S parameter, S_{11}, which is the same as the voltage reflection coefficient:

$$S_{11} = \rho$$

An antenna's input impedance, and thus its reflection coefficient, changes with frequency. This is the reason why there is always only a limited number of limited frequency ranges over which an antenna will be well matched to the system. Outside of these ranges, the antenna will be poorly matched and as such will be a poor radiator.

Another fairly common measure of antenna impedance matching is the voltage standing wave ratio (VSWR). VSWR is related to ρ as follows:

$$\text{VSWR} = \frac{1 + |\rho|}{1 - |\rho|} \tag{4.15}$$

4.6.2 Radiation Efficiency

The real part of Z_{in} is due to the fact that the antenna dissipates power (P_{in}). However, not all of the input power is radiated as electromagnetic waves. Some of the power is lost as heat in the antenna due ohmic losses in the conductors and losses in any dielectrics, if there are any. All the various losses are lumped together into a single term, P_{loss}. The total input power is then the sum of the radiated power and the lost power:

$$P_{in} = P_{rad} + P_{loss} \tag{4.16}$$

The antenna's input resistance (R_{in}), must take into account the power radiated by the antenna and the power lost by the antenna, and so it takes the form:

$$R_{in} = R_{rad} + R_{loss} \tag{4.17}$$

The greater the proportion of input power that is lost as heat, the lower the proportion of radiated power will be. An antenna's radiation efficiency, ξ, is defined [56] as the ratio of radiated power to its total input power:

$$\xi = \frac{P_{rad}}{P_{in}} = \frac{P_{rad}}{P_{rad} + P_{loss}} = \frac{R_{rad}}{R_{rad} + R_{loss}} \tag{4.18}$$

4.6.3 Bandwidth Plots

Since the bandwidth of an antenna refers to the range of frequencies over which the antenna can perform satisfactorily, it is usual to present its impedance match against frequency. The most commonly used measure of impedance match is the voltage reflection coefficient (ρ or S_{11}), as described above. It is also common practice to express S_{11} in decibels (dB).

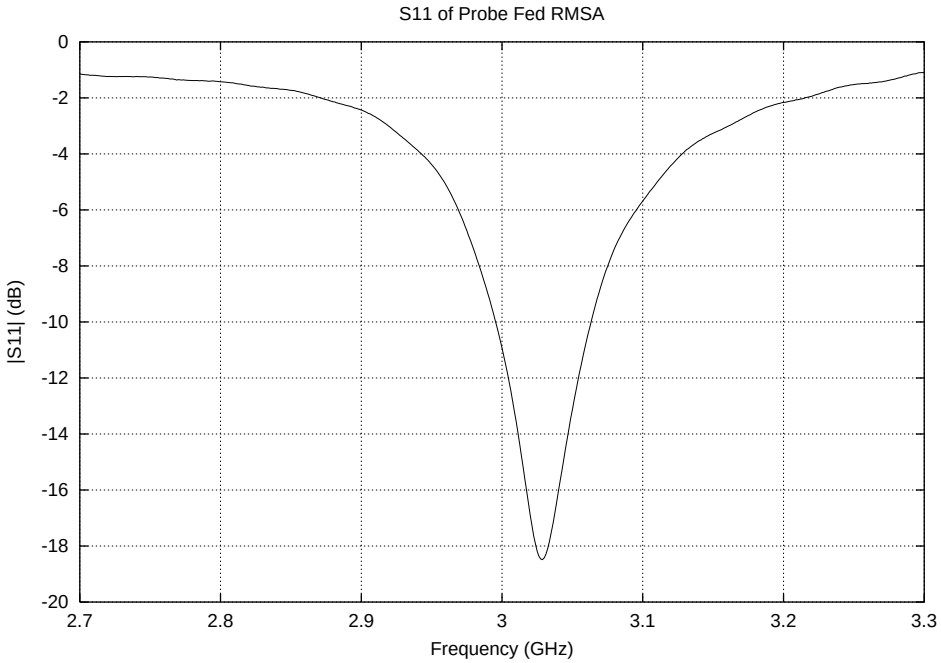


Figure 4.8: S_{11} of Rectangular Microstrip Antenna (RMSA).

When expressed in dB, an S_{11} of zero means that all of the incident power at the antenna's input is being reflected back towards the transmitter. This is the worst possible case as it means that no power will be radiated. In contrast, the more negative the S_{11} value is, the lower the proportion of power reflected is. A widely accepted measure of bandwidth is the 10 dB bandwidth. This is the frequency range over which an antenna's S_{11} is less than or equal to -10 dB. In Fig. 4.8 the 10 dB bandwidth is 0.070 GHz (70 MHz).

Bandwidths are often expressed as the ratio of the bandwidth, Δf to the centre frequency of the band f_0. This is known as fractional or percent bandwidth:

$$BW = \frac{\Delta f}{f_0}$$

For most antennas, the centre frequency of a band is usually the lowest (most negative) S_{11} value. In Fig. 4.8 the centre frequency is 3.029 GHz, and so the fractional bandwidth is:

$$BW = \frac{70}{3029} = 0.0231 = 2.31\%$$

4.6.4 Quality (Q) Factor

The Quality Factor, or Q as it is commonly known, of a resonant system is the ratio of stored energy to lost energy per cycle. More specifically, according to [57], the Q is defined to be 2π times the ratio of the maximum energy stored to the total energy lost per period. For an antenna, the following definition for Q is generally accepted:

$$Q = \frac{\omega W}{P_{rad}} \tag{4.19}$$

Where W is the total stored energy.

A system that stores a relatively high proportion of energy compared to the proportion of energy lost will have a relatively high Q factor. When energised with an impulse (a brief burst of energy) high Q systems resonate for many cycles because they do not lose their stored energy quickly. Conversely, the impulse response of low Q systems decays away relatively quickly.

When considering electrical/electronic systems, it is often more useful to consider the frequency response, i.e., when the system is fed with steady-state sinusoids over a given frequency range. When describing the performance of electrical/electronic filters, Q factor is a convenient concept. The higher the Q factor the narrower the frequency response becomes, as can be seen in Fig. 4.9.

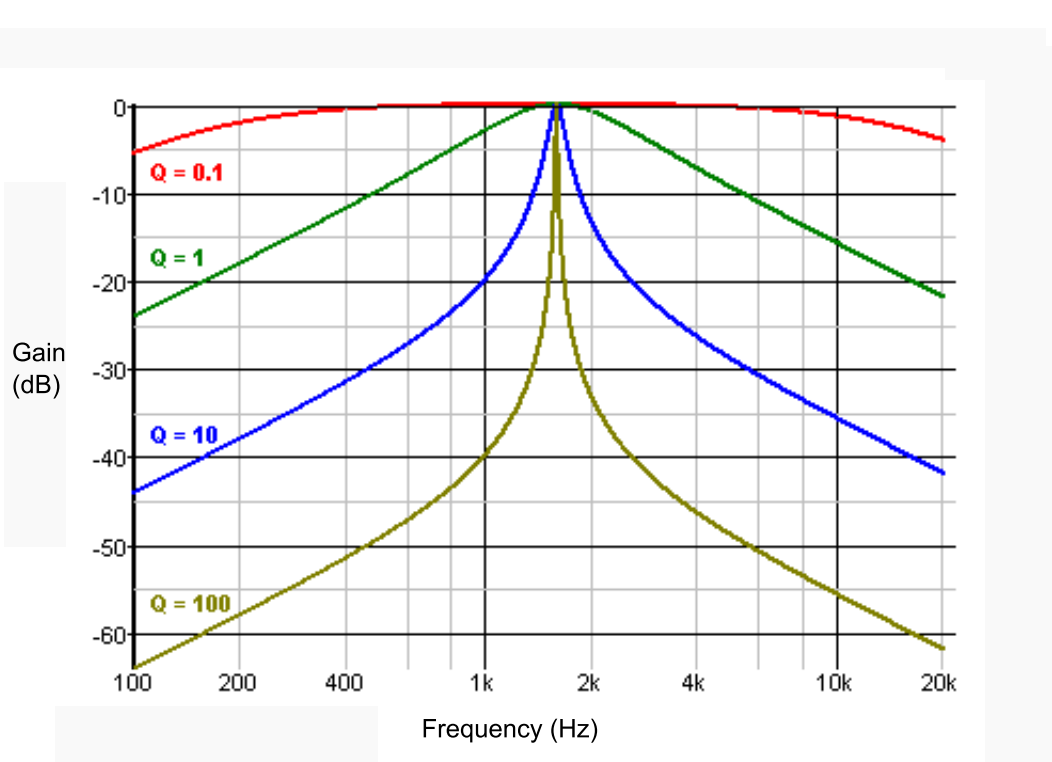


Figure 4.9: Frequency response of an electronic filter with variable Q. [10]

A general phenomenon of resonant systems is that the Q factor is the inverse of the bandwidth:

$$Q = \frac{1}{BW} = \frac{f_0}{\Delta f} \tag{4.20}$$

In the case of antennas, a more specific expression relating Q and bandwidth can be used [58] [59]:

$$BW = \frac{S-1}{Q\sqrt{S}} \tag{4.21}$$

where:

$$S = S : 1 \text{ VSWR}$$

A VSWR of 2 to 1, i.e., S = 2, corresponds to a voltage reflection coefficient (ρ) of -10 dB.

More specifically, S, in this context, can be thought of as the required VSWR of the antenna. It is described by [59] as:

The value of VSWR which can be tolerated then defines the bandwidth of the antenna. If this value is to be less than S, the usable bandwidth of the antenna is related to the total Q factor.

An example, that looks at this point from a slightly different angle, would be that of an antenna with a single return loss null centered at 2.44 GHz. The return loss of the antenna is -10 dB (S = 2) at 2.4 GHz and 2.48 GHz, with the return loss being more negative than -10 dB in between these frequencies and with it being more positive than -10 dB outside of these frequencies. The -10 dB bandwidth of the antenna is then 80 MHz. Therefore, the corresponding values that could be used in eqn. 4.21 would be: S = 2 and BW = 0.08 / 2.44 = 0.0328. If a different value of acceptable return loss is chosen (e.g. -6 dB) then S and thus BW will correspondingly change as well. In the case of a return loss of -6 dB, the fractional bandwidth (BW) will be higher as this is at a wider point in the null, i.e., further from the bottom of the null.

4.7 Polarisation

An antenna's polarisation refers to the polarisation of the electromagnetic waves that it radiates (when in transmission). The polarisation of an electromagnetic wave is the time-varying orientation of the electric field component. A more specific definition is given in [56]:

the polarisation of a wave describes the shape and locus of the tip of the $\boldsymbol{E}$ *vector (in the plane orthogonal to the direction of propagation) at a given point in space as a function of time.*

At different directions from an antenna, different polarisations may occur. Polarisation usually means the polarisation in the direction of maximum directivity. Due to reciprocity, when in reception an antenna will only pick up waves of the same polarisation that it would produce if it were transmitting.

There are three types of polarisation: linear, circular and elliptical. The most general is elliptical. linear and circular are special cases of elliptical polarisation.

4.7.1 Elliptical Polarisation

Consider an electromagnetic wave plane propagating in the z direction, the electric field phasor is:

$$\boldsymbol{E}(z,t) = \hat{\boldsymbol{x}}\, E_x(z,t) \;+\; \hat{\boldsymbol{y}}\, E_y(z,t) \tag{4.22}$$

where:

$$E_x(z,t) = E_{x0}\, e^{j(\omega t - kz + \phi_x)} \tag{4.23}$$

$$E_y(z,t) = E_{y0}\, e^{j(\omega t - kz + \phi_y)} \tag{4.24}$$

It is the relative phase difference between the E_x and E_y components that is important, so for simplicity the phase of E_x can be set to zero and the phase of E_y can be adjusted accordingly:

$$E_x(z,t) = E_{x0}\, e^{j(\omega t - kz)} \tag{4.25}$$

$$E_y(z,t) = E_{y0}\, e^{j(\omega t - kz + \phi)} \tag{4.26}$$

$$\text{where:} \qquad \phi = \phi_y - \phi_x \tag{4.27}$$

The instantaneous electric field, $\boldsymbol{e}(z,t)$, is the real part of the electric field phasor:

$$\boldsymbol{e}(z,t) = \Re\{\boldsymbol{E}(z,t)\} = \hat{\boldsymbol{x}}\, E_{x0}\, \cos(\omega t - kz) + \hat{\boldsymbol{y}}\, E_{y0}\, \cos(\omega t - kz + \phi) \tag{4.28}$$

The magnitude of the instantaneous electric field as the wave propagates through space and time is:

$$|\boldsymbol{e}(z,t)| = \sqrt{|E_{x0}|^2\, \cos^2(\omega t - kz) + |E_{y0}|^2\, \cos^2(\omega t - kz + \phi)} \tag{4.29}$$

4.7.2 Linear Polarisation

Linear polarisation is a special case of elliptical polarisation in which the ellipse collapses down to a straight line. This is because the E_x and E_y components are in phase ($\phi = 0$):

$$|\boldsymbol{e}(z,t)| = \cos(\omega t - kz)\, \sqrt{|E_{x0}|^2 + |E_{y0}|^2} \tag{4.30}$$

4.7.3 Circular Polarisation

Circular polarisation is another special case of elliptical polarisation. The magnitudes of the E_x and E_y components are the same and there is a quarter of a cycle phase shift between them:

$$\phi = \pi/2$$

$$|E_{x0}| = |E_{y0}| = |E_0|$$

$$\implies \qquad |\boldsymbol{e}(z,t)| = |E_0| \tag{4.31}$$

4.7.4 Visualising Polarisation

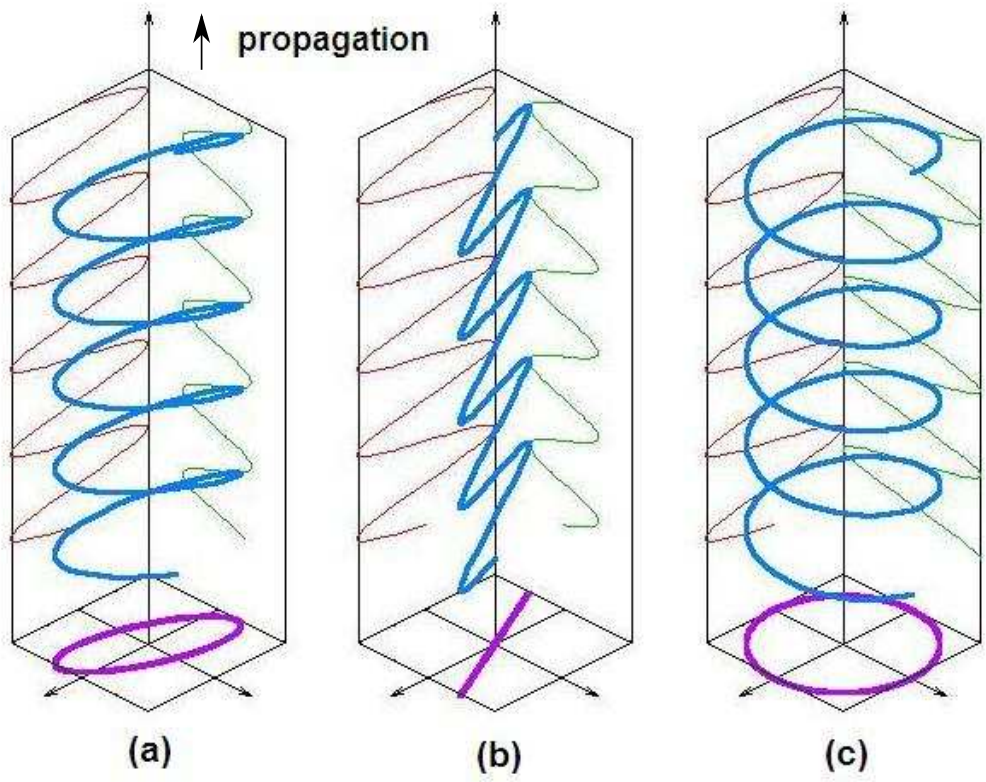


Figure 4.10: Polarisation: (a) elliptical, (b) linear, (c) circular. [11]

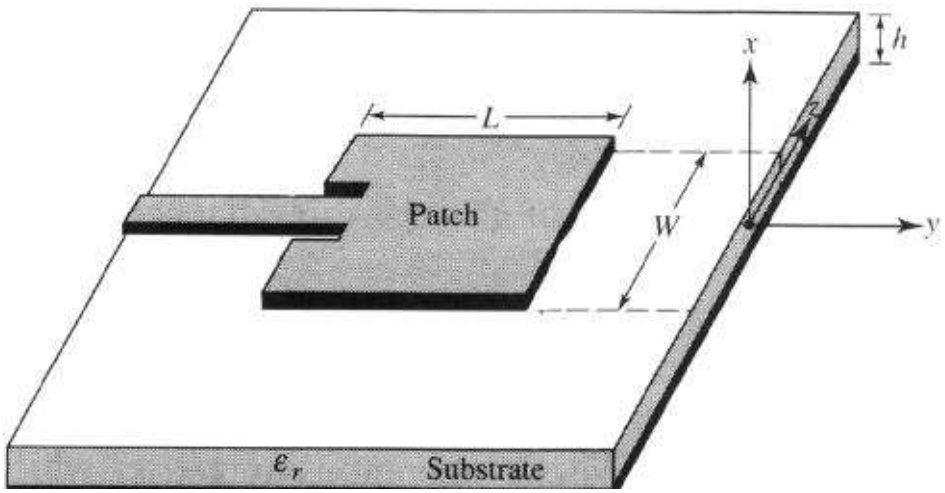
In Fig. 4.10 the pattern that the electric field vector creates as it propagates through space and time, in blue, can be seen for the three types of wave polarisation. The red and green components are the E_x and E_y components respectively. The projection of the electric field onto a fixed plane that is orthogonal to the direction of propagation is the shape in purple at the bottom of each figure.

All real antennas have elliptical polarisation. In the case of an antenna that is specified as being linearly polarised the E_x and E_y components will never actually be exactly in phase and so a long thin ellipse will result. In the case of circularly polarised antennas, the E_x and E_y components will never have exactly the same amplitude or exactly a $\pi/2$ phase shift between them or both. This will result in an imperfect circle.

Chapter 5

Introduction to Microstrip Antennas

Microstrip antennas (MSA) in their simplest form consist of a metal patch on one side of thin dielectric substrate with a ground plane on the other side. For obvious reasons, MSA are often referred to as patch antennas. The patch can be any shape, but in general, the simpler and more regular the shape the easier the design and analysis is. The simplest type of MSA is the rectangular MSA (RMSA), an example of which can be seen in figure 5.1.



L = patch length
W = patch width
h = substrate thickness
ϵ_r = relative permittivity of dielectric

Figure 5.1: Geometry of RMSA. [12]

5.1 MSA Patch Shapes

In theory there is no limit as to the shape that the patch may take. MSAs can contain more than one patch which may have differing sizes and/or shapes. MSA's containing a huge variety of patch shapes and configurations have been built for varying applications. Fig. 5.2 below, shows some of the more common patch shapes.

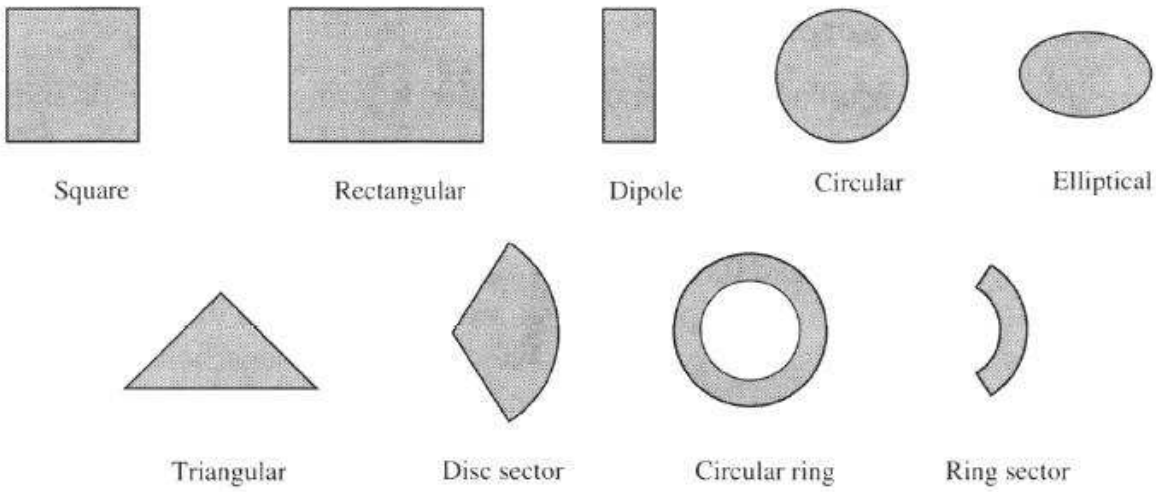


Figure 5.2: Common MSA patch shapes. [12]

5.2 MSA Feed Types

There are several ways in which MSAs can be fed, i.e., supplied with the signal that they are required to radiate. MSA feed types fall into one of two categories: contacting and non-contacting. In the contacting method, the input power is fed directly to the patch, i.e., conduction current can flow directly from the feed connector onto the patch. In the non-contacting scheme, the input power is transferred to the patch by electric and magnetic fields which couple between the feed structure and the patch.

5.2.1 Microstrip Line Feed

In this type of feed technique, a microstrip transmission line is connected directly to the patch, as in Fig. 5.3. The input impedance can be quite easily matched to the transmission line by controlling the size of the inset, see Fig. 5.1.

Microstrip feeding is easy to fabricate and straightforward to model computationally. However, as the substrate thickness increases the magnitude of surface waves and spurious radiation from the feed line increases. This reduces the efficiency and limits the bandwidth (typically $2-5\%$) [12].

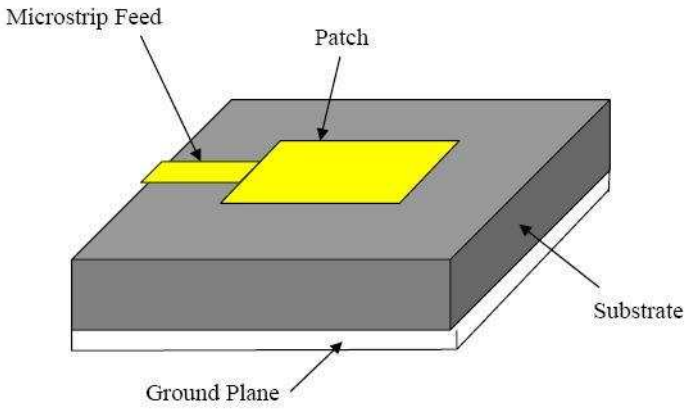


Figure 5.3: Microstrip line fed MSA. [13]

5.2.2 Probe Feed

In probe feeding, the central conductor of the coaxial connector is attached to the patch while the outer conductor is connected to the ground plane, see Fig. 5.4. This type of feeding is widely used because it is easy to fabricate. Its major disadvantage is low bandwidth. For thin substrates it is relatively easy to model and match. When the substrate becomes thicker ($h > 0.02\lambda_0$) is is more difficult to match because of the increased probe inductance.

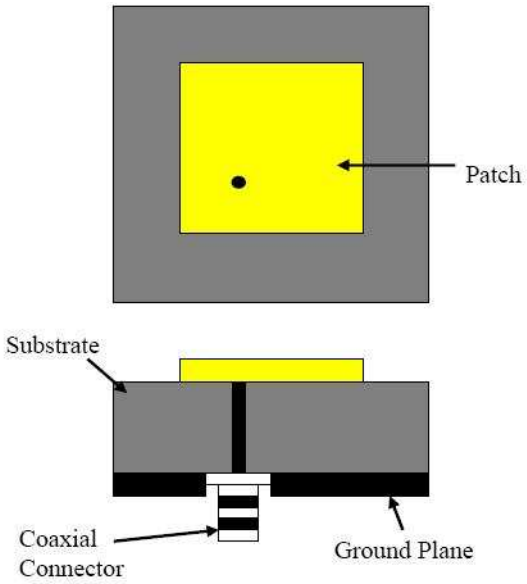


Figure 5.4: Probe fed MSA. [13]

5.2.3 Aperture Coupled Feed

The aperture coupled feed method consists of two substrates separated by a ground plane, as in Fig. 5.5. There is a microstrip transmission line on the bottom side of the lower substrate. Energy is coupled from this line to the patch, by electric and magnetic fields, through a slot in the ground plane. This type of feed is the hardest to fabricate and also has narrow bandwidth. However, it is relatively easy to model and has low spurious radiation [12]. Impedance matching is achieved by adjusting the width of the feed line and the length of the slot.

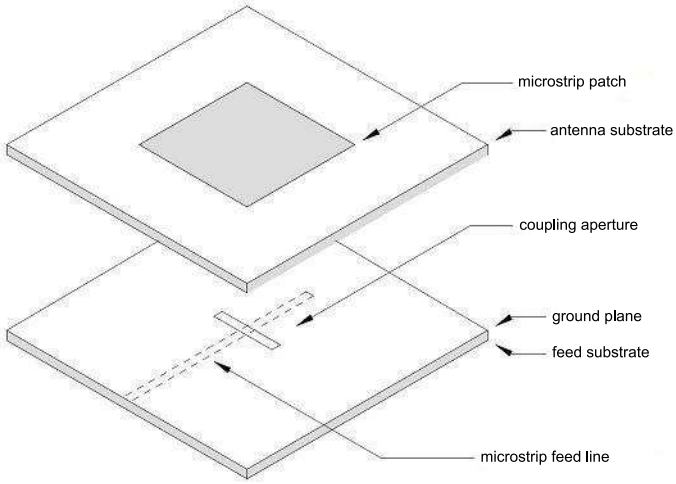


Figure 5.5: Aperture coupled fed MSA. [14]

5.2.4 Proximity Feed

As can be seen in Fig. 5.6, proximity feeding consists of a microstrip line in between two substrates and a patch on the top surface of the upper substrate. Electric and magnetic fields couple between the microstrip line and the patch. The main advantage of this feed technique is that it eliminates spurious feed radiation and provides very high bandwidth (as high as 13%) [13]. Matching is performed by controlling the length of the feed line and the width-to-line ratio of the patch. The two substrates need to be properly aligned, thus making the fabrication process more difficult.

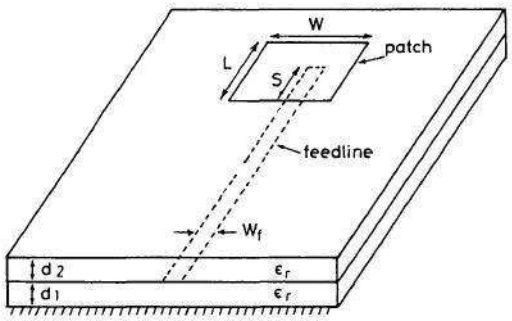


Figure 5.6: Proximity fed MSA. [15]

5.3 RMSA Operation

In this section, the main aspects of RMSA operation are described. These include: the radiation mechanism and impedance matching. The RMSA has been chosen because it is the simplest and most intuitive type of MSA.

5.3.1 Radiation Mechanism

All MSAs are resonant structures. This is because waves propagate within MSAs and are reflected at the patch edges. This leads to the formation of standing waves within MSAs. It is easiest to visualise this phenomenon when considering RMSAs.

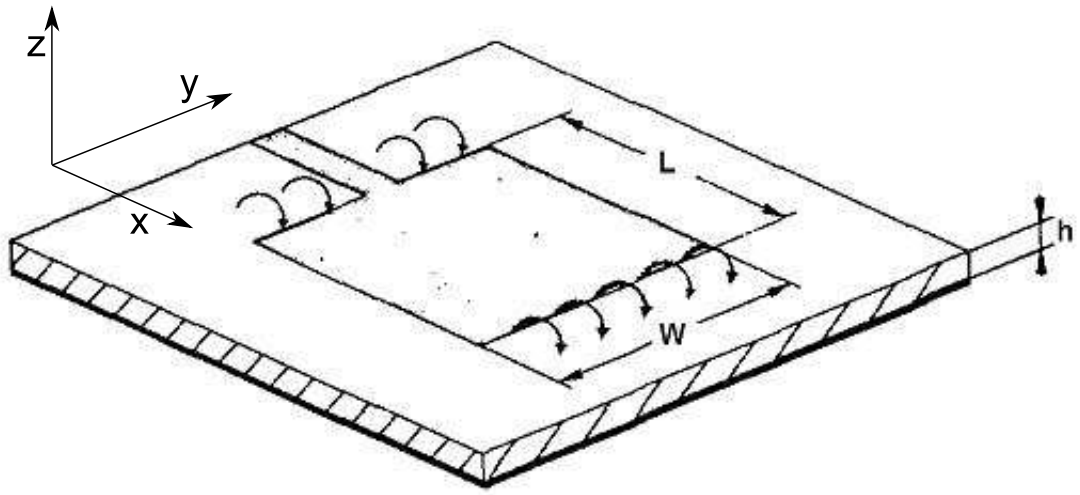


Figure 5.7: RMSA and its radiating slots. [16]

In between the patch and the ground plane are significant electric fields which are normal to the patch and ground plane. At the edges (lengthways) of the patch, the electric field *fringes* around the top surface of the patch, see Fig. 5.9. Where the fringing fields are in free space, there is effectively a constant, uniform direction (lengthways) electric field. These fields form what is known as the radiating slots. Their width is the same as that of the patch. It is these slots which radiate energy away from the antenna.

With reference to Fig. 5.7, within the radiating slots the electric field only has a single component (E_x). This leads to only a single electric field component in the far field (E_θ). Because only electromagnetic waves exist in the far field region, the only magnetic field component will be H_ϕ. This is the reason why RMSAs are linearly polarised. If an RMSA is orientated such that the lengthways direction (x in Fig. 5.7) is vertical then the RMSA will be vertically polarised. If it is orientated so that the lengthways direction is horizontal then it will be horizontal polarised.

Due to the fact that there are two slots, one at each end, the RMSA is effectively an array antenna consisting of two slots. Each slot can be modeled as an aperture antenna. The total radiation pattern of the RMSA can then be determined analytically by combining the radiation of each aperture using array antenna techniques.

5.3.2 Loss Mechanism

The two main loss mechanisms in dielectrics are known as *conduction* and *dielectric* loss. Conduction loss occurs when the material has some conductivity (σ) and thus current flows in the material when an electric field is applied to it. This results in energy being dissipated as heat in the material. Dielectric loss is caused by electrons (in the atoms of the material) moving in response to the applied electric field. There is some opposition to this movement which results in energy being lost as heat. The combined effects of conduction and dielectric loss can be incorporated into the *loss tangent* ($\tan\,\delta$), which is a commonly used metric of the overall loss of a dielectric.

$$\tan\,\delta = \frac{\omega\,\epsilon'' + \sigma}{\omega\,\epsilon'} \tag{5.1}$$

where the permittivity (ϵ) of the dielectric is expressed as:

$$\epsilon = \epsilon' + j\epsilon'' \tag{5.2}$$

In equation 5.2, ϵ'' is the reactive permittivity due to the dielectric loss. In most practical MSAs, dielectric loss is by far the dominant loss mechanism as most of the substrates used to make PCBs have extremely low conductivity.

Another way in which energy is lost in MSAs, rather than it being radiated, is via surface waves. Surface waves are generated in the cavity between the patch and the ground plane. They propagate in the substrate, i.e., between the ground plane and the dielectric/air interface due to total internal reflection. When the surface waves reach the edges of the substrate they are reflected, scattered and diffracted.

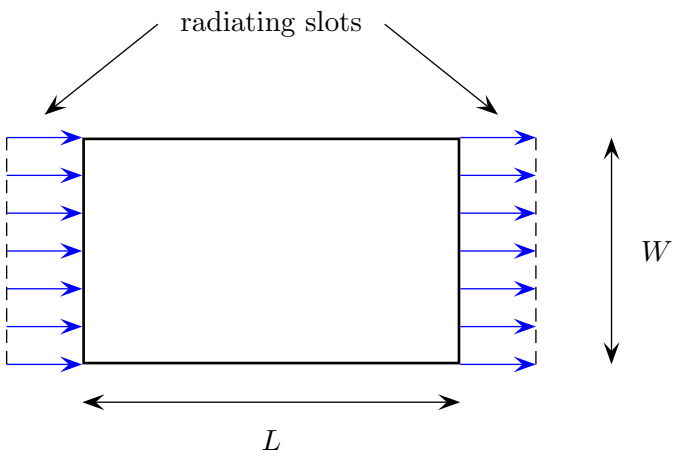


Figure 5.8: Top view of RMSA and its radiating slots.

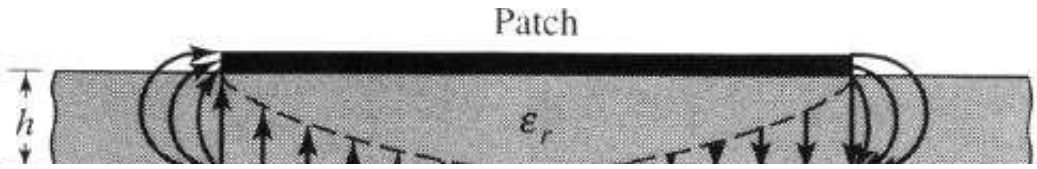


Figure 5.9: Side view of RMSA. [17]

5.3.3 Current, Voltage and Input Impedance

RMSAs are resonant structures. Their fundamental resonant mode is when the RMSA is half a wavelength long. Due to the fringing fields, the actual length of the patch is slightly less than half a wavelength.

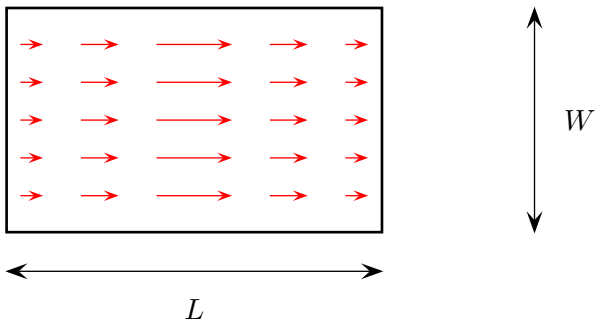


Figure 5.10: Top view of surface current distribution of RMSA.

The theoretical surface current distribution of the fundamental mode can be seen in Fig. 5.10. A simple but convenient way of looking at RMSAs is to assume that the patch and ground plane form a length of microstrip transmission line. As far as the current and voltage distribution on the transmission line is concerned, at each (radiating) end of the transmission line there is an open circuit. Since no current can flow into an open circuit, the current distribution is zero at each radiating edge. Additionally, the voltage is maximum at the open circuits and falls to zero in the centre of the patch.

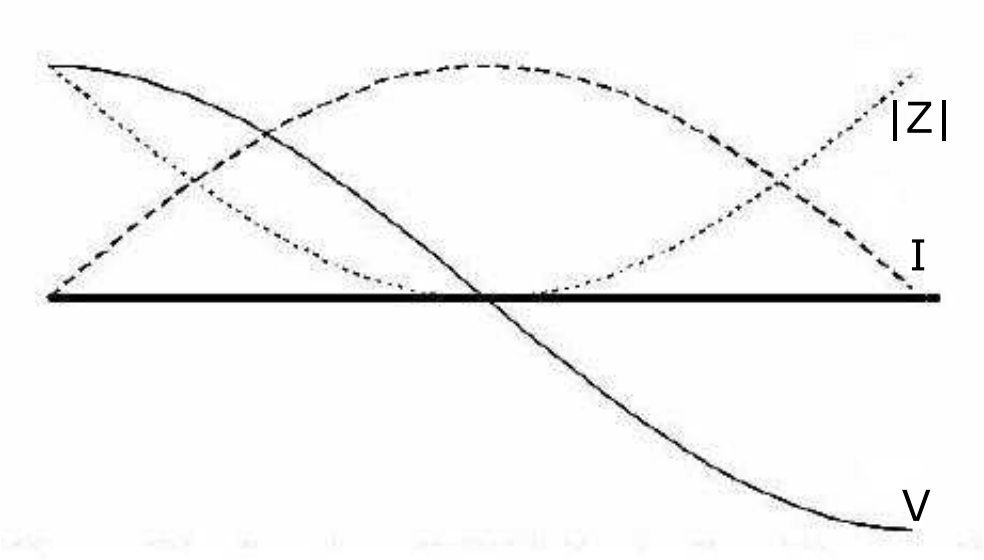


Figure 5.11: Current (I), voltage (V) and input impedance ($|Z|$) along the length of an RMSA. [18]

The input impedance that the feed 'sees' is simply the ratio of voltage to current at the feed point. The input impedance goes from a maximum at the radiating edges to zero in the centre of the patch. By placing the feed point in the appropriate position, the input impedance of the patch can be matched to the impedance of the system driving it (typically $50\,\Omega$).

5.4 Main Characteristics of MSAs

In general, MSAs, as compared to many other types of antennas, have relatively low bandwidths. The main reason for this is that they have relatively high Q factors. This is due to the fact that the ratio of stored energy to radiated energy is relatively high. Since bandwidth and Q are inversely related, MSA bandwidths are low.

The radiation pattern of an MSA depends on its size. In the case of an RMSA, for a given frequency and substrate, changing the patch width will change the radiation pattern, as in section 5.5.2.

MSAs are often used in mobile communications devices, such as mobile phones, because they have a low profile and sufficiently small footprint. Conveniently, this small footprint typically results in a simple radiation pattern, consisting of a single broad main lobe.

Due to the fact that RMSAs are linearly polarised, i.e., the electric field only having a single component in the lengthways direction of the patch, two orthogonal radiation pattern 'cuts' present themselves. These are the E plane and the H plane. The E and H planes are lengthways and width ways cuts respectively. With reference to Fig. 5.9, the E and H planes are:

E plane: (x-z plane)	$-180^\circ \leq \theta \leq 180^\circ$	$\phi = 0^\circ$
H plane: (y-z plane)	$-180^\circ \leq \theta \leq 180^\circ$	$\phi = 90^\circ$

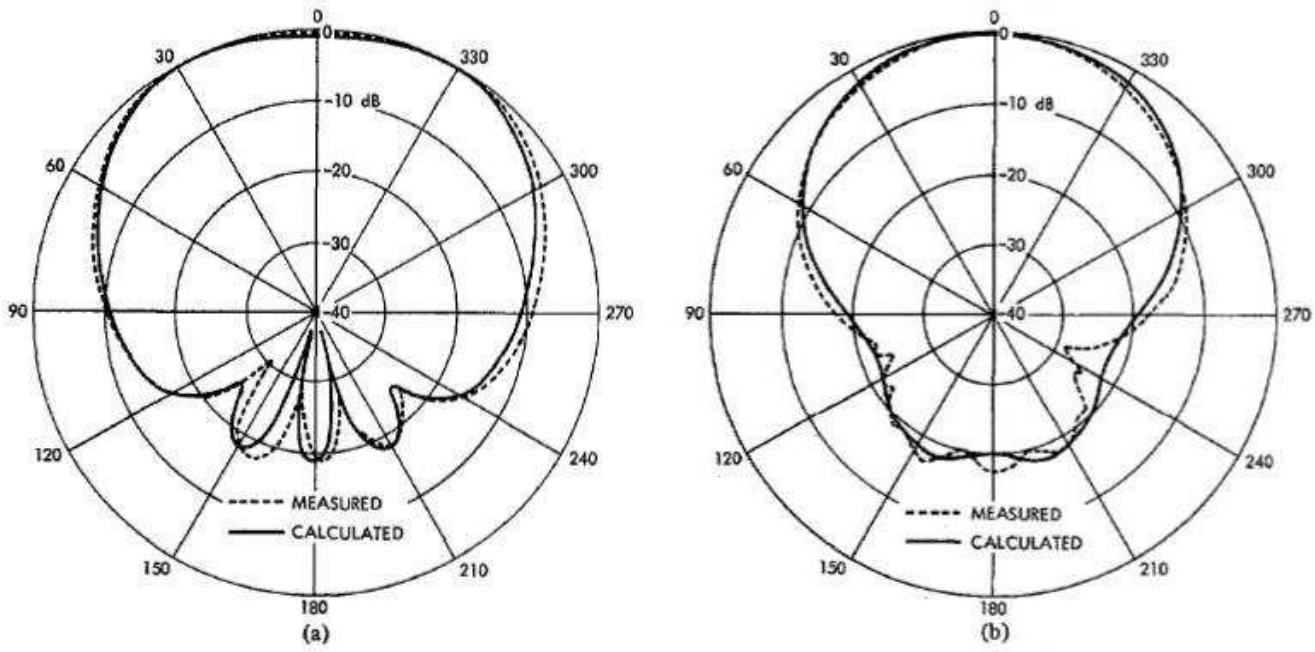


Figure 5.12: RMSA radiation patterns: (a) E plane, (b) H plane. [19]

In all real RMSAs the ground plane is obviously finite. As a result there is some degree of back radiation, i.e., behind the ground plane ($90° \leq \theta \leq 270°$). This can be seen in Fig. 5.12. As long as the ground plane extends past the edge of the patch by 6 times the substrate thickness (h) or more [21], then as far as the main characteristics are concerned (bandwidth, main lobe HPBW and input impedance), the MSA behaves as though the ground plane were infinite .

Since MSAs can be of virtually any shape, elliptical and circular polarisations are possible as well as linear. There are various ways of achieving circular polarisation, but all methods involve tight control of the dimensions of the patch, and of any slots cut into the patch.

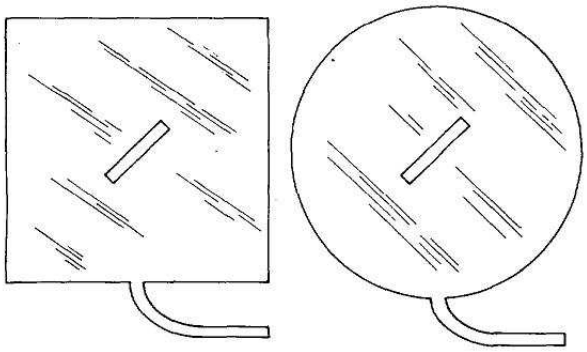


Figure 5.13: Circularly polarised square and circular MSAs with thin diagonal centre slots. [20]

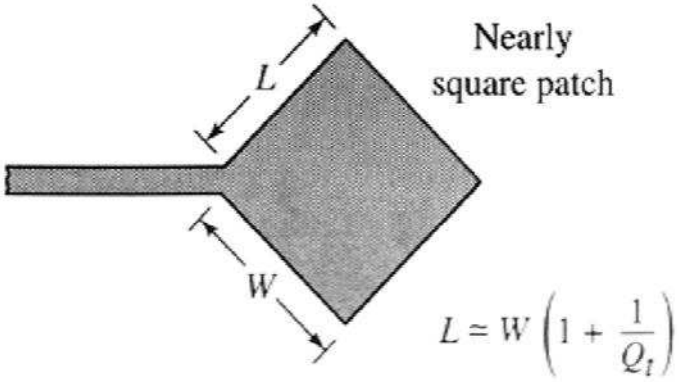


Figure 5.14: Nearly square patch circularly polarised MSA. [17]

In order to achieve circular polarisation, two orthogonal modes need to be excited. These modes must be equal in magnitude and must have a 90 degree phase difference between them. This situation can be achieved from a single feed point with a patch with the correct geometry. The antennas shown in Fig. 5.13 have slots cut into them in order to affect the current distribution in such a way that the two orthogonal modes that are necessary for circular polarisation are achieved. Best results occur when the two orthogonal modes resonate at very slightly different frequencies. This is why the antenna shown in Fig. 5.14 is not a perfect square. In Fig. 5.14, Q_t is the total Q factor of the antenna.

5.5 Parametric Analysis of RMSAs

In this analysis, various parameters of RMSAs, such as patch width and height, are varied in turn to see their effect on the performance. The operating frequency remains constant. The data in this section is from various researcher's scientific studies, i.e., one parameter was varied whilst keeping the others constant. The source of the relevant data is given in the corresponding sub-section.

5.5.1 Effect of Relative Permittivity (ϵ_r)

Increasing the relative permittivity of an MSA whilst keeping the operating frequency fixed, will result in the patch length, and optionally the patch width, decreasing. This is because, increasing ϵ_r, lowers the propagation velocity of waves in the RMSA and therefore also decreases the wavelength. As RMSAs are half a wavelength long, their length naturally decrease. Whether the patch width decreases or not is up to the designer. Maintaining the same patch aspect ratio, i.e., the length to width ratio, whilst increasing ϵ_r is a simple technique for miniaturising MSAs.

However, miniaturising MSAs by increasing ϵ_r affects both the bandwidth and directivity. Most importantly, it significantly reduces the bandwidth.

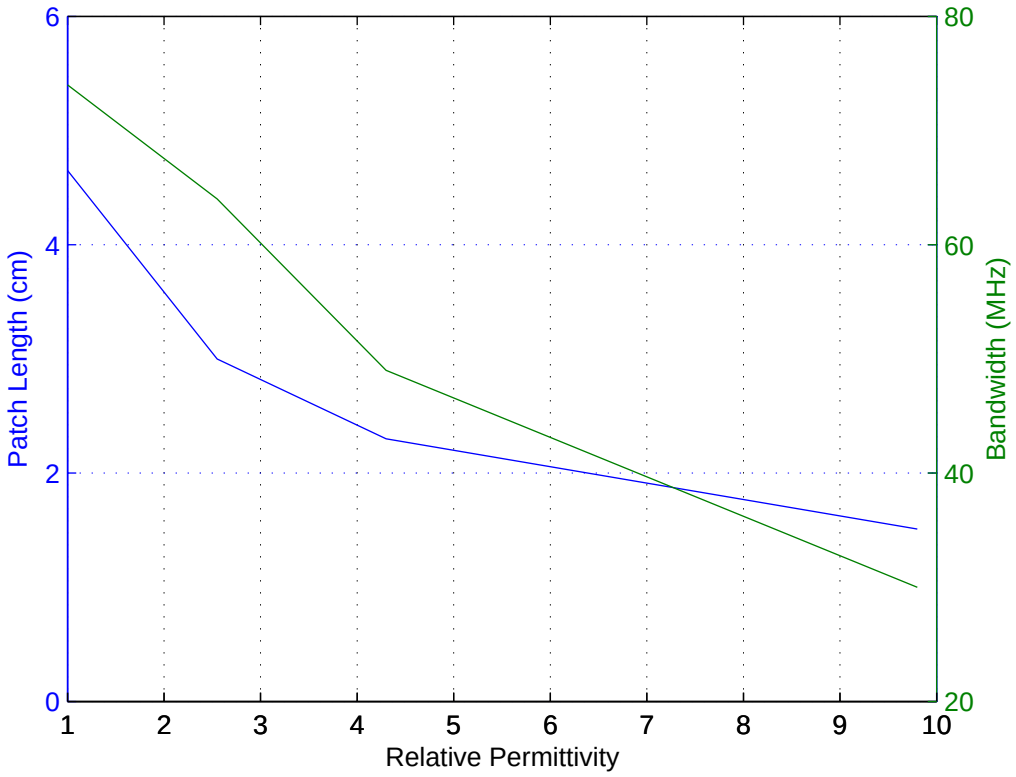


Figure 5.15: Effect of ϵ_r on patch length and bandwidth for fixed frequency RMSA (centre frequency is 1.8 GHz), data from [21]

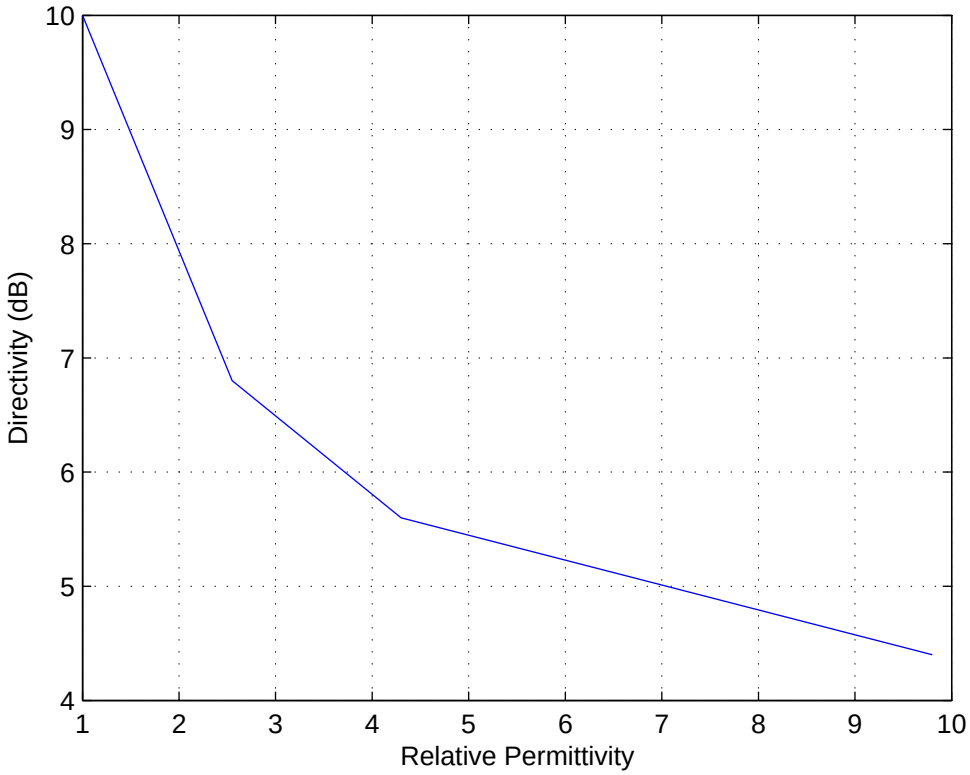


Figure 5.16: Effect of ϵ_r on directivity for fixed frequency RMSA (centre frequency is 1.8 GHz), data from [21]

Miniaturising MSAs by increasing ϵ_r reduces the bandwidth because the ratio of antenna size to free space wavelength decreases. According to the Chu-Harrington limit, which is discussed in more detail in later chapters, this will reduce the bandwidth. In addition, a general rule of antennas is that the smaller an antenna is electrically, i.e., compared to the free space wavelength, then the lower its directivity will be.

5.5.2 Effect of Patch Width (W)

As the patch width is increased, the RMSA becomes electrically larger and consequently its bandwidth and directivity increase.

More specifically, as W increases, the radiation pattern becomes more complicated, as can be seen in Fig. 5.17. The maximum magnitude of the main lobe (directivity) increases and the number of lobes and nulls also increases. The data for Fig. 5.17 comes from the analytical expression for the H plane radiation of an RMSA, as described in section 6.1.4.

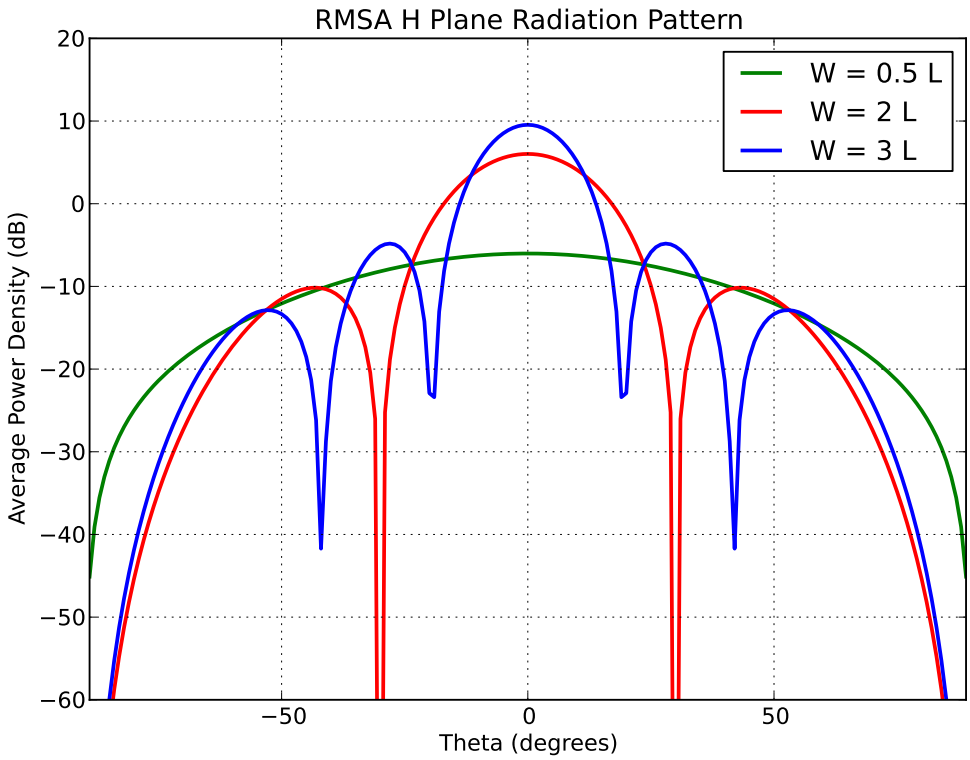


Figure 5.17: Effect of W on the H plane radiation pattern of RMSAs.

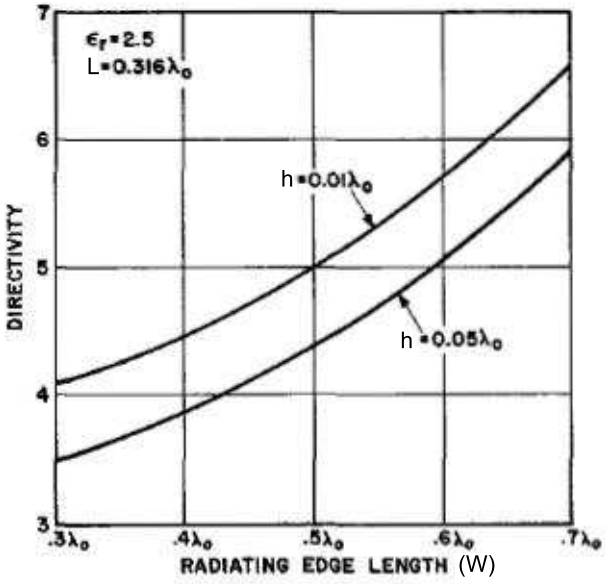


Figure 5.18: RMSA directivity against patch width. Directivity is in dB. [20]

5.5.3 Effect of Substrate Thickness (h)

When the normalised substrate thickness is small, i.e., where $h/\lambda_0 \leq 0.05$ (λ_0 = free space wavelength), bandwidth increases proportionally with h. This can be explained by the fact that the volume occupied by the antenna is increasing, whilst its footprint is not, and as indicated by the Chu-Harrington limit a corresponding increase in bandwidth will result.

The directivity also increases slightly because greater fringing results in a slightly increased effective area of the antenna.

As h increases further the radiation efficiency drops due to an increase in surface wave generation.

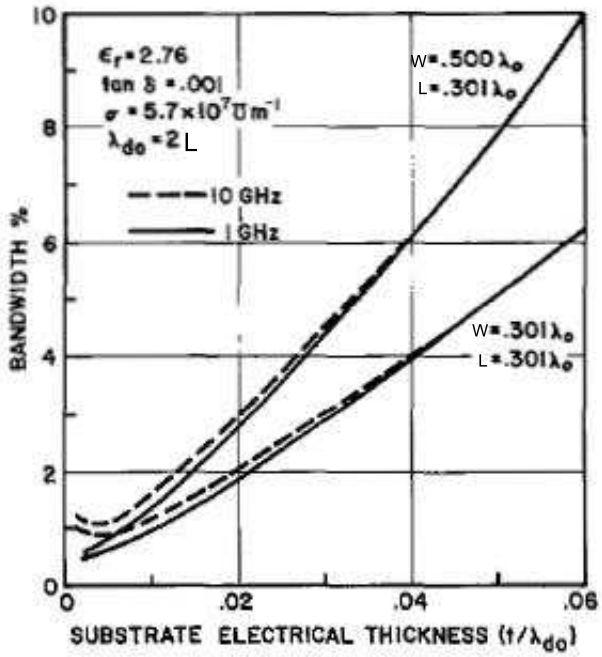


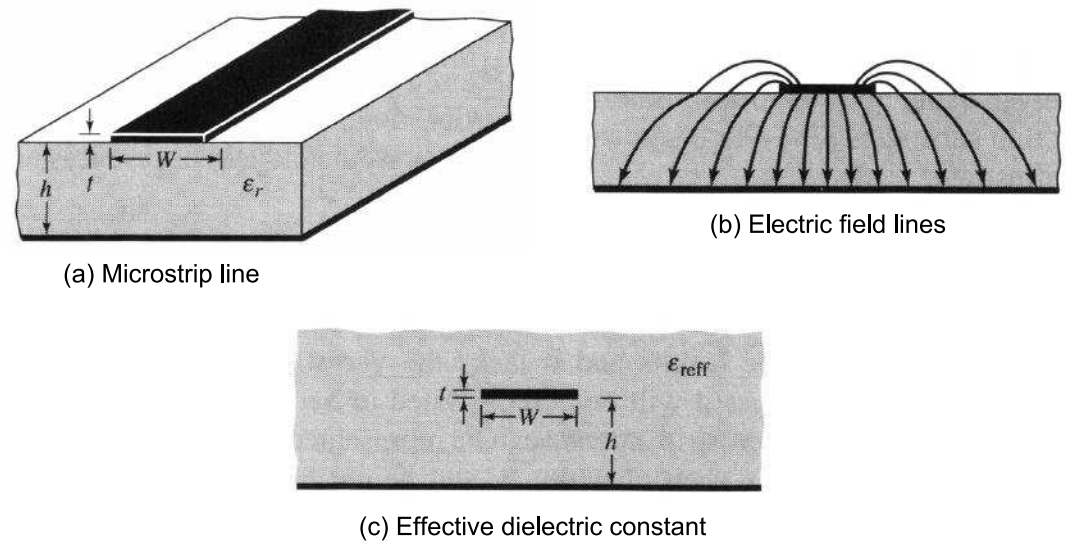
Figure 5.19: RMSA bandwidth against substrate thickness. [20]

Chapter 6

MSA Analytical Models

6.1 The Transmission Line Model

The *transmission line model* is only applicable to rectangular MSAs (RMSAs) [60] [16]. This is because the RMSA is modeled as consisting of a half wavelength long section of microstrip transmission line. At the two lengthways ends of the patch are radiating slots. These slots both store and radiate energy away from the patch and so are modeled as consisting of both a resistive and a reactive impedance. A microstrip transmission line is simply made up of a ground plane on the bottom surface of a dielectric and a track on the top surface, as in Fig. 6.1 (a). Due to their ease of manufacture and low cost, microstrip transmission lines are widely used on printed circuit boards (PCBs).



(a) Microstrip line

(b) Electric field lines

(c) Effective dielectric constant

Figure 6.1: Microstrip transmission line: (a) geometry, (b) electric field lines & (c) effective dielectric constant. [17]

As can be seen in Fig. 6.1 (b), the electric field lines of a microstrip transmission line reside in both the substrate and the air. For the majority of microstrip lines ($W/h >> 1$) most of the electric field lines concentrate mostly in the dielectric. Some of the electromagnetic waves traveling down the line travel in the air and some travel in the substrate. In order to simplify the modeling of microstrip transmission lines, the concept of *effective dielectric constant* ϵ_e is used.

This concept assumes that the microstrip line's trace, with its original dimensions and height above the ground plane is embedded in a single homogeneous dielectric, as shown in Fig. 6.1 (c). For a line with air above the substrate, the effective dielectric constant has values in the range of $1 < \epsilon_e < \epsilon_r$. For $W/h > 1$, effective dielectric constant is given by [12]:

$$\epsilon_e \approx \frac{\epsilon_r + 1}{2} + \frac{\epsilon_r - 1}{2}\left[1 + 12\,\frac{h}{W}\right]^{-1/2} \tag{6.1}$$

6.1.1 Patch Dimensions

The fringing at the radiating edges of an RMSA means that the effective length (L_e) of the patch is slightly greater than its physical length. The length of the patch is extended by an amount ΔL on each side (see Fig. 6.2):

$$L_e = L + 2\,\Delta L \tag{6.2}$$

The velocity v of the waves propagating in the dominant (lengthways) direction of the RMSA is:

$$v = \frac{c}{\sqrt{\epsilon_e}} \tag{6.3}$$

where:

$$c = \text{speed of light in free space}$$

The wavelength in the RMSA λ is:

$$\lambda = \frac{c}{f_0\sqrt{\epsilon_e}} = \frac{\lambda_0}{\sqrt{\epsilon_e}} \tag{6.4}$$

where:

$$\lambda_0 = \text{wavelength in free space}$$
$$f_0 = \text{centre frequency of operating band}$$

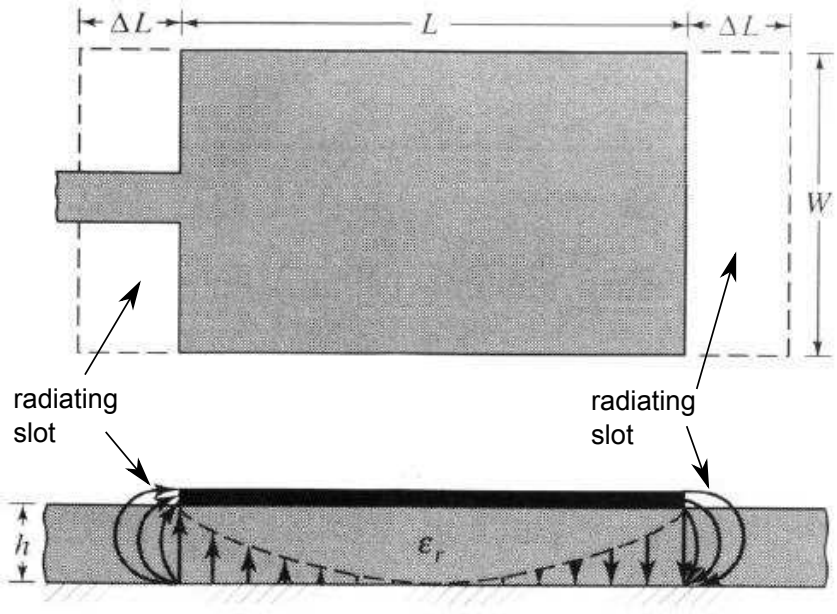


Figure 6.2: Physical and effective lengths of an RMSA. [17]

The effective length of the patch must be equal to half a wavelength at the frequency of operation:

$$L_e = L + 2\,\Delta L = \frac{\lambda}{2} = \frac{\lambda_0}{2\sqrt{\epsilon_e}} = \frac{c}{2\,f_0\sqrt{\epsilon_e}} \tag{6.5}$$

A commonly used equation for calculating the patch extension is [61]:

$$\Delta L = 0.412\,h\,\frac{(\epsilon_e\,+\,0.3)(W/h\,+\,0.265)}{(\epsilon_e - 0.258)(W/h + 0.8)} \tag{6.6}$$

The patch width (W) is arbitrary but there are practical constraints. If the width is too small, the patch will not radiate well. If it is too large, the low impedance of the line will make it hard to match and the more complicated radiation pattern, i.e., the appearance of nulls, may be undesirable. A practical width that leads to good radiation efficiencies is [21]:

$$W = \frac{c}{2\,f_0}\sqrt{\frac{2}{\epsilon_r\,+\,1}} \tag{6.7}$$

6.1.2 Radiating Slots

As can be seen in Fig. 6.2, the electric field under the patch encroaches into the air above the substrate at the (lengthways) ends of the patch. The regions in which this occurs are known as slots, and the electric field in the slots is tangential to the surface of the substrate. They are called slots because their field pattern is the same as that of a slot antenna.

The slots effectively increase the length of the patch by an amount ΔL at each (lengthways) end of the patch. The radiation comes from these slots, and as there are two of them, an RMSA is therefore an array antenna of two elements. The mechanism by which the slots radiate is discussed in more detail in section 6.1.4.

6.1.3 Input Impedance

With the length and width of the RMSA determined, it is then necessary to find the feed point position that will yield a good impedance match. This is done by considering the complete transmission line model of the RMSA, as in Fig. 6.3. This model consists of two sections of, typically different length, transmission line, with each section terminated in a complex load (slot).

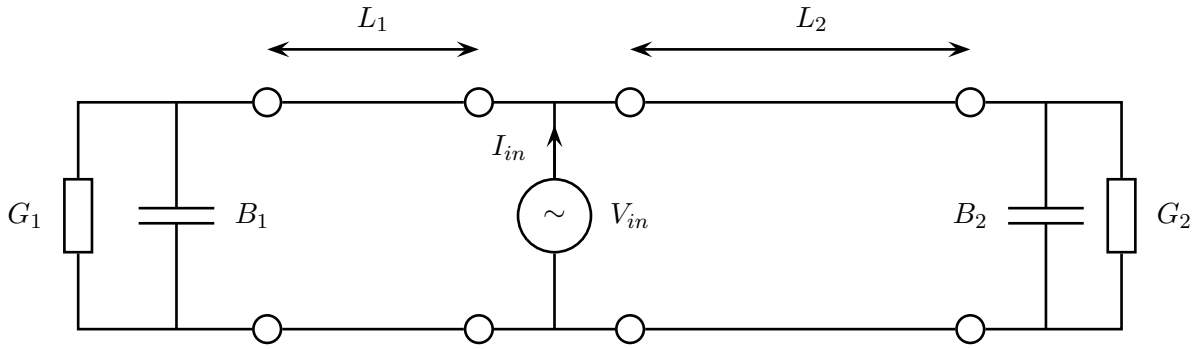


Figure 6.3: Transmission line model of an RMSA.

The input impedance is simply the ratio of voltage to current seen at the antenna's feed point:

$$Z_{in} = \frac{V_{in}}{I_{in}} = \frac{1}{Y_{in}} \tag{6.8}$$

As can be seen from Fig. 6.3, the input impedance is the parallel combination of the two slots, each transformed by their respective lengths of transmission line. As it is easier to add parallel loads when using admittance rather than impedance, it is the input admittance Y_{in} that is typically calculated first:

$$\begin{aligned} Y_{in} = Y_0 \left(\frac{G_1 \, + \, j(B_1 + Y_0 \tan \beta L_1)}{Y_0 \, - \, B_1 \tan \beta L_1 \, + \, jG_1 \tan \beta L_1} \right) \\ + Y_0 \left(\frac{G_2 \, + \, j(B_2 + Y_0 \tan \beta L_2)}{Y_0 \, - \, B_2 \tan \beta L_2 \, + \, jG_2 \tan \beta L_2} \right) \end{aligned} \tag{6.9}$$

where:

$$Y_0 = \text{characteristic impedance of transmission line}$$

$$\beta = 2\pi/\lambda$$

The slot conductances and susceptances ($G_1, B_1, G_2 \text{and} B_2$) can be found by performing spectral analysis on an arbitrary thin slot [12, Ch 12] [62, pp 180-183]. Alternatively, the slot conductances can be found by integrating the total radiated power from the antenna [12] [16]. As the two slots are identical, their respective conductances and susceptances also have the same magnitudes, i.e., $G_1 = G_2$ and $B_1 = B_2$. The values of the slot conductances and susceptances are given in several texts including [12, Ch 12] [62, pp 180-183] [61]:

$$G = \begin{cases} \left(\frac{W}{90\lambda_0}\right)^2, & W < 0.35\lambda_0 \\ \frac{W}{120\lambda_0} - \frac{1}{60\pi^2}, & 0.35 \leq W \leq 2\lambda_0 \\ \frac{W}{120\lambda_0}, & 2\lambda_0 \leq W \end{cases} \tag{6.10}$$

$$B = \frac{W}{120\lambda_0}\left[1 - 0.636\ln(k_0 h)\right] \tag{6.11}$$

where:

$$k_0 = 2\pi/\lambda_0$$

At resonance the input impedance is purely real, i.e., $Im\{Y_{in}\} = 0$. Consequently the input impedance approximates to [61]:

$$Z_{in} \approx \frac{\cos^2(\beta L_1)}{2G} \tag{6.12}$$

As can be seen in eqn. 6.12 the input impedance depends on the distance the feed point is from the edge of the patch (slot 1). By controlling this distance, the input impedance of the RMSA can be matched to that of the system that is driving it , which is typically 50Ω:

$$L_{feed} = L_1 \approx \frac{\cos^{-1}\sqrt{2\,Z_{in}\,G}}{\beta} \tag{6.13}$$

In the case of probe fed RMSAs, L_{feed} is the distance of the probe feed from the edge (radiating slot) of the patch. For inset fed RMSAs, it is the length of the inset.

6.1.4 Radiation Pattern

It is the slots at the radiating (lengthways) edges that radiate electromagnetic waves away from the RMSA. When modeling these slots analytically it is commonly assumed that, as the slots are narrow, the electric field across them is constant. In other words, the electric field across the slots does not vary in the either the lengthways or width ways directions. Furthermore, the electric field only has a lengthways component. With reference to Fig. 6.4, the electric field in the slot $\boldsymbol{E}_{slot}$ is:

$$\boldsymbol{E}_{slot} = \hat{\boldsymbol{x}}\, E_x = \hat{\boldsymbol{x}}\, \frac{V_{in}}{h} \tag{6.14}$$

It is because the electric field in the slots only has an x component that RMSA produces linearly polarised waves. The E plane is the $x-z$ plane and the H plane is the $y-z$ plane.

Using Huygen's field equivalence principle [12, Section 12.2], each slot is modeled as having an equivalent magnetic surface current density $\boldsymbol{M}_s$ it, i.e., one that produces the same fields:

$$\boldsymbol{M}_s = \hat{\boldsymbol{y}}\, 2\, E_x = \hat{\boldsymbol{y}}\, \frac{2\, V_{in}}{h} \tag{6.15}$$

For electrically thin dielectrics ($h << \lambda_0$), the magnetic surface current density produces fields in the far field region that are of the form [21] [61] [12, Ch 14] [16]:

$$E_\theta \approx j \frac{V_{in}}{\pi}\, \frac{e^{-jkr}}{r}\, \sin\theta\, \frac{\sin\left(\frac{k_0 W}{2}\, \cos\theta\right)}{\cos\theta} \tag{6.16}$$

Eqn 6.16 describes the radiation from each slot individually. The RMSA has two slots of course, so the RMSA is effectively an array of two slot antennas. As such, the radiation of both slots can be combined analytically using an array factor of two elements. This results in overall far field electric fields of the form:

$$E_\theta \approx j \frac{2\, V_{in}}{\pi}\, \frac{e^{-jkr}}{r} \left\{ \sin\theta\, \frac{\sin\left(\frac{k_0 W}{2}\, \cos\theta\right)}{\cos\theta} \right\} \cos\left(\frac{k_0 L_e}{2}\, \sin\theta\, \sin\phi\right) \tag{6.17}$$

For the E plane, and:

$$E_\theta \approx j \frac{k_0\, W\, V_{in}}{\pi}\, \frac{e^{-jkr}}{r} \left\{ \frac{\sin\left(\frac{k_0 h}{2}\, \cos\phi\right)}{\frac{k_0 h}{2}\, \cos\phi} \right\} \cos\left(\frac{k_0 L_e}{2}\, \sin\phi\right) \tag{6.18}$$

For the H plane.

In the far field, E_r and E_ϕ are negligible compared to E_θ, thus giving linear polarisation.

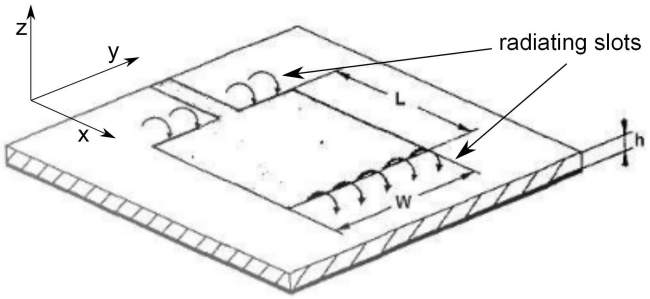


Figure 6.4: Microstrip line fed RMSA and its radiating slots. [16]

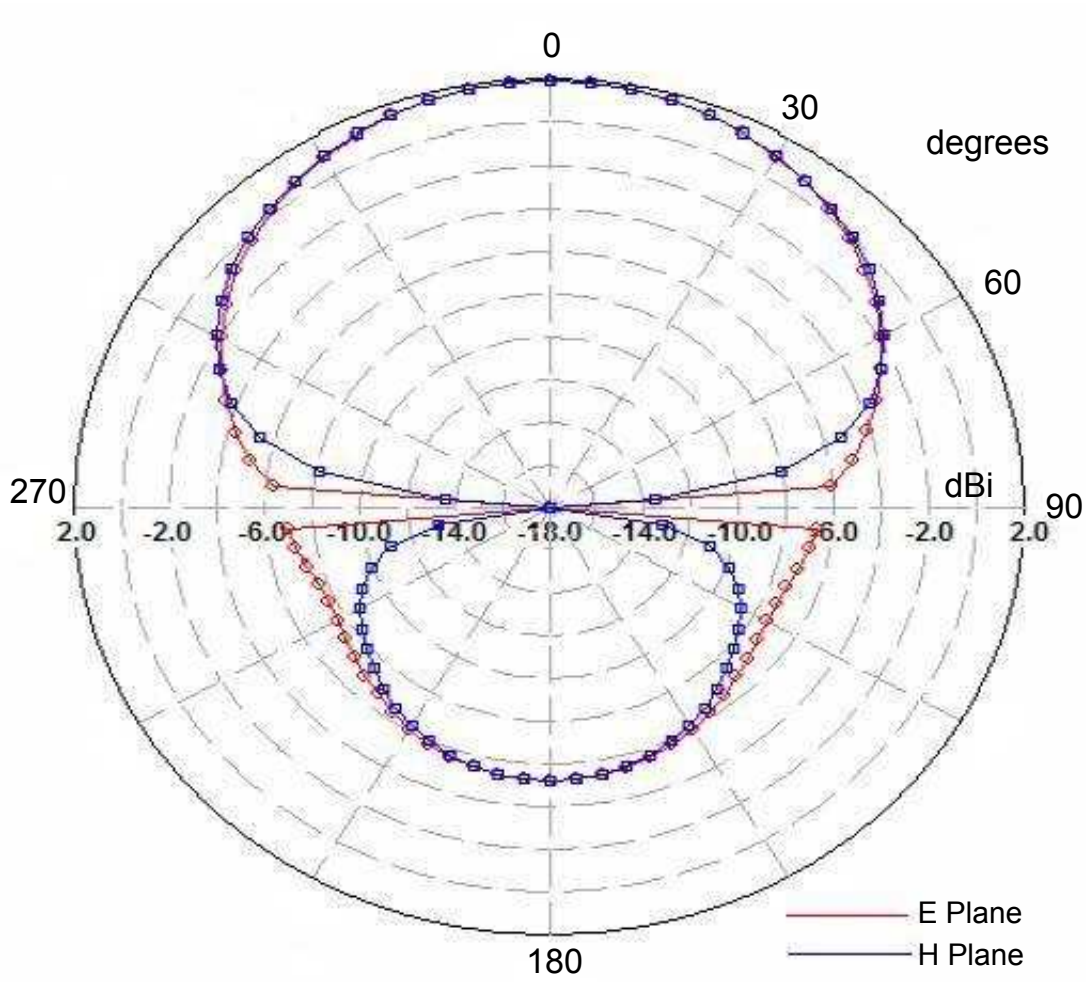


Figure 6.5: RMSA radiation pattern. [22]

6.2 The Cavity Model

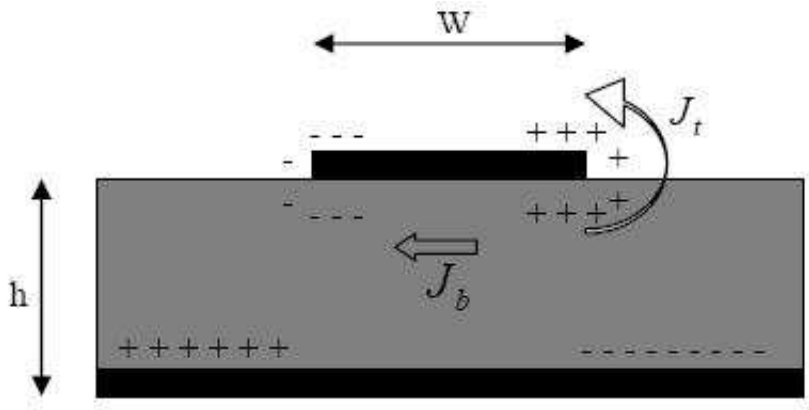


Figure 6.6: Charge distribution and current density on microstrip patch [12].

The *cavity model* is a more rigorous model than the transmission line model because it is developed from a more fundamental starting point, i.e. from more fundamental physics than the transmission line model.

The cavity model was developed independently by several researchers and started to appear in the widely available literature in the late 1970s and early 1980s [63] and [64]. It is also discussed in several antenna theory text books to a greater or lesser extent [61].

In the transmission line model, only a single dominant resonant mode is assumed. However, the cavity model mathematically proves that there are many more resonant modes inside the cavity than just the one at the desired resonant frequency.

As its name suggests, the antenna is modeled as a cavity, in which several modes are able to propagate. The fields within the cavity are evaluated first and then the exterior (radiated) fields can be found from the interior ones.

6.2.1 Overview

As can be seen from figure 6.6, when the antenna is driven by a source, a charge distribution appears (i.e. areas of overall positive or negative charge) on the top and bottom surfaces of the patch and on the top surface of the ground plane. It should be noted that what is shown in figure 6.6 is a instantaneous snapshot in time. For instance, half a cycle later from that seen in figure 6.6, the positively charged areas will be negatively charged and vice versa.

Opposite charges on the bottom of the patch and on the ground plane will attract each other and will tend to maintain the charge concentration on the bottom side of the patch. However, opposite charges on the bottom of the patch will be attracted to each other thus creating current flow $\boldsymbol{J}_b$ on the bottom side of the patch.

At the same time like charges on the bottom side of the patch will tend to push some of this charge round the sides of the patch and onto the the patch's top surface thus creating a current flow on the top surface of the patch $\boldsymbol{J}_t$.

Most practical microstrip antennas have a width (W) much larger than their height (h). This means that the charge distribution is dominated by the attractive mechanism. As such most of the charge and current is on the bottom surface of the patch, and so the current flowing round the edges to the top is relatively small.

Due to the fact that the current density on the top surface is so relatively small, the tangential magnetic field ($\boldsymbol{H}_{dielectric}$) at the edges of the patch is also negligible. This means that the whole structure can be modeled as a cavity extending down from the bottom surface of the patch to the ground plane. The side walls of the cavity are modeled as if made of a perfectly magnetic conductor because the tangential field at the surface is virtually zero. The top and bottom surfaces of the cavity, i.e. the patch and the ground plane directly beneath the patch, are modeled as perfect electric conductors because they are made of metal.

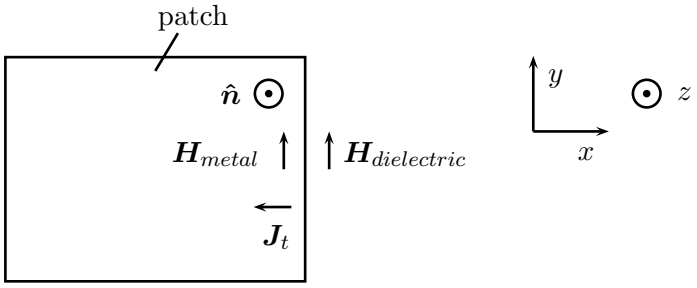


Figure 6.7: Top surface of patch from above.

Figure 6.7 is the view of the top surface of the patch in figure 6.6 when looking from above. The vector that is normal to the surface if the patch ($\hat{\boldsymbol{n}}$) is in the same direction as the z axis.

At edge of the patch:

$$\hat{\boldsymbol{n}} \times (\boldsymbol{H}_{dielectric} - \boldsymbol{H}_{metal}) = \boldsymbol{J}_t \tag{6.19}$$

$$\boldsymbol{H}_{metal} = 0 \tag{6.20}$$

$$\leftrightarrow \qquad \hat{\boldsymbol{n}} \times \boldsymbol{H}_{dielectric} = \boldsymbol{J}_t \tag{6.21}$$

$$\rightarrow \qquad |\boldsymbol{H}_{dielectric}| = |\boldsymbol{J}_t| \tag{6.22}$$

However:

$$|\boldsymbol{J}_t| \approx 0 \qquad \rightarrow \qquad |\boldsymbol{H}_{dielectric}| \approx 0 \tag{6.23}$$

Equations 6.19 to 6.22 form the standard tangential magnetic field boundary condition for a metal-dielectric interface. This boundary condition states that the magnitude of the tangential magnetic field, just above a metal surface, is equal to the current density in the metal at that point.

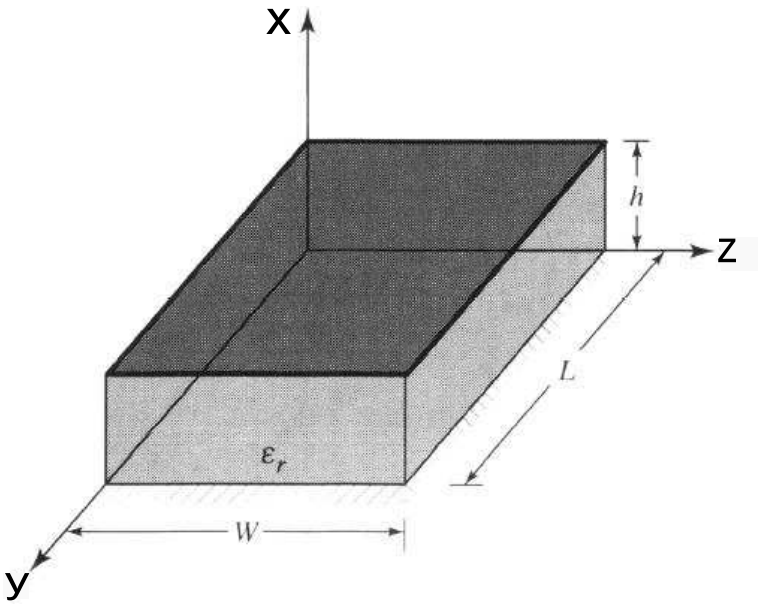


Figure 6.8: Cavity geometry [17].

However, as stated above, the current density on the top surface of the patch is negligible, because of the typically very small height to width ratio of the antenna. This means that the magnetic field at the patch edge, in the direction of the patch edge, will also be negligible. This is the justification of why the side walls of the cavity can be modeled as if they were made of a perfect magnetic conductor. A perfect magnetic conductor can have no tangential magnetic field at its surface.

6.2.2 Fields Inside The Cavity

In the light of the material properties of the boundary of the cavity now being known (e.g. the side walls being made of a perfect magnetic conductor) and the fact that the height to width ratio is small, certain assumptions about the fields in the cavity can be made.

If the cavity is orientated as in figure 6.8 then currents (on the patch and ground plane) will only flow in the y and/or z directions, and thus only y and/or z direction magnetic fields will be generated. In addition there will be no tangential electric field, i.e. no y or z components, at the surface of the patch and ground plane. Since the electric fields within the cavity are generated at these points this means the x component of the electric field will be the dominant one inside the cavity. This knowledge of the fields inside the cavity leads to the ability to use the transverse magnetic mode for the x direction (TM^x). In short, using the TM^x mode ensures that waves inside the cavity will travel in the y or z directions only and that the magnetic field will have no x component.

In addition to using the TM^x mode, it is also assumed that there will be no variation for the electric field in the x direction because the height of the cavity is so much smaller than its length or width.

Assumptions about the fields inside the cavity:

$$\frac{\partial E_x}{\partial x} = 0 \tag{6.24}$$

$$E_x \gg E_y\,, \qquad E_x \gg E_z \tag{6.25}$$

$$\boldsymbol{H} = \hat{\boldsymbol{y}} H_y + \hat{\boldsymbol{z}} H_z \tag{6.26}$$

In many analytical electromagnetic problems, instead of calculating the electric and/or magnetic field components directly from a source variable, an intermediate potential variable is first determined. In [12, Chapter 14] the solution to the cavity's interior electric and magnetic fields is determined using the magnetic vector potential $\boldsymbol{A}$.

Inside the cavity there exists only the dielectric, which for the sake of simplification, is assumed to be at this stage to be lossless, i.e., no conduction current can flow within it. All field variables, i.e., the electric and magnetic fields and the various scalar and vector potential functions that are based on them, must obey the wave equation.

The magnetic vector potential is no exception, and because the cavity is assumed to be lossless, its wave equation is homogeneous:

$$\nabla^2 \boldsymbol{A} + k^2 \boldsymbol{A} = 0 \tag{6.27}$$

$$\leftrightarrow \qquad \nabla^2 A_x + k^2 A_x = 0 \tag{6.28}$$

Faraday's law together with the definition of the magnetic vector potential give an expression for the electric field in terms of the magnetic vector potential and the electric scalar potential, more commonly referred to as voltage (ϕ_e):

$$\nabla \times \boldsymbol{E} = -j\omega\mu \boldsymbol{H} \tag{6.29}$$

$$\boldsymbol{H} = \frac{1}{\mu} \nabla \times \boldsymbol{A} \tag{6.30}$$

$$\leftrightarrow \qquad \boldsymbol{E} = -\nabla \phi_e - j\omega \boldsymbol{A} \tag{6.31}$$

The Lorentz gauge which makes use of redundant degrees of freedom in field variables, can be written as:

$$\phi_e = j \frac{\nabla \cdot \boldsymbol{A}}{\omega \mu \epsilon} \tag{6.32}$$

This leads to an expression for the electric field that depends on the magnetic vector potential only:

$$\boldsymbol{E} = -j\omega\boldsymbol{A} - j\frac{1}{\omega\mu\epsilon}\nabla(\nabla\cdot\boldsymbol{A}) \tag{6.33}$$

In order to satisfy equations 6.26 and 6.30 the magnetic vector potential can only have an x component:

$$\boldsymbol{A} = \hat{\boldsymbol{x}}A_x \tag{6.34}$$

Expanding equation 6.33 using equation 6.34 gives:

$$\boldsymbol{E} = -\hat{\boldsymbol{x}}j\omega A_x - j\frac{1}{\omega\mu\epsilon}(\hat{\boldsymbol{x}}\frac{\partial^2 A_x}{\partial x^2} + \hat{\boldsymbol{y}}\frac{\partial^2 A_x}{\partial x\partial y} + \hat{\boldsymbol{z}}\frac{\partial^2 A_x}{\partial x\partial z}) \tag{6.35}$$

$$\rightarrow \qquad E_x = -j\frac{1}{\omega\mu\epsilon}(\frac{\partial^2}{\partial x^2} + \beta^2)A_x \tag{6.36}$$

$$\beta^2 = \omega^2\mu\epsilon \tag{6.37}$$

Expressions for E_y, E_z, H_y and H_z can be generated from 6.33 in a similar manner:

$$E_y = -j\frac{1}{\omega\mu\epsilon}\frac{\partial^2 A_x}{\partial x\partial y} \tag{6.38}$$

$$E_z = -j\frac{1}{\omega\mu\epsilon}\frac{\partial^2 A_x}{\partial x\partial z} \tag{6.39}$$

$$H_y = \frac{1}{\mu}\frac{\partial A_x}{\partial z} \tag{6.40}$$

$$H_z = \frac{1}{\mu}\frac{\partial A_x}{\partial y} \tag{6.41}$$

$$H_x = 0 \tag{6.42}$$

Now a solution for $\boldsymbol{A}$ is needed in order to find $\boldsymbol{E}$. The separation of variables technique is used which assumes:

$$A_x(x,y,z) = f(x)g(y)h(z) \tag{6.43}$$

This has the solution:

$$A_x(x,y,z) = [A_1\cos(k_x x) + B_1 sin(k_x x)][A_2\cos(k_y y) + B_2 sin(k_y y)] \cdot[A_3\cos(k_z z) + B_3 sin(k_z z)] \tag{6.44}$$

where k_x, k_y and k_z are the wave numbers along the x, y and z directions.

At the bottom side of the patch and the top of the ground plane there is no tangential electric field, i.e., the electric field here will have no x or y component. This is because these surfaces are made of metal which is almost a perfect electric conductor and always has a zero tangential electric field. Consequently, certain boundary conditions can be applied. When equation 6.44 is substituted into equation 6.38 with the following boundary conditions:

$$E_y(x = 0, 0 \leq y \leq L, 0 \leq z \leq W) = 0 \tag{6.45}$$

$$E_y(x = h, 0 \leq y \leq L, 0 \leq z \leq W) = 0 \tag{6.46}$$

it follows that in equation 6.44 $B_1 = 0$.

By applying appropriate boundary conditions, i.e., zero tangential magnetic field at the perfectly magnetically conducting walls, to equations 6.40 and 6.41 it can be shown that in equation 6.44 $B_2 = 0$ and $B_3 = 0$.

This results in the final form of A_x within the cavity as:

$$A_x(x, y, z) = A_{mnp} \cos(k_x x) \cos(k_y y) \cos(k_z z) \tag{6.47}$$

A_{mnp} is an amplitude term for each mode, and the wave numbers are:

$$k_x = \left(\frac{m\pi}{h}\right), \quad m = 0, 1, 2, \ldots \tag{6.48}$$

$$k_y = \left(\frac{n\pi}{L}\right), \quad n = 0, 1, 2, \ldots \tag{6.49}$$

$$k_z = \left(\frac{p\pi}{W}\right), \quad p = 0, 1, 2, \ldots \tag{6.50}$$

where:

$$k_x^2 + k_y^2 + k_z^2 = k^2 = \beta^2 = \omega^2 \mu\epsilon \tag{6.51}$$

Substituting equation 6.47 into equation 6.36 gives the final form of the x, i.e., dominant, component of the electric field:

$$E_x = -j \frac{1}{\omega\mu\epsilon} (k^2 - k_x^2) \, A_{mnp} \cos(k_x x) \cos(k_y y) \cos(k_z z) \tag{6.52}$$

The total electric field is of course the summation of all the various modes in the cavity. Due to the assumption that the E_x component does not vary with x, the E_x component is then the summation of the modes in the y and z directions:

$$E_x = \sum_{n=0}^{\infty} \sum_{p=0}^{\infty} A_{np} \cos(k_y y) \cos(k_z z) \tag{6.53}$$

where A_{np} are the amplitude coefficients.

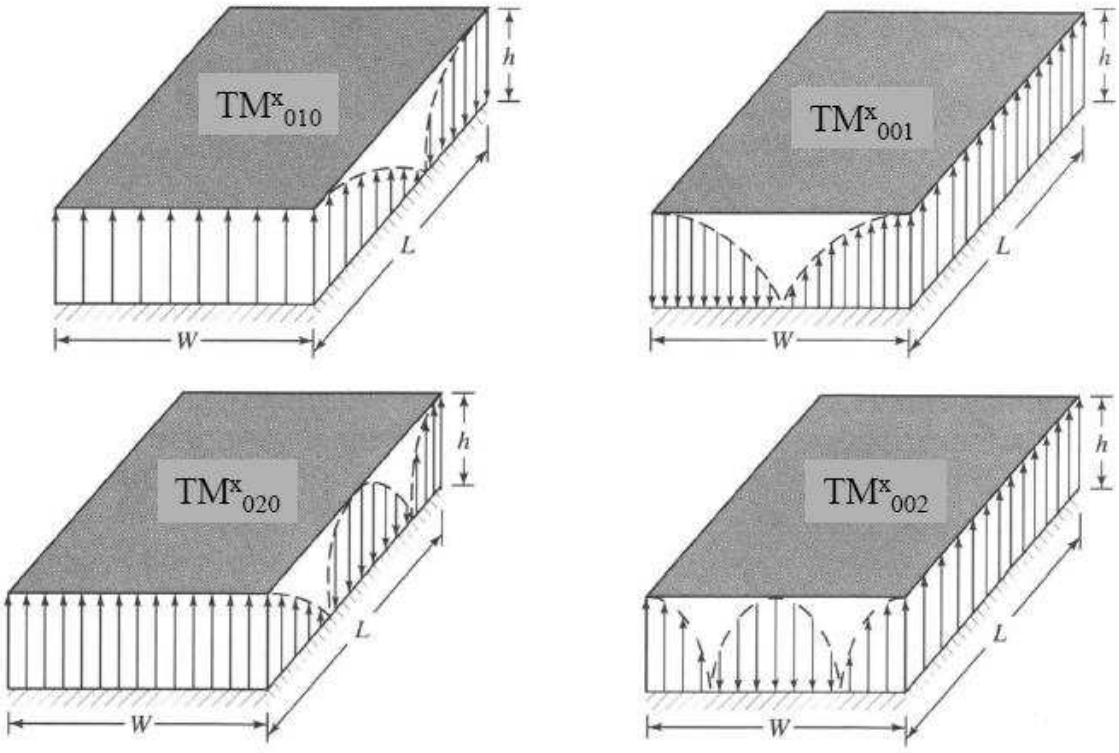


Figure 6.9: Electric field configurations (modes) inside the cavity [17].

Similarly substituting equation 6.47 into equations 6.40 and 6.41 gives:

$$H_y = -\frac{k_z}{\mu} A_{mnp} \cos(k_x x) \cos(k_y y) \sin(k_z z) \tag{6.54}$$

$$H_z = \frac{k_y}{\mu} A_{mnp} \cos(k_x x) \sin(k_y y) \cos(k_z z) \tag{6.55}$$

The waves inside the cavity experience total reflection, i.e., 100% of their power is reflected. This is because all the walls of the cavity are made of perfect conductors, i.e., electric or magnetic. For each individual mode, a standing wave pattern is formed, due to the incident and reflected fields summing at each point is space (superposition).

Figure 6.9 shows some of the possible modes of equation 6.52. Each point in the cavity varies sinusoidally with time, but spatially there is the standard fixed sinusoidal standing wave pattern.

The TM^x_{010} mode is the dominant mode because the wave is traveling up and down the length of the antenna (in the y direction) and there is no variation in the z (width) direction. When in operation, the antenna should be driven at a frequency in which this mode is well established. The TM^x_{010} resonant frequency is:

$$f_{010} = \frac{c}{2L\sqrt{\epsilon_r}} \tag{6.56}$$

where c is the speed of light in free space and ϵ_r is relative permittivity of the dielectric. In reality, there will be some fringing of the non-radiated fields outside of the cavity. For the dominant mode, the effective length of the cavity can be calculated and used, which is the same as that calculated by the transmission line model.

The amplitude coefficients of each mode A_{np} in eqn. 6.53 are obtained by multiplying by ψ^*_{np} ($*$ = complex conjugate) and integrating over the area of the patch [65]:

$$A_{np} = \frac{j\,\omega\mu}{k^s\ -\ k^2_{np}}\ \frac{\iint J_x\,\psi^*_{np}\,dx\,dy}{\iint \psi_{np}\,\psi^*_{np}\,dx\,dy} \tag{6.57}$$

where:

$$\psi_{np} = \cos(k_y y)\cos(k_z z) \tag{6.58}$$

$$k^2_{np} = \left(\frac{n\pi}{L}\right)^2 + \left(\frac{p\pi}{W}\right)^2 \tag{6.59}$$

J_x = current density in input probe

6.2.3 Input Impedance

Calculating the input impedance using the cavity model approach is relatively involved. There are several slightly different approaches [64] [61] [65].

All of these methods involve calculating the radiated energy P_{rad}, the energy stored in fields under the patch W and the lost energy P_{loss}. Most methods involve combining these terms into a single effective loss tangent term δ_{eff}, which is inversely related to the Q of the antenna.

$$\delta_{eff} = \frac{1}{Q} = \frac{P_{rad}\ +\ P_{loss}}{\omega W} \tag{6.60}$$

The input impedance is simply the ratio of input voltage to input current:

$$Z_{in} = \frac{V_{in}}{I_{in}} \tag{6.61}$$

For a probe fed RMSA of thin dielectric, with the feed point at (y_0, z_0):

$$V_{in} = E_z(y_0,\, z_0)/h \tag{6.62}$$

The input current can be found by integrating the probe current density across the cross sectional area of the feed:

$$I_{in} = \iint J_x\,ds \tag{6.63}$$

One particular technique [64], involves modeling the patch as a parallel RLC circuit and calculating the component values using the effective loss tangent.

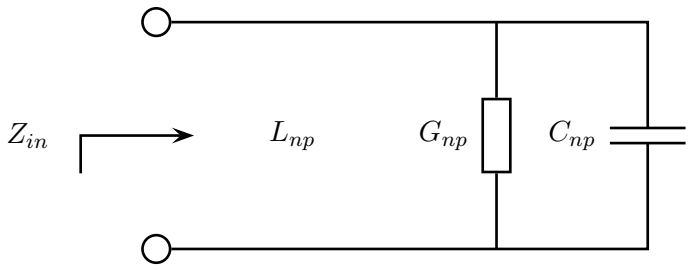


Figure 6.10: Equivalent circuit model of an RMSA.

The equivalent component values (L_{np}, C_{np} and G_{np}) depend on the particular mode. Their actual derivation is straightforward but time consuming, so are not included here. As many discrete modes are excited simultaneously within the cavity, the input impedance is the sum of the effects of all the various modes [64]:

$$Z_{in} = \sum_{n=0}^{\infty} \sum_{p=0}^{\infty} \frac{j\omega L_{np}}{1 - \omega^2 L_{np} C_{np} + j\omega L_{np} G_{np}} \tag{6.64}$$

6.2.4 Radiated Fields

Within the cavity the electric field (E_x) is dominant, i.e., its magnitude is significantly greater than that of the magnetic field (H_y and H_z). This means that it is only necessary to consider the electric field when determining the radiation from the RMSA, i.e., the radiation can be regarded as being generated solely by the electric field.

In all of the modes, the electric field has its greatest magnitude at the magnetic (side) walls of the cavity, as can be seen in Fig. 6.9. As the electric field dominates, it can be replaced by an equivalent magnetic current density $\boldsymbol{M}_s$ at the side walls, using Huygen's field equivalence principle. Furthermore, as the side walls are thin and are close to an effectively infinite ground plane, the resulting equivalent magnetic current density has twice the magnitude of the electric field:

$$\boldsymbol{M}_s = -2\hat{\boldsymbol{n}} \times \boldsymbol{E} \tag{6.65}$$

where $\hat{\boldsymbol{n}}$ is a unit vector pointing outwards from each side wall. The predominant radiation mechanism of the antenna is via these side walls which are known as slots. The dominant mode (TM^x_{010}) has the greatest electric field magnitude, and so results in the greatest amount of radiation.As can be seen in Fig. 6.11. the two lengthways slots of the dominant mode have equivalent magnetic current densities of the same magnitude and phase. As such these two slots do not cancel each other out and thus perform most of the radiation. They are therefore known as the *radiating slots*. Due to the fact that the height of the slots is electrically small , i.e., $h << \lambda_0$, $\boldsymbol{M}_s$ can be regarded as consisting of a single thin strip of current. These strips of magnetic current are in virtually exactly the same place as those of the slots of the transmission line model. As such, the same technique for determining the radiated fields from these slots can be used.

As can be seen in Fig. 6.12, the width ways slots of the dominant mode have current densities of equal magnitude but opposite phase on the opposite cavity walls.

Analytically it can be shown that these two slots largely cancel each other out [12], and thus the radiation from them is negligible when compared to that of the radiating slots. They are therefore known as the *non-radiating slots*. When only the radiating slots of the dominant mode are considered, the radiation pattern derived from the cavity model is exactly the same as that from the transmission line model.

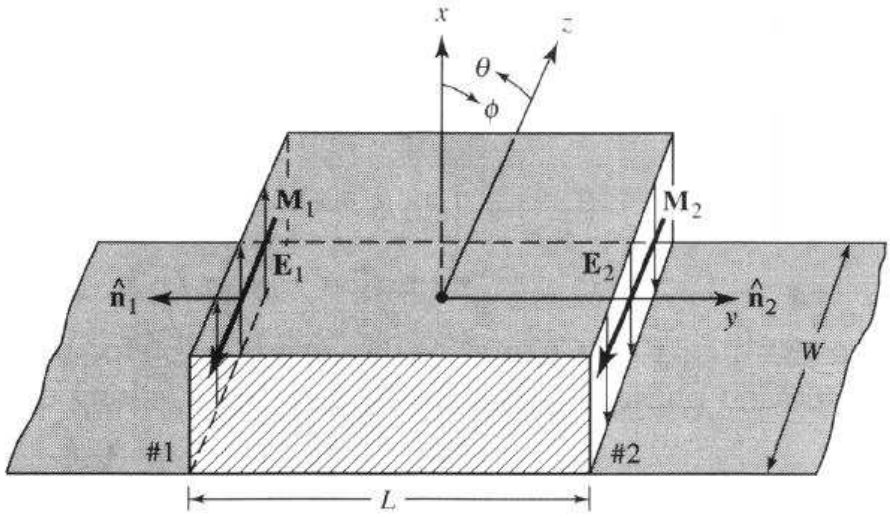


Figure 6.11: Radiating slots of RMSA for the dominant mode [17].

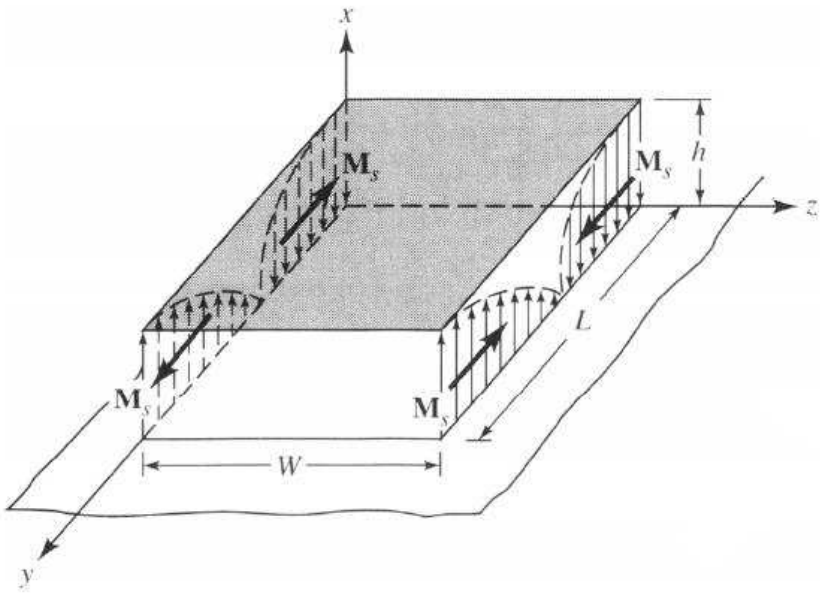


Figure 6.12: Non-radiating slots of RMSA for the dominant mode [17].

6.2.5 Other Patch Shapes

The cavity model has been used to analyse many different patch shapes, including: circles (disks), circular rings, ellipses and right angle triangles. The procedure always involves solving the wave equation for the fields inside the cavity and applying the boundary conditions. Different coordinate systems can be used to describe the cavity geometry. For example, when analysing circular patches it is most convenient to use cylindrical coordinates. As the patch shapes become more complicated, the equations for the modes also become more complicated. Beyond a certain patch complexity, the use of the cavity model becomes practically, if not actually, impossible.

6.3 Conclusion

The transmission line model and the cavity model are not the only analytical models of MSAs. There are several others including: the multiport network model (MNM) [59] and an analytical technique based on the method of moments (MoM) [66]. All these models have different starting points but often there is some overlap between them.

In particular, is the issue of mutual coupling, which is necessary in the modeling of multiple patches in array antennas. It can also be used to model the behaviour of parasitic patches ,i.e., patches that are not directly driven form the feed.

The transmission line model and the cavity model have been discussed in this chapter because they are the most widely used, and hence the most widely covered in the literature. Furthermore, they are the two models, in which their basic starting principles and assumptions can be developed into complete models of the antenna's behaviour, in the the most straightforward way.

As expected, as the complexity of the patch geometry and the number of patches increases, so does the complexity of all the analytical models. The complexity does not have to be considerable before the use of analytical modeling becomes impractical. However, when modeling simple geometries, analytical techniques can offer hugely significant insight into the behaviour of antennas.

Chapter 7

Compacting and Multi-Banding MSAs

7.1 Introduction

Modern mobile communications devices, such as mobile phones and personal digital assistants, typically require antennas that are physically compact. Increasingly these devices also require antennas that can operate at several non-harmonically related frequencies. This is because the functionality of mobile devices is ever increasing and this requires the use of an increasing range of communications protocols and standards. For example, a device may need to utilise mobile phone protocols (GSM, GPRS & UMTS etc.), personal area network protocols (bluetooth etc.), wireless LANs (IEEE 802.11b/g) and global positioning (GPS).

Integrating this ever increasing amount of wireless functionality into mobile devices is a challenging problem. One solution is to have a separate antenna for all of the individual wireless systems. There are several problems with this approach. Firstly there is the problem of space. Secondly all the various antennas will be electrically near each other and will therefore interact, i.e., couple, with each other. This is highly likely to significantly affect their performance. In particular it will affect (skew) their operating frequencies (de-tune).

The most pragmatic solution is to have a single antenna that is capable of operating at all of the required frequency bands.

7.2 Conventional Methods for Compacting and Multi-Banding MSAs

Conventional in this context refers to the fact that these methods have been discovered by engineers and researchers applying insight and experience in order to enhance the performance of MSAs. Much of the insight into MSA operation will have come from analytical models of MSAs. Un-conventional methods would be those that have arisen from the application of, for example, computational evolutionary techniques.

Techniques for both compacting and multi-banding MSAs are considered here at the same time. This is because several of these techniques do both these things.

Multi-band antennas have several, usually narrow, operating bands. Broad-band antennas, on the other hand, usually have a single wide operating band. It is generally easier to design multi-band antennas than broad-band antennas. If the single wide band of a broad-band antenna covers the several bands that a device uses than the broad-band antenna can be used just as well as a multi-band antenna. For this reason, broad-banding techniques are considered here as well as multi-banding techniques.

7.2.1 Changing the Properties of the Dielectric

One of the simplest methods of compacting MSAs is simply by increasing the relative permittivity (ϵ_r) of the dielectric. Significant reductions in size can be obtained but there is always a corresponding reduction in bandwidth. As such this technique is only suitable for when very small bandwidths are required.

Another simple technique for increasing the bandwidth of MSAs is to increase the height (h) of the dielectric. When the dielectric is electrically thin ($h << \lambda_0$) the bandwidth is proportional to h. However, as h increases further, the characteristics of the feed start to become affected and thus the matching is affected, usually adversely. For example, in the case of probe fed MSAs, the increases probe length results in increased probe inductance which makes the impedance match increasingly worse [21]. Furthermore, surface waves start to become excited and propagate in the dielectric, due to total internal reflection between the top and bottom surfaces of the dielectric, until they reach the edges of the substrate where they are transmitted and scattered [21]. This adversely affects the polarisation and the efficiency of the antenna.

7.2.2 Including Shorting Posts and Strips

A shorting post is basically a straight piece of wire that joins the patch and the ground plane. A shorting strip, on the other hand, is a solid continuous strip of metal that joins the patch and the ground plane. Using a row of shorting posts or a shorting strip, a half wavelength RMSA can be reduced in length to a quarter of a wavelength.

An example of an actual shorted quarter wavelength RMSA is given in [21]. A half wavelength RMSA with resonant frequency of 1.47 GHz is reduced to half its length whilst keeping all the other parameters the same (W, h, ϵ_r etc.). The bandwidth changes only marginally, i.e., from 10 to 9 MHz.

The antenna's other characteristics, such as radiation pattern, also undergo negligible changes. A drawback of this technique is that the sensitivity of the input impedance to the feed position is greatly increased.

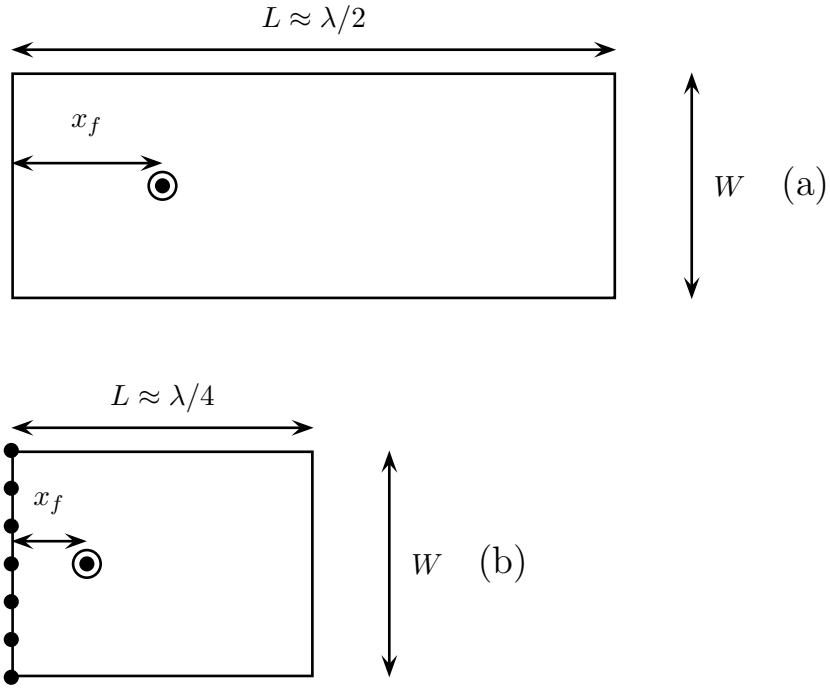


Figure 7.1: Top view of probe fed (a) half wavelength RMSA and (b) quarter wavelength shorted RMSA.

the maximum amount of size reduction can be achieved when a single shorting post is used [67]. ItF has been demonstrated that the dimensions of an RMSA can be reduced by a factor of 3.6 for a given centre frequency and bandwidth [67]. Significant reductions in size whilst maintaining performance have also been achieved for aperture coupled MSAs [68].

7.2.3 Slits in the Patch

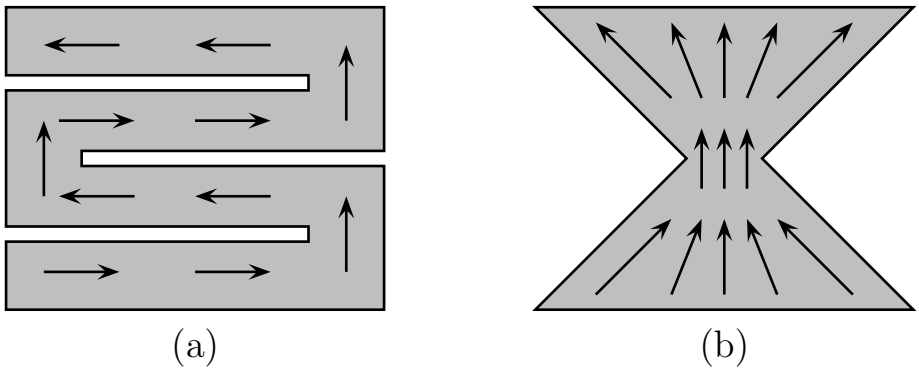


Figure 7.2: Surface current distributions for RMSAs with (a) long narrow slits and (b) triangular notches. [23]

Cutting slits in the patch has been shown to be an effective compacting and broad/multi-banding technique [23]. The reason for this is that the slots create meandered, and therefore longer, current paths. For a given patch footprint, this can result in a greatly reduced resonant frequency and thus a greatly reduced electrical size. This technique can also result in several resonances due to several different current paths. As such it can result in multi-band operation.

Fig. 7.2 shows how different shaped slits cut into an RMSA result in different surface current distributions.

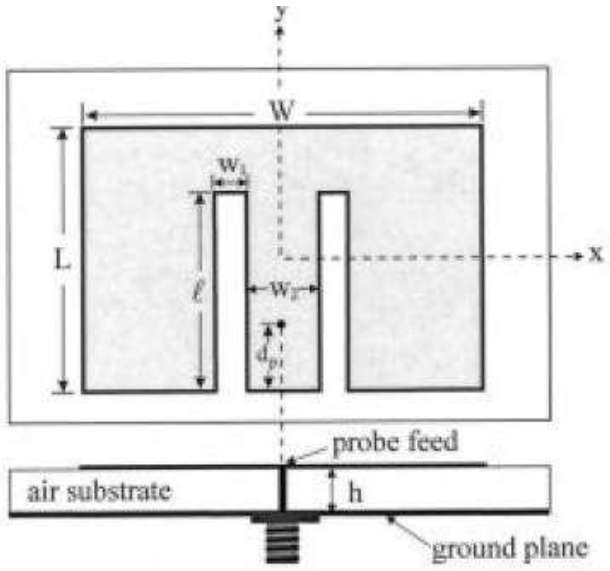


Figure 7.3: Geometry of RMSA with slits. [24]

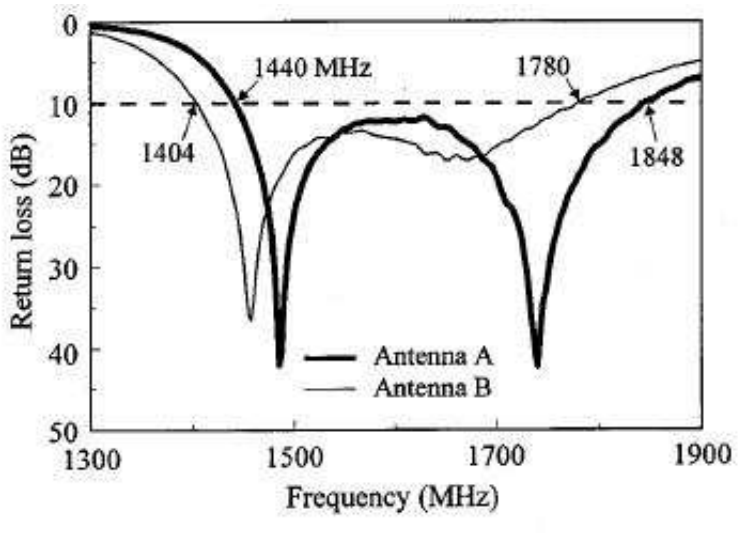


Figure 7.4: Measured S11 of RMSA with slits. [24]

A good example of using slits to improve the bandwidth of an RMSA is given in [24]. By using only two slits, as in Fig. 7.3 , dramatic results can be obtained. Fig. 7.4 shows the S11 of two MSAs (A and B) that have slightly different slit dimensions. An RMSA with the same parameters but without the slits would have approximately a 50 MHz bandwidth centered at 1.78 GHz. As can be seen in 7.4 the bandwidths of both antennas is much greater then 50 MHz. Additionally, most of the band, if not all, of each of the antennas is below 1.78 GHz. As these frequencies are lower, and thus the wavelengths are longer, these antennas are electrically smaller than the basic rectangular patch. This illustrates that cutting slits into the patch can be both a multi/broad-banding technique as well as a compacting technique. In Fig. 7.4, both antennas have a single wide band, i.e., the part of the response that is below -10 dB. This is due to the fact that, as can be seen more clearly with antenna A, there are two resonance which overlap. If the two resonances were further apart then multi-band, rather than broad-band, antennas would result.

7.2.4 Parasitic Patches

Parasitic patches are ones which are not directly driven by the MSAs's feed. Instead they are excited, i.e., receive energy, from neighboring patches via coupling of electric and/or magnetic fields. In the case of probe fed MSAs, parasitic patches are simply ones which do not have a direct connection to the probe feed. For electromagnetically coupled MSAs, such as aperture coupled, parasitic patches are ones which are usually a relatively long distance from the feed and so do not receive much energy directly from the feed.

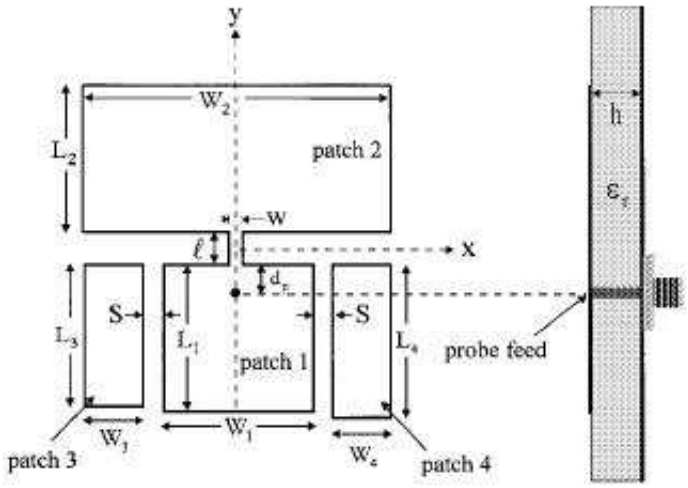


Figure 7.5: Geometry of MSA with directly coupled and parasitic patches. [25]

The use of parasitic patches to enhance the bandwidth of an MSA is given in [25]. The geometry of the antenna can be seen in Fig. 7.5. This MSA consists of 4 patches, with the two directly coupled ones (1 & 2) being joined together. Each patch has its own resonant modes. Patches 3 & 4 (parasitic) have the same dimensions and so have the same resonant modes.

Current will flow from patch 1 to patch 2 and so together these two patches form a resonant structure. Due to the frequencies of many of these resonances overlapping a much broader than normal frequency band results, as can be seen in Fig. 7.6. The fractional bandwidth of this antenna is 12.7 %, which is significantly greater than the 2 to 3 % of a standard RMSA. If all the gaps between the patches were filled in, i.e., forming a standard RMSA with the same footprint area, then it would have a centre frequency of 2710 MHz. As can be seen in Fig. 7.6, this is close to the lower cutoff frequency of the antenna's band. Consequently, in this example, the use of parasitic patches has successfully been used to enhance the bandwidth of the antenna but not to compact it. By necessity, parasitic patches have to be smaller than the overall footprint of the MSA. As such they often resonate at frequencies that are considerably higher than that of a single patch that covers the footprint area. Thus they have primarily been used for multi/broad banding purposes. However, parasitic patches of more complex shapes, such as with slits, could resonate at lower frequencies than their overall footprint size might suggest. Consequently, such patches could be used to create both compact and multi/broad band MSAs.

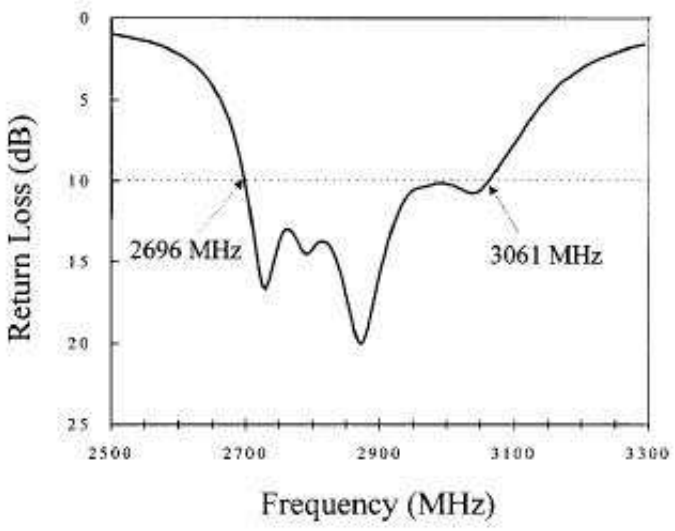


Figure 7.6: Measured S11 of MSA with directly coupled and parasitic patches. [25]

7.3 Computational Optimisation Of MSAs

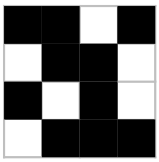
As described in section 7.2 manipulating the substrate parameters of MSAs (ϵ_r & h) is severely limited and is often not practical. Other techniques, such as the use of slits and parasitic patches have produced extremely good results when used to compact and multi/broad band MSAs. The vast majority of the use of such techniques has been done based on conventional design methods. In other words, engineers and researchers, largely due to their experience and insight from analytical models, have designed geometries that have desirable characteristics. Due to the nature in which they were designed, most of these geometries are highly regular.

For instance, parasitic patches and slits are often rectangular, triangular or circular. Although such structures have been shown to work well, there could well be many potentially beneficial geometries that are highly irregular. It is unlikely that conventional design approaches could ever discover such geometries. On the other hand, computational optimisation techniques could well discover advantageous irregular geometries. This is because they do not rely on any guidelines, heuristics, principles and rules etc. Indeed, they have no 'intelligence' or knowledge of antennas at all. As such, computational optimisation techniques are not constrained by any knowledge. They are free to try any possible geometry without bias. Consequently, they can potentially discover beneficial structures that can not be discovered by conventional design approaches.

7.4 Boolean Grid Representation

In order for computational optimisation techniques to be able to discover beneficial geometries, a way of describing MSA patch geometry (representation) is required that will allow the optimisation algorithm to work effectively. A representation that allows the creation of slits and parasitic patches is that of a Boolean grid. The footprint of the MSA is simply divided into a two dimensional grid as in Fig.7.7. Each cell in the grid can either be metalised or non-metalised. This can be represented, and therefore manipulated, by an algorithm with a string of bits. A 1 indicates that a particular cell should be metalised and a 0 that it should not.

Grid:



Bit String:

1101011010100111

Figure 7.7: 4 x 4 grid and its corresponding bit string.

In practice, larger grids, i.e., grids with more cells, would need to be used than that shown in Fig.7.7 in order to achieve practically useful results. The resolution of the grid (cell size) would depend on the printing resolution of the circuit board making process. The overall grid size would depend on the required MSA's characteristics (bandwidth & directivity etc.).

The Boolean grid representation has many advantageous characteristics, the main ones being that it is simple and robust. Fig. 7.7 shows how straightforward it is to convert a bit string to a corresponding grid. This technique is clearly capable of generating complex antenna geometries such as meandering current paths and irregular parasitic patches. Additionally, many computational optimisation techniques (GAs and CGP etc.) have been shown to work extremely well with Boolean bit strings. It is for these reasons that the Boolean grid was chosen as the form the antennas in this study would use.

7.5 Applicability of Computational Techniques to Boolean Grid Optimisation

The techniques mentioned in this section have previously been described in chapter 3. When optimising Boolean grids, the search variables are the bits in the grid bit string. In the example of Fig. 7.7 there are 16 search variables, i.e., the search space has 16 dimensions.

7.5.1 Inapplicable Techniques

Gradient methods and particle swarm optimisation can only optimise problems that have continuous search variables. They therefore can not be applied to problems that have Boolean variables.

Simulated annealing (SA) can optimise problems with Boolean variables. The starting solution and candidate new solutions can be randomly created by simply randomly creating new bit strings. However, a significant problem with SA is that it is an inefficient optimisation technique. This is because its performance is extremely sensitive to its control parameters, i.e., its starting temperature and temperature decrement. Many runs of an SA based system are often necessary in order to find sufficiently good control parameters for the given problem [49]. In other words it is necessary to optimise the optimisation. Since other techniques (GAs etc) are known to not suffer so badly from this problem, they were used instead of SA in this investigation.

7.6 Genetic Algorithms (GA)

GAs can optimise Boolean problems, indeed, it is in this area that they are best known. The two main reproductive operators of GAs: crossover and mutation, can manipulate bit strings with ease. An example of mutation and single point crossover on 8-bit strings is shown below:

- Mutation

$$10110100 \longrightarrow 10010110$$

- Crossover

$$\begin{array}{c} 10110100 \\ 01001011 \end{array} \longrightarrow \begin{array}{c} 10110011 \\ 01001100 \end{array}$$

7.7 Genetic Programming (GP)

GP can be easily applied to Boolean problems by giving the algorithm a set of instructions that enable it to create patterns on the grid. This is best illustrated with an example, which is given below. In this example, a list of instructions describes a journey around the grid. As the journey progresses, various cells are metalised and de-metalised. Here, the list length is 10 instructions.

The 'go forward N' instructions require an upper limit on N, which in this case is 3. When the current position in the journey 'hits' the side wall of the grid, there can either be a reflection or the journey can continue in the same direction on the opposite side of the grid.

As long as there is a consistent rule, either option can be used. In this example, the journey continues in the same direction on the opposite side of the grid.

Having blank (no operation) instructions in the instruction set enables fixed length lists to be effectively variable length up to a maximum length.

Instruction set:
no operation
turn right
turn left
go forward N cells making cells '0'
go forward N cells making cells '1'
go forward N cells doing nothing
go forward N cells inverting cells

7.7.1 Example i

instruction list:
Turn left
go forward 2 cells making cells '1
Turn left
go forward 1 cell inverting cells
no operation
Turn right
go forward 1 cell doing nothing
go forward 2 cells making cells '0
Turn left
go forward 3 cells making cells '1

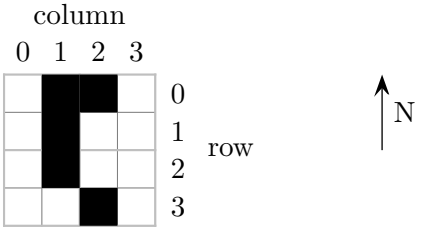


Figure 7.8: Resulting pattern of GP list on 4 x 4 grid.

Fig. 7.8 shows the pattern generated by the example GP list given above. More specifically, it is the pattern generated when the starting cell is (0,0), i.e., top left, the starting direction is S (south), and the grid is blank at the start, i.e., all the cells are 0 (de-metalised).

In the example given above there is no redundancy in the representation. This is because all of the instructions in the list are active. Apart from 'no operation' instructions, all of the instructions in a list have an effect on the final pattern on the grid. Although of course, instructions near the start of a list have a higher chance of their effect being overwritten by a subsequent instruction. Redundancy could be incorporated into this representation by giving each instruction an additional Boolean variable. The value of this Boolean variable would indicate whether or not the particular instruction is currently active or not.

As well as the optimisation algorithm mutating the individual instructions, it would also be able to mutate the Boolean redundancy variables. This would allow instructions to become activated and de-activated as the optimisation progresses. This form of redundancy would hence be *single-point* because instructions would de/re-activated individually.

7.7.2 Example ii

instruction list:

Turn left	(active)
go forward 2 cells making cells '1	(inactive)
Turn left	(inactive)
go forward 1 cell inverting cells	(active)
go forward 3 cells doing nothing	(active)
Turn right	(active)
go forward 1 cell doing nothing	(inactive)
go forward 2 cells making cells '0	(active)
Turn left	(active)
go forward 3 cells making cells '1	(active)

Fig. 7.9 shows the resulting grid pattern of the immediately above instruction list when the start cell is (0,0), the start orientation is south and the grid is blank. When reading the list, the inactive instructions are completely ignored.

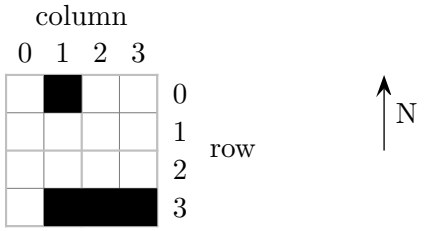


Figure 7.9: Resulting pattern of GP list with redundant instructions on 4 x 4 grid.

7.8 Cartesian Genetic Programming (CGP)

In most respects CGP is the same as GP. The key difference is that in CGP the instructions are stored in graphs rather than trees or lists. This enables large parts of the graphs to be redundant, which can increase the efficiency of the search.

CGP can be easily applied to Boolean grid problems by making each vertex in a graph contain an instruction. The graphs have a start vertex and an end vertex. They are also directed and acyclic. This means there is only a single path through a graph, i.e., from the start vertex to the end vertex. As with GP, a given instruction set is required.

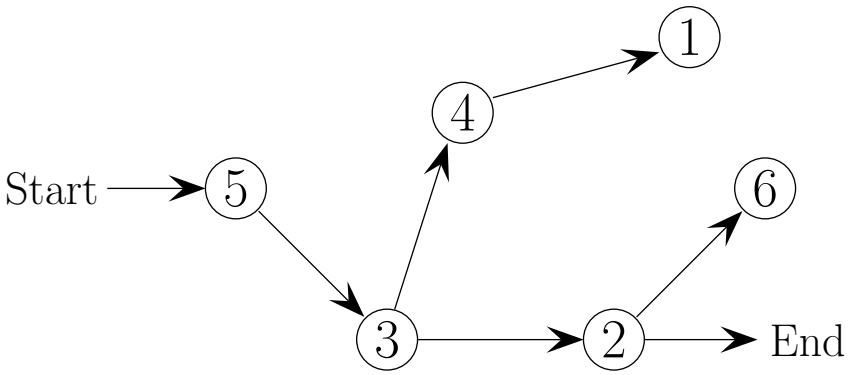


Figure 7.10: 6 vertex graph.

Vertex numbers and instructions:

1 go forward 2 cells making cells '1'
2 turn right
3 turn right
4 go forward 1 cells inverting cells
5 go forward 2 cells doing nothing
6 turn left

Active instruction sequence:

go forward 2 cells doing nothing
turn right
turn right

In Fig. 7.10 the active instruction sequence is 5-3-2. Vertices 1, 4 and 6 are redundant. The optimisation algorithm has the ability to mutate the instructions within each vertex, the structure of the graph and the start vertex number. Mutating a graph's structure and its start vertex number enables relatively large sections of the graph to be re/de-activated. This can result in a large proportion of the vertices being redundant. As described earlier this can greatly improve the efficiency of the technique.

7.9 Previously Computationally Optimised MSAs

In the literature there are several examples of the computational optimisation of MSAs. This section describes the key features of a few notable examples.

7.9.1 Dual Band MSA Optimised by GAs using Parallel Computation

In this study [26] the goal was to achieve a dual band MSA that would be suitable for hand held wireless communication devices. The two bands were centered at 1.9 and 2.4 GHz. The antennas were represented by a grid and a corresponding bit string, as shown in Fig. 7.7. The GA has a population size of 260 and ran for 200 generations. There was 90% elitism, i.e., only 10% of the population was replaced every generation. As such the GA was considerably steady-state in nature. Tournament selection was used and the mutation rate was a constant 5%.

The electromagnetic solver used for the fitness evaluation employed the method of moments (MoM) together with the much used electric field integral equation (EFIE). A supercomputing cluster was used to speed up the optimisation. All of the nodes in the cluster worked on the same fitness evaluation at the same time. Each node had the same model as the others but each individual node evaluated it for a unique frequency.

In this study there were two optimisation objectives: bandwidth and polarisation. Bandwidth was determined using VSWR. Polarisation was evaluated using the ratio of the co-polar (wanted) component of the radiated electric field to that of the cross-polar (unwanted) component. The two optimisation objectives were combined together to form a single overall fitness function.

The basic fitness term for bandwidth was:

$$E_m = \sqrt{\frac{1}{N_f} \sum_{n=1}^{N_f} \left(\frac{VSWR^{target} - VSWR_n}{VSWR^{target}} \right)^2} \tag{7.1}$$

where N_f is the number of frequencies and $VSWR^{target}$ is the desired VSWR value for all the frequencies of interest. As the target VSWR value would typically be set to a relatively low value ($1 < VSWR^{target} < 2$), the observed VSWR value at a particular frequency $VSWR_n$ would almost invariably approach the target value from above. This results in high, i.e., poor, VSWR values leading to a high value for the E_m term. VSWR values close to the target value will result in a small value for E_m. Poor VSWR values are thus penalised because they contribute to a large E_m value. As the VSWR values improve, i.e., approach the target value, the penalty decreases. When the target VSWR value is met at all of the frequencies of interest, there is no penalty.

The basic fitness term for polarisation was:

$$E_p = \frac{|E_{cross}^{max}|}{|E_{co}^{max}|} \tag{7.2}$$

The E_p term is simply the ratio of the maximum magnitudes of the co and cross polar electric field components taken a at single pre-defined frequency. Typically this frequency is the centre frequency of one of the desired bands. Ideally the maximum value of the cross component E_{cross}^{max} would be zero and thus the E_p term would also be zero.

The basic bandwidth and polarisation terms were then normalised:

$$F_m = \frac{1}{1 + E_m} \qquad F_p = \frac{1}{1 + E_p} \tag{7.3}$$

$$\longrightarrow \quad 0 \leq F_m \leq 1 \qquad 0 \leq F_p \leq 1 \tag{7.4}$$

The two normalised bandwidth and polarisation terms were then combined into a single overall fitness term:

$$F_{total} = \alpha\, F_m \; + \; (1 - \alpha)\, F_p \qquad 0 \leq \alpha \leq 1 \tag{7.5}$$

This resulted in a poor fitness function having a value close to zero and a good fitness value being close to 1. By varying the value of α in eqn. 7.5 the relative importance of bandwidth and polarisation to each other as optimisation goals can be adjusted.

An important constraint placed on the antennas was that there should be one axis of symmetry. This was done in order to assist in the achievement of a desirable co/cross polarisation ratio.

One of the antennas produced can be seen in Fig. 7.11

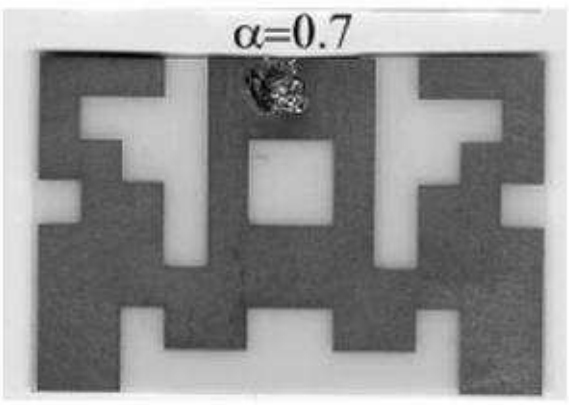


Figure 7.11: Geometry of MSA ($\alpha = 0.7$). [26]

It can be seen from Figs 7.12 and 7.13 that there is a single main lobe in the radiation pattern and in the direction of this main lobe the the co-polar component dominates the cross-polar component by about 10 dB. In addition to achieving a good cross-polarisation ratio, the antenna is well matched in the two desired frequency bands (1.9 and 2.4 GHz), as can be seen in Fig. 7.14. Fractional bandwidths ($S_{11} \leq$ -10 dB)of 5.3% and 7.0% were acheived for the bands centered at 1.9 and 2.4 GHz respectively. These bandwidths are impressive when compared to the 2-3% of the single band of an RMSA.

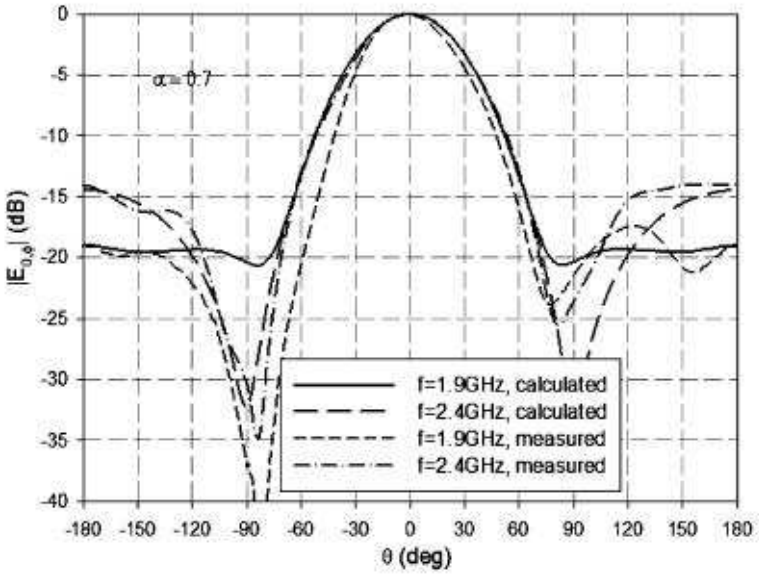


Figure 7.12: Normalised Co-polar component of radiated electric field ($\alpha = 0.7$). [26]

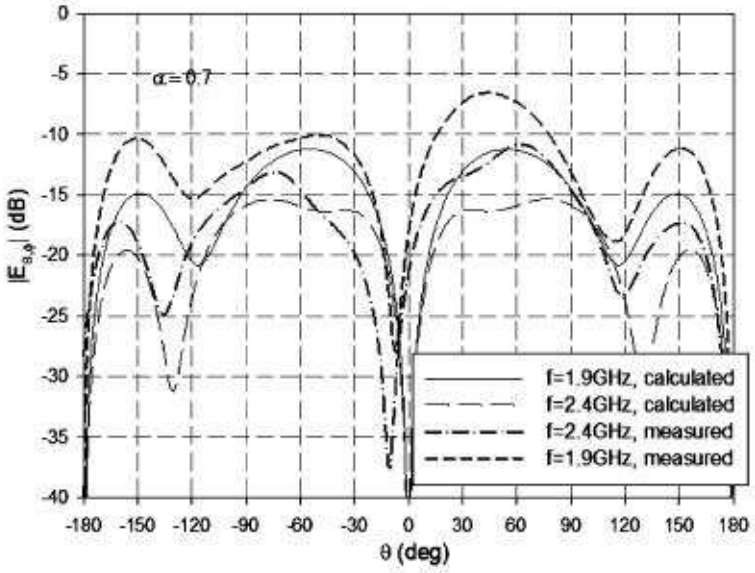


Figure 7.13: Normalised Cross-polar component of radiated electric field ($\alpha = 0.7$). [26]

By exploiting the power of a supercomputing cluster, a considerable reduction in the time taken to optimise the antennas was achieved. To optimise the antenna of Fig. 7.11 using just a single node of the cluster would have taken approximately 47 days. However, using the 26 node cluster, it actually took only 43 hours.

In conclusion, this study has shown that dual objective optimisation of MSAs can yield impressive results. The two objectives of bandwidth and polarisation undoubtedly compete with each other. However, it has been shown that solutions which have acceptable performance with regards to both objectives can be found effectively and efficiently.

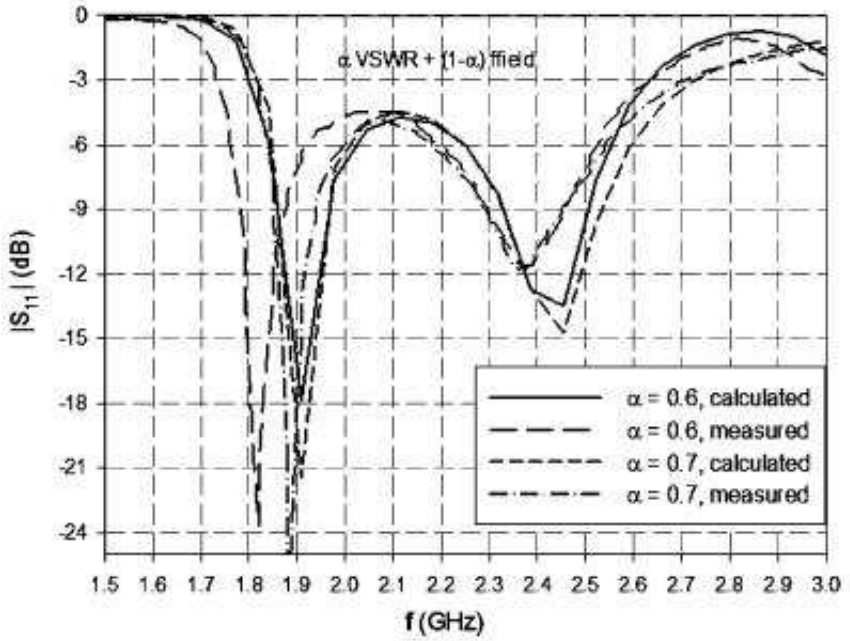


Figure 7.14: S_{11} of optimised MSA ($\alpha = 0.7$). [26]

7.9.2 Dual Band MSA Evolved with Increased Efficiency MoM

The main features of this study [27] are essentially the same as that of the previous section [26]. Indeed, Professor Yahya Rahmat-Samii, a well established figure in the field of evolved antennas, is a co-author on both these papers. For example, in this study, the MSAs are again represented as Boolean grids. The key difference is that, by exploiting the way MoM works, a significant increase in fitness evaluation time was achieved.

As in [26], MoM combined with the EFIE was used to model the antennas. The increase in evaluation time was done use of a *mother* structure together with *direct matrix manipulation* (DMM). The mother structure is simply the antenna that is formed when all of the cells in the grid are metalised, i.e., when all of the bits in the bit string are 1. The mother structure is then just a basic RMSA (Fig. 7.15).

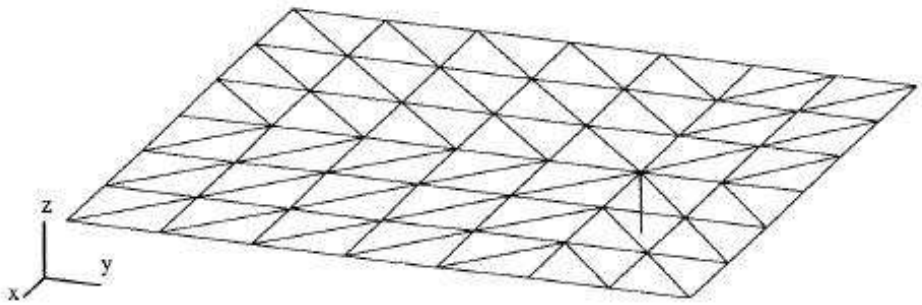


Figure 7.15: Mother structure MSA. [27]

Due to the choice of basis functions used, each square cell in the antenna's grid is actually modeled as two triangles, as can be seen in Fig. 7.15. In the well established EFIE/MoM technique a set of linear equations is solved using matrices. The matrix equation is:

$$ZI = V \tag{7.6}$$

where Z is the impedance matrix, whose elements specify the relationship between the various surface patches and wire segments in the problem and are calculated using the EFIE. V is a vector accounting for sources, and I is a vector containing the unknown currents on the structure that are to be found. The standard MoM procedure involves creating the Z and V matrices, inverting the Z matrix and then pre-multiplying V by Z^{-1} to give I.

The mother structure is itself significant because all other possible solutions are substructures of the mother structure. Removal of the metal from a particular cell in a grid, i.e., putting a 0 in the corresponding bit position, is equivalent to forcing the currents on that portion of the mother structure to be zero. This, in turn, is equivalent to setting the corresponding elements in the mother structure Z matrix to zero leaving rows and columns in the Z matrix filled with zeros. In practice, these zero rows and columns in the Z matrix make the matrix singular (non-invertible). By removing the rows and columns from the mother structure Z matrix that are due to the removed metal, a non-singular matrix is produced. The increase in speed of this technique is due to the fact that the mother structure Z matrix only needs to be created once, i.e., at the start of the optimisation. The calculation of all the individual elements of the mother structure Z matrix using the EFIE can take a significant time. The various substructure Z matrices can be relatively quickly created from the mother structure Z matrix as and when they are required during the course of the optimisation. This is the DMM technique.

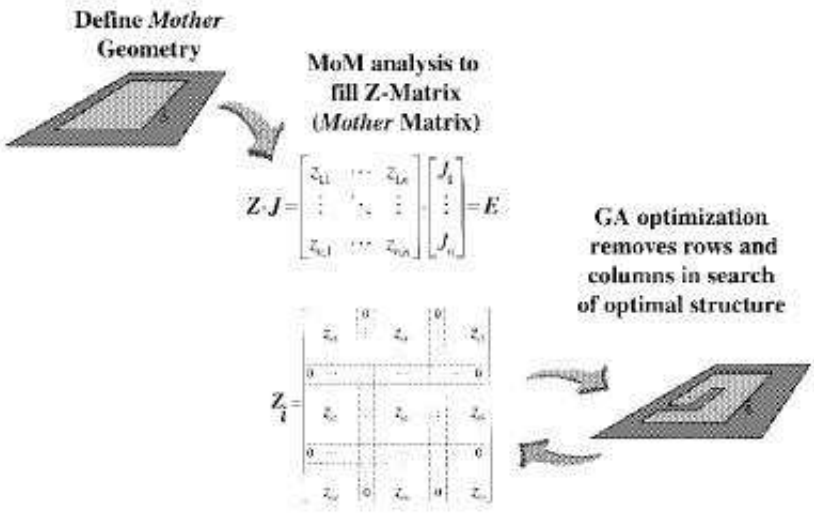


Figure 7.16: Creation of mother structure Z matrix and substructure Z matrix. [27]

As far as the GA was concerned there was only one objective - bandwidth. The goal was to produce a dual band MSA which had a good match at 3.0 and 4.0 GHz.

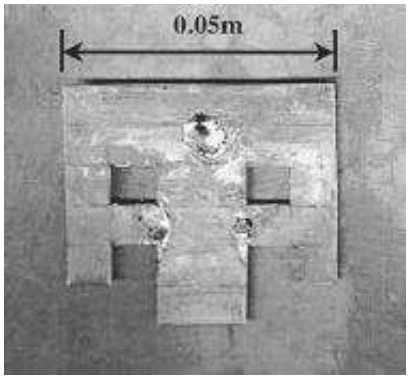


Figure 7.17: Geometry of evolved dual band MSA. [27]

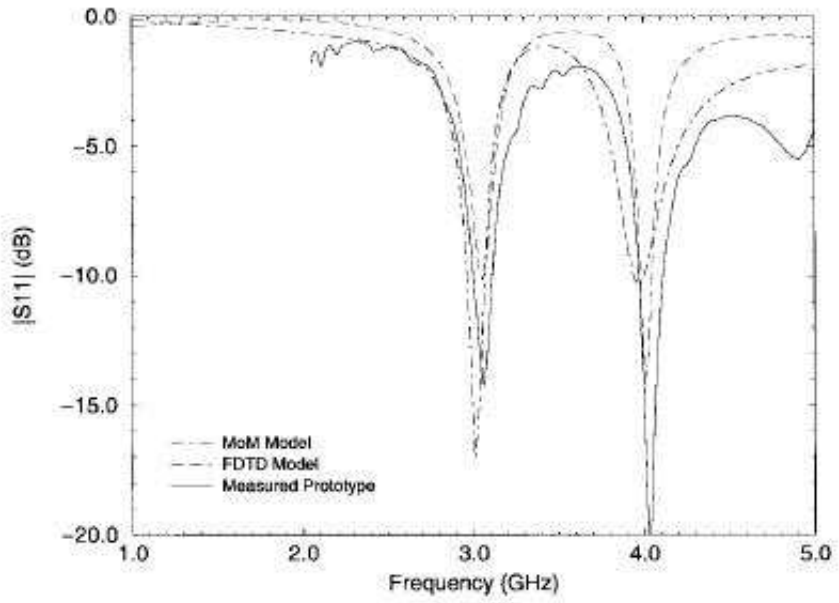


Figure 7.18: S_{11} of evolved dual band MSA. [27]

As in [26] there was the constraint of one plane of symmetry. This was to improve the likelihood of good linear polarisation. It also had the advantageous effect of further reducing the simulation time. Bandwidths (-10 dB) of approximately 100 MHz were achieved at the two required bands. Speed-up factors of approximately 4 were quoted when the DMM technique was used.

In conclusion, this study has shown that knowledge of the antenna simulation method (MoM) together with a certain key attribute of all the potential antennas (mother structure) has enabled the optimisation's efficiency to be significantly increased.

7.9.3 Circularly Polarised MSA

The goal of this optimisation [28] was to create a circularly polarised microstrip line fed MSA that operated at 11.0 GHz. The antenna simulation software used the MoM combined with Green's functions. A key difference from [26] and [27] was that only a limited part of the antenna's patch was able to be optimised. As can be seen in Fig. 7.19 only two diagonally opposing corners of the rectangular patch were allowed to be optimised. These regions had a grid structure which were represented by corresponding bit strings.

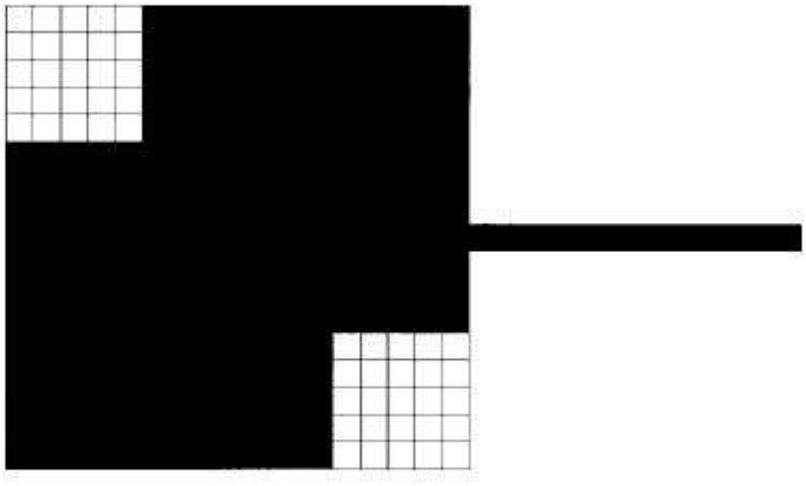


Figure 7.19: Regions of MSA able to be optimised. [28]

The main structure, i.e., the part not able to be optimised, was optimised manually, before the computational optimisation, to be matched at 11.0 GHz. A GA was then used to attempt to create a circularly polarised MSA. The fitness function was based on the axial ratio of the far field waves radiated by the antenna. The best possible axial ratio was 1 (0 dB), i.e., when the two orthogonal components of the radiated electric field have the same magnitude.

There are commonly used, single feed point, conventionally designed circularly polarised MSAs that are square in shape but have two diagonally opposing corners that have been modified in some way. One of the simplest methods involves simply trimming the edges in question, as shown in Fig. 7.20.

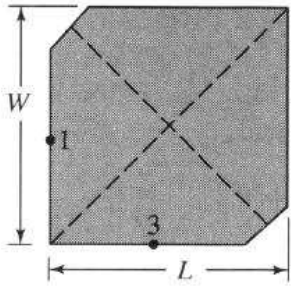


Figure 7.20: Trimmed square circularly polarised MSA. [17]

The patch shape of Fig. 7.20 produces circularly polarised waves because two orthogonal modes, with a 90° phase shift between them, are excited between the patch and the ground plane. However, the axial ratio varies with the direction from the antenna.

The purpose of this study was to try and evolve structures at the two diagonally opposing corners of the patch that would result in an axial ratio close to 1 for all directions from the antenna. Fig. 7.21 shows one of the resulting antenna patches and Fig. 7.22 is its axial ratio in a fixed plane. As shown in Fig. 7.22, the axial ratio is often close to 0 dB and is never greater than 1 dB.

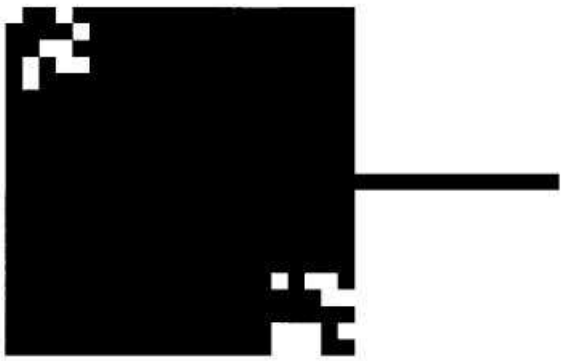


Figure 7.21: Geometry of patch of circularly polarised MSA. [28]

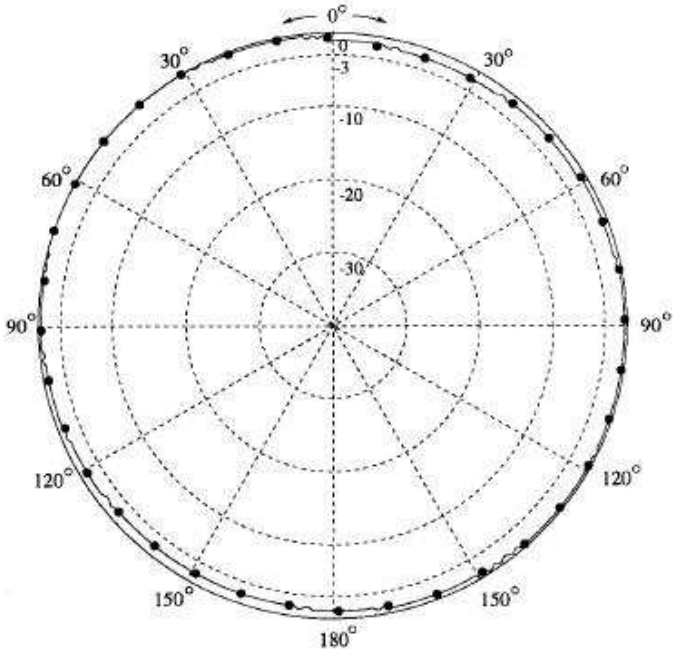


Figure 7.22: Measured axial ratio of circularly polarised MSA. [28]

In conclusion, this study has shown that skillful application of domain knowledge (circularly polarised MSAs) can be used to constrain the possible solutions in order to maximise the chances of achieving the desired goal.

Chapter 8

MSAs for Mobile Communication: Size and Environment

8.1 MSA Size and Bandwidth

8.1.1 General Relationship Between Antenna Size and Bandwidth

Generally speaking, an antenna's bandwidth is proportional to its volume. This is not an exact law but is more of a general rule. This relationship is known as the *Chu-Harrington limit* because its mathematical analysis is based on work by, firstly, Chu [69] and then Harrington [70]. The Chu-Harrington limit was later revised and made more accurate by McLean [71]. The Chu-Harrington limit uses the concept of Q factor (see section 4.6.4) to link antenna size and bandwidth. It also uses the convenient concept of the *radiansphere* [72]. The radiansphere of an antenna is the the smallest sphere which completely encloses the antenna.

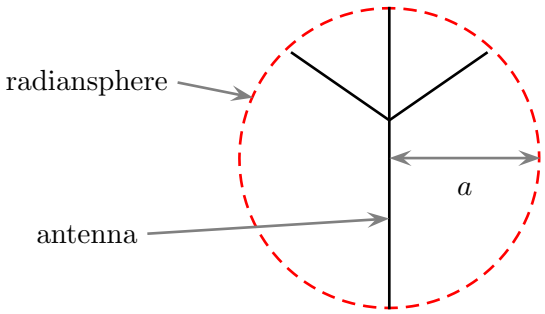


Figure 8.1: Radiansphere of an antenna.

Q and fractional bandwidth (BW) are inversely related, as can be seen in eqn. 8.1:

$$BW = \frac{\Delta f}{f_0} = \frac{S-1}{Q\sqrt{S}} \tag{8.1}$$

where:

$S = VSWR : 1$ ratio (e.g. for - 10 dB, $VSWR$ = 2:1, $S = 2$)

The lower an antenna's Q factor is, the greater its bandwidth will be. The theoretical minimum possible Q factor for an antenna depends on the radius of its radiansphere [71]:

$$Q_{min} = \frac{1}{(k\,a)^3} + \frac{1}{(k\,a)} \tag{8.2}$$

where:

$$k = 2\pi/\lambda_0 \quad (\lambda_0 = \text{free space wavelength})$$
$$a = \text{radius of radiansphere} \tag{8.3}$$

Q_{min} is only a theoretical minimum so in practice it is only ever approached but never equalled. The maximum possible bandwidth, BW_{max}, of a given radiansphere size is therefore a function of Q_{min}:

$$BW_{max} = \frac{S-1}{Q_{min}\sqrt{S}} \tag{8.4}$$

8.1.2 Relationship Between MSA Size and Bandwidth

In order to be as close as possible to the theoretical minimum Q, and thus the theoretical maximum bandwidth, for a given radiansphere size, the antenna must utilise the volume of the radiansphere as efficiently as possible [12, page 570]. Due to their low profile, MSAs generally 'waste' much of the radiansphere. This is why MSAs typically have such relatively low bandwidths. An antenna that fits inside the same size radiansphere as a given MSA, but that uses the volume of the radiansphere more effectively than the MSA can achieve significantly greater bandwidths.

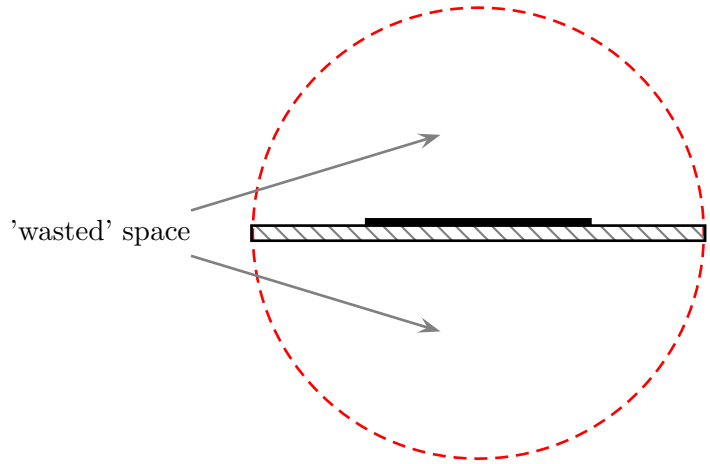


Figure 8.2: Side view of an MSA and the wasted space in its radiansphere.

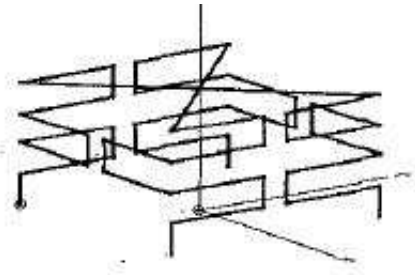


Figure 8.3: Meander line folded monopole. [29]

The folded wire antenna of Fig. 8.3 has a significantly greater bandwidth than a conventional MSA that would fit inside the same radiansphere. This is because the folded wire antenna utilises its radiansphere so much more effectively. In other words, it radiates waves, from surface currents, from a much greater proportion of the radiansphere than an MSA would.

This is the reason why compacting MSAs is not a trivial matter. Increasing the relative permittivity of the dielectric of basic MSAs (e.g. rectangular) does make them smaller but also results in a corresponding loss of bandwidth. As such, conventional techniques for compacting MSAs that do not result in a significant loss of bandwidth, as described in section 7.2, are more complex than simply increasing the relative permittivity of the dielectric.

8.1.3 Size of Grid Based MSA and Bandwidth

When computationally optimising antennas it is necessary to enable the resulting antennas to be of sufficient size in order to have the required bandwidth. If the maximum available space the antennas can occupy is too small then the required bandwidth may never be reached, no matter how long the optimisation is run for. On the other hand, if the maximum available space the antennas can occupy is excessively large then there is a high chance that the resulting antenna(s) will be far larger than is necessary and will thus be a waste of space. It is therefore important to ascertain the approximate size of the required antenna, before the optimisation starts, so that appropriate constraints can be placed on its size. In the case of grid based MSAs, this means determining the appropriate grid size, i.e., the size of each cell and the number of rows and columns in the grid.

As previously mentioned in section 8.1.2, MSAs do not fully utilise their radianspheres. This means that their Q values are much higher the minimum possible Q values (Q_{min}) for the same size radiansphere. As such, their bandwidths are significantly lower than the maximum possible bandwidth of the same size radiansphere. Indeed, MSAs have a bandwidth of between 40 and 60 times smaller than the maximum possible bandwidth of their radianspheres. Consequently, when designing MSAs, the antenna must have a markedly larger radiansphere than that would be the case if the MSA was operating at the Chu-Harrington limit (eqns 8.2 and 8.4). Thus, for MSAs:

$$BW_{sphere} = k_{msa} * BW_{req} \tag{8.5}$$

where:

BW_{sphere} = maximum possible fractional bandwidth of radiansphere enclosing MSA

BW_{req} = required fractional bandwidth of MSA

k_{msa} = bandwidth scaling factor $\qquad 40 \leq k_{msa} \leq 60$

An increase in the thickness (h) of a microstrip antenna will correspondingly result in an increase in the antenna's volume. At the same time the volume of the antenna's radiansphere will not increase significantly. This will be true as long as h remains much smaller than L and W. As such the ratio of the antenna's volume to that of its radiansphere will increase and thus it is likely that the bandwidth of the antenna will increase as well.

However, in the case of probe fed MSAs, increasing the substrate thickness does not result in a corresponding increase in bandwidth. This is because the increased bandwidth is actually cancelled out by an increased input impedance mis-match due to the increased inductance of the probe.

As such, probe fed MSAs do not have the advantage of having higher bandwidths with increased substrate thicknesses. Rather, their bandwidths appear largely constant with respect to h. There is a further disadvantage, as far as calculating the required grid size is concerned, in that a simple approach based on the antenna volume to radiansphere volume can not be used, because the bandwidth of a probe fed MSA is not a function of h.

Eqn. 8.5 gives the maximum possible bandwidth of an antenna that could just fit within the required radiansphere. This is the radiansphere that is required to enclose the grid based MSA that is to be computationally optimised. The minimum Q factor of a given radiansphere can easily determined from the radiansphere's maximum possible bandwidth using eqn. 8.4. S_{req} is the required VSWR of the antenna, i.e., the antenna must equal or better this value for the whole of its required band. The minimum Q of the required radiansphere is:

$$Q_{min} = \frac{S_{req} - 1}{BW_{sphere}\sqrt{S_{req}}} \tag{8.6}$$

Using eqn. 8.2, a cubic equation involving the radius of the required radiansphere (a) can be formed:

$$\frac{1}{(k\,a)^3} + \frac{1}{(k\,a)} = \frac{S_{req} - 1}{BW_{sphere}\sqrt{2}} \tag{8.7}$$

let:

$$m = \frac{1}{(k\,a)} \qquad B = \frac{S_{req} - 1}{BW_{sphere}\sqrt{2}}$$

$$\implies \quad m^3 + m - B = 0 \tag{8.8}$$

The roots of eqn. 8.8 can be simply found using mathematical software such as Matlab. There is only one purely real (i.e. no imaginary part) root of eqn. 8.8. It is this root that is used to find a.

For grid based MSAs, a is the distance from the centre of the grid to the edge of the radiansphere. In order for the antenna to function properly the ground plane must extend outwards a certain amount from the edge of the grid. Indeed, it must extend six times the substrate thickness (h) or more to ensure correct operation [21]. This distance is known as the *ground plane extension* (*gpe*). There is no need to make the *gpe* too large, as this would only result in the antenna occupying an excessive area. If the grid is square, i.e., the number of rows and columns are the same, then a will be equal to half the grid dimension (d) plus *gpe*.

Fig. 8.4 shows the geometry of a square grid based MSA and its radiansphere. As long as *gpe* is of sufficient size, the corners of the ground plane do not need to be included in the radiansphere.

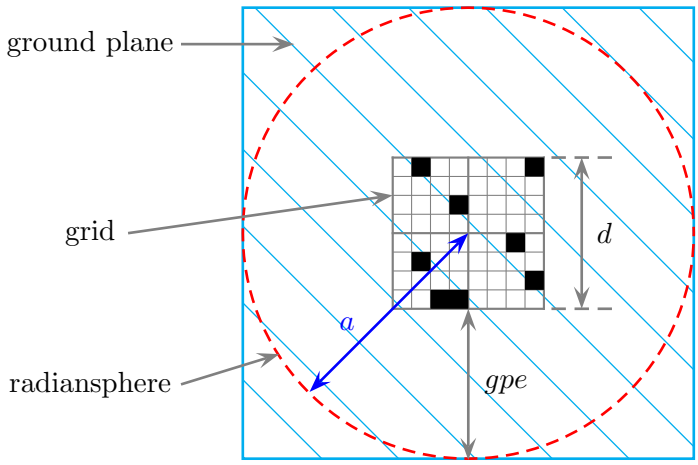


Figure 8.4: View from above of a square grid based MSA and its radiansphere.

The ground plane extension (gpe) should be limited according to the substrate thickness (h):

$$6h \leq gpe \leq 10h \tag{8.9}$$

As can be seen in Fig. 8.4, for a square grid, a is given by:

$$a = \frac{d}{2} + gpe \tag{8.10}$$

The dimension of the grid (d) is then:

$$d = 2\,(a - gpe) \tag{8.11}$$

This analysis has shown that it is straightforward to determine the approximate necessary grid dimension(s) that will enable the target bandwidth to be achieved. It is approximate because the radiansphere scaling factor (k_{msa}) is not exact.

8.1.4 Determination of Grid Size Example

The principle described above is perhaps illustrated best by an example. In this example the antenna requires a bandwidth of 100 MHz (Δf) which is centered at 2.45 GHz (f_0). For the whole of the 100 MHz band, the antenna must have a return loss of -10 dB or better (more negative). This corresponds to a VSWR value (S_{req}) of 1.92. A bandwidth scaling factor (k_{msa}) of 50 is to be used. The thickness of the substrate (h) is 1.6mm, and the ground plane extends 10 times h from the edges of the grid.

Input variables:
$\Delta f = 100$ MHz
$f_0 = 2450$ MHz
$S_{req} = 1.92$
$k_{msa} = 50$
$h = 1.6$ mm

intermediate variables:

$$\lambda_0 = c/f_0 = 0.122m \tag{8.12}$$

$$k_0 = 2\,\pi/\lambda_0 = 51.5 \tag{8.13}$$

$$gpe = 10h = 16mm \tag{8.14}$$

$$BW_{req} = \Delta f/f_0 = 0.0408 \tag{8.15}$$

$$BW_{sphere} = k_{msa} * BW_{req} = 2.04 \tag{8.16}$$

$$Q_{min} = \frac{S_{req} - 1}{BW_{sphere}\sqrt{S_{req}}} = 0.325 \tag{8.17}$$

cubic equation:

$$m^3 + m - 0.325 = 0 \tag{8.18}$$

real root of cubic equation:

$$m = 3.00 \tag{8.19}$$

radius of antenna's radiansphere:

$$a = \frac{1}{k_0\,m} = 0.065m \tag{8.20}$$

Grid side dimension:

$$d = 2(a - gpe) = 2(0.065 - 0.016) = 0.098m \tag{8.21}$$

8.2 Antenna Size and Directionality

8.2.1 General Relationship Between Antenna Size and Directionality

A radiation pattern is an interference pattern. An interference pattern describes how waves produced by various different sources interact with each other. Typically all of the sources are radiating waves of the same frequency, which also usually means all of the waves have the same wavelength. At each point in space, the overall field strength is simply the summation of all the individual waves at that point. When two waves, at a particular point, are in phase, their magnitudes will add together to produce a greater field magnitude. This is known as *constructive interference*. Two waves that are out of phase will cancel each other out. This is known as *destructive interference*.

The simplest radiation patterns are produced by two coherent, scalar sources of the same magnitude. In other words, the two sources are of the same frequency, phase and magnitude, and they are each radiating a scalar field. This is also a convenient way of visualising the dependence of the characteristics of the interference pattern on the distance between the sources.

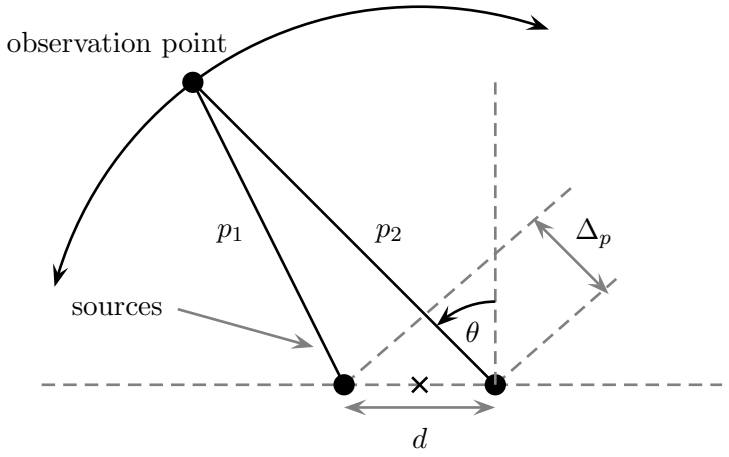


Figure 8.5: Path difference geometry of two source interference pattern.

In Fig. 8.5 the observation point remains a constant distance from the centre of the two sources. This is consistent with the measurement and simulation of antenna radiation patterns. As such, the loci of the observation points form an arc around the centre of the two sources. For a given observation point position, the path difference (Δ_p) between the two source points and the observation point can be easily determined.

Constructive interference occurs when the path difference is an integer number of wavelengths. This is because the the signals from the two sources will be in phase at the observation point because the two sources are in phase. When the path difference is an integer number of half wavelengths, destructive interference occurs. If the two sources were of opposite phase then the opposite situation would result.

$$\Delta_p = |p_1 - p_2| = d \sin\theta \tag{8.22}$$

$$\begin{aligned} \text{when:} \quad \Delta_p &= n\lambda \quad \longrightarrow \quad \text{constructive interference} \\ \text{when:} \quad \Delta_p &= \frac{(2n-1)\lambda}{2} \quad \longrightarrow \quad \text{destructive interference} \\ n &= 0,\ 1,\ 2\ \ldots \end{aligned}$$

As can be seen in Fig. 8.6, as the source separation increases, i.e., moving left to right from a) to d), the interference pattern becomes more complicated. More specifically, the number of lobes and nulls increases as the source separation increases.

In an antenna, the sources that radiate are the different areas of current on the antenna's surface. Although the example of the two coherent sources is an extremely basic one, it clearly illustrates a general rule of antennas that, as an antenna's size increases relative to the wavelength, its radiation pattern becomes more complex.

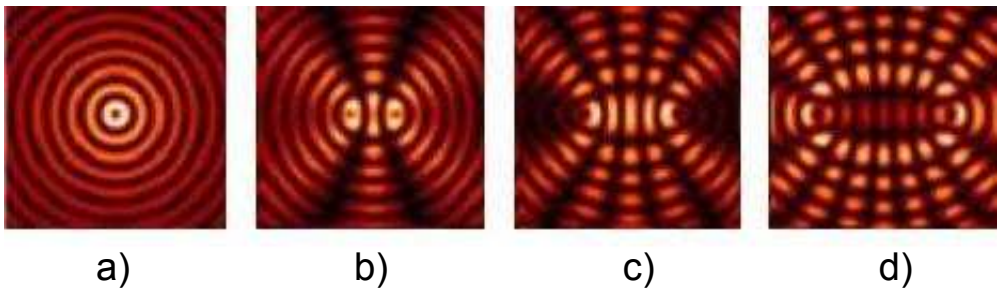


Figure 8.6: Interference patterns from two coherent sources for increasing source separations. [30]

The relationship between antenna size and radiation pattern was further investigated by using a model that assumed that an antenna is simply a group of point sources. In reality, all antennas have a continuous surface current distribution. The radiation from an antenna in transmission is from this continuous surface current distribution and so it is clearly an approximation to model the radiation as coming from individual discrete sources. This model also ignores other such phenomena as refraction. However, it is not being used to accurately determine an individual antenna's radiation pattern but rather to explore the general relationship between the size of a radiating structure and the overall complexity of the resulting radiation pattern. This technique also employed a statistical approach in that for each configuration, several runs were conducted and the results were averaged.

The technique involves determining the overall scaler field magnitude at each observation point. The source points are all relatively close to the origin whilst the observation points are at relatively large, constant distance from the origin. As such, the observation points form an arc around the cluster of source points.

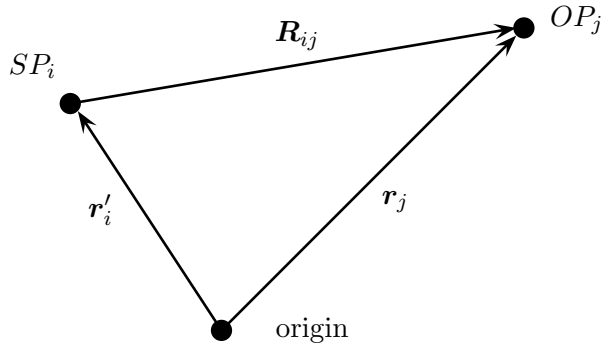


Figure 8.7: Geometry of ith source point and jth observation point.

Each source is modeled as a scalar point source. Therefore each source only produces a scaler field. Fig. 8.7 illustrates the geometry between the ith source point and jth observation point. Each individual source also has its own magnitude and phase. To determine the total field at a given observation point (jth), the contribution from each of the individual (ith) source points must first be found. In this model, each source is a fixed frequency, steady-state source. With reference to Fig. 8.7, the parameters involved in determining the field at OP_j due to SP_i are:

A_i = magnitude of ith source point
α_i = phase of ith source point
$\boldsymbol{R}_{ij}$ = vector from SP_i to OP_j
$R_{ij} = |\boldsymbol{R}_{ij}|$
$k_0 = 2\,\pi/\lambda_0$
λ_0 = free space wavelength

The field at OP_j due to SP_i is then:

$$A_i(\boldsymbol{r}_j) = \frac{A_i}{R_{ij}^2}\, e^{\,j\,(\alpha_i - k_0\,R_{ij})} \tag{8.23}$$

The total field at OP_j is then simply the summation of the contributions from all of the individual source points:

$$A(\boldsymbol{r}_j) = \sum_{i=1}^{N_s} \frac{A_i}{R_{ij}^2}\, e^{\,j\,(\alpha_i - k_0\,R_{ij})} \tag{8.24}$$

where N_s is the number of sources.

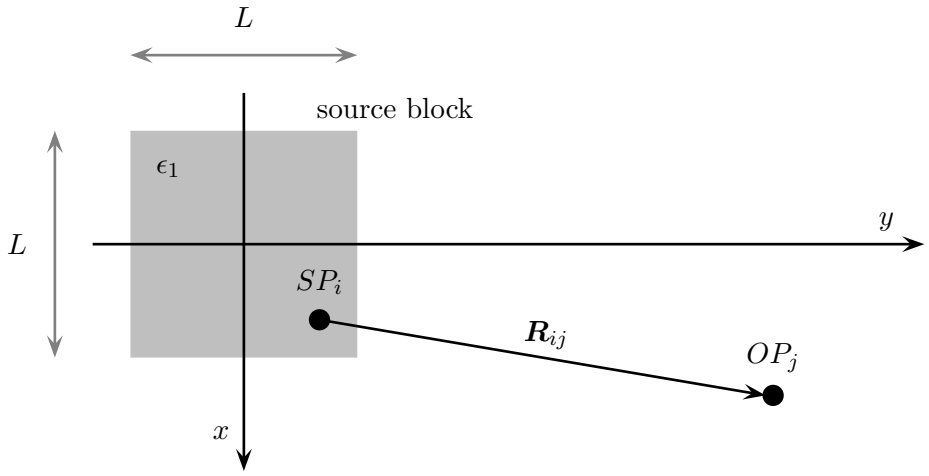


Figure 8.8: Geometry of *ith* source point and *jth* observation point.

The model was enhanced by placing the source points within a block of dielectric material which was centered at the origin. The geometry can be seen in Fig. 8.8. The block is a cube of side length L. Its permittivity is ϵ_1 ($\epsilon_0 \leq \epsilon_1$). Outside of the block is free space. In order to calculate the total electric field at a given observation point, the propagation constants (k_0 & k_1) inside and outside of the block must be known:

$k_0 = 2\,\pi/\lambda_0$
$\lambda_1 = \lambda_0/\sqrt{\epsilon_1}$ = wavelength in source (dielectric) block
$k_1 = 2\,\pi/\lambda_1$

Part of the path from the *ith* source point to the *jth* observation point is within the source block. The other part of the path is in free space. When $1 < \epsilon_1$, the propagation velocity in the source block will be slower than that of free space. This must be taken into account when calculating the contribution from a given source point. This is done by splitting the vector $\boldsymbol{R}_{ij}$ into two parts: $\boldsymbol{R}^1_{ij}$ & $\boldsymbol{R}^2_{ij}$.

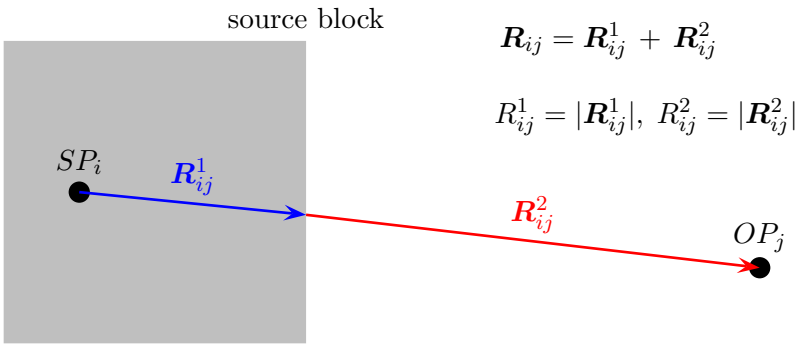


Figure 8.9: Path from *ith* source point to *jth* observation point.

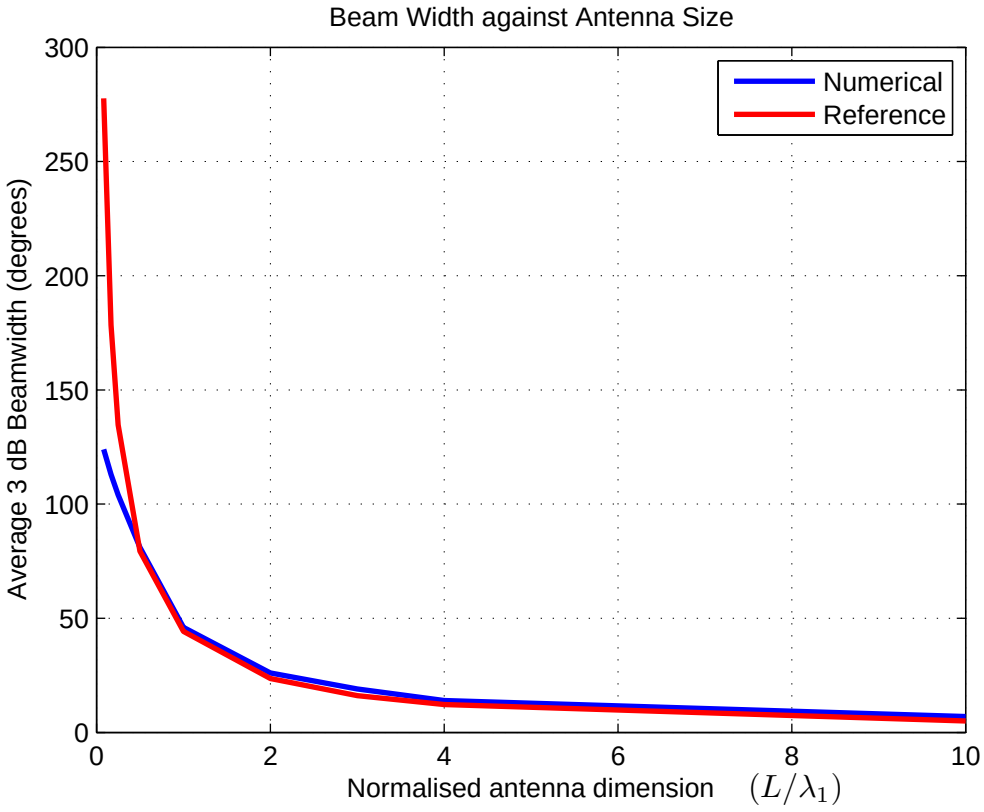


Figure 8.10: Average 3 dB beamwidth of main lobe against antenna (source block) dimension.

The total field at the jth observation point is then:

$$A(\boldsymbol{r}_j) = \sum_{i=1}^{N_s} \frac{A_i}{R_{ij}^2} \, e^{\,j\,(\alpha_i - (k_1\,R_{ij}^1 + k_0\,R_{ij}^2)} \tag{8.25}$$

This technique was straightforward to implement in software. For each source block size, 100 independent runs, each with 1000 randomly placed source points (within the source block) were performed. The magnitude (A_i) and phase (α_i) of each source were assigned randomly. A_i could vary by a factor of 1000 (60 dB) and α_i was between $+180°$ and $-180°$. Several observation points, all of a fixed distance from the origin, were restricted to two planes: θ variation and ϕ variation. For each run, the main lobe (highest magnitude) was identified and its half power (3 dB) beamwidth was determined. For each source block size, the average 3 dB beamwidth of all 100 runs was calculated.

The results of the average half power (3 dB) beamwidths against source block size can be seen in Fig. 8.10. In Fig. 8.10, the numerical curve was obtained using the above computational technique, i.e., from eqn. 8.25.

In Fig. 8.10, the reference curve is derived from [73], and is:

$$HPBW = \frac{90}{\sqrt{\pi \left(\frac{L}{\lambda_1}\right)^2 + \frac{L}{\lambda_1}}} \qquad \text{degrees} \tag{8.26}$$

This investigation has shown that it is the size of the source block (L) the in terms of the wavelength that the sources radiate into (λ_1) that is the key in determining the main lobe beamwidth. All of the sources are embedded in the dielectric block and so the path between any two sources is entirely within the block. Changing the permittivity of the block (ϵ_1), but keeping all of the sources in the same position will result in an entirely different radiation pattern. This is because the distance between the sources in terms of the wavelength will have changed.

It has also shown that the reference curve can be used to accurately obtain a quick analytical indication of the beamwidth for a given antenna size.

8.2.2 RMSA Size and Directionality

Before considering the relationship between the size of grid based MSAs and directionality, it is useful to consider that of RMSAs, in order to gain insight into relevant phenomena.

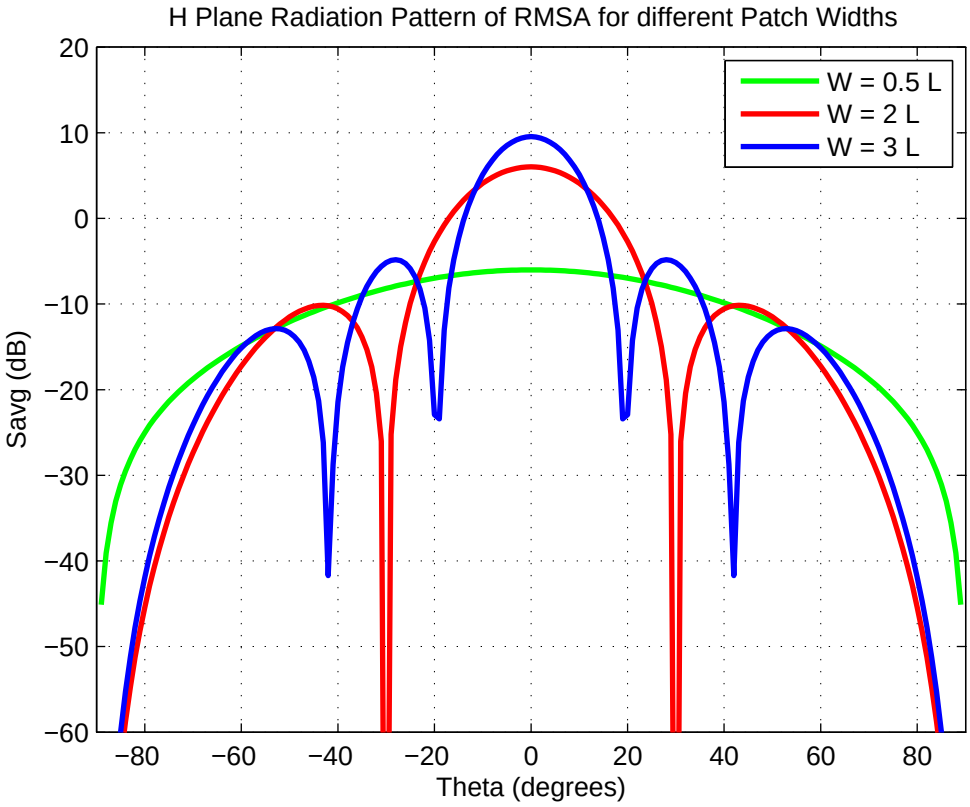


Figure 8.11: RMSA radiation pattern.

It can be seen from Fig. 8.11 that as the patch width (W) increases, the radiation pattern becomes more complicated. More specifically, the number of lobes and nulls increases. Furthermore, the maximum radiation intensity of the main lobe increases, and so the directivity of the antenna also increases. These findings are consistent with those from section 8.2.1, in that the main lobe beamwidth decreases as the antenna size increases (relative to the wavelength). It must be noted that in Fig. 8.11, the patch length (L) is constant and so all of the RMSAs are operating at the same frequency and thus the same wavelength.

In the case of RMSAs, the point source numerical technique, described in the previous section, was compared against some published results. The published results gave directivity (D) instead of 3 dB beamwidth. Assuming that the radiation pattern is dominated by a single, approximately rotationally symmetrical, main lobe, then the approximate directivity can be easily calculated from the 3 dB beamwidth. As long as the antenna size is not too large, i.e., the longest antenna dimension is not longer than approximately 4 times the wavelength in the antenna, then this assumption is valid.

$$D \approx \frac{4\,\pi}{HPBW^2} \tag{8.27}$$

The published results in Fig. 8.12 are from [21, Table 2, Page 46]. The effective relative permittivity is used to characterise the RMSAs because it is a composite term which takes into account: the patch width (W), the patch height (h) and the relative permittivity of the dielectric (ϵ_r). To make a fair comparison of the antennas, these three key parameters must be considered. With reference to Fig. 8.12, as ϵ_e increases, the propagation velocity in the dielectric decreases and consequently so does the wavelength in the dielectric (λ). This means that in terms of the free space wavelength (λ_0), which is fixed, the antenna is becoming smaller and therefore less directional.

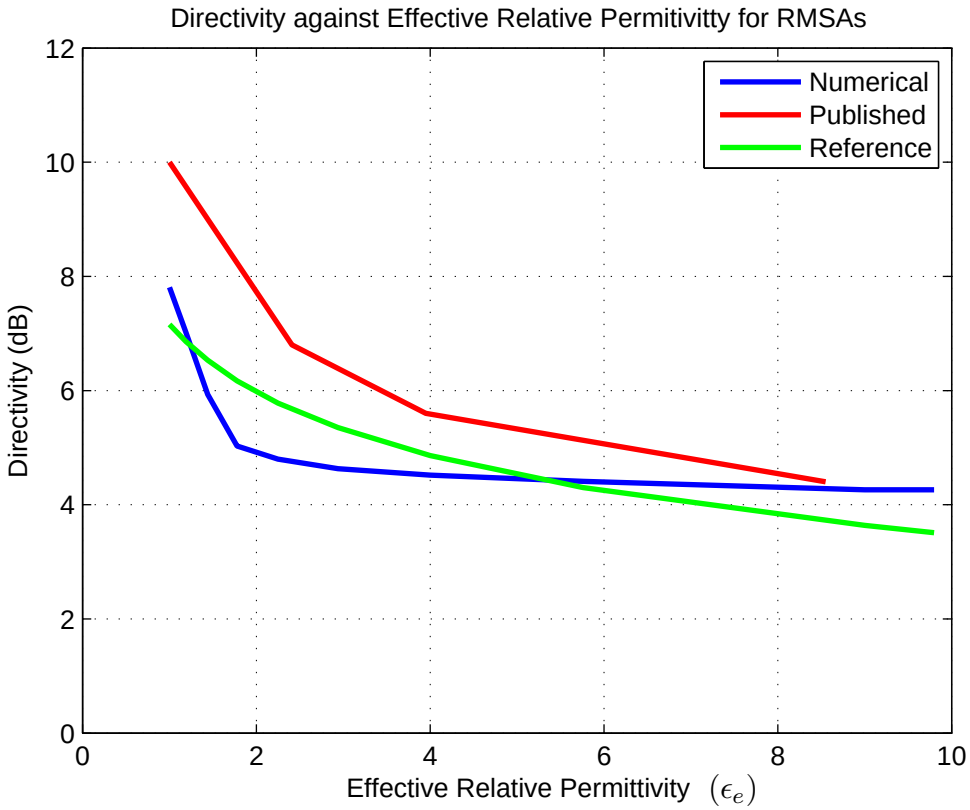


Figure 8.12: RMSA directivity against ϵ_e.

In Fig. 8.12, the reference curve is:

$$D \approx \frac{4\,\pi}{HPBW^2} \tag{8.28}$$

where:

$$HPBW = \frac{90}{\sqrt{\pi \left(\frac{W}{\lambda_0}\right)^2 + \frac{W}{\lambda_0}}} \qquad \text{degrees} \tag{8.29}$$

In the context of antenna size and directivity, ϵ_e can be regarded as a composite term that acts a measure of the miniturisation of the antenna, i.e., the larger ϵ_e is then the more miniaturised the antenna is. Fig. 8.12 clearly shows the key issue that the more an antenna is miniaturised (made smaller relative to its operating wavelength) the lower its directivity becomes. In other words, as an antenna gets smaller it gets more omni-directional.

The key difference between eqn. 8.26 and eqn. 8.29 is that in the latter, it is the free space wavelength (λ_0) that is used, rather than the wavelength in the antenna. This is significant because the wavelength that resonates in the cavity of an RMSA (λ) is typically significantly smaller than the free space wavelength:

$$\lambda = \frac{\lambda_0}{\sqrt{\epsilon_e}} \tag{8.30}$$

where:

$$\frac{\epsilon_r + 1}{2} \leq \epsilon_e \leq \epsilon_r \tag{8.31}$$

An RMSA can be modeled using the Coulomb point source model (section 8.2.1) as shown in Fig. 8.13. As long as the ground plane extends sufficiently beyond the edges of the patch, the ground plane will appear approximately infinite. This means that image theory can be used which means that the sources on the patch have mirror images on the other side of the ground plane.

As shown in in Fig. 8.13, the patch (front) sources radiate directly into free space. The path between any two front sources is entirely within free space. On the other hand, the ground plane sources radiate into the dielectric. However, there is only the distance of twice the dielectric thickness ($2h$) before the wave from a given ground plane source reaches its first patch source. As the dielectric thickness is significantly smaller than the wavelengths in both the dielectric and free space, then so is twice this distance, i.e., $2h \ll \lambda < \lambda_0$. The ground plane sources therefore only have to radiate through a thin sheet of dielectric before their radiation reaches free space. The patch sources are actually in free space and the ground plane sources are effectively in free space. The result of this is that an RMSA can be regarded as radiating directly into free space. This is why the half power beamwidth, and thus the directivity, is a function of λ_0, as in eqn. 8.29.

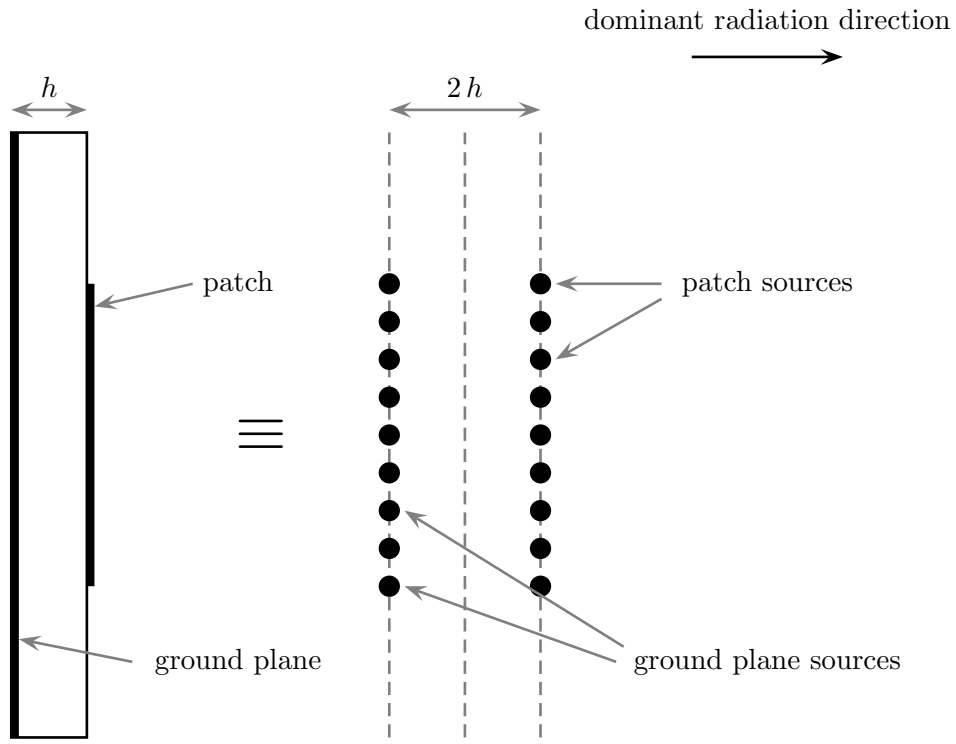


Figure 8.13: Equivalent RMSA sources.

8.2.3 Grid based MSA Size and Directionality

The Coulomb point source model (section 8.2.1) was used to determine the average 3 dB beamwidth of square grid based MSAs. The sources were arranged as two grids of 101 * 101 sources, similar to that seen in Fig. 8.13. This gave 204022 sources in total. The patch grid was a small distance ($2h$) in front of the ground plane grid. The magnitude of each patch (front) grid source was randomly generated and was between 1000 and 1000000 (60 dB). The phase of each patch grid source was also randomly generated and was between $+180°$ and $-180°$. The magnitude and phase of each ground plane (back) source was appropriately linked to that of the patch grid source that was directly in front of it. 20 independent runs were performed. The grid dimension (d) is the length of the side of the square grid as shown in Fig. 8.4. The results are shown in Fig. 8.14.

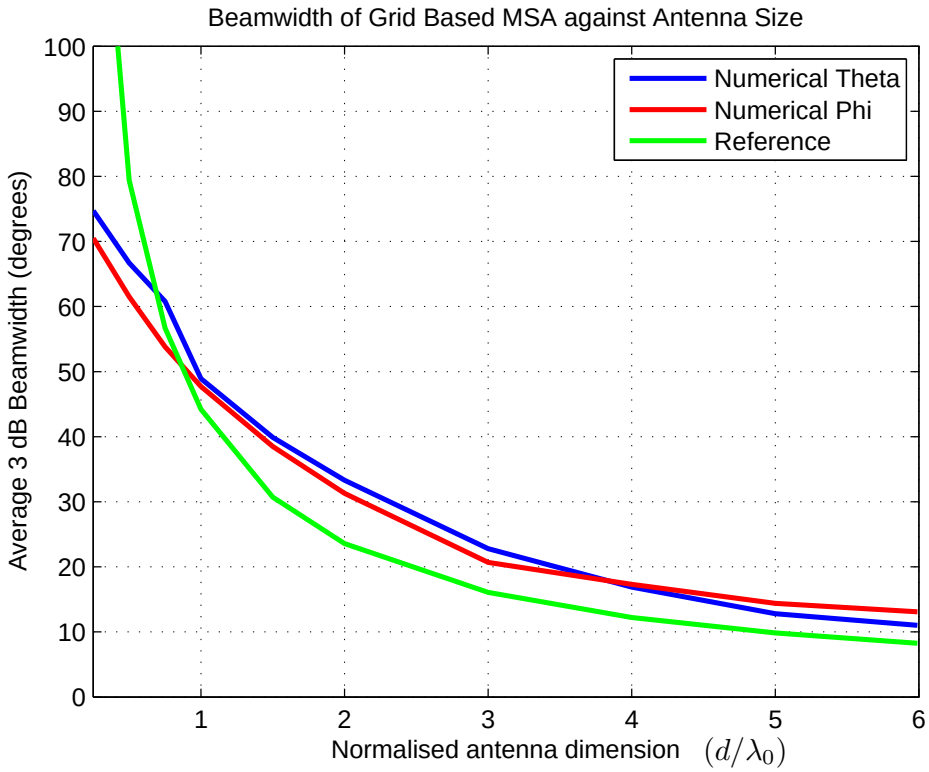


Figure 8.14: Beamwidth of grid based MSA against grid size.

In Fig. 8.14, the reference curve is:

$$HPBW = \frac{90}{\sqrt{\pi \left(\frac{d}{\lambda_0}\right)^2 + \frac{d}{\lambda_0}}} \qquad \text{degrees} \tag{8.32}$$

The approximate directivity (D) is given by eqn. 8.27. In Fig. 8.14, the HPBW was calculated for two orthogonal planes (θ and ϕ variation). Due to the randomness of the sources, the average results for these two planes were the same. The reference function of eqn. 8.32 proved to be reliable for grid dimensions greater than about $\lambda_0/2$.

8.2.4 Conclusion of Antenna Size and Directionality Analysis

This investigation has shown how the approximate beamwidth size, and thus directivity, that can be expected from an antenna varies with the antenna's size. Regarding grid based MSAs, a key implication is that as long as the grid size (d) is kept relatively small (i.e. $0.25 \leq d/\lambda_0 \leq 0.75$) the antenna should have a low directivity. More specifically, it should have a single, approximately symmetrical, main lobe. Of course, if the antenna size is made too small then it is likely to have a detrimental effect on other areas of the antenna's performance such as bandwidth.

8.3 Propagation in Mobile Communications: Depolarisation

This section has been included in this thesis, despite perhaps not being strictly necessary, in order to convey the fact that a detailed analysis of propagation was conducted, and that the conclusion drawn from it is valid.

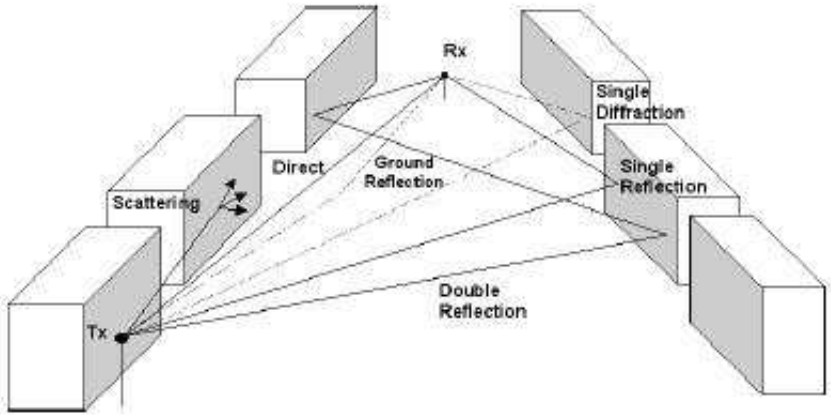


Figure 8.15: Typical ray geometry in urban streets. [31]

In a typical urban/sub-urban environment there are several paths by which a signal can travel between a mobile communications device (e.g. mobile phone) and a base station. Fig. 8.15 shows the main different path types.

8.3.1 Signal Strength Variation: Path Loss

Transmitted radio waves naturally spread out as they travel through space and so the further the receiver is from the transmitter, the lower the magnitude of the received signal will be. This phenomenon is known as *path loss*. In addition to the natural spreading out of wave energy, there are other disrupting mechanisms, as shown in Fig. 8.15. This means that the propagation from transmitter to receiver in a typical urban/sub-urban environment is usually fairly complex. There are two main types of path loss model: *empirical* and *deterministic*.

Empirical models are based on measurement data. They usually consist of one or more equation(s) and have a random (statistical) element to them. As such they are not particularly accurate, but rather give an approximate indication of the expected path loss. They are however, fairly quick and easy to use.

Deterministic models use an accurate geometric model of the environment together with techniques such as: Geometric Optics (GO) [74] and the Uniform Theory of Diffraction (UTD) [75]. Deterministic models are site specific and are significantly more computationally expensive than empirical models. However, they do generally give more accurate results.

Empirical models are much better suited for modeling outdoor propagation, especially in urban/sub-urban environments. Outdoor environments are usually far too large and complicated for a sufficiently accurate geometric model to be made of them. In addition, the variations due to rain and moving traffic further complicate the situation.

Deterministic models have been used to model propagation in certain indoor environments more accurately than empirical models [76]. Path losses are conventionally expressed in dB. A basic empirical model has the form:

$$PL(d)_{dB} = PL(d_0)_{dB} + 10\,n\,\log(d/d_0) + S \tag{8.33}$$

where:

$PL(d)_{dB}$ = path loss at distance d in dB
$PL(d_0)_{dB}$ = path loss at distance d_0 in dB = 20 log $\left(\frac{4\,\pi\,d_0}{\lambda}\right)$
d_0 = starting distance (typically around 100m)
n = path loss exponent
S = normally distributed random number: mean = 0, standard deviation = σ

The $PL(d_0)_{dB}$ term assumes that between the transmitter and d_0, the path loss is entirely due to free space, i.e., there are no signal disrupting mechanisms in this region. The path loss exponent determines how quickly the path loss increases with distance. For free space, n= 2. For outdoor urban/sub-urban environments n is typically between 3 and 5. Inaccuracies in n can contribute significantly to inaccuracies in the overall path loss.

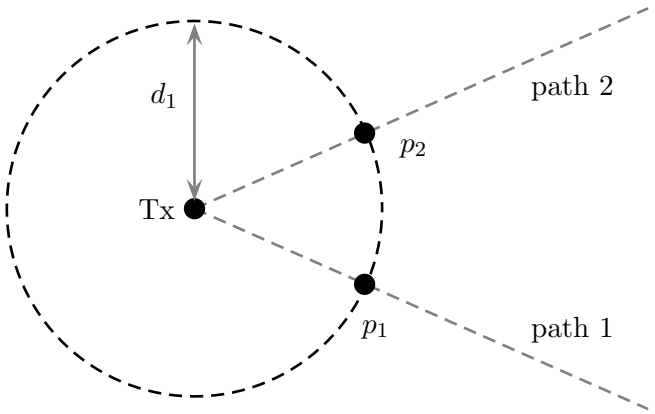


Figure 8.16: Path loss example geometry.

At a given constant distance from the transmitter, in an urban environment, there will typically be a significant change in the path loss as the receiver's position is moved. In Fig. 8.16 this is equivalent to moving from p_1 to p_2. This is why there is the S term in the empirical path loss model of eqn. 8.33. This term accounts for the random variation in the path loss that occur when the distance from the transmitter is constant.

In addition to the constant distance random variation, is the difference in path loss increase rate with distance for different paths. With reference to Fig. 8.16, moving from p_1 along path 1, away from the transmitter, a different path loss increase rate (i.e. n) will probably be observed to that when moving from p_2 along path 2. However, as can be seen in eqn. 8.33, only a single value of n is used. This is why empirical models usually only have a limited accuracy.

The S term in eqn. 8.33 is also known as the *fading error*. Its value is in dB and it has a normal distribution. This is known as a *log normal* distribution.

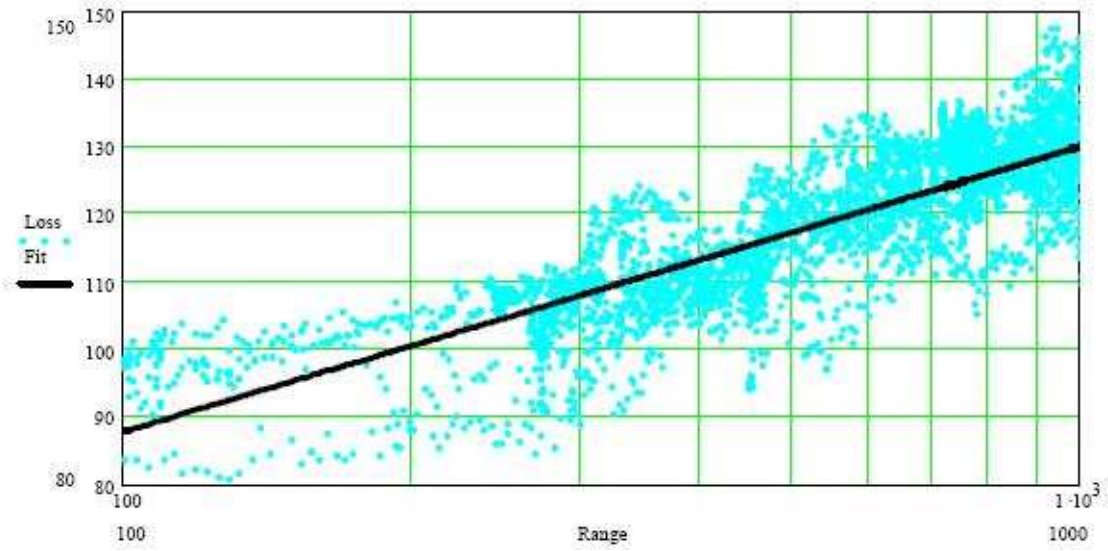


Figure 8.17: Measured path loss and least squares approximation versus range for sample data set. [32]

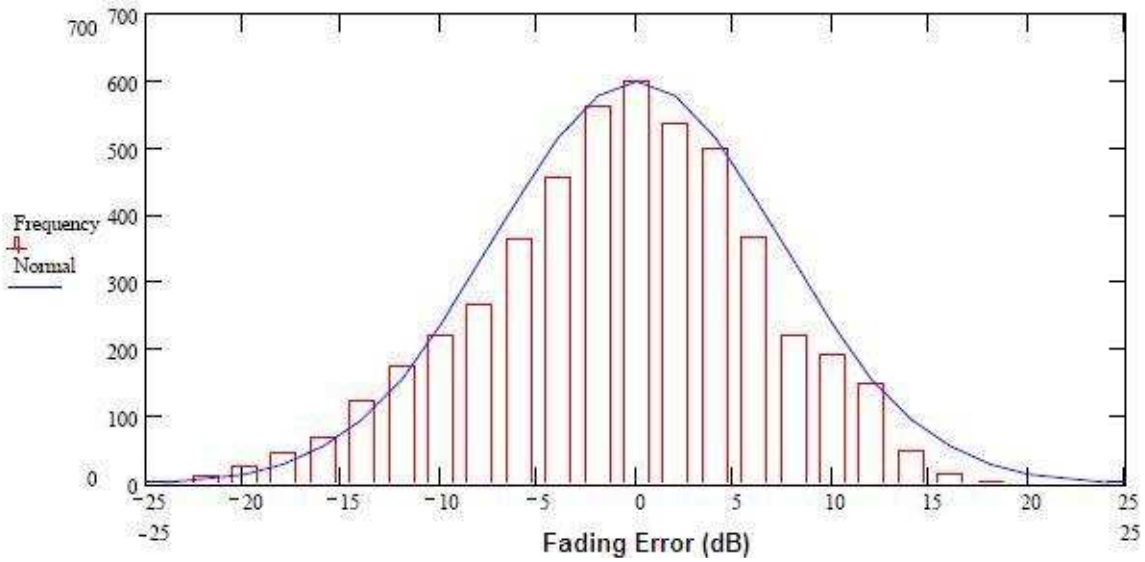


Figure 8.18: Distribution of measured fading error compared with normal distribution. [32]

It can be seen from Fig. 8.18 that the probability distribution of the measured fading error, for an urban area in the UK, closely matches a Gaussian distribution. In a Gaussian probability distribution, 95 % of the solutions lie within ± 2 standard deviations (σ) of the mean. In Fig. 8.18, approximately 95 % of the solutions lie within ± 20 dB of the mean (0 dB). The standard deviation of the data set of Figs. 8.17 and 8.18 is therefore close to 10 dB. In general the standard deviation of the fading error in urban/sub-urban environments is between 5 and 10 dB [77] [76] [78]. To summarize, this means that in a fairly typical urban/sub-urban area, moving a short distance ($\sim$ 10 m) can result in signal strength variations of between 20 and 40 dB.

8.3.2 Localised Urban Propagation Effects

In urban areas specific structures and objects can have a dramatic effect on the propagation of electromagnetic waves. A particularly good example of this can be found in [33]. The view from the receiver and the geometry of this example can be seen in Figs. 8.19 and 8.20.

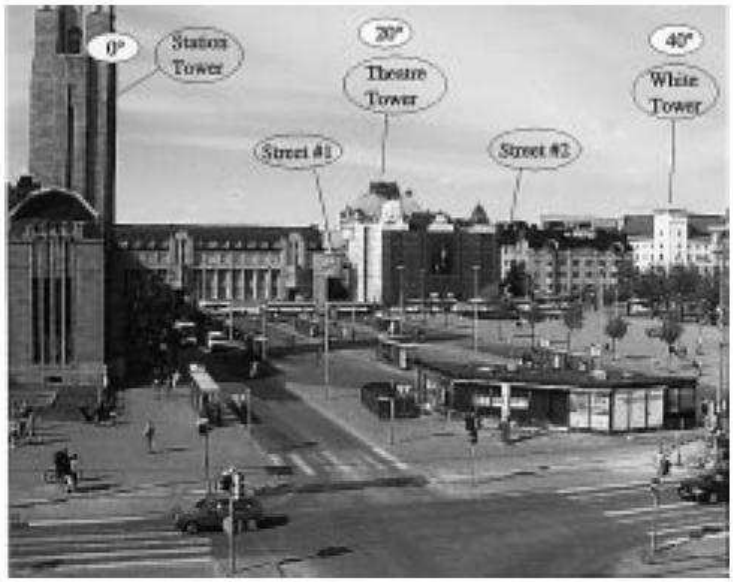


Figure 8.19: View from the receiver . [33]

In this investigation, a physical, planar array of elements was combined with appropriate post-processing software to produce a synthetic array. This array was capable of determining the direction of arrival and magnitude of electromagnetic waves. The measurements were performed in the centre of Helsinki.The transmitter was at a height of 1.5m and was omnidirectional in the azimuthal plane. Its elevation half power beamwidth was 87°. The receiver was on the third floor (10m) of a building.

In this example, the dominant propagation path is via the two street canyons. Tall buildings either side of a street can act as a lossy waveguide [33]. The waves propagate down the canyon by reflecting off the building walls and the street. As long as the wavelength is not too small, the surfaces of the walls and street will appear fairly smooth and will thus be reasonably good reflecting surfaces.

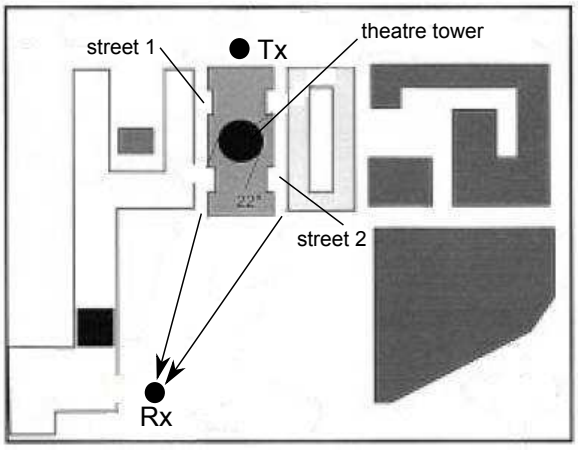


Figure 8.20: Top view of measurement site location. [33]

As can be seen from Figs. 8.21 and 8.22, the signals propagating down the two street canyons (streets #1 & #2) are of significantly higher magnitude than that of signals received at other orientations. There is a park just north of the transmitter, which reflects waves back down the two streets, which exaggerates the spreading in both elevation and delay for a given azimuth. If the two streets did not exist then the received signal would be extremely small due to the shadowing effect of the theatre building. A less dramatic, but still important, propagation path is that due to the dome on the theatre tower. It is likely that waves diffract around this structure to some degree.

This example has shown the dramatic effect that individual propagation paths can have on the magnitude of the received signal at different orientations.

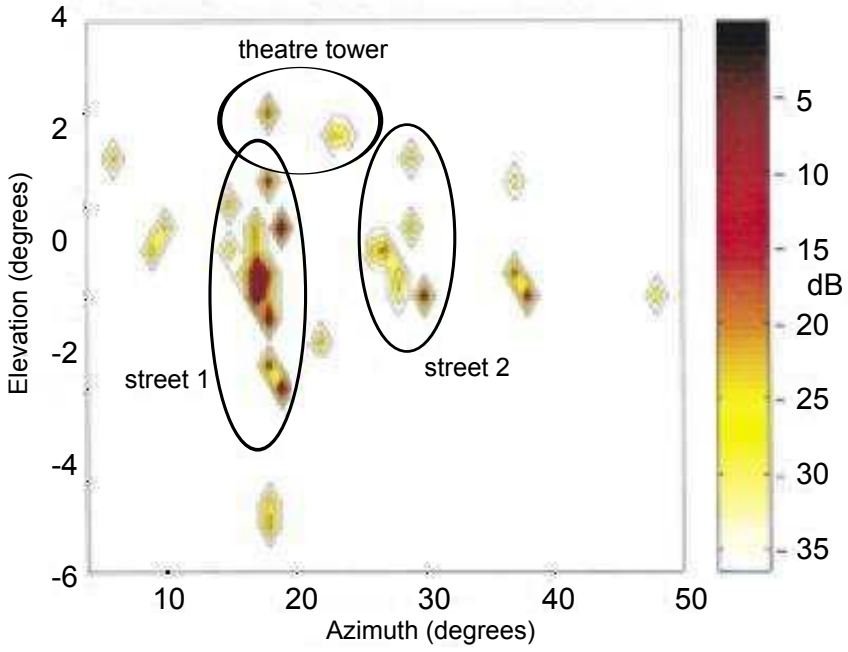


Figure 8.21: Azimuth-elevation plane of received signal. [33]

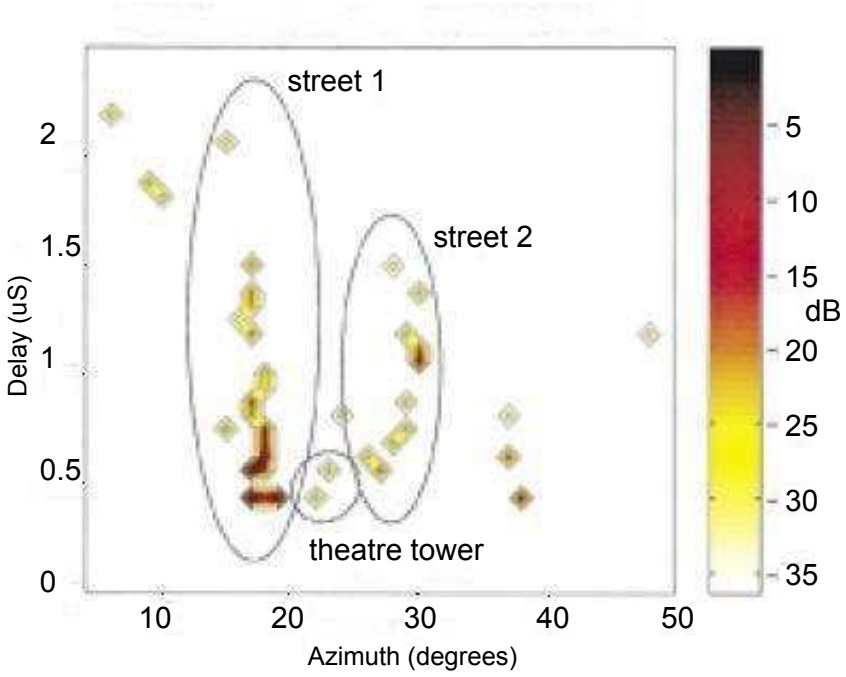


Figure 8.22: Azimuth-delay plane of received signal. [33]

8.3.3 Sources of Depolarisation

In section .8.3.2 the dramatic effect that individual structures can have on the magnitude of the received signal was demonstrated. In a related investigation [79], the effect that individual structures can have on the polarisation of the received signal was investigated.

In this investigation, a spherical array of dual-polarised elements was used in order to determine the direction of arrival of the signal and its polarisation. The horizontal and vertical components of the received signal were able to be measured separately. The transmitted signal was vertically polarised. Several different locations for both the transmitter and receiver were used. In order to express the polarisation state of the received signal, the cross-polarisation ratio (XPR) was used:

$$XPR_{dB} = 20 \log \left(\frac{|\boldsymbol{E}_{total}^{vertical}|}{|\boldsymbol{E}_{total}^{horizontal}|} \right) \tag{8.34}$$

As the receiver could determine the direction of arrival and the polarisation state of the received signal, it could therefore determine the depolarising effect that individual structures had on the signal. The effect that different types of structure have on depolarising the signal can be seen in table 8.1. It can be seen that individual structures can have a significant depolarising effect. It has also been shown that certain structures, such as the dome of a tower, can result in the received signal having a greater cross-polar component than its co-polar component.

Scattering Source	XPR (dB)
Corner of building	7.1 to 16.1
Wall of building	10.8 to 15.8
Roof of building	7.9 to 9.9
Tower dome	-3.3

Table 8.1: Depolarising effect of various urban structures.

8.3.4 Conclusion of Depolarisation Analysis

It is highly likely that in a typical urban environment there will be several depolarising structures, of the type mentioned in table 8.1, affecting the signal. Furthermore, in addition to the relatively large structures that were looked at in the investigation of section 8.3.3, there are usually many much smaller depolarising objects in the signal path. These include: lamp posts, vehicles, cables, window/door frames, pipes etc. The cumulative effect of all of these objects in depolarising the signal is considerable. It is therefore reasonable to assume that the received signal will most likely have an XPR in the region of -20 dB to 20 dB. As the receiver moves, the XPR will certainly vary considerably. For example, when using a mobile phone, moving just a few metres will often result in a sizable change in the polarisation state of the signal.

It is consequently impossible to predict the polarisation state of the signal received by a mobile communications device when it is being used in a urban environment. Due to reciprocity, it is likewise impossible to predict the polarisation state of the signal received by a base station, which was transmitted by a mobile communications device, in an urban environment. More specifically, 'impossible to predict' in this context means that the XPR of the signal is effectively random and can be anywhere between -20 dB and 20 dB.

In section 8.3.1 it was noted that the path loss for a constant distance (d) from a transmitter is likely to be somewhere between -20 dB and 20 dB of the average path loss at that distance. This means that, if at a particular location the path loss was -20 dB from the average and the signal's XPR was also -20 dB, the co-polar component of the received signal would be 40 dB below the average signal strength for that distance. This situation undoubtedly occurs relatively often in urban environments. However, the performance of mobile communications systems in urban environments, i.e. the signal strength, is almost invariably sufficient. This is due mainly to the fact that the handset and base station receivers are sufficiently sensitive and that the size of the macro/micro cells is sufficiently small.

This phenomenon of the XPR of the signal being anywhere between -20 dB and 20 dB, can actually be regarded as a good thing because it means that mobile communications devices do not need to be orientated to the same direction as the transmitted signal. For instance, mobile phones do not have to be held vertically in order to be used. Rather, they can be held in any way and will still operate perfectly normally. This also means that mobile communications device designers effectively do not have to consider the polarisation of the antenna within the device. Similarly, when computationally optimising antennas for mobile communications applications, polarisation does not have to be considered.

8.4 Grid Resolution

The resolution of the grid refers to how many rows and columns it is divided up into. For a given grid dimension (d), which has been determined already with regard to bandwidth and directionality considerations, the number of rows and columns also determines the grid cell size.

The cell size should not be too small for two main reasons. The first is due to the fact that the antennas will be made using standard PCB printing methods. These methods can print accurately to a certain margin of error. For example, in the Department of Electronics at the University of York, structures can be printed to 0.1mm. In other words, there is a $\pm$ 0.05mm margin of error. If the grid cell size is too small then the printing process will be unable to produce the final design accurately.

The second reason is due to the fact that the software used to fitness test the antennas will need to discretise the antenna models in order to simulate them. The solver software will need to use at least one of its own cells for each grid cell. If the grid cells are excessively small, then the solver will need to use an extremely large amount of its own cells when simulating the antennas. This will lead to the simulation time (fitness testing) of each antenna being excessively long and thus the the whole optimisation process would be impractical.

In addition to the grid resolution not being too fine, it should also not be too coarse. If the grid is divided into too few cells then the optimisation will not have adequate flexibility when it comes to creating potential solutions. The resolution of the grid should be sufficiently fine as to allow potentially beneficial structures such as meandered current paths and parasitic patches to be created.

8.5 Conclusion

The various analyses described in this chapter have shown that the bandwidth of an MSA depends on its size. In order for a computationally optimised grid based MSA to have a high probability of achieving a given bandwidth, the size of its grid must be sufficiently large. Similarly, the directionality of grid based MSAs, like all antennas, is closely linked to the antenna's size. As long as the antenna's grid is sufficiently small, then the antenna will have adequately wide coverage, which is required in mobile communications applications. Consequently, there is a trade off between bandwidth requirements and guaranteeing a wide coverage. Fortunately, the typical fractional bandwidths required by mobile communications applications (approximately 3 to 6 %) can be acheived by sufficiently small MSA sizes that guarantee wide coverage.

It has also been shown that the polarisation of signals used by mobile communications applications in urban/suburban environments is random. This phenomenon does not affect the performance of such systems.

The overall conclusion of this chapter is that only bandwidth needs to be optimised for when optimising grid based MSAs for mobile communications applications. This results in the optimisations being single-objective, rather than multi-objective, and thus conceptually simpler and easier to implement.

Chapter 9

Optimisation of Fitness Evaluation

Fitness testing is an essential part of all computational optimisation techniques. The optimisation must know how good, i.e., how close to the required specification, the antennas in the current population are. This is paramount because it enables the optimisation to reliably choose the appropriate individuals with which to form the next generation from the current generation.

The process of fitness testing used by the computational optimisation techniques of this study involved simulating the antennas using the Finite Difference Time Domain (FDTD) method. FDTD is described in more detail in chapter 10.

Fitness testing must be both accurate and efficient. If it is not accurate then there will be no point in performing the optimisation in the first place. This is because it will only result in the creation of an individual that is nothing like what it is believed to be, and therefore, more than likely does not meet the required specification.

If fitness testing is accurate but not efficient then the optimisation will take so long to run that it will be practically useless. In general, the more accurate the computational modeling of an electromagnetic structure (e.g. an antenna) is, then the longer the simulation will take to run. It is therefore of paramount importance to find the optimum trade off point between accuracy and efficiency.

It can be seen from Fig. 9.1 that the simulation run time increases linearly with the number of cells in the FDTD problem space and the number of time steps. An example of an FDTD simulation parameter having a marked impact on the accuracy and run time of the simulation is that of the gap between the object being simulated and the edge of the problem space. This gap is referred to in this thesis as the *boundary*. The size of the boundary can significantly affect the accuracy of the simulation. However, increasing it equally on all sides, i.e. in all of the three dimensions, will result in the number of cells in the problem space increasing cubicly. Thus a small increase in the boundary can greatly increase the run time of the simulation. This is why the optimum trade off point between accuracy and efficiency has to be found.

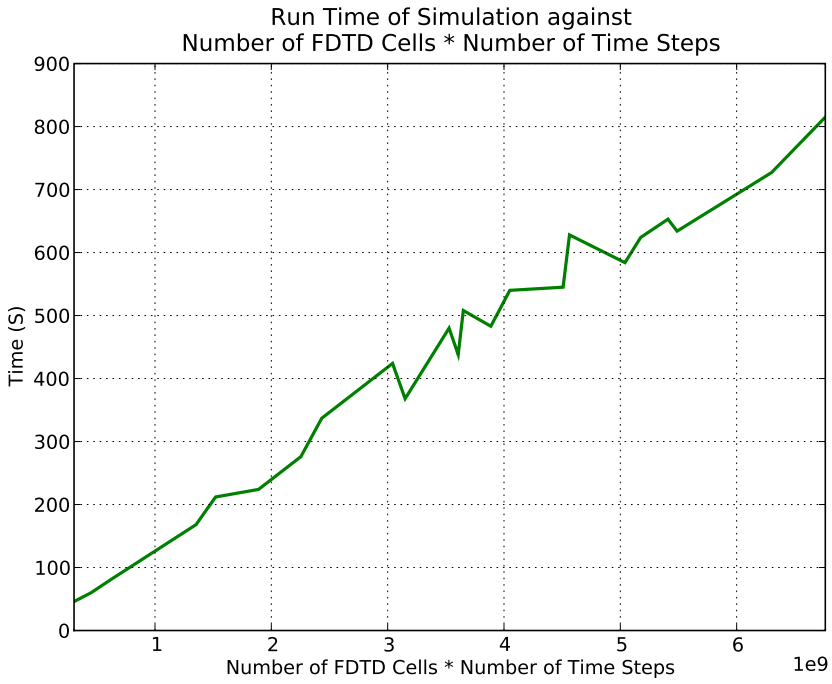


Figure 9.1: FDTD simulation run time against number of FDTD cells * number of time steps (simulations were performed on a windows desktop PC).

The analysis of fitness testing for this study can be divided into two main parts. Firstly there is the investigation into individual FDTD modeling variables that affect the accuracy and efficiency of the model. These include phenomena such as the size of the boundary (in FDTD cells) around the object being modeled in the FDTD problem space.

Secondly, there is the modeling of complete structures including those based on microstrip transmission lines and RMSAs.

9.1 Analysis of FDTD Modeling Parameters

9.1.1 Microstrip Transmission Lines

Microstrip transmission lines are a good place to start when looking at how individual parameters can affect the performance of an FDTD model. This is because, they are relatively simple structures whose theory of operation and characteristics are well documented. As the name suggests, there is much in common with their operation and that of microstrip antennas. A model of a microstrip transmission line with a resistive load can be fairly simply created in Falcon, which was the FDTD solver that was used in this study. Falcon is described in more detail in chapter 10. Equations for the characteristic impedance of a microstrip transmission line as a function of its geometry and permittivity of the substrate are widely available in the literature. The resistive load was made up of a block of material of a given size and conductance that resulted in the desired resistance.

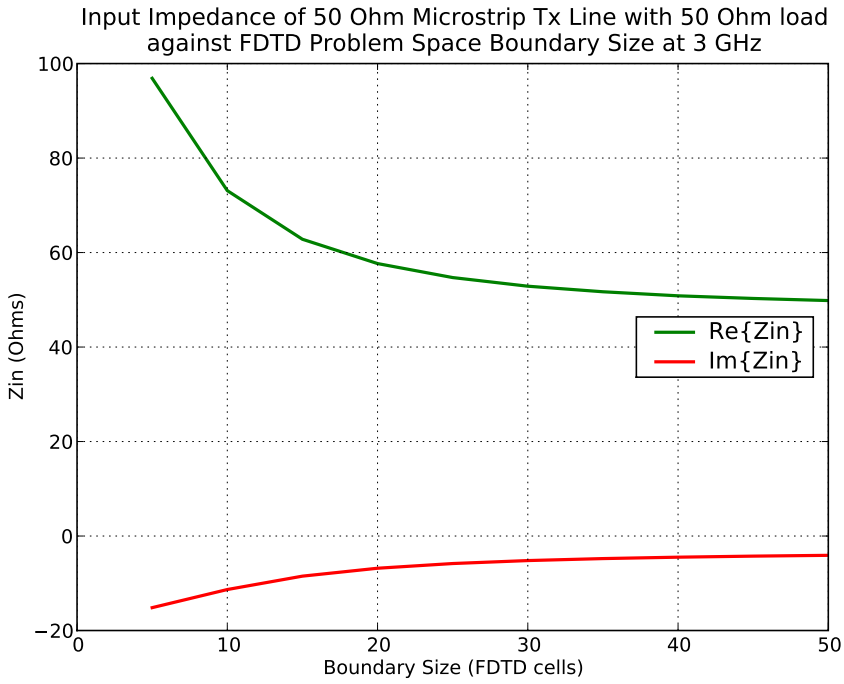


Figure 9.2: Input impedance of 50 Ohm transmission line with matched load against boundary size.

The first modeling parameter to be investigated was that of the size of the boundary (in FDTD cells) between the model and the edge of the problem space. As can be seen in Fig. 9.2, the input impedance of the microstrip transmission line with an matched resistive load converges asymptotically on the most accurate value. For the model used to generate Fig. 9.2, the optimum trade off point between accuracy and efficiency would be a boundary size of around 30 FDTD cells.

The second modeling parameter to be investigated was that of the number of time steps to be used in the FDTD simulation. The same model that was used to investigate the effect of the boundary size was used to look at that of the number of time steps. The results can be seen in Fig. 9.3. The optimum trade off point between accuracy and efficiency would be a number of time steps around 1500.

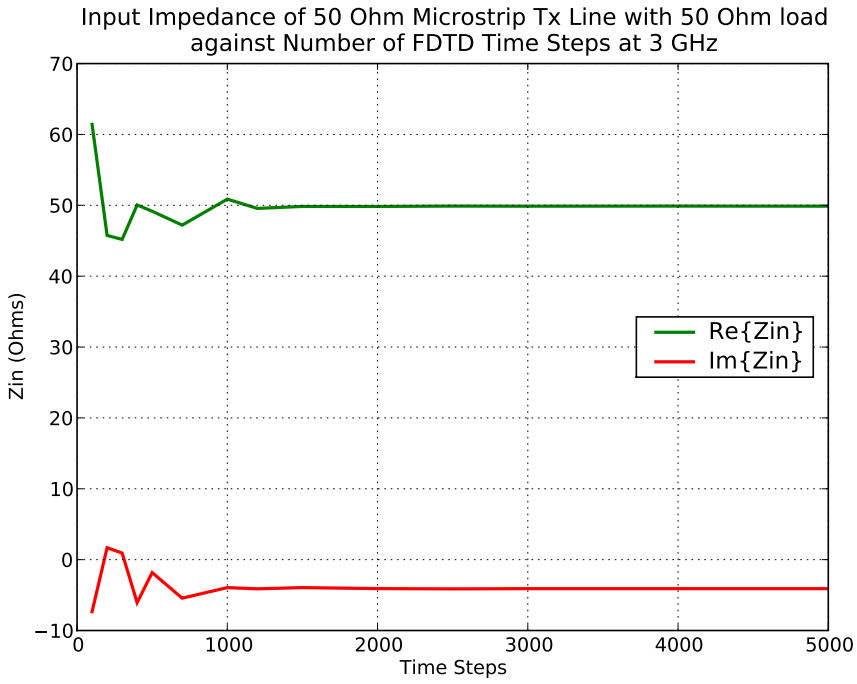


Figure 9.3: Input impedance of 50 Ohm transmission line with matched load against time steps.

9.1.2 RMSAs

RMSAs were considered to be the next logical step, after microstrip transmission lines, for investigating FDTD modeling parameters. As with microstrip transmission lines, there is much in the literature regarding the design and operation of RMSAs. RMSAs with 50 Ohm input impedance, with both probe and inset feeds, were designed (using conventional methods) and built on 1.6mm thick FR4 circuit board. FDTD models of the same antennas were also created.

The effect of the size of the FDTD boundary was investigated and in Fig. 9.4 the input impedance of a probe fed 3GHz RMSA against boundary size can be seen. It can be seen that the model continually gets more accurate as the boundary increases rather than converging asymptotically. The S11 (return loss) of the same model can be seen in Fig. 9.5. This is the minimum S11 of the antenna, i.e., the S11 at the bottom of the antenna's null. In general, the depth of an antenna's S11 null is not nearly as important at the width of the null. For this reason, there is no need to have a particularly large boundary size, but rather one that will result in a reasonably accurate S11 whilst still being relatively efficient.

The effect of the number of time steps on the accuracy of the RMSA model can be seen in Fig. 9.6. In this figure, it can be seen that the reactive (imaginary) part of the input impedance settles close to its final accurate value much sooner than the real part, which, by contrast, converges more asymptotically. The model used to generate Fig. 9.6 is slightly different from the model used to generate previous figures in this section in that the radius of the feed wire (probe) was different.

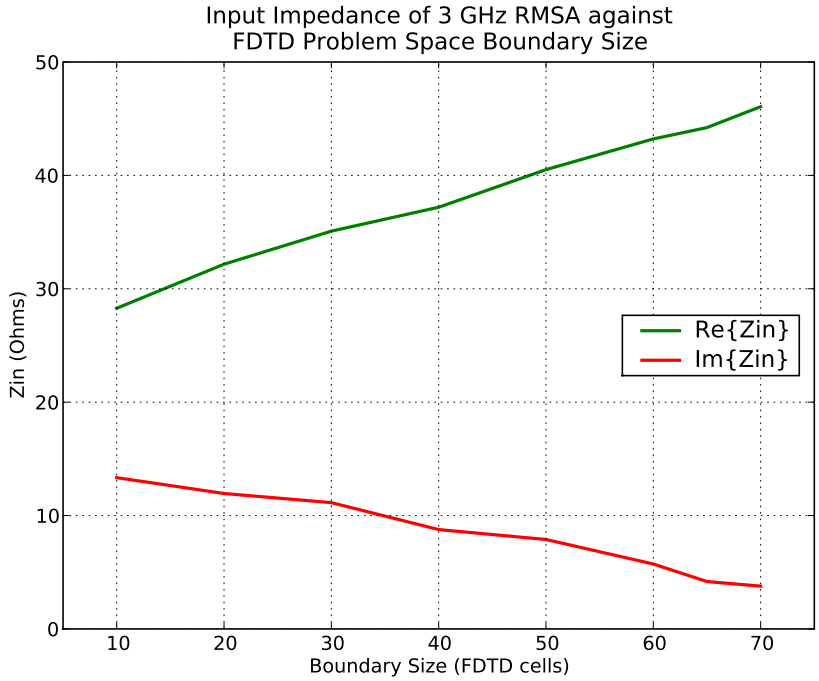


Figure 9.4: Input impedance of 50 Ohm probe fed RMSA against boundary size.

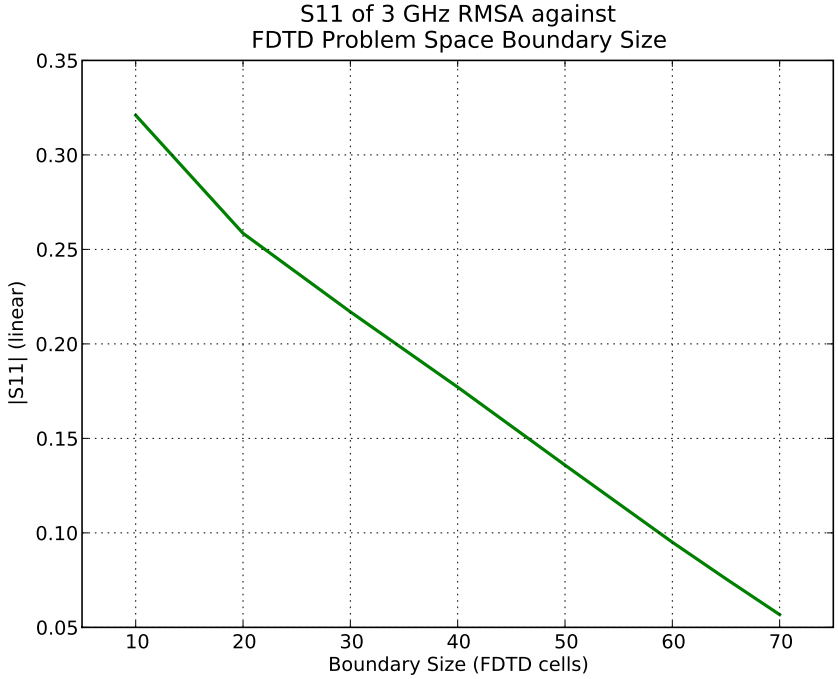


Figure 9.5: S11 of 50 Ohm probe fed RMSA against boundary size.

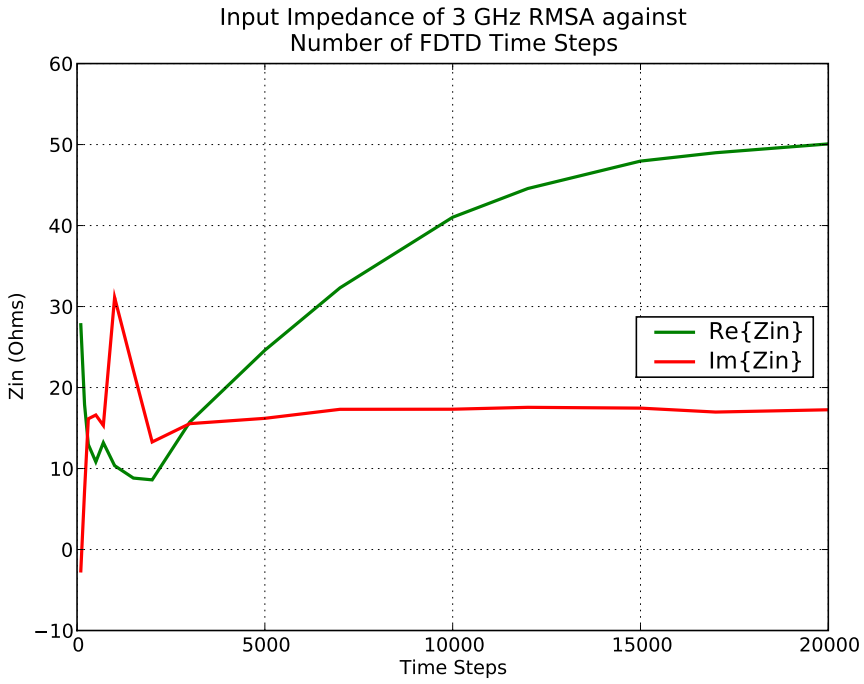


Figure 9.6: Input impedance of 50 Ohm probe fed RMSA against time steps.

The radius of the feed probe was another parameter that was investigated. Changing this did not affect the length of time that the model took to run (efficiency) but it did affect the accuracy of the model. All of the probe fed antenna models that were built and tested in this study had a 1mm diameter feed probe. The radius of the feed wire of the FDTD model was varied until it gave the best match with the built and tested real antennas. The most accurate feed wire radius of the FDTD model was found to be 0.4mm. Changing the feed wire radius changes the inductance of the feed probe. This affect was accurately observed in the FDTD models of RMSAs. The model used to generate Fig. 9.6 has a smaller radius feed wire than the model used to generate Fig. 9.4. This smaller feed wire radius results in a greater probe inductance and thus a greater positive magnitude of for the reactive part of it input impedance.

9.2 Complete Structure FDTD Modeling

In this section, some results from the modeling of complete structures are shown. These simulated results were compared against theoretical or measured results to ascertain their accuracy.

9.2.1 Microstrip Transmission Line Base Structures

A structure that can be easily modeled and compared with theory is that of a microstrip transmission line with an unmatched load. The results of such a model can be seen in Fig. 9.7.

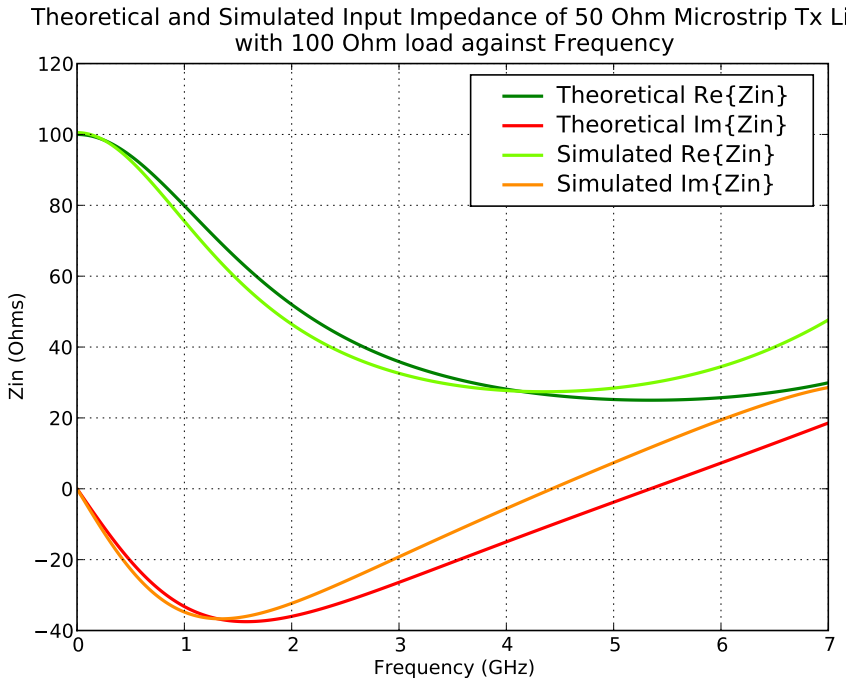


Figure 9.7: Input impedance of 50 Ohm transmission line with 100 Ohm load.

As can be seen in Fig. 9.7, the simulation closely matches theory. Indeed, as with the simulation, a real microstrip transmission line and load will not have an exact theoretical match. Other, more complex, microstrip transmission line based structures were also simulated. These were a quarter wave transformer and several single stub matching structures. These models used values for the boundary size and number of time steps determined from the analysis described in section 9.1. All of these models yielded results that matched closely with theory. The S11 of a microstrip transmission line with a single matching stub, designed to match at 3GHz, can be seen in Fig. 9.8. It can be seen in this figure that the stub provides a good match at the desired operating frequency.

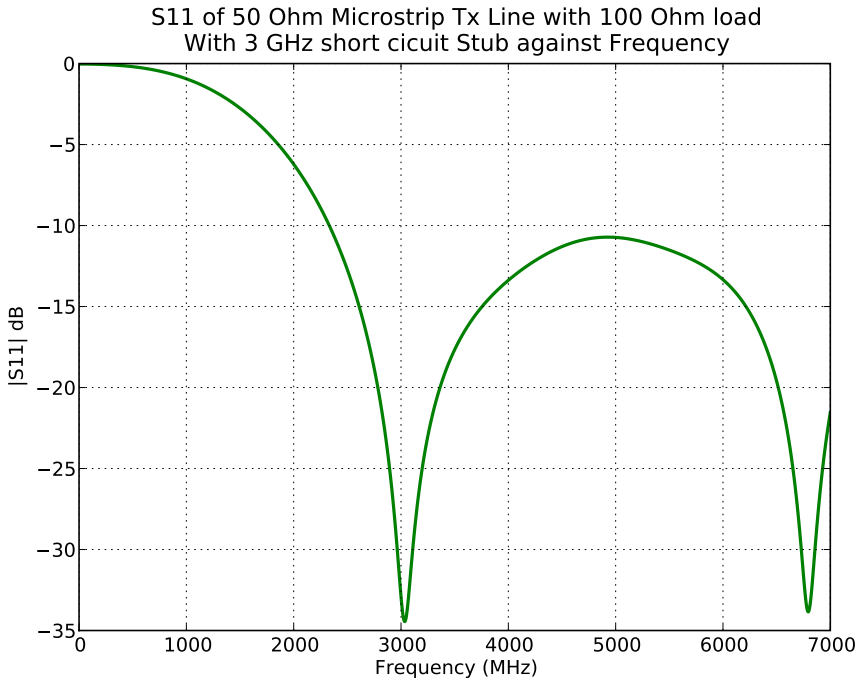


Figure 9.8: Input impedance of 50 Ohm transmission line matching stub and 100 Ohm load.

9.2.2 RMSAs

Several RMSAs, designed to operate at different frequencies, and with different feed types were built and had their return losses measured. Computational FDTD models were also made of the same antennas, and the measured and simulated results were compared. As can be seen in Figs. 9.9 and 9.10, the simulated results closely match the measured results. In Fig. 9.9, the depth of the simulated null is much deeper than that of the measured. This, however, is not particularly important. The width of the null at lower S11 levels, such as -10 dB, is typically of much more relevance to antenna engineers because this determines the bandwidth of the antenna. In Fig. 9.9 it can be seen that the width of the simulated and measured nulls are very close at -10 dB and so this shows that the computational model is accurate.

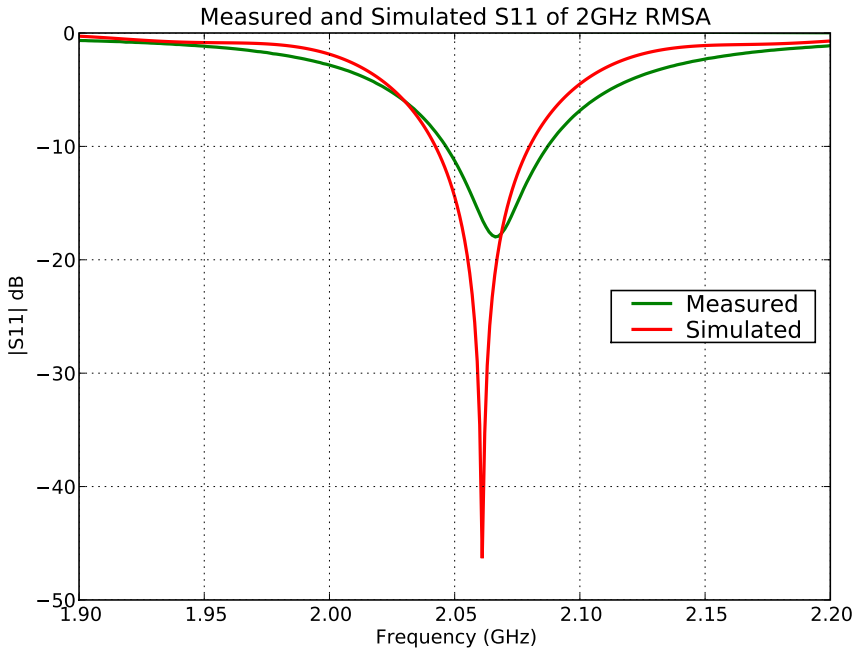


Figure 9.9: S11 of 50 Ohm 2GHz probe fed RMSA.

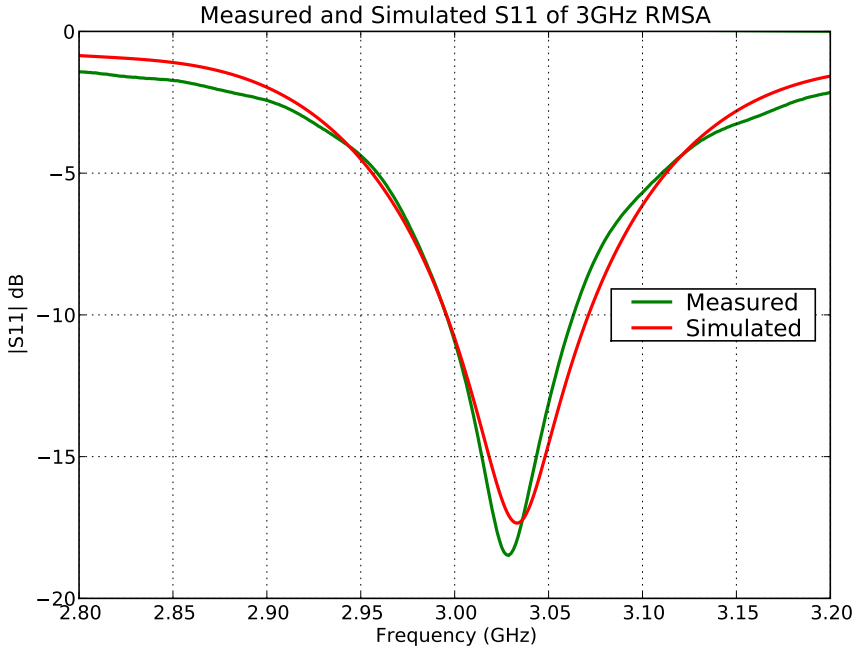


Figure 9.10: S11 of 50 Ohm 3GHz probe fed RMSA.

9.2.3 Random Grid MSAs

The antennas that were produced by the computational optimisation techniques used in the empirical study, as described in this thesis, are grid based MSAs, as described in chapter 7. These antennas will naturally be fairly random in appearance, i.e., there will probably not be any obvious well-defined uniform structures or obvious symmetry. Consequently, in order to be confident of the accuracy of the modeling of such antennas, several FDTD models of random grid based antennas were created and simulated. Some of the models with 'interesting' S11 plots, i.e., ones with several nulls, were then built and measured and so that a comparison could be made.

Figure 9.11: Random grid based MSA.

The S11 plot of the antenna shown in Fig. 9.11 can be seen in Fig. 9.12. The S11 plot of another random grid MSA. whose geometry is not shown, can be seen in Fig. 9.13. The FDTD models used to simulate these antennas used values for the boundary size and number of time steps ascertained from the analysis that is documented in section 9.1. The S11 plots of these random grid MSAs show that the modeling is sufficiently accurate whilst at the same time also being efficient as well.

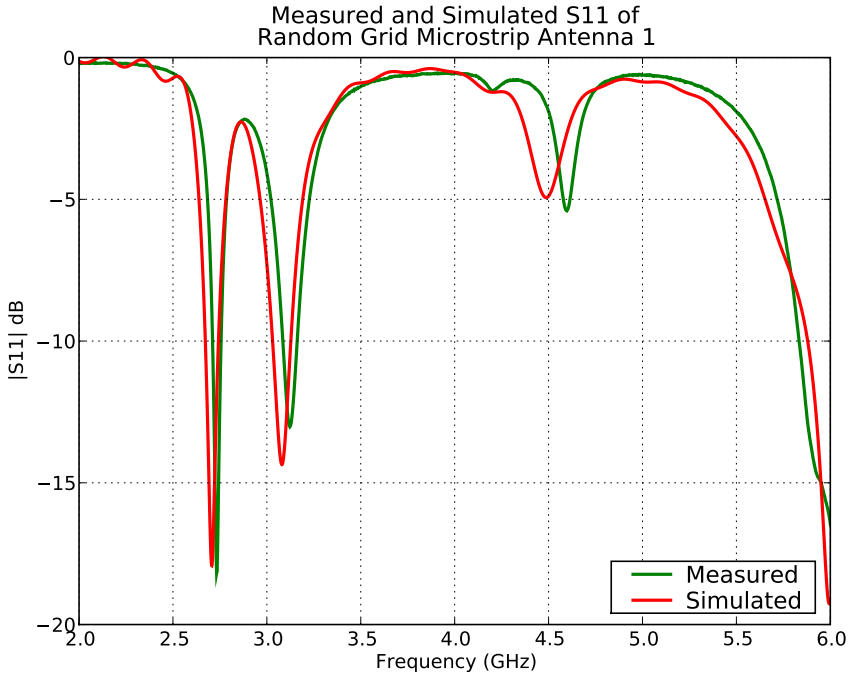


Figure 9.12: S11 of Random grid based MSA (Fig. 9.11).

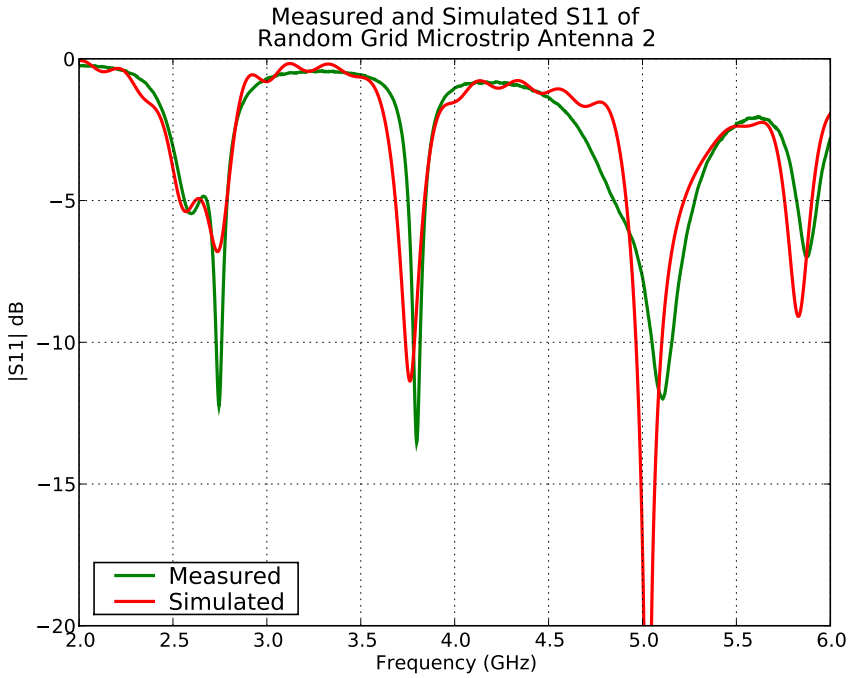


Figure 9.13: S11 of another Random grid based MSA.

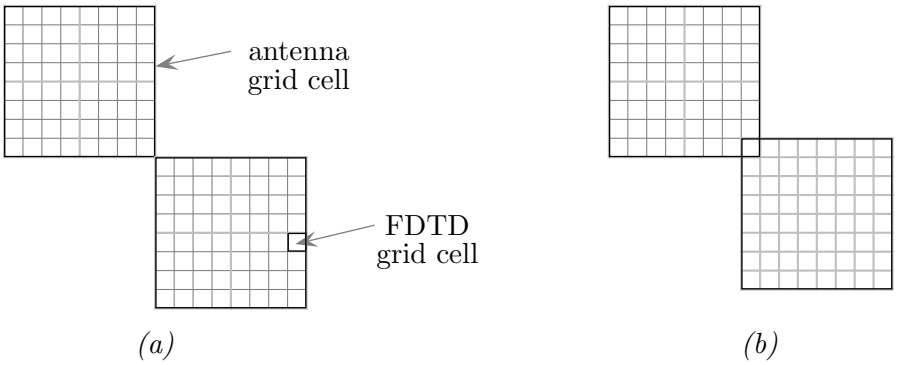


Figure 9.14: Illustration of antenna grid cell overlap: a) no overlap and b) 1 FDTD cell overlap.

Another issue that is of paramount importance to the modeling accuracy of grid based MSAs is that of the grid cell resolution. Each cell in the antenna's grid is actually made up of several FDTD cells. As part of the investigation into random grid MSA modeling accuracy, the number of FDTD cells that make up each antenna grid cell was varied. The size of each grid cell remained the same so when the number of FDTD cells per grid cell was reduced, the FDTD mesh size was increased to compensate. It was found that the best results were achieved by having each antenna grid cell consist of 8 FDTD cells.

A key issue related to grid cell resolution is grid cell overlap. This refers to the extent to which, i.e., how many FDTD cells, diagonally adjacent antenna grid cells should overlap each other. Grid cell overlap is illustrated in Fig. 9.14. Some degree of overlap is required between diagonally adjacent grid cells in order to guarantee a good (low impedance) electrical connection between them. An overlap of one FDTD cell yielded good results and so was adopted for the study.

9.3 Conclusion

The extensive analysis of FDTD modeling, as described in this chapter, resulted in the modeling to be used by the computational optimisation techniques of the study, being both accurate and efficient. Consequently, there was a high degree of confidence that the antennas produced during the course of the study would perform as expected when built and measured.

Chapter 10

Empirical Study

10.1 Justification of the Empirical Study

Computational optimisation techniques (GAs, GP etc.) are, by their very nature, stochastic processes. In other words, a significant degree of randomness is central to their operation. For this reason, the mathematical analysis of computational optimisation techniques is probabilistic in nature. This is not a problem in itself, however, the analysis of these techniques is severely limited for other reasons. The main problem is that these analyses yield very few practically useful conclusions. For example, schema theory and its related phenomena (implicit parallelism etc.), as described in section 3.5.6, provide an overall justification as to why GAs are an efficient search technique. However, they do not provide any particularly useful findings as far as implementing GAs is concerned. For instance, they do not give any indication as to how to determine the optimal values for the GA's control parameters (mutation rate, population size etc.). Even later, more mathematically rigorous, analyses of GAs, such as [80], do not provide much, if any, practically useful conclusions.

The only truly valid means of comparing the efficiency of different computational optimisation techniques and their control parameters, is to compare them empirically. This approach has been used several times by different researchers. A particularly notable example is that of Julian Miller [81]. In this study, a large number of independent runs of CGP were performed, with the purpose of trying to evolve certain logic functions (e.g. 2 bit multiplier). The population size was systematically varied in order to ascertain its effect on the efficiency of CGP. As well as showing the effect of the population size, it also convincingly showed that CGP outperforms GP when trying to evolve certain logic functions. This study is a clear example that when attempting to investigate the efficiency of various computational optimisation techniques, empirical comparison is both effective and is the only viable option. The main disadvantage of an empirical comparison is the large amount of computational time and/or power that is required.

It is therefore clear that an empirical study presents the only viable option for comparing the efficiencies of different computational techniques for the optimisation of microstrip antennas. Such a study was indeed performed and it forms the basis of this PhD investigation.

10.2 Computational Techniques Used

As described in chapter 3, there are several distinctly different computational optimisation techniques. However, as previously mentioned in section 7.5.1, some techniques are not suitable for optimising Boolean problems. Gradient methods and particle swarm optimisation can only optimise problems that have continuous search variables. They therefore can not be applied to problems that have Boolean variables. Simulated annealing has been shown to be significantly less efficient when compared to GAs [49]. For this reason, it was not chosen to be used in this study.

GAs, GP and CGP were the techniques that were used in this study. GAs have been previously used many times in many different types of optimisation. They have a proven record of being a robust and efficient search technique. GP is based on GA but has a distinct difference. This is that instructions are used rather than direct variables. GP has been less well applied than GA. Two main reasons for this are likely to be that GP is both newer and less well known than GA. CGP is GP based but has a distinctive difference. This is in the way that it uses redundancy. The only significant application of CGP has been in the evolution of combinational logic circuits, where it has proven to be extremely effective.

It was therefore concluded that comparing the efficiencies of these three techniques in evolving Boolean grid based MSAs would be a useful and intriguing endeavor.

10.3 Optimisation Specification

In order to reliably compare the various optimisation techniques used in this study, a predetermined optimisation specification was required. It was decided that this specification should reflect the requirements of MSAs that are used in current and emerging mobile communications applications. As far as bandwidth was concerned, the MSA's target was to have two bands centered at 3.5 and 4 GHz, with each band having a -10 dB, or less, return loss bandwidth of 100 MHz. A dual band antenna was deemed appropriate because modern mobile communications devices typically utilise more than one band. 3.5 GHz was chosen because it is already used for some WiMAX (IEEE 802.16) applications. 4 GHz is not currently used, but is sufficiently far from 3.5 GHz and most importantly it is not harmonically related to 3.5 GHz. The actual frequencies themselves are not of major importance as the purpose of the study was to compare the efficiencies of different techniques in a general context. Bandwidths of 100 MHz (-10 dB return loss) are typical of those required by modern devices.

The target specification only involved bandwidth because of reasons described in chapter 8. In short, the reason why directionality is not involved is that, as long as the MSA is sufficiently small then it is guaranteed to have sufficient coverage (i.e., 3 dB beam width $\geq 60^{\circ}$). The polarisation of the antenna is not evolved for because in an urban/sub-urban environment the polarisation of the signals traveling between a mobile device and various base stations is linear but almost universally random in orientation.

Indeed, it does not matter what the polarisation of the resulting antennas is because they will certainly be able to communicate with other linearly polarised antennas.

To conclude, the antennas in this study were only assessed in terms of bandwidth. The optimisations were therefore single-objective. The directional properties and polarisation of the MSAs were not considered because this was unnecessary.

10.4 Practical Considerations

10.4.1 Electromagnetic Simulation Software

A crucial process in all computational optimisation techniques is fitness testing. Fitness testing must be both accurate and efficient in order for the technique to be viable. In this study, fitness testing involves simulating the various antenna models that are created and assessing their performance. As mentioned previously in section 10.3, the antennas in this study were only assessed in terms of bandwidth.

The fitness testing software that was chosen for this study is called Falcon. Falcon is a 3D Finite Difference Time Domain (FDTD) solver, which was written by Dr Stuart Porter at the University of York. Falcon was chosen because it was the only solver available at the time which could both determine a grid based MSA's bandwidth and was easily automatable. Every time an antenna needs to have its fitness evaluated, the optimisation algorithm runs the simulation software. Thus, the simulation software needs to be able to be reliably integrated into the optimisation algorithm as a whole.

FDTD was first proposed by K. S. Yee in 1966 [82] and it has become a widely used and documented technique. FDTD uses Maxwell's two curl equations and a discretisation of space and time. This results in two grids: electric field and magnetic field. The two grids are offset from each other by both half a grid cell spatially and by half a time step temporally.

For a given time step, the electric and magnetic fields are calculated at each point in turn, using iterative equations that involve various field points around that point and some field values from previous time steps. When all of the field values have been evaluated for the current time step, the process is repeated for the next time step.

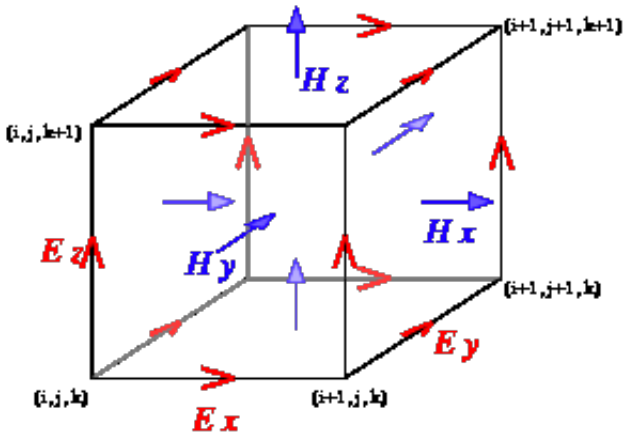


Figure 10.1: Yee cell. [34]

Fig. 10.1 shows the standard Yee cell. The electric field points, in red, are located at the edges of the cell. The magnetic field points, in blue, are located in the centre of the faces of the cell. This is how the spatial offset is achieved. For 3D, the iterative update equations are fairly lengthy, and can be obtained from many sources, and so are not included here.

As bandwidth depends on input impedance, FDTD needs to be able to determine the input impedance of antennas. It does this by using Ampere's law to determine the current flowing from the source point. By approximately integrating (i.e. summing) the magnetic field in a loop around the source point, the input current can be found. As the input voltage is set beforehand, the input impedance is simply the ratio of input voltage to input current. The input impedance can be recorded for each time step. After the FDTD simulation has finished, a Fourier transform can be used to generate the input impedance as a function of frequency. The return loss (S_{11}) as a function of frequency, and thus the bandwidth, can be easily determined from this.

10.4.2 Substrate Parameters

The purpose of this study is to compare the efficiencies of various computational optimisation techniques. The creation of practical antennas for particular applications is not its concern. As such it was decided that the substrate that the antennas would use would be FR4. FR4 is a general purpose PCB material. It is not specifically tailored to carry high frequency signals. It is generally considered to have satisfactory performance when carrying signals of up to 1 GHz. It has become the standard PCB material in the electronics industry primarily because it is low cost and easy to work with, i.e., its etching process in reliable and straightforward. For these reasons, FR4 glass reinforced plastic circuit board is widely used by the Electronics department at the University of York, and so it was readily available as a substrate for this study. In the wider electronics industry it is simply referred to as FR4. A substrate thickness (h) of 1.6mm was chosen because this thickness was readily available.

The main disadvantage of using FR4 at high frequencies, i.e., greater than 1 GHz, is that the relative permittivity (ϵ_r) is usually not constant. More specifically, ϵ_r changes (usually falls) as the frequency increases and also the ϵ_r of FR4 varies from piece to piece. Preliminary modeling showed that the ϵ_r of the FR4 used by the Electronics department at the University of York varied between 3.9 and 4.0 for the frequency range of 3 to 4 GHz. This variation is not particularly significant. It would only result in slight frequency shift in the return loss of the antennas. As this study is concerned with comparing optimisation efficiencies rather than creating practical antennas, this would not be a problem.

AS FR4 is a substrate which has not been specifically designed for high frequency applications, then potentially its loss (due to its conductivity) could have been a problem. However, preliminary modeling that loss of the FR4 to be used in this study was virtually negligible for the frequency range of 3 to 4 GHz. If a substrate has some loss, it is because the material's conductivity is not zero. When electric fields are in such materials, conduction current will be induced in the material which will be in phase with the electric field. As such, power will be dissipated in the material, in the form of heat. A substrate with zero conductivity will have no conduction current within it. Instead, the electric field will only induce what is known as *displacement current*. A commonly used measure of substrate lossiness is the loss tangent ($\tan \delta$). This is simply the ratio of conduction current to displacement current in the material, see equation 5.1.

10.4.3 Grid Size and Resolution

In order to enable the required bandwidth (100 MHz) of the two bands (3.5 and 4 GHz) to be physically possible to be evolved, the dimension of the grid (d) needed to be of an adequate size. In terms of fractional bandwidth, the lower band (3.5 GHz) has the highest bandwidth (2.8 %). Using this frequency, the Chu-Harrington limit, as described in section 8.1.1, was applied to help determine a judicious grid size. It was deemed that a grid size of approximately 2cm would be expedient. This grid size was on the edge of what was considered feasible according to the Chu-Harrington limit. It is also a convieniently small size that could comfortably fit into mobile communications devices. This grid size therefore presented the various optimisation techniques with a difficult but feasible task.

It was decided that the grid should be divided up into 12 rows and 12 columns. Each cell in the grid was therefore approximately 2mm * 2mm. This meant that the grid cells would be able to be printed accurately with the available PCB construction method, which has a resolution of $\pm$ 0.05 mm. It also meant that resolution of the grid would be sufficiently fine as to allow potentially beneficial structures such as meandered current paths and parasitic patches to be created.

10.4.4 Number of Runs and Evaluations per Run

The preliminary findings of grid based MSA modeling showed that a model with parameters as described above, i.e., a 2cm * 2cm (12 by 12) grid on 1.6mm FR4, could be simulated with sufficient accuracy in between 10 and 25 minutes. The simulation time predictably depended on the speed of the computer being used. The computers available included: desktop PCs with Intel Core2Duo processors. The speed of a given optimisation was therefore entirely dependent on the time per evaluation and the number of evaluations. An optimisation that involved 1000 evaluations would take between one and three weeks to complete. In order for the comparison of the different optimisation techniques to be valid, all of the optimisations were to have the same number of fitness evaluations. This number was set at 1000 because it was considered to be enough to give each technique a reasonable chance to meet the specification and also would not take a prohibitively long time.

In addition to ensuring a fair comparison was made between different techniques, several independent runs of each technique would be required. The more runs that would be performed, the more reliable the comparison would be. However, an inexhaustible amount of time was not available so it was decided that 10 runs of each technique would be performed.

10.5 Optimisation Technique Control Parameters

The purpose of this study is to compare the efficiencies of various computational optimisation techniques and the key variations within each technique, such as, the difference between generational and steady-state GAs. It was therefore necessary to determine the various control parameters that would produce optimal performance for each type of technique. In addition, the values of various control parameters would have to ensure that there were significant differences between the different types of technique.

As mentioned previously, one of the key aspects of a GA is whether it can be regarded as being generational or steady-state. There are not two distinct types, rather, there is what can be regarded as a continuous spectrum, ranging from distinctly generational to distinctly steady-state. One of the simplest GAs, which is also particularly generational in nature, is the $1 + \lambda$ algorithm, see also section 3.7.2. In this algorithm, the population size is $1 + \lambda$, where λ is an integer. A brief description of this algorithm follows.

A population of $1 + \lambda$ individuals (λ is an integer) are randomly created and all of them have their fitnesses evaluated. All of the population are discarded apart from the fittest individual (parent). λ exact copies (children) of the parent are made and are randomly mutated. The fitnesses of all of the children are then evaluated. The cycle is then repeated, i.e., all of the population except the fittest individual are discarded and λ children are made from it. As is common practice in GAs, the algorithm repeats the above cycle until a pre-determined fitness level has been reached or a pre-determined number of generations have been completed.

The $1 + \lambda$ algorithm is extremely generational because only one individual, the parent, survives unchanged from one generation to the next. A notable feature of it, is that it only uses mutation in order to create new individuals. The $1 + \lambda$ algorithm is simple to implement and it has yielded impressive results in the past [81] [83]. It is for these reasons that the $1 + \lambda$ algorithm was used as the generational GA type in this study. As only one individual survives from one generation to the next, the $1 + \lambda$ algorithm is most efficient with small population sizes. Having a large population results in a lot of effort being wasted in every generation. Julian Miller, who uses the $1 + \lambda$ algorithm in conjunction with CGP to evolve logic functions, has found that population sizes around 5 (i.e. $\lambda = 4$) yield optimum performance. In an investigation by Robert Woodhouse, the author of this thesis, a population size of 2 (i.e. $\lambda = 1$) gave optimal performance when evolving wire antennas. As such, it was decided that primarily a population size of 2 would be used for the techniques that employed the $1 + \lambda$ algorithm. In the event of extra time being available, different population sizes would be tried.

The steady-state GA to be used in this study, needed to be significantly different in nature than the 1 + λ algorithm. It was decided that it should use tournament selection to select parents, as well as both crossover and mutation in order to create the next generation. Crossover would be the dominant reproductive operator. The steady-state GA would also use a significant amount of elitism. Elitism has no cost and preserves a proportion of the fittest individuals from one generation to the next.

As both GP and CGP are based on GAs (the key difference is the form of the representation), the algorithms employed by the GP and CGP systems would be the same as those used by the GAs. In other words, the same generational and steady-state algorithms (e.g. 1 + λ) and their associated control parameters were to be use by the GP and CGP techniques. However, in the case of CGP, only a generational (1 + λ) algorithm was to be used. This was because, applying crossover to graphs which code for Boolean grid patterns, is particularly difficult to implement. However, this should not be a limitation because the the high efficiency achieved by Julian Miller whilst using CGP was obtained using the 1 + λ algorithm. Indeed, Miller attributes CGP's efficiency not to the algorithm that CGP is used with but rather, to CGP's use of graphs and their implicit redundancy.

The principal features of the optimisations systems that were used in this study are listed in table 10.2. The key control parameters of the two main system types (generational and steady-state) are shown in table 10.1.

10.5.1 Fitness Function

A vital consideration of any computational optimisation system is the fitness function. The fitness function needed to be the same for all of the techniques employed in the study, to ensure a fair comparison. The fitness function that was used is:

$$f = \frac{1}{N_f} \sum_{i=1}^{N_f} A_i \tag{10.1}$$

$$\text{where} \quad A_i = \begin{cases} 1 - |\rho_i|, & |\rho_i| > |\rho_{spec}| \\ 1, & |\rho_i| \leq |\rho_{spec}| \end{cases} \tag{10.2}$$

N_f = number of frequencies
$0 \leq |\rho_i| \leq 1$
$\rho_{spec} = 0.316$ (-10 dB)

$\implies \quad 0 \leq f \leq 1 \quad (f_{worst} = 0,\ f_{best} = 1)$

As the antennas in this study were to be only optimised for bandwidth, the fitness function required some measure of input matching of the antennas. The electromagnetic simulation software that was used (Falcon) provides the voltage reflection coefficient (ρ) for each pre-determined frequency. The input impedance bandwidth of the antenna can be eas-

ily determined form this. The above fitness function was used because it would be simple to implement and would give a clear indication of the quality of an antenna's performance as regards the specification. The specification was simply that the voltage reflection coefficient should be less than 0.316 (-10 dB) for each frequency of interest, i.e. for each frequency within the two required bands (3.5 & 4 GHz). A frequency step of 1 MHz was used for the two 100 MHz bands. This meant that 202 frequencies (3.45 GHz to 3.55 GHz and 3.95 GHz to 4.05 GHz) were evaluated during each simulation of an antenna. The fitness function results in the worst possible fitness of 0 and a best possible fitness of 1.

As far as mutation was concerned, it was decided to use an adaptive linear mutation function:

$$MR = MR_{start}\,(1 - f_{max}) \tag{10.3}$$

where:

MR = current mutation rate
MR_{start} = mutation rate at start of optimisation
f_{max} = current highest fitness achieved so far in the optimisation ($0 \leq f_{max} \leq 1$)

This mutation scheme was the same for all of the techniques employed in the study, to ensure a fair comparison. Every generation the current mutation rate is calculated. It is dependent on the best fitness achieved so far in the optimisation. When the best fitness is poor, i.e., close to 0, the mutation rate will be high, i.e., close to MR_{start}. As the optimisation progresses and the best fitness improves, the mutation rate will fall towards 0. This enables large changes to individuals to be made when the population is of poor fitness. As the fitness of the population improves, ever smaller changes will be possible, in order to aid the populations's convergence on an acceptable solution.

This mutation scheme was used because it would be simple to implement and would, more than likely, not skew the performance of the various optimisation techniques by being too complicated. Indeed, the purpose of this study is to compare different techniques (GA, GP & CGP) rather than looking at mutation schemes.

Redundancy, as described in sections 2.8 & 3.7.2, is an important phenomenon of computational optimisation. As such, the effect of redundancy was investigated in this study. Whether or not a particular computational optimisation system can exploit redundancy depends on the representation that it uses. In this study, the GA based techniques that use a bit map representation can not use redundancy. This is because there is a one to one correspondence between a particular bit in a given individual's bit string and the corresponding cell on the grid.

Systems that use GA and CGP can use redundancy because some of a particular individual's instructions can be inactive and an antenna geometry can still be created. Table 10.2 summarises the key features of the systems that were used in this study. As can be seen, the two types of GP system (generational & steady-state) each have a version that does and does not use redundancy. The type of redundancy used by the GA systems was *explicit*, as described in section 7.7.

The GA based systems that do not use redundancy create antenna geometries in the same way as described in the example of section 7.7.1. Similarly, the systems that do use redundancy create antenna geometries in the same way as described in the example of section 7.7.2.

CGP based systems exploit *implicit* redundancy. This redundancy arises due to the fact that CGP uses graphs to store the instructions, as described in 7.8.

System Type	**Population Size**	**Elitism Group Size**	**Number of Generations**	**Starting Mutation Rate**
Generational	2	1	1000	50%
Steady-state (small population)	10	4	165	5%
Steady-state (large population)	60	12	20	5%

Table 10.1: Summary of key control parameters.

10.5.2 Implementation of Neutral search

Neutral search, like redundancy, is another potentially powerful phenomenon of computational optimisation. It can be applied to any computational optimisation technique (GA, GP etc.). It has been successfully applied previously to CGP by Julian Miller, as described in section 3.7.2.

The basic principle behind neutral search is that the currently best individual can be explicitly replaced by a new individual of the same fitness. As in many algorithms, such as the $1 + \lambda$ algorithm, the area currently being searched within the search space is centered around the current best individual. As such, neutral search gives the algorithm the opportunity to 'jump' to a completely different area of the search space. More specifically, it enables this without accepting a degradation in fitness.

Neutral search was implemented in this study by enabling a new individual to become the current best individual even if the new individual had a worse fitness. However, this worse fitness had to be within a certain pre-determined range of the current best fitness. This range was known as the *neutral search margin*. The neutral search margin was adopted because the fitness values were floating point values, and so two different fitness values had virtually no chance of ever being exactly the same. It was decided that as there were relatively, i.e., compared to other optimisation studies in the literature, so few fitness evaluations per run, that every time a new individual that was within the neutral search margin of the current best fitness was discovered, this new individual should become the best individual. If this

was not the case then it would have been probable that the neutral search mechanism would hardly have been used and so its effect could not have been determined.

In this study, neutral search was applied to CGP only. The reason for this was that there was not sufficient time to apply it to all of the various types of technique. CGP was chosen because of its previous impressive results when using neutral search.

System Type	**Representation**	**Selection Method**	**Reproductive Operators**	**Comments**
Generational GA	bit map	fittest individual becomes parent	mutation	$1 + \lambda$ algorithm
Steady-state GA	bit map	tournament selection	mutation & crossover	
Generational GP - no redundancy	instruction list	fittest individual becomes parent	mutation	$1 + \lambda$ algorithm, blank instructions for variable list length
Steady-state GP - no redundancy	instruction list	tournament selection	mutation & crossover	blank instructions for variable list length
Generational GP - with redundancy	instruction list	fittest individual becomes parent	mutation	$1 + \lambda$ algorithm, instructions explicitly active or non-active
Steady-state GP - with redundancy	instruction list	tournament selection	mutation & crossover	instructions explicitly active or non-active
Generational CGP (implicit redundancy)	instruction graph	fittest individual becomes parent	mutation	$1 + \lambda$ algorithm, instructions activated /de-activated due to structure of graphs
Generational CGP (implicit redundancy) - with neutral search	instruction graph	fittest individual becomes parent	mutation	$1 + \lambda$ algorithm, instructions activated /de-activated due to structure of graphs

Table 10.2: Summary of system types and their attributes.

10.6 Instruction Sets

The GP and CGP based systems use sequences of instructions to create MSA geometries. A sequence of instructions effectively describes a journey around the grid. As the journey progresses the MSA's geometry is built up. There were a number of important considerations when these systems were implemented.

Firstly, there was the question of what would happen if the path of the journey being described by the sequence of instructions went off one of the sides of the grid. The solution that was adopted was simply that the path should continue on the opposite side of the grid. This meant that that the edges of the grid were effectively continuous.

Secondly, there was the issue of exactly which instructions should be in the possible instruction sets. As part of the study, it was intended that different instruction sets should be tried, to see their impact on the effectiveness of the particular optimisation technique. The available instructions that were chosen are:

no operation (blank)
turn right
turn left
go forward N_1 cells making cells '0'
go forward N_1 cells making cells '1'
go forward N_1 cells doing nothing
go forward N_1 cells inverting cells
set N_2 cells in vicinity to '0'
set N_2 cells in vicinity to '1'

The no operation (blank) instruction was used by the GP systems that did not use redundancy. All of the GP based systems used a fixed length instruction list. The blank instruction effectively enabled variable length lists to be created when no redundancy was used. The GP based systems that did use redundancy did not require blank instructions because variable length sequences of instructions could be formed by activating/de-activating instructions.

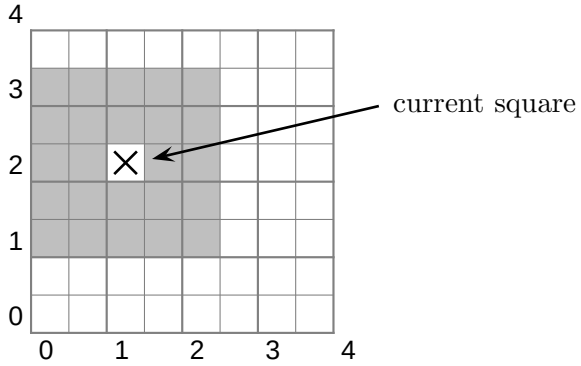


Figure 10.2: Example of a 'set 2 cells in vicinity' instruction.

It can be seen from the list of available instructions that there are two variables, N_1 & N_2, that are associated with the various types of instructions. N_1 is simply the number of cells that are in the forward direction in front of the current cell. N_2 determines the number of cells, around the current cell, that are to be metalised ('1') or non-metalised ('0'). In Fig. 10.2, the grey cells around the current cell are the ones that affected by the set cells in vicinity instruction ($N_2 = 2$).

10.6.1 Instruction Set Types and List/Graph Length

The various optimisation techniques use all or a sub-set of the above instructions. This leads to two general types of instruction set: *constructive only* and *constructive & destructive*. Constructive only instruction sets only have instructions that result in the metalisation of grid cells. In other words, they do not contain instructions that non-metalise cells. There is an important consideration when using constructive only instruction sets and that is that an optimal structure size needs to be determined and used.

This consideration is that of determining an effective instruction list length. In the case of constructive only instruction sets, using a list length that is too short could often result in an insufficient amount of the grid being traversed. This would result in not enough metal being put down on the grid to enable potentially beneficial structures to be created. On the other hand, having a list length that is too long, could likely result in the grid becoming too metalised such that distinctive structures are unlikely to be created. In order to avoid both situations, it is necessary to determine the list length, for the given instruction set and its associated variables (N_1 & N_2), that on average results in a near optimal degree of grid metalisation.

This list length was determined by creating a large number (10000) of random instruction sets of a particular length. The average number of metalised cells, and thus non-metalised cells, generated was then calculated. The list length that resulted in approximately 50% grid metalisation was determined as being the most optimal list length for the given instruction set. These instruction lists were random, i.e. not optimised in terms of the specification, so 50% metalisation is the most appropriate starting point for determining the optimal list length. This is because as the computational optimisation proceeds, i.e. as the grid patterns evolve to meet the specification, the proportion of grid metalisation will vary. As 50% is equally half way between no metalisation and full metalisation, the instruction lists will have equal flexibility in gaining or losing metal.

As far as constructive and destructive instruction sets are concerned, there is a similar issue regarding the optimum instruction list length for a particular instruction set and its associated variables. As with constructive only instruction sets, having too short a list length will result in an insufficient amount of the grid being traversed. Having too long a list length is not such a serious disadvantage as with constructive only instruction sets. Having too long a list length will probably result in all, or parts of, the grid being overwritten several times.

This is not an advantage in any way but it is also not particularly disadvantageous. This is because the constructive and destructive instructions can both add and remove metal and so it is unlikely that the grid will become prohibitively metalised. Using a list length that is excessively long, i.e., one that overwrites the grid very many times, is inefficient in terms of computer memory and thus should be avoided. It could also lead to the amount of time spent converting each list into its corresponding grid geometry becoming significant in terms of the fitness evaluation time. If this were the case then the overall time needed for the whole optimisation could be markedly increased. However, when optimising antennas, the fitness evaluation time is almost always very much longer than any other task. This is certainly true for this study.

In order to determine the optimum list length for constructive and destructive instruction sets, a similar technique to that used in the case of constructive only instruction sets was used. In order to determine the optimum list length a similar technique to that used in the case of constructive only instruction sets was used. For several list lengths, a large number (10000) of random instruction sets were generated. The average number of cells traversed was calculated. The list size that was determined as being optimal was the one that traversed twice the number of cells in the grid. This is because it is likely that the whole grid will be traversed by the active instruction sequence but there will not be a significant degree of over-writing.

The CGP based techniques do not require blank instructions because significantly large parts of the graph are redundant, resulting in a variable length sequence of active instructions. This is an intrinsic phenomenon of the directed acyclic graph structure. The method for determining the optimum graph size for a particular instruction set and its associated variables, was the same as that for the GP based techniques when using constructive and destructive instruction sets.

Chapter 11

Empirical Study Statistical Results

11.1 Results Overview

In this section, the main statistical results of the empirical study, as shown in table 11.1, are discussed. The generation of these statistical results is the main purpose of the study, rather than the generation of actual antenna designs. A statistical approach is the only meaningful way to compare the efficiencies of the various computational optimisation techniques used in the study.

All of the computational optimisation systems used in the study used the same fitness function, 10.5.1. This fitness function returns values in the range of 0 to 1. The worst possible fitness is 0 and the best possible fitness is 1. As described in chapter 12, 10 runs of each system were performed. The average (mean) highest fitness and average standard deviation were then calculated for each system.

The highest average fitness was achieved by steady-state GAs with a small population size (row 2 of table 11.1). Several of the techniques produced at least one antenna with a fitness of 0.95 or higher. Some of the antennas were built and measured, as can be seen in chapter 12. The best possible result that a system could have would be for all of its 10 runs to have had a fitness value of 1.

The average standard deviation was close to 0.1, i.e., 10% of the fitness function range. In a Gaussian distribution, 95 % of the solutions lie within $\pm$ 2 standard deviations of the mean. Consequently, if the results of this study are assumed to have an approximate Gaussian distribution, then they will be contained within a window of width 0.4 which is centered at the mean. The standard deviation of most of the techniques being close to 10% of the possible fitness range suggests that confidence in the mean fitness values is justified. In other words, the fitness values are fairly well clustered around the mean value for the vast majority of the techniques.

Several techniques have a mean fitness of 0.8 or higher. Although this is not as good as having an average fitness of 1, it is still a good indicator of strong performance because the mean fitness is within 2 standard deviations of the highest possible fitness.

As the mean approaches 1, due to several of the runs having a fitness of 1 (max), then the Gaussian distribution will become distorted somewhat. When this happens the standard deviation will become less reliable as a measure of the spread of the distribution. Therefore, when different techniques have means close to 1, they can only be compared using their means.

Row	**System Type**	**Population Size**	**Generations**	**Elitism Group Size**	**Instruction Set**	**List/Graph Length**	N_1	N_2	**Average Highest Fitness**	**Standard Deviation**	**Comments**
1	Generational GA	2	1000	1	n/a	n/a	n/a	n/a	0.81	0.14	
2	Steady-State GA	10	165	4	n/a	n/a	n/a	n/a	0.94	0.08	
3	Steady-State GA	60	20	12	n/a	n/a	n/a	n/a	0.91	0.11	
4	Generational GP	2	1000	1	constructive & destructive	250	4	2	0.64	0.07	
5	Generational GP	2	1000	1	constructive	250	4	2	0.89	0.09	
6	Steady-State GP	10	165	4	constructive	250	4	2	0.78	0.08	
7	Steady-State GP	60	20	12	constructive	250	4	2	0.79	0.09	
8	Generational GP with redundancy	2	1000	1	constructive	500	4	2	0.78	0.06	
9	Steady-State GP with redundancy	10	165	4	constructive	500	4	2	0.87	0.10	
10	Steady-State GP with redundancy	10	165	4	constructive & destructive	500	4	2	0.77	0.13	
11	Steady-State GP with redundancy	10	165	4	constructive	1000	1	1	0.67	0.08	
12	Steady-State GP with redundancy	10	165	4	constructive & destructive	1000	1	1	0.66	0.07	
13	Steady-State GP with redundancy	60	20	12	constructive	500	4	2	0.86	0.09	
14	Generational CGP	2	1000	1	constructive	30000	2	1	0.69	0.10	
15	Generational CGP	2	1000	1	constructive	100000	2	1	0.63	0.04	
16	Generational CGP	2	1000	1	constructive & destructive	1000	3	3	0.77	0.18	
17	Generational CGP	2	1000	1	constructive & destructive	1000	2	1	0.86	0.12	
18	Generational CGP with neutral search	2	1000	1	constructive & destructive	10000	3	3	0.82	0.12	neutral search margin = 0.05
19	Generational CGP with neutral search	2	1000	1	constructive & destructive	10000	3	3	0.73	0.13	neutral search margin = 0.025
20	Generational CGP with neutral search	4	333	1	constructive & destructive	10000	3	3	0.76	0.19	neutral search margin = 0.05

Table 11.1: Summary of key control parameters and their corresponding results.

11.2 Results Description

In table 11.1, the various parameters are described in sections 10.5 and 10.6. As far as the GA based techniques are concerned, there is a fairly distinct improvement in fitness when going from generational (row 1) to steady-state (rows 2 & 3). Both of the steady-state GA systems performed well, with the small population size (row 2) achieving the highest average fitness in the whole study.

For the GP based systems that did not use redundancy (rows 4 to 7), the best performance was achieved by a generational system with a constructive instruction set (row 5). The average fitness of this technique was only just second to that of the best technique in the whole study. Comparing the two generational GP systems that did not use redundancy (rows 4 & 5), there is marked difference between the constructive & destructive instruction set and the constructive only instruction set. The constructive only instruction set performed very much better than the constructive & destructive instruction set. It was for this reason, combined with the fact that only a limited amount of time for the empirical study was available, that constructive only instruction sets were used for the remaining runs of the GP based systems that did not use redundancy. The steady-state GP based systems that did not use redundancy (rows 6 & 7) performed noticeably worse than the corresponding generational systems (rows 4 & 5). This finding is the reverse of that of the GA systems, i.e., for the GA systems, generational was worse than steady-state.

As far as the GP based systems that did use redundancy (rows 8 to 13) are concerned, some interesting findings were observed. An initial finding was that when using constructive only instruction sets, it was found that steady-state (row 9) out performed generational (row 8) by a significant, although not huge, degree. This finding is in agreement with that of GAs but not with that of GP based systems that did not use redundancy. When moving to a constructive and destructive instruction set (row 10), it can be seen that the fitness of the steady-state system falls back to that of the generational system (row 8). However, the standard deviation for the constructive and destructive instruction set (row 10) is markedly worse (double) than that of the generational system (row 8). This means that the fitnesses values are more spread and thus the technique is less reliable.

An interesting comparison is that between rows 9 & 10 and rows 11 & 12. The results from all these four rows were generated by steady-state GP with redundancy. Indeed, rows 9 & 11 and rows 10 & 12 have identical parameters respectively apart from the list length and maximum instruction variable values (N_1 & N_2). With smaller N_1 & N_2 values, the list length must be correspondingly increased because each instruction has less scope, i.e., it can modify fewer cells on the antenna's grid. The clear difference in the performance of the smaller list length (rows 9 & 10) and the longer list length (rows 11 & 12) is that the smaller list length clearly out performs the longer list length.

When the smaller list length (500) is resumed but the population size is increased (row 13), the performance returns to that of the best case for GP based systems that did use redundancy, i.e., row 9. Again, this was achieved using a constructive only instruction set. There was insufficient time to run this set of parameters with constructive and destructive instruction sets.

All of the CGP based systems used the $1 + \lambda$ algorithm (as described in section 3.7.2) which is generational. There were several reasons why CGP was only implemented with $1 + \lambda$ algorithm. The main reason was that CGP has been previously very successfully implemented and used, most notably by Julian Miller [81], in the evolution of Boolean logic functions. Miller argues that CGP's striking success is due to its use of graphs (as opposed to trees etc.) rather than to the actual algorithm type (generational/steady-state) and its associated parameters. This fact, combined with that of limited time, resulted in the CGP based systems of this study utilising the $1 + \lambda$ algorithm only.

Although all the CGP implementations of this study were generational, there were two significantly different sub-types. These were due to the fact that neutral search could be either enabled or disabled. As there was the not the important comparison of generational vs steady-state for CGP in this study, there were less trials (different parameter sets) that needed to be performed. It was for this reason that CGP was the technique that was chosen to investigate the effect of neutral search. Additionally, when CGP has been explicitly combined with neutral search in the past [5], it again yielded impressive results. To recap, neutral search was implemented in the $1 + \lambda$ algorithm for CGP in this study, by enabling a new individual to become the parent even if it had a slightly lower (worse) fitness than the current parent. In the standard $1 + \lambda$ algorithm, an individual can only become the new parent if its fitness is better than the current parent. In this study, if an individual had a lower fitness than the current parent, but that fitness was within a pre-determined margin (neutral search margin), then the individual in question could become the new parent with a given probability. As there were relatively very few (compared to the literature) evaluations per each independent trial run, this probability was set to 1. Otherwise there would not be sufficient likelihood that the neutral search mechanism would have a chance to work.

Overall, the performance of CGP in this study was relatively poor. This was the case for when neutral search was both enabled and disabled. The best performing CGP run was that of row 17, i.e., when neutral search was not used. In this instance, CGP's performance was very close to that of the best of the GP implementations (both with and without redundancy). The average highest fitness was virtually identical, but the average standard deviation was slightly higher, meaning that the fitness values were less tightly clustered around the mean. Indeed, apart from the case of the particularly large graph size (row 15), the average standard deviations for all of the CGP parameter sets was relatively large.

The first runs of CGP (without neutral search) used constructive only instruction sets (rows 14 & 15) and fairly large graph sizes. Constructive only instruction sets were chosen because they had performed well when used with previous techniques in the study. Large graph sizes were used in order to produce complex graphs with much redundancy. However, these large graph sizes produced significantly worse results than those of dramatically smaller graph sizes (rows 16 & 17). It must be noted that the smaller graph sizes (rows 16 & 17) used constructive & destructive instruction sets rather than constructive only, so a straight comparison is invalid. A dramatic change in parameters, i.e., going from those of row 15 to those of row 16 was justified at the time because of the markedly poor performance of

the first two parameter sets (rows 14 & 15). There was never going to be enough time to perform a more systematic and thus more scientifically rigorous study than was actually performed.

By the time the CGP implementation that used neutral search was to be run (rows 18 to 20), there was not very much run time left for the study. As such it was felt that using an intermediate graph size (10000) would be a reasonable choice. Likewise, as constructive & destructive instruction sets had performed better for CGP without neutral search, it was decided to use them. The parameter set of row 18 performed the best for CGP with neutral search. Here the neutral search margin was 0.05, which was 5% of the fitness range. When the neutral search margin was halved to 0.025 (row 19), the average highest fitness dropped significantly. Lastly, the population size was increased (row 20), but the fitness only improved slightly and the standard deviation became significantly worse.

11.3 Results Analysis

The purpose of the empirical study was to generate statistics that could be used to reliably compare the efficiencies of various computational optimisation techniques. This aim has largely been achieved despite the very significant time limitation due to the relatively very long, i.e., when compared to studies in the literature, fitness evaluation time. The most marked finding, as described in the previous section, is that, in general, the simpler techniques (e.g. GAs) out performed the more complex techniques (e.g. CGP).

As the fitness evaluation time was so long, only a relatively small number (1000) of fitness evaluations could be performed during each computational run. When compared to other studies concerning computational optimisation in general in the literature, this number is particularly low. It is likely that this low number of evaluations is a significant reason why the more complex techniques used in this study, such as CGP, did not yield better results than they actually did. Phenomena such as redundancy and neutral search need to be given a certain amount of time before they actually start to have a beneficial impact on the progression of the optimisation. Techniques that utilise these phenomena therefore need to be run for a minimum amount of time, i.e., minimum amount of evaluations, before any advantage that may arise from utilising such a phenomena becomes apparent. In other words, these more complex techniques may well need to be run for a longer time than other simpler techniques before their advantages can be seen. If run time was unlimited and all of the techniques were run for many more evaluations than they actually were in this study, then it is likely that the simpler techniques would have stagnated relatively early on and the more complex techniques would have gradually continued to make fitness improvements. This is because, when correctly applied, phenomena such as redundancy and neutral search, reduce the likelihood of stagnation.

Computationally optimising antennas will always involve relatively long fitness evaluation times and thus the amount of fitness evaluations will be significantly limited. This is a good reason for using a relatively basic technique rather than one with more complex features.

Another reason to avoid the more complex techniques when the number of fitness evaluations is strictly limited, is that there are more control parameters for the user to choose and therefore more chance that the optimisation's performance will be markedly affected for the worse by a sub-optimal value. As can be seen in the previous section, the choice of a particular control parameter can have a dramatic effect on a technique's performance. Additionally, the effect that a particular control parameter can have on the performance can be particularly counter-intuitive. Consequently, the more complex a technique is, then generally the more control parameters it will require. Furthermore, the likelihood that the technique's performance is dramatically sensitive (i.e. non-linear) to one, or a combination of, of these control parameters increases. This finding, combined with the observation that the choice of certain control parameters can be markedly counter-intuitive, is a good reason for avoiding the more complex techniques when run time is strictly limited.

11.4 Results Conclusion

As previously mentioned, the highest average fitness was achieved by steady-state GAs with a small population size (row 2 of table 11.1). The other, more intrinsically complicated techniques (GP & CGP), did not perform as well. This is believed to be due to the fact that such a small number of fitness evaluations was used relative to other computational optimisation studies [50] [51] [52].

Many of the techniques produced at least one antenna with a fitness of 0.95 or higher. In other words, several of the techniques produced antennas that met, or very nearly met, the target specification. These antennas had highly irregular geometries that could not have been designed using conventional antenna design techniques. Therefore, the use of these computational optimisation techniques has resulted in completely novel antenna designs. Some of these antennas were physically built and measured and can be seen in chapter 12.

Chapter 12

Evolved Antennas from the Empirical Study

The main purpose of the empirical study is to compare the efficiencies of various different computational techniques when optimising MSAs. During the course of the study, it was expected that several antenna designs would be created which would completely, or very nearly, meet the optimisation specification. Nearly all of the optimisation techniques used in the study produced at least one antenna that met the specification. This chapter is concerned with the analysis of some of those antennas.

12.1 Antenna 1

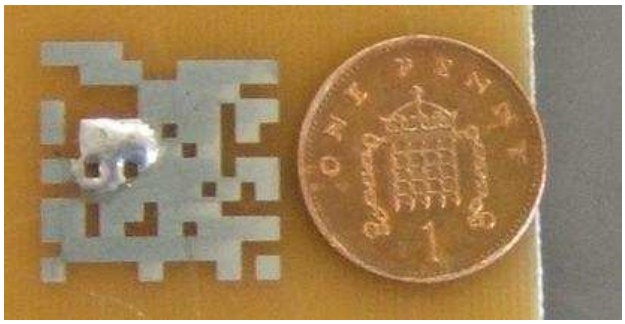


Figure 12.1: Top view of antenna 1.

12.1.1 Bandwidth

This antenna was evolved by the generational GA system when using a small population size. It can be seen from Fig. 12.2 that when the antenna was physically built, its measured S_{11} closely matched its simulated S_{11}. Even with a higher resolution computational model, it would be virtually impossible to make the measured and simulated S_{11} curves exactly line up. One of the main reasons for this is due to the fact that the permittivity and loss of FR4 varies slightly across the same piece of FR4 and also with frequency. However, apart from the built antenna having slightly less bandwidth than the computational model, it is very close to perfectly meeting the specification.

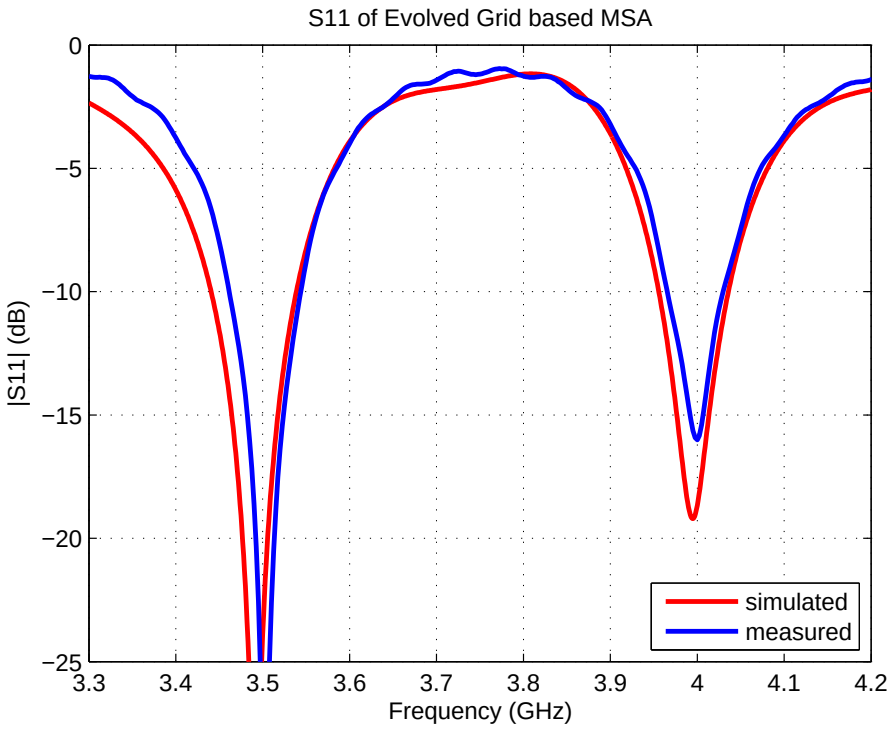


Figure 12.2: Return loss of antenna 1.

12.1.2 Radiation Pattern

The radiation patterns in this chapter were generated by modeling the antennas with a method of moments full wave electromagnetic solver called concept. This was because the FDTD software (falcon) used by the optimisation techniques of the study did not determine radiation patterns.

Figs. 12.4 and 12.5 show the far field electric field magnitude, i.e., the radiation pattern, of antenna 1 for two separate orthogonal planes. It can be seen that, in both planes, the dominant direction of radiation is perpendicular to the orientation of the substrate. In other words, if the MSA is lying flat, then the direction of maximum radiation is straight up. This is the same as for RMSAs. It can also be seen that there is a single main lobe which has a 3 dB beam width of approximately 90°. The radiation pattern for the same two orthogonal planes at 4 GHz was almost identical to that at 3.5 GHz. The only difference of any note was a slightly narrower 3 dB beam width.

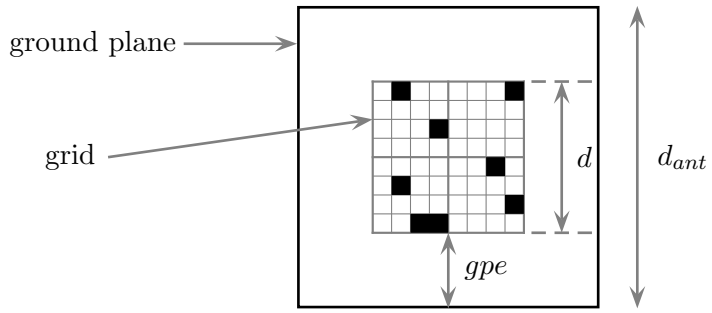


Figure 12.3: View from above of a square grid based MSA and its ground plane.

With reference to Fig. 12.3, when a suitable ground plane extension is considered, the observed 3 dB beam width is close to that predicted by theory . In this case the ground plane extension is 6 times the substrate thickness (h). The analysis is as follows:

$f = 3.5$ GHz
$\lambda_0 = 8.57$ cm
$h = 1.6$ mm
d = grid dimension = 2 cm
$\longrightarrow \quad gpe$ = ground plane extension = 6 h = 0.96 cm
$d_{ant} = d + 2\,gpe \approx 4$ cm
$d_{ant}/\lambda_0 \approx 0.5$
$HPBW_{theory} = f(d_{ant}/\lambda_0) \approx 75^\circ$

It is perfectly reasonable to include the ground plane extension because the antenna requires it in order to function correctly. In addition, much of the radiation from grid based MSAs is likely to come from the fringing fields at the edge of the grid. These fringing fields at the grid edge effectively make the grid electrically larger.

As can be seen from Fig. 8.14, when the MSA's dimension is around half the free space wavelength, the 3 dB beam width should be around 75°. This is close to that of the observed 3 dB beam width ($\sim 90^\circ$).

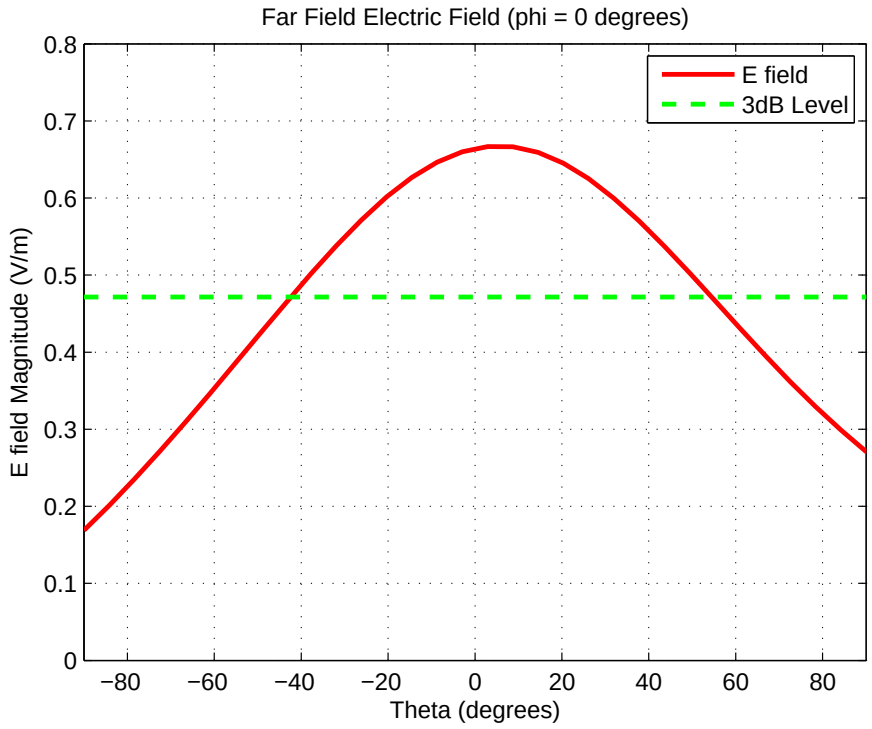


Figure 12.4: Radiation pattern of antenna 1 at 3.5 GHz (plane 1).

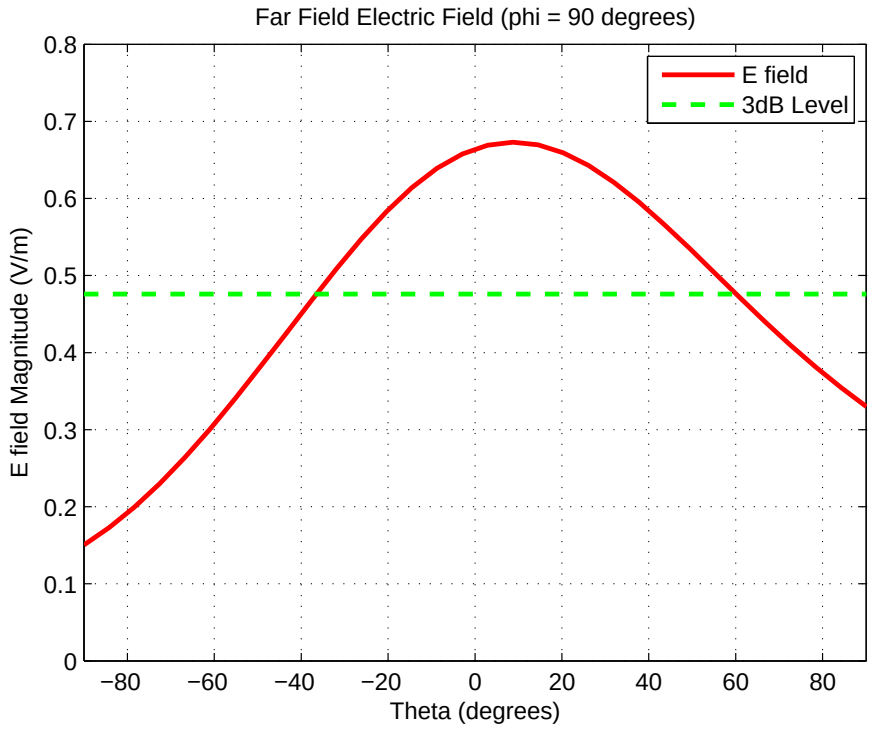


Figure 12.5: Radiation pattern of antenna 1 at 3.5 GHz (plane 2).

12.1.3 Surface Current Distribution

Considering the surface current distribution of an MSA can provide some explanation as to the functioning of the antenna. As can be seen from Fig. 12.6, there is a dominant resonant path on the right hand side of the grid (as it is orientated in Fig. 12.6). In Fig. 12.1, this corresponds to the bottom side of the grid. This dominant pathway can be seen to actually split into two paths , as they go around an 'L' shaped hole (non-metalised area) in the grid. The majority of the radiation from the antenna is likely to come from the grid edges that are at the ends of this dominant resonant path.

The speed of waves traveling across the grid depends on the effective relative permittivity (ϵ_e). Unlike microstrip transmission lines and RMSAs, the ϵ_e of an irregular grid based MSA changes with position. This is because the width of the track that guides the waves changes with position. However, the fact that ϵ_e always lies between two known extreme cases can be used to perform an approximate calculation of the resonant frequency of a particular path.

$\epsilon_e^{max} = \epsilon_r = 4.0$
$\epsilon_e^{min} = (\epsilon_r + 1)/2 = 2.5$
$\longrightarrow \quad \epsilon_e^{min} \leq \epsilon_e \leq \epsilon_e^{max}$

The mean effective relative permittivity can be used in the resonant frequency calculation:
$\bar{\epsilon_e} = (\epsilon_e^{max} + \epsilon_e^{min})/2 = 3.25$

The average velocity on the grid is:

$\bar{v} = c/\sqrt{\bar{\epsilon_e}} = 1.65 * 10^8 ms^{-1}$

As the length (L) of a resonant path is half a wavelength (for the fundamental frequency), the resonant frequency is then:

$f = \bar{v}/\, 2L$

The length of the dominant resonant path of antenna 1 for the 3.5 GHz band, as shown in Fig. 12.6, is approximately 14 cells long. Each cell is a square of 1.6 mm dimension. The approximate resonant frequency of the dominant current pathway is then:

$f = 1.65 * 10^8/(1.6 * 10^{-3} * 2 * 14) = 3.68$ GHz

This approximate resonant frequency is remarkably close to the actual target (and measured) value of 3.5 GHz.

When then antenna is operated at 4 GHz, there is a noticeable change in the surface current distribution, as can be seen in Fig. 12.7. There is no longer a clearly dominant current path. Rather, the whole of the largely metalised central area of the grid has a significant current density. However, the left hand side of the grid (Fig. 12.7), has some what greater magnitude currents than other areas of the grid. This current path is approximately 13 cells long. Using the same technique as used for the 3.5 GHz situation, this current path can be shown to approximately resonate at 4 GHz. It is therefore highly likely that the grid edges at the ends of this current path significantly contribute to the radiation from the antenna.

It can also be seen from Figs. 12.7 and 12.6 that some areas of the grid are hardly utilised at all in either of the operating bands. For instance, the area of the grid at the bottom left hand corner has hardly any current density and thus contributes very little to the antenna's radiation. Similarly, the small parasitic patches, consisting of one or two cells, hardly contribute. It is therefore probable that these areas could be removed without significantly adversely affecting the performance of the antenna.

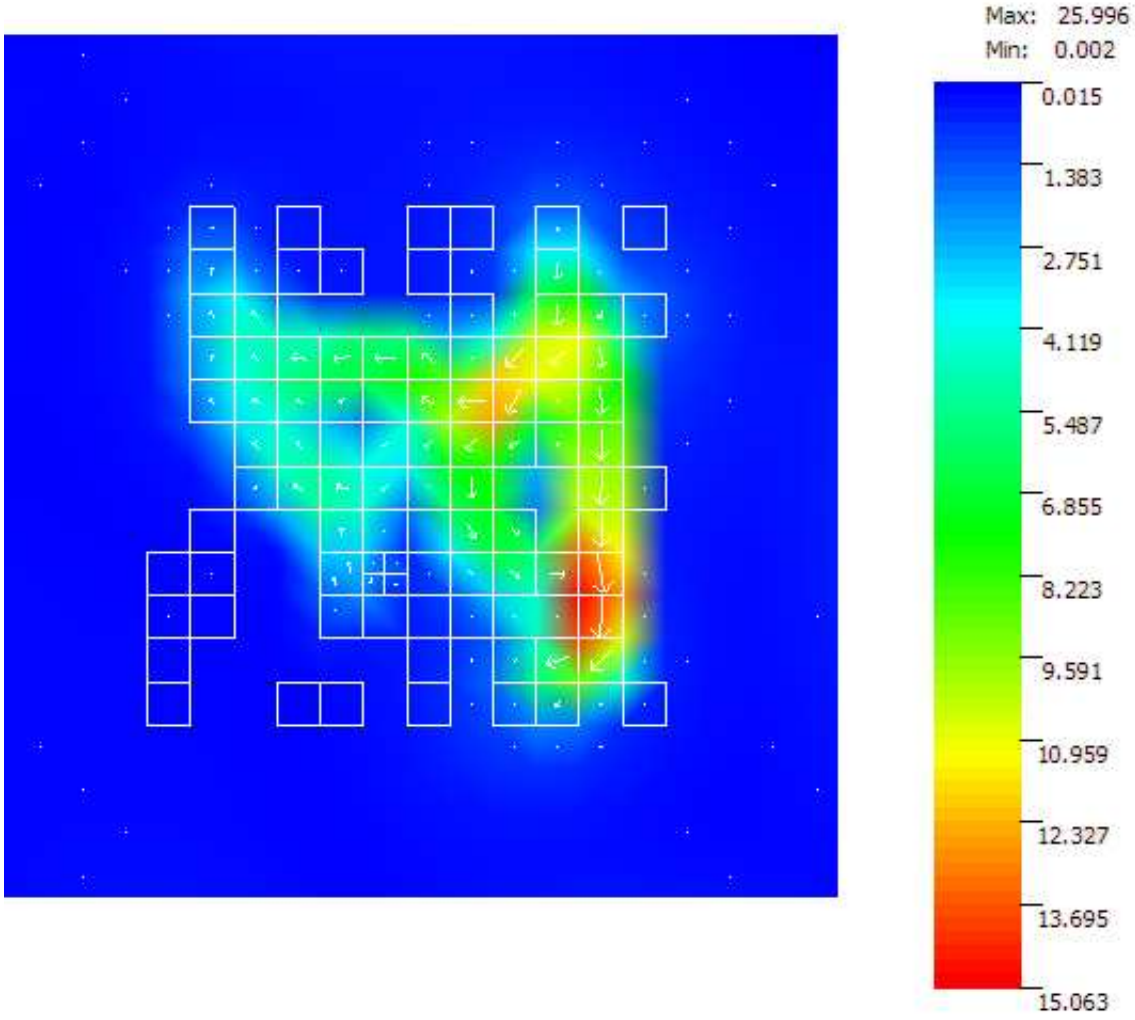


Figure 12.6: Surface current distribution of antenna 1 at 3.5 GHz.

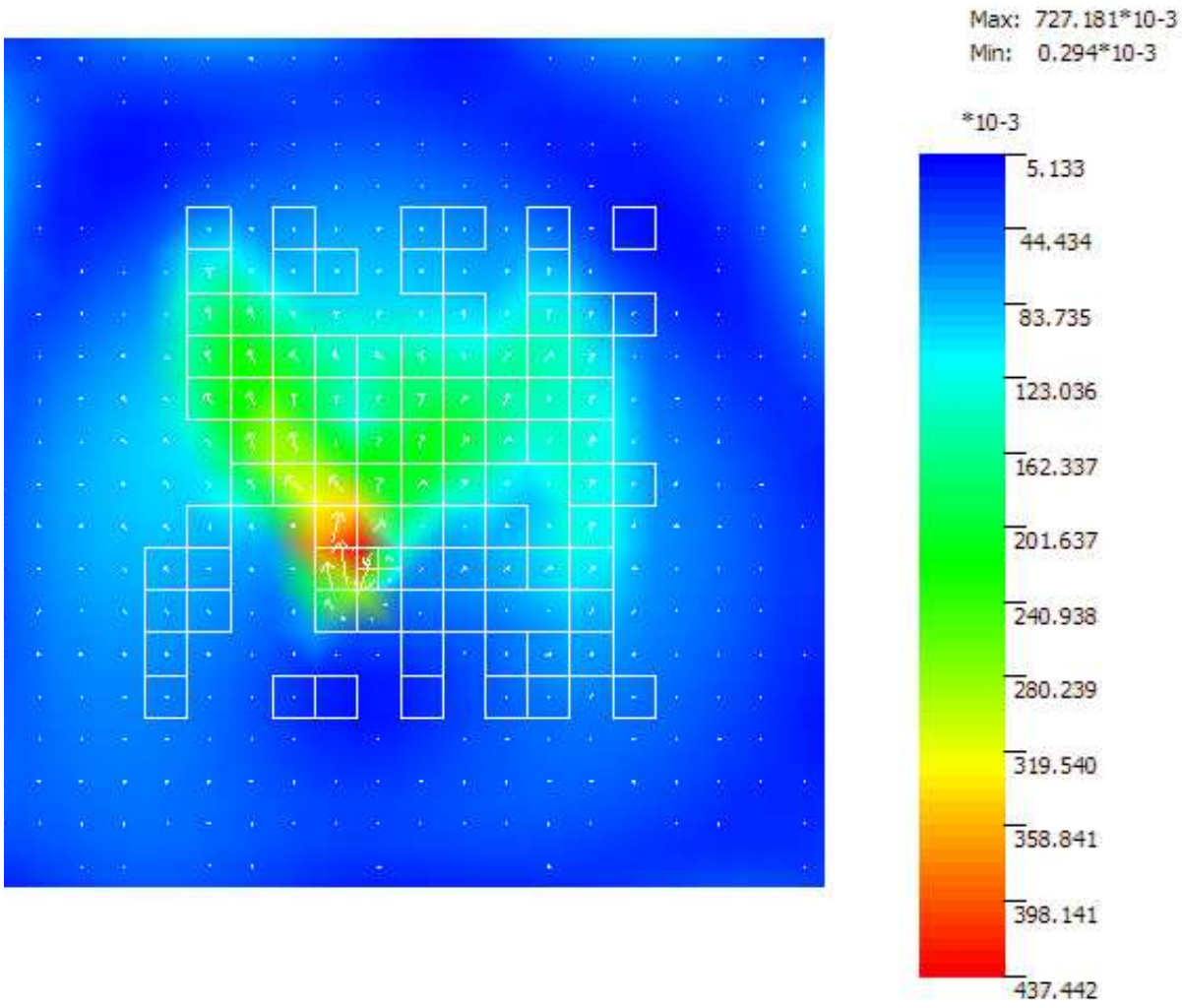


Figure 12.7: Surface current distribution of antenna 1 at 4 GHz.

12.2 Antenna 2

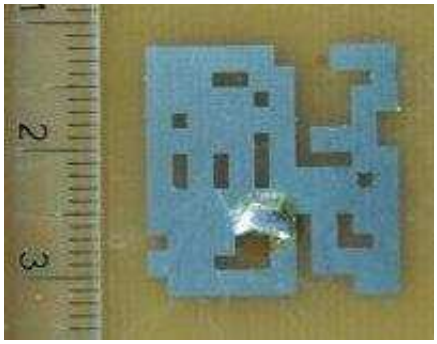


Figure 12.8: Top view of antenna 2 (scale is in cm).

12.2.1 Bandwidth

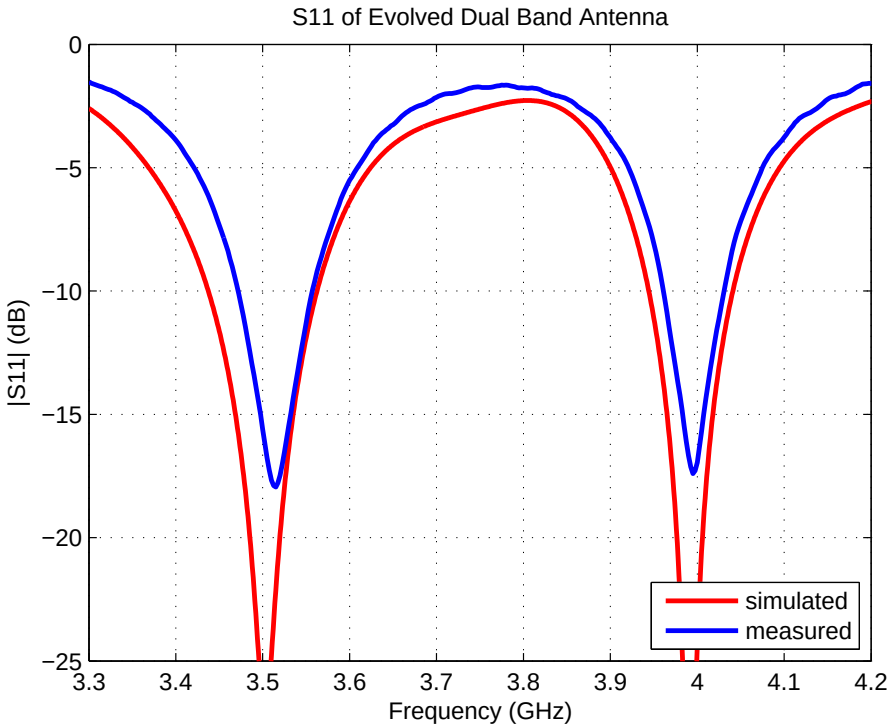


Figure 12.9: Return loss of antenna 2.

This antenna was evolved by the steady-state genetic programming technique that utilises explicit redundancy. As with antenna 1, the measured and simulated S_{11} of antenna 2 match closely, as shown in Fig. 12.9.

12.2.2 Radiation Pattern

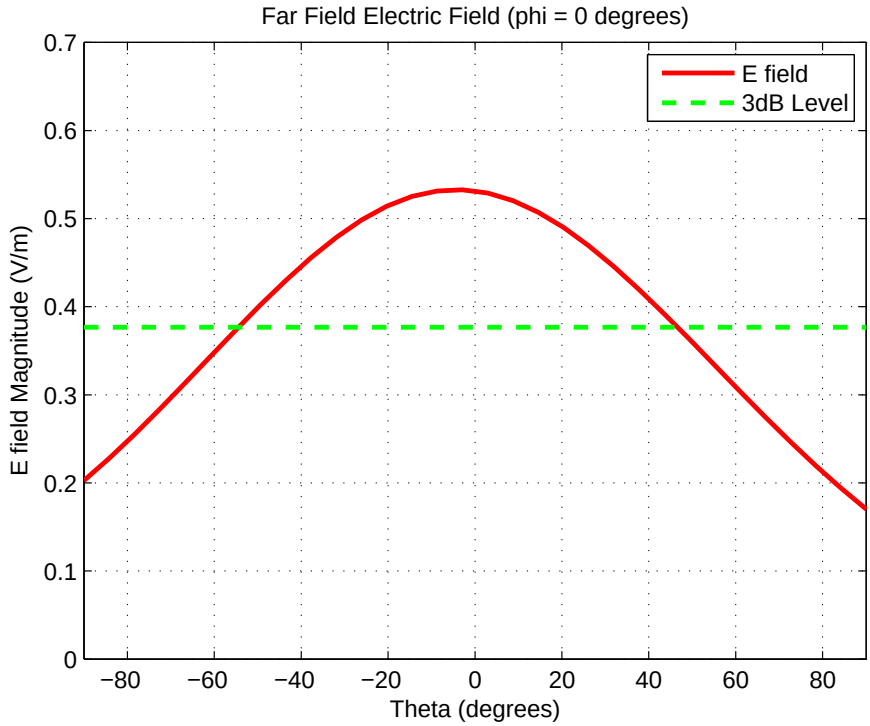


Figure 12.10: Radiation pattern of antenna 2 at 3.5 GHz (plane 1).

The observed 3 dB beam width of antenna 2, as shown in Figs. 12.10 and 12.10, is around 100°. This value is only slightly larger than that predicted by theory (section 12.1.2).

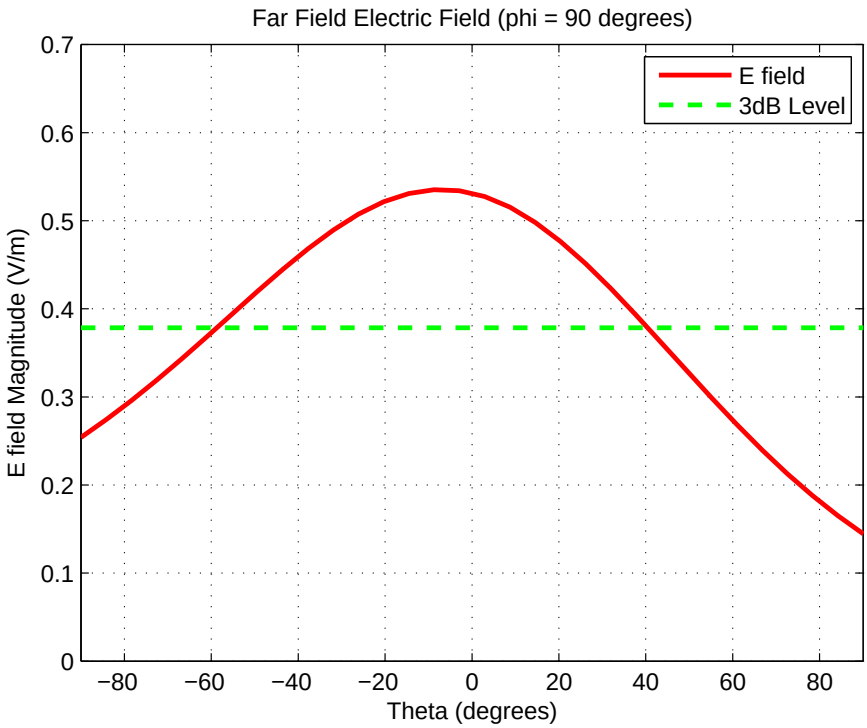


Figure 12.11: Radiation pattern of antenna 2 at 3.5 GHz (plane 2).

12.2.3 Surface Current Distribution

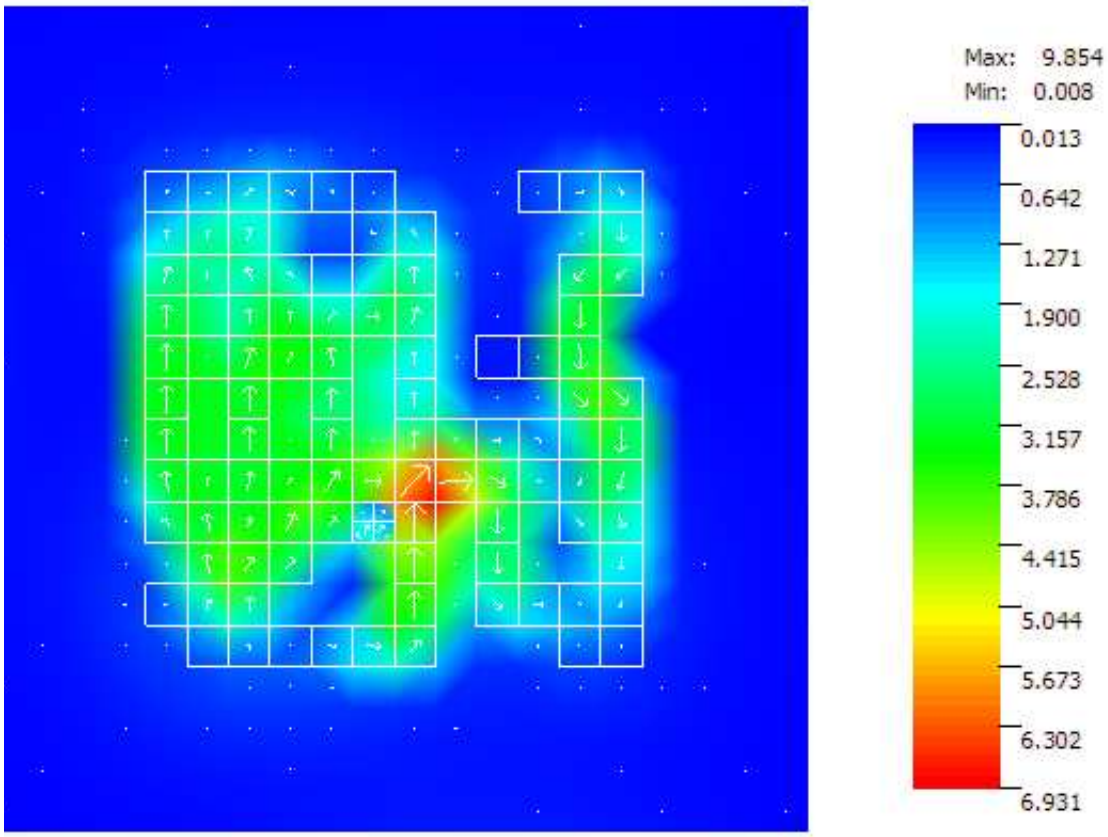


Figure 12.12: Surface current distribution of antenna 2 at 3.5 GHz.

As can be seen from Fig. 12.12, there is no clearly identifiable single dominant resonant current path for antenna 2 at 3.5 GHz. The whole left hand side of the grid resonates as a single structure. The narrow 'arm' on the right hand side of the grid also resonates at this frequency.

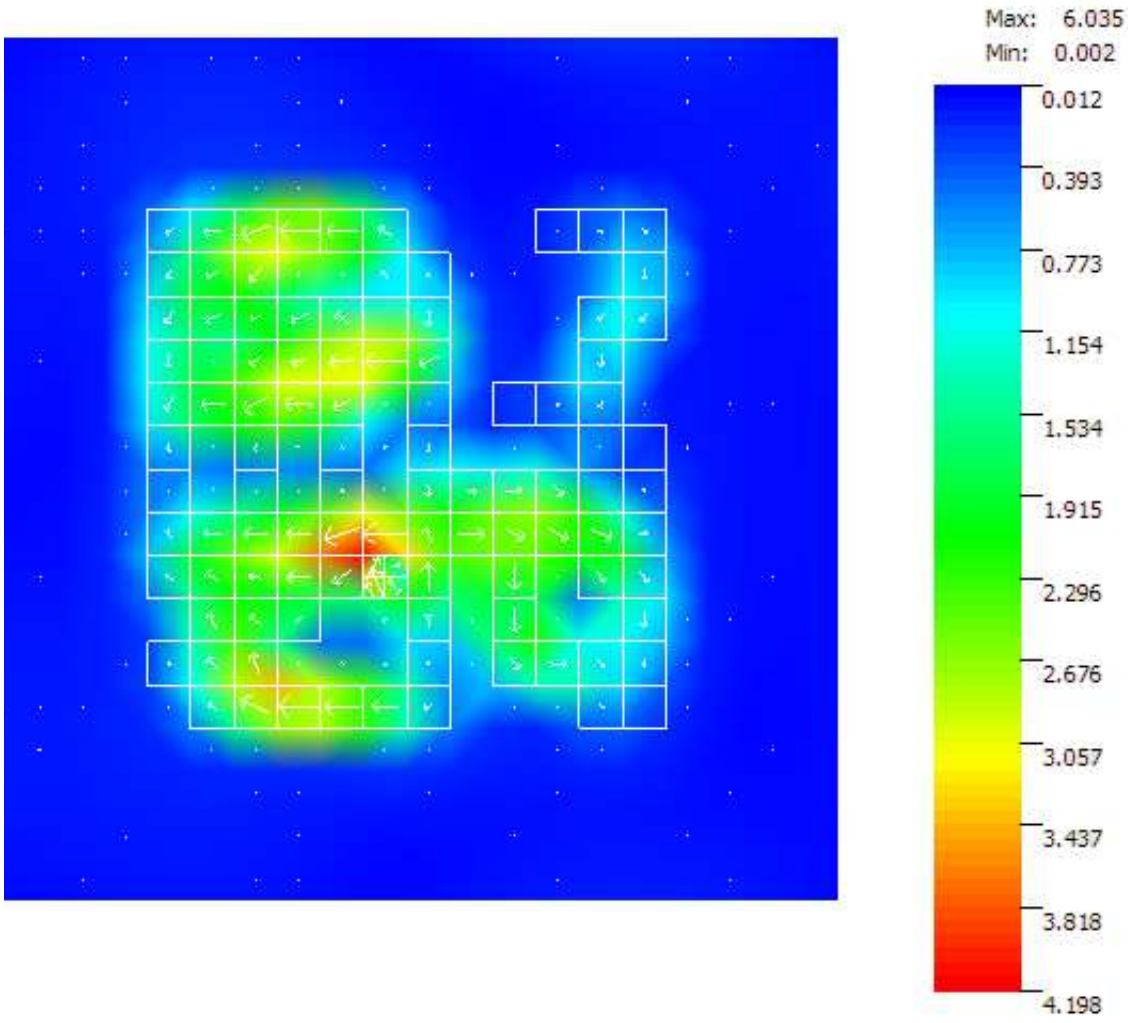


Figure 12.13: Surface current distribution of antenna 2 at 4 GHz.

It can be clearly seen from Figs. 12.12 and 12.13 that antenna 2 resonates in an entirely different way at 4 GHz than it does at 3.5 GHz. For instance, the meandered current path at the top right corner is of the grid is no longer significant at 4 GHz. A key difference is that at 3.5 GHz the resonances are vertically orientated, but at 4 GHz they are horizontally orientated.

12.3 Conclusion

Several different antenna designs, of which two are described in this chapter, were created during the course of the empirical study. As can be seen in Figs. 12.1 and 12.8, the geometries of the two antennas are particularly asymmetric and they have a markedly random appearance. These geometries could not have been created with conventional MSA design techniques. Thus, the use of certain computational optimisation techniques has resulted in novel MSA designs.

As far as the antenna's S_{11} plots are concerned, in every case there was a close (within 3 dB) agreement between the measured and simulated values. However, for all of the antennas, the width of the nulls when measured was always either equal to, or slightly less than that of the nulls when simulated. This can be seen in Figs. 12.2 and 12.9. This is believed to be due to the fact that the size of the grid is right on the limit of Chu-Harrington limit, as described in section 8.1.2. In other words, according to the Chu-Harrington limit an MSA of this size is only just capable of achieving the desired bandwidth (100 MHz at 3.5 GHz). The antenna is electrically smaller at the lower (3.5 GHz) band and so this is the band that is more critical as regards the relationship between antenna size and bandwidth.

Another observed feature of some of the antenna's S_{11} plots was that of frequency shifting. As mentioned above, for all of the antennas that met the specification and were then built and tested, the measured and simulated S_{11} curves match closely. However, in some cases there is a significant frequency shift (up to 10 MHz) between the measured and simulated curves. This is due to the relative permittivity of the FR4 varying from piece to piece. The relative permittivity of the built and tested antennas varied from 3.9 to 4.05. The simulated value was 4.0.

An additional, but far less dramatic phenomenon that was sometimes observed was a varying shift between the measured and simulated S_{11} plots of a given antenna. For instance, the the measured and simulated S_{11} of a particular antenna could line up closely for one of the bands, whilst there could be a significant frequency shift between them at the other band. This phenomenon is due to the relative permittivity of FR4 varying with frequency, it is not due to the accuracy of the computational model. The FDTD software that was used by the techniques of the study used a single fixed value for the permittivity of the dielectric. There was therefore no way for the software to take into account the fact that the relative permittivity of the FR4 substrate changes with frequency. When single band antennas and other narrow band structures were modeled, some of which can be seen in chapter 9, the simulated results were very accurate when the relative permittivity value used by the software was correct. When it was not correct, a frequency shift would be observed between the simulated and measured results.

Due to its properties, FR4 is not specifically intended to be an RF substrate and as such it is not ideal for use at the frequencies of interest in this study. However, as producing practical antennas to be used for a specific application was not the goal of this study, the small frequency shift of the antenna's S_{11} is not a problem. The goal of the study was to compare the efficiencies of various different computational optimisation techniques for evolving microstrip antennas.

Chapter 13

Conclusion

This study has shown that the computational optimisation of MSAs is viable and can lead to antenna designs that could not have been created using conventional techniques. It has shown that the simplest technique that was used, i.e. GAs, is the most efficient.

As described in chapter 1, this study is novel in several different ways. To summarise these again, the main source of novelty is that an empirical study of computational optimisation techniques for any type of antenna does not appear in the literature. A further source of novelty is that a comprehensive and complete analysis of the phenomena that affect the performance of grid based MSAs was carried out as part of this study. Such a complete and rigorous study does not appear in the literature. Lastly, an investigation into which computational optimisation techniques can be applied, and how, to the optimisation of grid based MSAs was also performed, which again can not be found in the literature.

In spite of the fact that, this study has shown that the computational optimisation of MSAs is a viable and practical technique, it will always, at best, be a complementary tool for antenna engineers. It will never remove the need for competent antenna engineers. The reason for this is that domain knowledge (antenna theory and design) is a pre-requisite for effective computational optimisation of antennas. There are several reasons for this. Two of the most important are described below.

Firstly, there is the requirement to ask the computational optimisation system to find a solution that is physically possible. An understanding of the physical limits of antennas is essential for this purpose. For instance, an understanding of the relationship between antenna size and bandwidth is necessary for determining the amount of space (volume) that the optimisation can use when trying to achieve a given bandwidth. Someone who is not familiar with this relationship could well ask the system to try achieve an impossible target.

Secondly, there is the requirement to achieve optimal use of the electromagnetic simulation software. More specifically, the optimal point that balances accuracy against efficiency needs to be found. Significant knowledge and experience of computational electromagnetic modeling is necessary in order to be able to achieve this.

Another important observation is that the computational optimisation of MSAs is definitely a practical endeavor. This is evident from the fact that given a middle of the range (circa 2006) desktop computer and a couple of weeks run time, MSAs of the type generated by this study, can be produced. As such, the computational optimisation of MSAs, and indeed that of many other types of antenna, is a viable and practical tool that can be used in conjunction with, and in parallel with, the conventional design of antennas.

Chapter 14

Further Work

As has been shown in this study, the computational optimisation of microstrip antennas can be a genuinely useful technique. Much of the time, highly irregular geometries are produced that could not have been created using more conventional methods. It would be relatively straightforward to apply some or all of the computational optimisation techniques used in this study to the optimisation of other antenna/RF problems.

14.1 MSAs with Higher Bandwidth Feeds

There are several ways of driving MSAs, some of which are described in section 5.2. Probe fed MSAs were used in this study because they are particularly simple to both computationally model and physically build. However, other types of feed, usually more complicated, can result in increased bandwidth. A particularly striking example of this is that the of proximity feed. Fractional bandwidths of over 10 % can be acheived with this type of feed. As such, it would be a worthwhile endeavour to computationally optimisatise MSAs that have higher bandwidth feeds when high bandwidths are required. An essential prerequisite would be the accurate computational modeling of the chosen feed type. This would undoubtedly take some time to perfect.

14.2 Printed Monopoles

Printed monopoles, as with MSAs, are commonly used in wireless data links utilised by consumer electronic equipment for applications such as WiFi and Bluetooth. Printed monopoles differ from MSAs in that there is no ground plane directly under the radiating part of the antenna. However, Like MSAs, resonance is still the dominant aspect of the radiation mechanism. Printed monopoles are sometimes favored over MSAs because they are more 'wire-like' than MSAs, and so can take up less of a PCB's surface area. Most commercial designs have highly regular geometries. Included in this category, for simplicity, are antennas such as the inverted-f printed antenna, which, may or may not technically be a monopole, but has most of its key characteristics in common with a printed monopole.

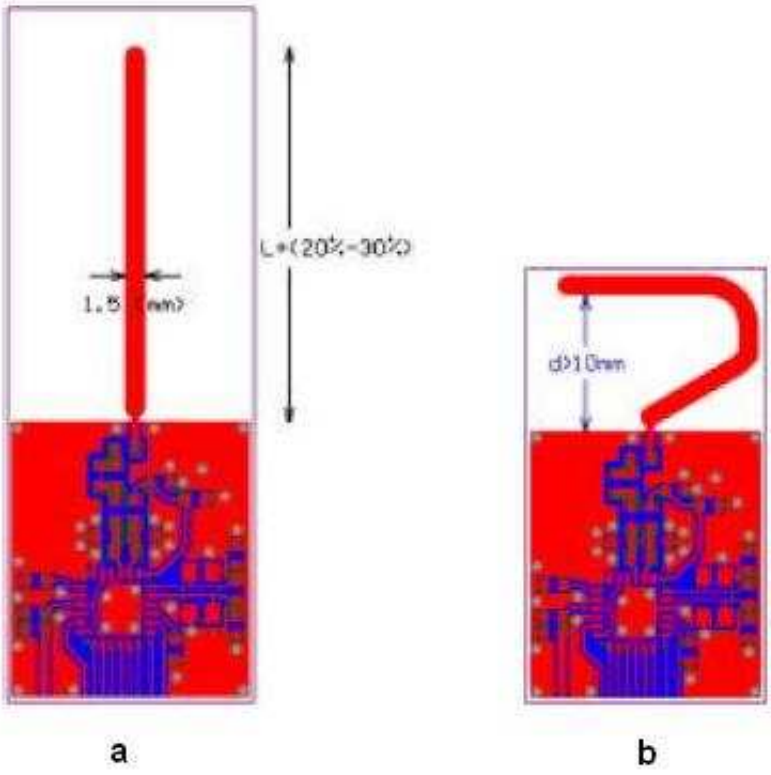


Figure 14.1: Printed monopoles: (a) straight & (b) bent. [35]

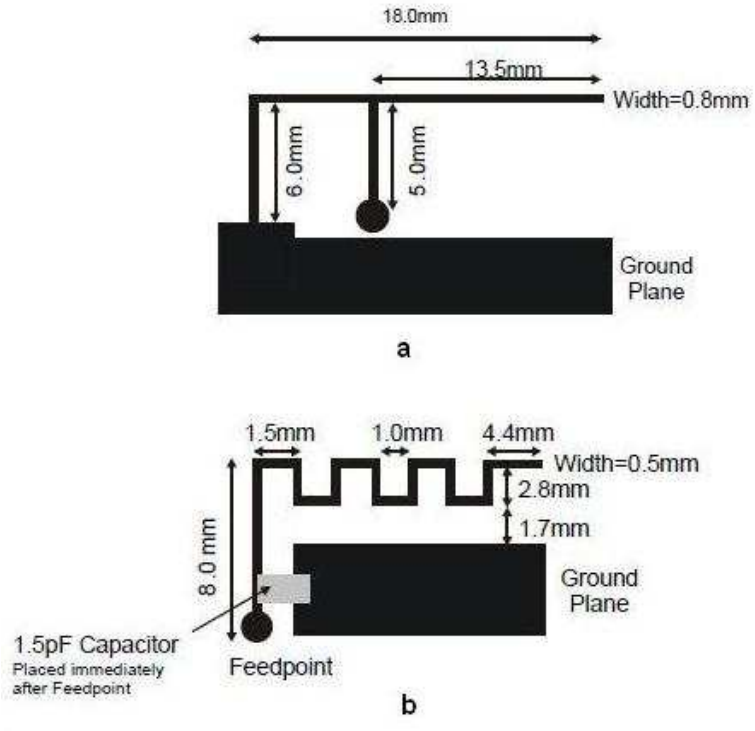


Figure 14.2: (a) inverted-f antenna & (b) meander line antenna. [36]

All of the antennas shown in Figs. 14.1 & 14.2 are intended to be printed on FR4 and are for use in the 2.4 GHz ISM band. The longest free space wavelength of interest is 12.5 cm, which corresponds to 2.4 GHz. A quarter of this wavelength is 3.125 cm, which is the approximate length of a straight monopole for use at this frequency. The antennas of Figs. 14.1 & 14.2 have been significantly compressed in size, relative to a straight monopole. However, all of their geometries are highly regular. It is probable that the use of computational optimisation techniques could result in still further miniaturisation, whilst maintaining acceptable performance.

14.3 Wire Antennas

Not all wireless communication applications require very low profile antennas. A good example of this is the internal WiFi antenna inside a laptop computer. The main body of a laptop is often at least 2cm high. By using antennas that are 3 dimensional rather than 2 dimensional (planar), it is possible that the problem of limited bandwidth could be better addressed. The relationship between antenna size and bandwidth is described in section 8.1.1. Planar antennas generally have low bandwidths because they have low volumes relative to the volume of the sphere that encloses them (radiansphere). Better utilisation of an antenna's radiansphere can lead to a higher bandwidth. A particularly good example of evolved wire antennas is that of Derek Linden. Computationally optimising wire antennas would require electromagnetic software that could accurately model complex wire structures. The *Numerical Electromagnetics Code (NEC)* would be very well suited to this purpose.

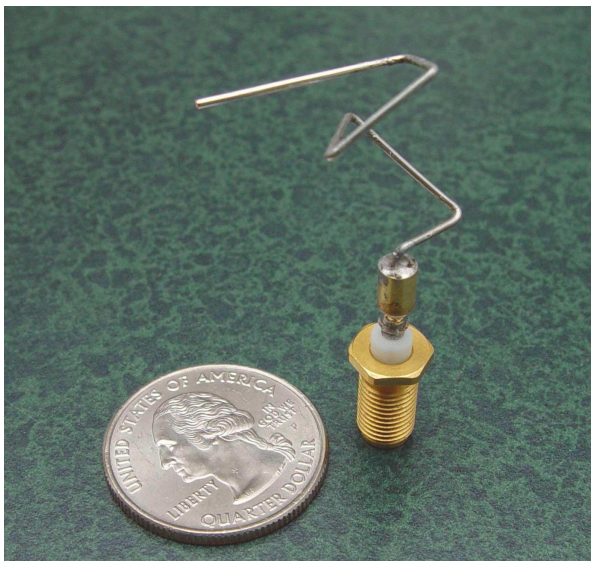


Figure 14.3: Evolved antenna for NASA's ST5 satellite. [37]

14.4 Planar Microwave Structures

There are several types of printed structure that are regularly used in RF circuits. These include: filters, baluns, passive components (e.g. inductors) and directional couplers. As is the case with printed antennas, their geometries are usually highly regular, as can be seen in the example of Fig. 14.4. The computational optimisation of planar microwave structures could well lead to novel geometries, with potentially improved characteristics. Accurate modeling, and extraction from the software, of the relevant parameters, such as S parameters (S_{11}, S_{21}, S_{12} & S_{22}), would be a prerequisite.

Figure 14.4: Edge coupled microstrip Tx line band pass filter. [38]

References

[1] Dr Bo Yuan, "The web page of max-set of gaussians landscape generator," http://boyuan.global-optimization.com/boyuan/lgmvg/.

[2] Wikipedia (http://en.wikipedia.org), "Gradient descent," .

[3] Virtual Institute of Applied Science (www.vias.org), "Optimization methods: Gradient descent," *Teach/Me Data Analysis - Multivariate Data - Optimization - Survey of Methods - Gradient Descent Methods.*

[4] Wikipedia (http://en.wikipedia.org), "Genetic programming," .

[5] V. K. Vasselin and J. F. Miller, "The advantages of landscape neutrality in digital circuit evolution," *Proceedings of the Third International Conference on Evolvable Systems: From Biology to Hardware. Lecture Notes in Computer Science*, vol. 1801, pp. 252 – 263, 2000.

[6] Julian F. Miller and Stephen L. Smith, "Redundancy and computational efficiency in cartesian genetic programming," *IEEE Transactions on Evolutionary Computation*, vol. 10, no. 2, pp. 167–174, April 2006.

[7] Molecular Nano-Optics and Spins Group, "Light waves and photons web page," *Molecular Physics Faculty, Leiden University*, http://www.molphys.leidenuniv.nl/monos/smo/index.html.

[8] Wikipedia (http://en.wikipedia.org), "Steradian," .

[9] John N. Sahalos, *Orthogonal Methods for Array Synthesis: Theory and the ORAMA Computer Tool*, John Wiley, 2006.

[10] The Beis Family Homepage (http://www.beis.de), "Universal filter," .

[11] Wikipedia (http://en.wikipedia.org), "Polarization," .

[12] C. A. Balanis, *Antenna Theory*, John Wiley, 2 edition, 1997.

[13] Thomas Okoth Moses, "A survey of antennas for wireless communication systems (masters thesis)," *Department of Electrical and Computer Engineering, Florida State University*, 2005, http://etd.lib.fsu.edu/theses/available/etd-04102004-143656/unrestricted/Chapter3.pdf.

[14] David M. Pozar, "A review of aperture coupled microstrip antennas: History, operation, development, and applications," *http://www.ecs.umass.edu/ece/pozar/aperture.pdf*.

[15] "http://microstrip-antennas.blogspot.com/2008/06/feeding-methods.html," .

[16] Anders G. Derneryd, "Linearly polarized microstrip antennas," *IEEE Transactions on Antennas and Propagation*, pp. 846–851, November 1976.

[17] Prof. Dirk Heberling, "Planar antennas lecture notes," 2003, http://www.ihf.rwth-aachen.de.

[18] D. Orban and G.J.K. Moernaut, *The Basics of Patch Antennas*, Orban Microwave Products (www.orbanmicrowave.com).

[19] John Huang, "The finite ground plane effect on the microstrip antenna radiation patterns," *IEEE Transactions on Antennas and Propagation*, vol. AP-31, no. 4, pp. 649–653, July 1983.

[20] Keith R. Carver, "Microstrip antenna technology," *IEEE Transactions on Antennas and Propagation*, vol. AP-29, no. 1, pp. 2–24, January 1981.

[21] Girish Kumar and K. P. Ray, *Broadband Microstrip Antennas*, Artech House, 2003.

[22] Punit S. Nakar, *Design of a compact microstrip patch antenna for use in wireless/cellular device*, 2004, Masters Thesis Report.

[23] Kin-Lu Wong, *Compact and Broadband Microstrip Antennas*, John Wiley, 2002.

[24] Kin-Lu Wong and Wen-Hsiu Hsu, "A broad-band rectangular patch antenna with a pair of wide slits," *IEEE Transactions on Antennas and Propagation*, vol. 49, no. 9, pp. 1345–1347, September 2001.

[25] Chun-Kun Wu and Kin-Lu Wong, "Broadband microstrip antenna with directly coupled and parasitic patches," *Microwave and Optical Technology Letters*, vol. 22, no. 5, pp. 348–349, September 1999.

[26] Frank J. Villegas et al, "A parallel electromagnetic genetic-algorithm optimization (ego) application for patch antenna design," *IEEE Transactions on Antennas and Propagation*, vol. 52, no. 9, pp. 2424–2435, September 2004.

[27] J. Michael Johnson and Yahya Rahmat-Samii, "Genetic algorithms and method of moments (ga/mom) for the design of integrated antennas," *IEEE Transactions on Antennas and Propagation*, vol. 47, no. 10, pp. 1606–1614, October 1999.

[28] Alatan et al, "Use of computationally efficient method of moments in the optimization of printed antennas," *IEEE Transactions on Antennas and Propagation*, vol. 47, no. 4, pp. 725–732, April 1999.

[29] Steven R. Best, "The performance properties of electrically small resonant multiple-arm folded wire antennas," *IEEE Antennas and Propagation Magazine*, vol. 47, no. 4, pp. 13–27, August 2005.

[30] Wikipedia (http://en.wikipedia.org), "Interference," .

[31] Athanasios G. Kanatas and Philip Constantinou, "A propagation prediction tool for urban mobile radio systems," *IEEE Transactions on Vehicular Technology*, vol. 49, no. 4, pp. 1348–1355, July 2000.

[32] M.C. Walden and F.J. Rowsell, "Urban propagation measurements and statistical path loss model at 3.5 ghz," *IEEE Antennas and Propagation Society International Symposium 2005*, vol. 1A, pp. 363–366, July 2005.

[33] Juha Laurila et al, "Wide-band 3-d characterization of mobile radio channels in urban environment," *IEEE Transactions on Antennas and Propagation*, vol. 50, no. 2, pp. 233–243, February 2002.

[34] Wikipedia (http://en.wikipedia.org), "Finite-difference time-domain method," .

[35] Nordic Semiconductor, *Quarter Wave Printed Monopole Antenna for 2.45GHz White Paper*, www.nordicsemi.no, 2005.

[36] CSR, *2.4GHz Inverted-F and Meander Line Antennas Application Note*, www.csr.com, 2007, CS-101512-ANP2_2.4.

[37] Derek S. Linden & Jason D. Lohn Gregory S. Hornby, Al Globus, "Automated antenna design with evolutionary algorithms," *AIAA Space, San Jose, California, 19-21 Sept.*, 2006.

[38] Thomas H. Lee, *Planar Microwave Engineering: A Practical Guide to Theory, Measurement, and Circuits*, Cambridge University Press, 2004.

[39] Edwin K. P. Chong and Stanislaw H. Zak, *An Introduction to Optimization*, John Wiley, 2001.

[40] Wikipedia (http://en.wikipedia.org), "Particle swarm optimization," .

[41] Nanbo Jin and Yahya Rahmat-Samii, "Parallel particle swarm optimization and finite difference time domain (pso/fdtd) algorithm for multiband and wide-band patch antenna designs," *IEEE Transactions on Antennas and Propagation*, vol. 53, no. 11, pp. 3459–3468, November 2005.

[42] C. D. Gelatt S. Kirkpatrick and M. P. Vecchi, "Optimization by simulated annealing," *Science*, vol. 220, no. 4598, pp. 671–680, May 1983.

[43] J. H. Holland, *Adaptation in Natural and Artificial Systems*, Univerisity of Michigan Press, Ann Arbor, MI, 1975.

[44] Yahya Rahmat-Samii and Eric Michielssen, *Electromagnetic Optimization by Genetic Algorithms*, John Wiley, 1999.

[45] Alex J. Champandard, *AI Game Development: Synthetic Creatures with Learning and Reactive Behaviors*, New Riders, 2003.

[46] D. E. Goldberg, *Genetic Algorithms in Search, Optimization and Machine Learning*, Addison-Wesley, Reading, MA, 1989.

[47] Melanie Mitchell, *An Introduction to Genetic Algorithms*, MIT Press, 1998.

[48] J. Michael Johnson and Yahya Rahmat-Samii, "Genetic algorithms in engineering electromagnetics," *IEEE Antennas and Propagation Magazine*, vol. 39, no. 4, pp. 7–21, August 1997.

[49] Ben Kemp, *Evolutionary Optimisation for Electromagnetics Design*, 2000, PhD thesis.

[50] J. R. Koza, *Genetic Programming: On the programming of computers by means of natural selection*, MIT Press, 1992.

[51] J. F. Miller and P. Thomson, "Cartesian genetic programming," *Third European Conference on Genetic Programming. Lecture Notes in Computer Science*, vol. 1802, pp. 121–132, 2000.

[52] JF Miller et al, "Principles in the evolutionary design of digital circuits - part 1," *Genetic Programming and Evolvable Machines*, vol. 1, pp. 7–35, 2000.

[53] Andy Marvin, *Antennas Course Lecture Notes*, Department of Electronics, University of York.

[54] Charles Capps (Delphi Automotive Systems), "Near field or far field?," *Electronic Design News (www.edn.com)*, August 2001.

[55] John D. Kraus, *Antennas*, McGraw-Hill, 2 edition, 1988.

[56] Fawwaz T. Ulaby, *Fundamentals of Applied Electromagnetics*, Prentice Hall, 1999 edition edition, 1999.

[57] James S. McLean, "A re-examination of the fundamental limits on the radiation q of electrically small antennas," *IEEE Transactions on Antennas and Propagation*, vol. 44, pp. 672–676, May 1996.

[58] Randy Bancroft, *Fundamental Dimension Limits of Antennas*, Westminster, Colorado, http://www.cs.berkeley.edu/ culler/AIIT/papers/radio/antennalimits.pdf.

[59] J.R. James and P.S. Hall, *Handbook of Microstrip Antennas*, Peter Peregrinus/IEE, 1989.

[60] Robert E. Munson, "Conformal microstrip antennas and microstrip phased arrays," *IEEE Transactions on Antennas and Propagation*, pp. 74–78, January 1974.

[61] KazuHiro Hirasawa and Misao Haneishi, *Analysis, Design and Measurement of Small and Low-Profile Antennas*, Artech House, 1992.

[62] Roger F. Harrington, *Time-Harmonic Electromagnetic Fields*, John Wiley, 1961.

[63] D. Solomon Y. T. Lo and W. F. Richards, "Theory and experiment on microstrip antennas," *IEEE Transactions on Antennas and Propagation*, vol. AP-27, pp. 137–145, March 1984.

[64] Daniel D. Harrison Willian F. Richards, Yuen T. Lo, "An improved theory for microstrip antennas and applications," *IEEE Transactions on Antennas and Propagation*, vol. AP-29, pp. 38–46, January 1981.

[65] Ramesh Garg, *Microstrip Antenna Design Handbook*, Artech House, 2001.

[66] R. A. Abd-Alhameed, "Procedure for analysis of microstrip patch antennas using the method of moments," *IEE proceedings. Microwaves, antennas and propagation*, vol. 145, no. 6, pp. 455–459, December 1998.

[67] S. D. Targonski Rod B. Waterhouse and D. M. Kokotoff, "Design and performance of small printed antennas," *IEEE Transactions on Antennas and Propagation*, vol. 46, no. 11, pp. 1629–1633, November 1998.

[68] Ikmo Park and R Mittra, "Aperture-coupled quarter-wave microstrip antenna," *Antennas and Propagation Society International Symposium. AP-S. Digest*, vol. 1, pp. 14–17, July 1996.

[69] L.J. Chu, "Physical limitations of omni-directional antennas," *Journal of Applied Physics*, vol. 19, pp. 1163–1175, 1948.

[70] Roger Harrington, "Effects of antenna size on gain, bandwidth, and efficiency," *Journal of the National Bureau of Standards*, vol. 64D, no. 1, pp. 1–12, January-February 1960.

[71] James S. McClean, "A re-examination of the fundamental limits on the radiation q of electrically small antennas," *IEEE Transactions on Antennas and Propagation*, vol. 44, no. 5, pp. 672–675, May 1996.

[72] Harold Wheeler, "The radiansphere around a small antenna," *Proceedings of The I.R.E. (IEEE)*, vol. 47, no. 8, pp. 1325–1331, August 1959.

[73] Roger F Harrington, "On the gain & beamwidth of directional antennas," *IRE (IEEE) Transactions on Antennas and Propagation*, vol. 6, no. 3, pp. 219–225, July 1958.

[74] Eugene Hecht, *Optics*, Addison Wesley, 2 edition, 1987.

[75] C. W. I. Pistotius D. A. McNamara, *Introduction to the Uniform Geometrical Theory of Diffraction*, Artech Print on Demand, 1 edition, 1990.

[76] S. Saunders, *Antennas and Propagation for Wireless Communication Systems*, John Wiley, 2000.

[77] Sylvain Ranvier, "Course notes: Physical layer methods in wireless communication systems (s-72.333)," November 2004, http://www.comlab.hut.fi/opetus/333/2004_2005_slides/Path_loss_models.pdf.

[78] Vinko Erceg et al, "An empirically based path loss model for wireless channels in suburban environments," *IEEE Journal on Selected Areas in Communications*, vol. 17, no. 7, pp. 1205–1211, July 1999.

[79] Pertti Vainikainen Lasse Vuokko and Jun ichi Takada, "Clusters extracted from measured propagation channels in macrocellular environments," *IEEE Transactions on Antennas and Propagation*, vol. 53, no. 12, pp. 4089–4098, December 2005.

[80] Gunter Rudolph, "Convergence analysis of canonical genetic algorithms," *IEEE Transactions on Neural Networks*, vol. 5, no. 1, pp. 96–101, January 1994.

[81] Julian F. Miller, "An empirical study of the efficiency of learning boolean functions using a cartesian genetic programming approach," *Proceedings of the First Genetic and Evolutionary Computation Conference (GECCO 99)*, pp. 1135–1142, 1999.

[82] K. S. Yee, "Numerical solution of initial boundary value problems involving maxwell's equations in isotropic media," *IEEE Transactions on Antennas and Propagation*, vol. 14, pp. 302307, 1966.

[83] R. A. Woodhouse, *Evolved Antennas*, 2005, MEng Project Report.