A NOVEL NETWORK REPRESENTATION FOR MODELLING THE ELECTRONIC WAVEFUNCTION IN TWO DIMENSIONAL QUANTUM SYSTEMS

By
Jeremy Pepin

A Thesis
presented to Glasgow University in partial fulfillment of the requirements for entry to the degree of Doctor of Philosophy.
Department of Electronic and Electrical Engineering

To the use of science to fulfill human need, not profit.

CONTENTS Page

LIST OF FIGURES

ACKNOWLEDGEMENTS

I Would Like to thank Professor John Barker for his help and collaboration throughout the course of this work and for his confidence in me during periods of difficulty. I also thank Dr Mike Finch and Mike Laughton for useful discussion on this work and related areas. The presentation of this Thesis would have been a far greater task without the use of a word processor located in Professor Steve Beaumont's laboratory. This work was partially funded by the Science and Engineering Research Council.

ABSTRACT

An overview of quantum phenomena associated with nanoelectronic structures is presented, including resonant tunnelling and mini-band formation in vertical transport devices and channel conductance quantization and interference in lateral devices. The method of construction of these structures is briefly described.

Methods of calculating the transmission coefficient are reviewed. In one dimension the transfer matrix method is described and also two derivatives of the approach for circumventing the numerical instability encountered when calculating the wavefunction. In two dimensions an un-coupled matching states method and an asymptotic time dependent method are described.

As an alternative to the above methods a coupled network theory is presented for the first time which genuinely represents the 2D time independent electronic wavefunction. Nodes on the network are described by a unitary scattering matrix from which a 2D transfer matrix is derived, connecting lines on the network. The scattering matrix for the whole system is created by combining the 2D scattering matrices for each line, themselves derived from the transfer matrices. The use of the scattering matrix is necessary to ensure numerical stability and current conservation.

It is shown that the bandstructure of the network is essential to creating a genuine 2D model whilst at the same time introducing a perturbing influence on the manifestation of physical phenomena. The advantages over other models is the complete absence of restriction on the potential profile considered and no requirement to separate the scalar energy and potential quantities into x and y components. Also no problem with current continuity has been encountered. A major disadvantage is the large time required to calculate wavefunctions compared with the un-coupled matching states method.

The network is shown to reproduce the channel conductance quantization recently observed experimentally and is in good agreement with both a 1D analytic model and a 2D un-coupled model.

The network is applied to channels containing single and double barriers. In the

latter case the resonances are found not to coincide with those predicted by a 1D model. Also the wavefunction on resonance resembles one of the quasi states of the well but with a phase shift.

When applied to waveguides involving an interface between channels of different widths the network reveals a tendency for the wavefunction to relax to its original transverse state as it gets further from the interface. This tendency is most pronounced for a tapered junction at low energy (energy of the order of the first transverse eigenvalue). The transmission coefficient for an abrupt junction displays unusual dips above the quantization threshold of the narrow channel. Scattering into higher modes is reduced both by reducing the ratio of channel widths and by reducing the absolute lengths of the device.

Finally circle and ring devices are studied, results displaying similarities with Finch's time dependent calculations. In particular scattering into the arms of the ring is observed to be mainly into the first mode if the energy is low and mainly into the third mode if the energy is of the same order as the third transverse eigenvalue of the channel. The tendency to relax into the original transverse state still operates over the whole device.

1 INTRODUCTION

1.1 Historical Background

The purpose of this chapter is to locate the topic of this thesis in the general context of semiconductor physics by way of reviewing the historical route to contemporary technology in this field and drawing attention to the phenomena involved in understanding the operation of todays nanoelectronic devices.

The history of the miniaturisation of electronics began in the 1940's. The mechanical operation (relays) and memory (ferrite cores) was expensive to manufacture and unreliable in service. The impetus for change was thus economic as well as functional; of course such a macroscopic implementation prevented significant advances in speed or memory size.

This dual incentive drove the industry through thermionic valves and magnetic disk memory before taking a leap forward with the introduction of the discrete transistor in the late 50's. This quantitative trend continued with the planar processing techniques becoming available in the 60's, allowing integration of devices and miniaturisation on a far greater scale. It was realised as early as this, by Landauer [1961], that this quantitative change could become a qualitative one, predicting a limit to the extent of the scaling down by reference to the power dissipation and the minimum number of 'particles' required to represent information.

This qualitative difference could be summed up as a move into the quantum regime where geometry of devices was defined on a scale of the order of the de Broglie wavelength of the electrons and the discreet nature of the world could no longer be ignored. This 'problem' was double - edged. A new understanding of integrated circuits became necessary and the design of VLSI became a discipline in itself [see e.g. Mead and Conway 1977]. On the other hand the new phenomena now available for scrutiny were embraced by physicists.

Since the first demonstrable resonant tunnelling devices in the early 70's [e.g. Esaki and Chang 1970] quantum effects have been predicted and realised by the use of band - gap engineering and more recently by lateral confinement of a two - dimensional electron gas (2DEG). It is at this point that the subject of the present work is introduced to the story. We are now in a position where 2DEG formation, with geometry defined on a scale less than the mean free path of the electron, allows ballistic transport to be studied in practice. The purpose of the network model proposed here is to interpret and predict the behaviour of these devices.

1.2 Quantum Phenomena

1.21 Introduction

In this section we briefly review the novel phenomena that are encountered in the ultra - small structures presently being produced. The devices can be grouped into two categories, according to whether transport is in the direction of growth (vertical) or parallel to it (lateral). We first describe the method of construction of the vertical devices and then the phenomena associated with vertical transport (section 1.22) and then similarly for lateral devices (section 1.23).

1.22 Vertical Transport Architecture and Associated Quantum Phenomena

The growth of epitaxial layers of semiconductor by Molecular Beam Epitaxy (MBE) or Metal-Organic Chemical Vapour Deposition (MOCVD) allows devices to be built up of alternating material and therefore alternating band-gap. This has the effect of imposing a synthesized potential profile which may be defined to within a single layer of molecules of either component semiconductor. The two components of this Heterostructure must have roughly equal lattice spacing and GaAs/$Al_xGa_{1-x}As$ is a common choice although others are in use. There is argument about the potential discontinuity appearing in the new conduction band edge; Dingle [1975] originally proposed 85% of the total band - gap difference but more recently a figure of about 65% has attained a large degree of acceptance [e.g. Okumura et al 1985, Miller et al 1984].

The phenomenon of primary interest here is that of resonant tunnelling, first

proposed in 1951 by Bohm [1951] but not seriously addressed until the early 70's [Esaki and Tsu 1970, Lebwohl and Tsu 1970] when the construction of suitable devices became feasible. We give a mathematical treatment of 1D resonant Tunnelling in chapter 2 and restrict ourselves here to a physical description.

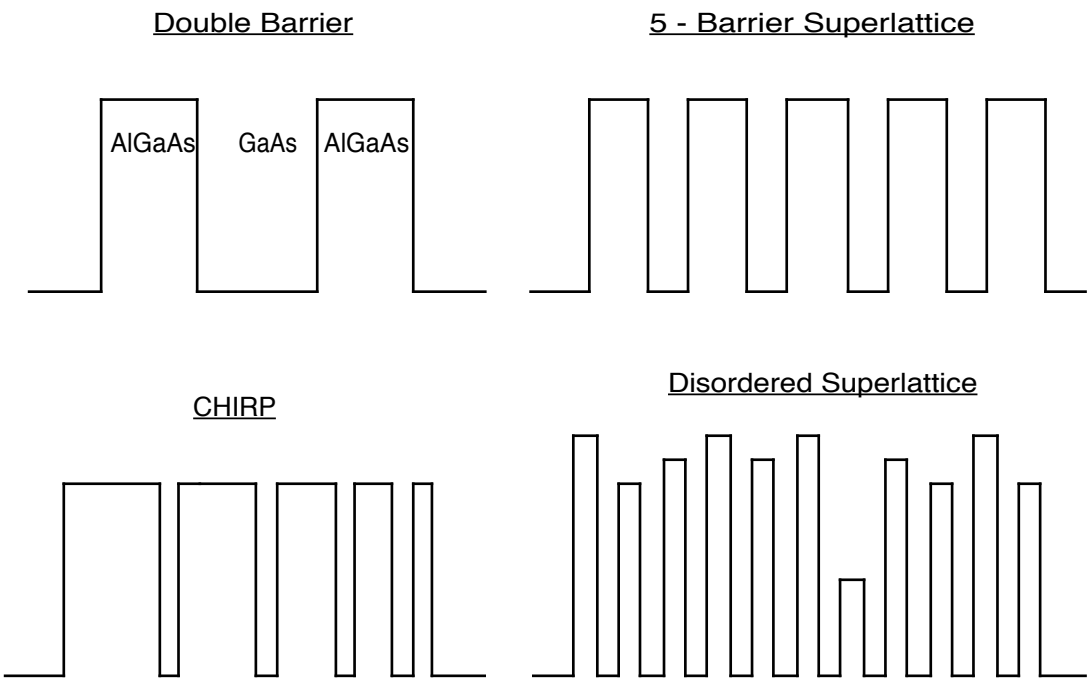
The minimum device for resonant tunnelling is the double - barrier heterostructure, constructed typically by a layer of GaAs sandwiched between two layers of AlGaAs (there are of course contact regions at both ends of the device). A typical transmission versus energy curve is shown in Fig 1 for a pair of 50Å wide barriers separated by a 25Å well. The barrier height is 250meV. The resonant peak in the transmission at $\approx$ 170meV - below the barrier height - is due to coupling between the quasi - bound state in the well and continuum states in the contacts, occurring when the Fermi energy of the latter coincides with the bound - state energy. At energies between E_1 and E_2 the device displays Negative Differential Resistance (NDR), that is an increase in voltage across the device results in a decrease in current.

An important extension of the double - barrier is the superlattice; constructed with any number of barriers and wells. The double - barrier may be said to be the shortest member of the superlattice family. In the superlattice the quasi - bound states of each well couple with each other to form extended states across the device which, by the exclusion principle, separate in energy to form a mini - band. Of course for a finite number of wells the energy levels in the mini - band are discreet.

The literature on superlattices is comprehensive but we note here a few variations on the theme. The width of the barriers can be varied across the device to produce what is known as a CHIRP structure [Summers et al 1987] which has advantages for use as an energy filter. Another use is as millimetre wave sources [Grondin and Wang 1986], or as infrared detectors [Levine et al 1987]. Bloch Oscillations are suggested by Grondin et al [1985]. Cappasso [1985] gives a review of other variations . Another very important and extensively researched area is the study of disorder, the periodicity of the superlattice being destroyed in some way leading to possible localization of the wavefunctions. There is a vast amount of literature on this subject; one of the early workers in the field was Borland [1963]. A more recent review is given by Lee and Ramakrishnan [1985] and Pendry has been one of the more prolific contributors to the subject [e.g. Pendry 1987].

Some of the variations on the potential profile are shown in fig 2.

Fig 2 One-Dimensional Potential Profiles



A few of the problems involved in modelling these heterostructures are noted. The non - parabolic correction to the effective mass in the mini - bands is discussed by williams [1983] and by Welch Wicks and Eastman [1984] and the value of the bulk effective mass by williams et al [1983] and by Schulman [1985]. The accuracy of the effective mass approximation comes under scrutiny from Marsh and Inkson [1984] and Collins Lowe and Barker [1985] and self-consistency is addressed by Ricco [1986].

Modulation doping, where the doping concentration is varied in the growth direction can be used as an alternative method for synthesizing the potential profile [e.g. Esaki 1986].

1.23 Lateral Transport Architecture and Associated Quantum Phenomena

This group is really constructed using a combination of vertical and lateral geometry. The long mean free path required is obtained via the High Electron Mobility Transistor (HEMT) structure. We point out a few salient points here; for a full review see Hill [1986].

A Narrow quantum well is manufactured using MBE at the heterojunction between, for instance, un-doped GaAs and n-doped AlGaAs. The fermi level continuity assures band bending (Fig 3), since it is near the conduction band edge in the AlGaAs and in the middle of the band gap for the GaAs. If the well is narrow enough the electrons will occupy only the first subband and therefore form a 2DEG. This narrowness can be achieved by increasing the Al content of the AlGaAs but a limit of 45% is imposed by the changeover to an indirect - gap semiconductor. Mobility is enhanced by leaving a narrow layer of undoped AlGaAs adjacent to the interface, thus spatially separating electrons in the well from donor scattering centres. The order of improvement in the mean free path is from nanometres in bulk GaAs to microns in the 2DEG. The factors affecting mobility are numerous as Hill Explains, but the net result is to increase carrier concentration, as with doping, but without introducing significant scattering centres.

The potential profile in the plane of the 2DEG may be engineered to give similar structures as in the vertical transport devices. For instance a double barrier is achieved by adding two metal gates to the surface of the AlGaAs. The depletion regions associated with these will depend on the voltage applied to them and thus the effect in the 2DEG and therefore the potential barrier can be controlled. This is shown schematically in Fig 4.

Fig 3 Band - Bending at a Heterojunction

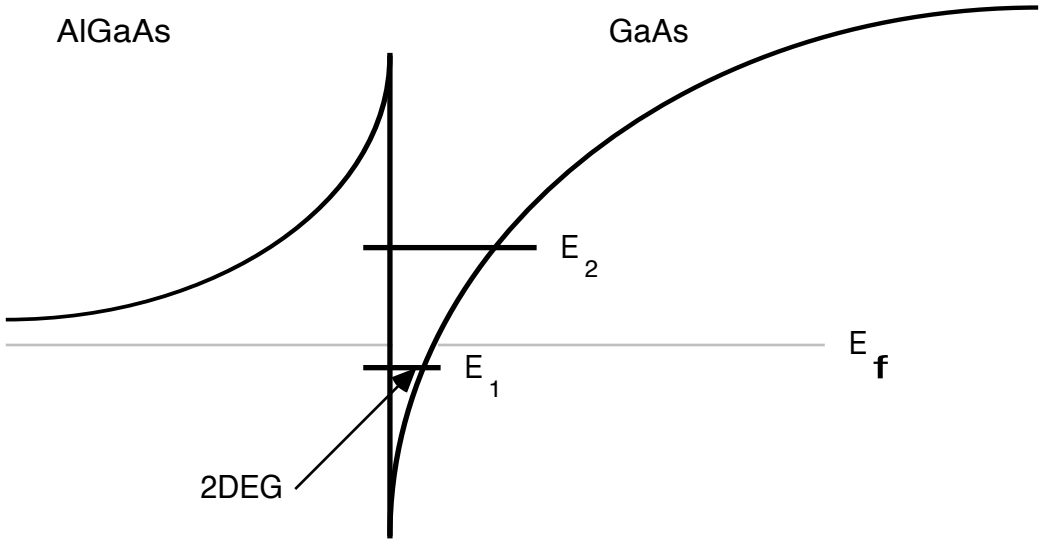
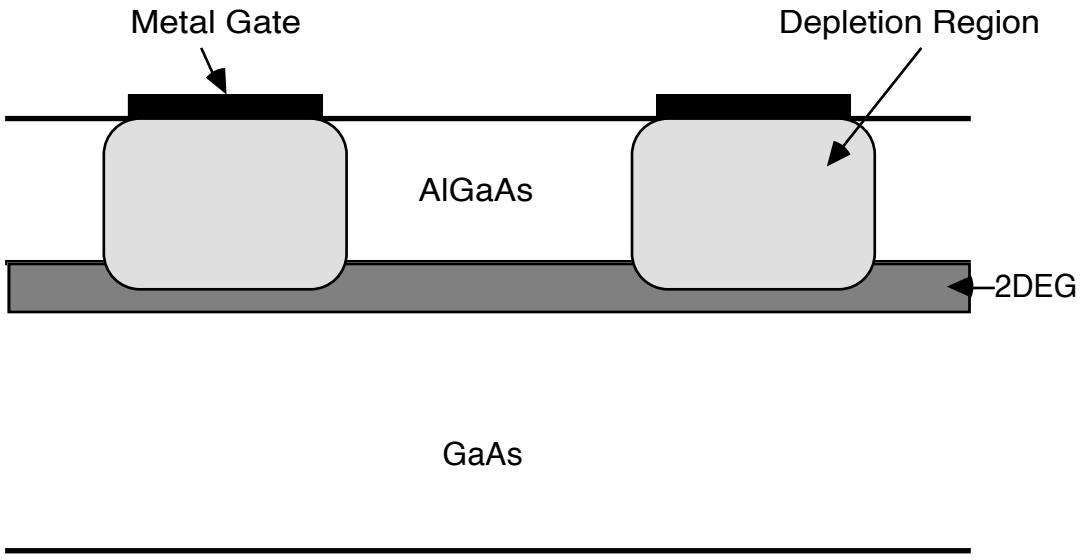


Fig 4 Gate Controlled Depletion Regions

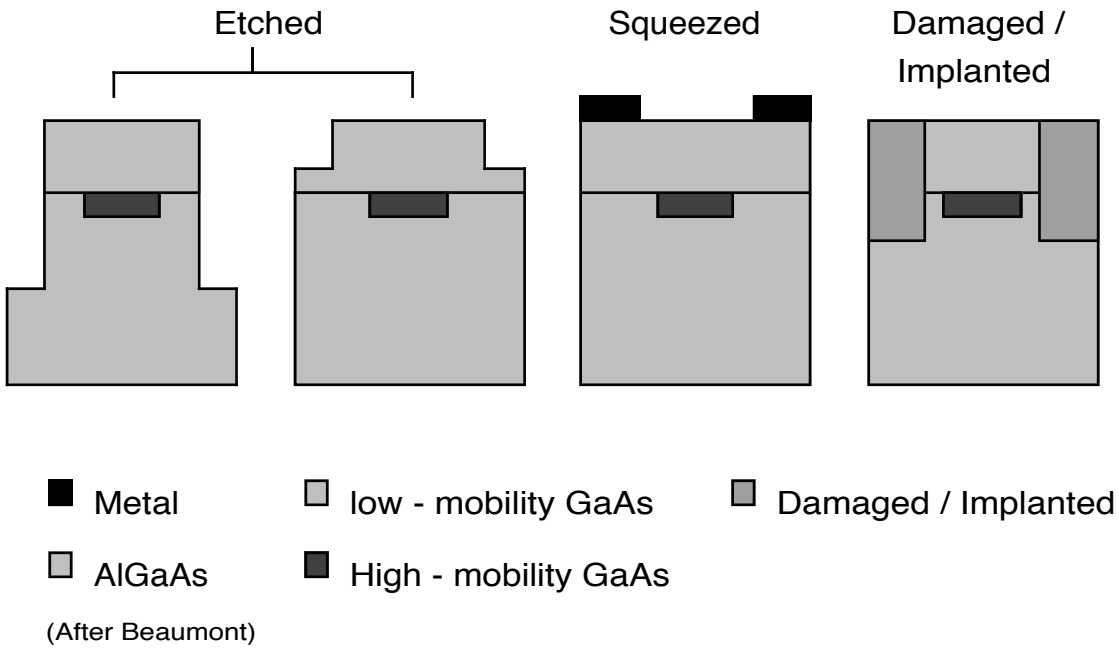


The lithographic processes are well described in the literature [e.g. Wilkinson and Beaumont 1986]. Two basic methods exist, both requiring the irradiation of a resist to be spatially controllable on a scale of the geometry of the device. The first

method employs resist to protect the material underneath it, thus in Fig 4 the whole area between the source and the drain would be covered in metal followed by resist. The resist would then be exposed, over the gate regions with a negative resist or away from the gate regions for positive resist. Etching would then remove both resist and metal from unwanted areas and finally the resist remaining over the metal gates would be dissolved away. In the alternative method, lift - off, the resist is patterned by exposure and etching first to remain only in areas where metal is not required. A layer of metal is then put down over the whole area before the remaining resist is dissolved away, leaving metal contacts on the AlGaAs surface.

Lift - off processing is favoured for fine detail and the radiation employed needs to be electrons, the limit of longer wavelength light being about 0.5 microns resolution. This accuracy is important for the construction of quasi-one-dimensional channels or wires. Beaumont et al [1988] detail three ways of doing this and put the limits of resolution, under ideal conditions, at 10nm linewidth and 50nm pitch (centre to centre distance between lines). The three methods of construction are explained in Fig 5. A problem with all of them is to determine the real width of the channel created in the 2DEG.

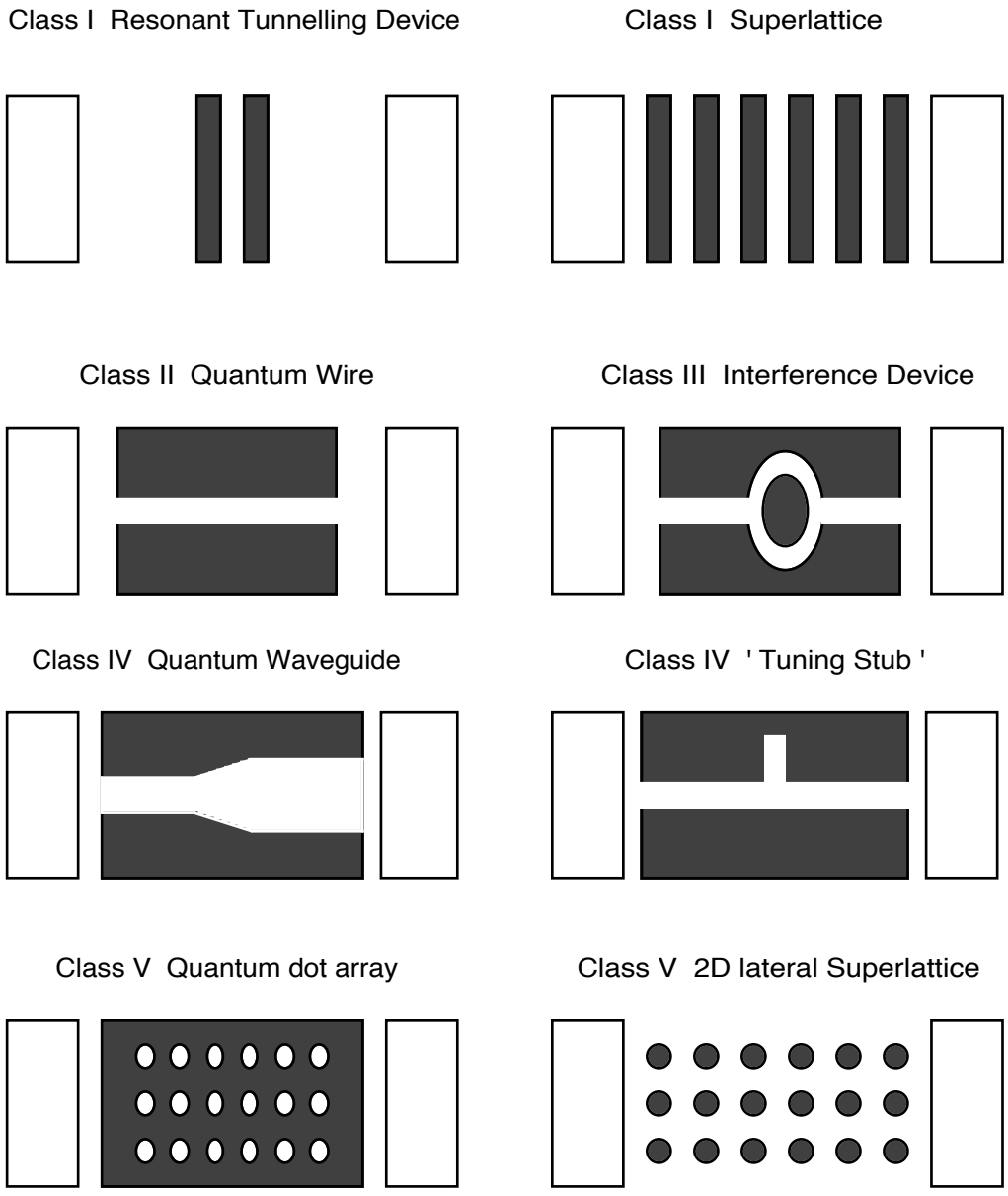
Fig 5 Three methods of Construction of One-Dimensional Channels



(After Beaumont)

Devices in the lateral group may be further subdivided into classes as shown in Fig 6.

Fig 6 Lateral Devices



(after Barker)

The class I devices display the same phenomena as the vertical devices, the dimension transverse to the direction of transport being large enough to ignore quantization in that direction. Class II devices are subject to the lateral quantization in the channel and are one - dimensional or quasi - one - dimensional. These

definitions require explanation. The general preference in the literature seems to be that a channel is 1D if the quantization is strong enough in two directions that it may be assumed that only the ground state in those directions is occupied. If in one other direction more than one state is occupied but level separations are nevertheless significant then the channel is said to be quasi - 1D. However we feel that these definitions are inadequate. It is possible to have a narrow channel where only one lateral mode is occupied but the potential profile ensures that only a 2D model will correctly describe the wavefunction. For instance a channel with a double barrier is not expected to behave exactly as the 1D resonant tunnelling device since in general structure in the transmission curve will be influenced by the lateral confinement. In this sense it is the very wide channels, where the lateral motion is effectively free, that display 1D characteristics.

Lateral quantization of the conductance was first shown experimentally by Thornton et al [1986] and more recently by Van Wees et al [1988] and by Wharam et al [1988] who both relate it to a variation of the Landauer conductance formula [Landauer 1961, Stone and Szafer 1988], finding quantization in units of $2e^2/h$. Both groups use the electrostatic squeezing between metal gates on the surface of the AlGaAs to define their channels. Varying the gate voltage alters the effective width of the channel and as the subband levels are increased with a decrease in width they are successively lifted above the Fermi energy and depopulated. This is the cause of the stepwise decrease in conductance. The point contact geometry employed by Van Wees was first proposed as early as 1965 by Sharvin [1965] for the study of ballistic transport in metals but the advantage of being easily able to control the width is only possible via the electrostatic squeezing technique available in the semiconductor heterostructure. Wharam follows Sharvin's classical treatment of ballistic transport (i.e. transport in which no scattering of any kind is encountered). The current due to a single channel (e.g. one subband) is simply given by:

$$I_m = n\, e\, \delta v \qquad (1.1)$$

where n is the number of carriers per unit length and δv is the increase in velocity acquired during transit. This leads to the multi - channel conductance:

$$G = \sum 2e^2/h \qquad (1.2)$$

where the summation is over all occupied subbands. Van Wees points out that this

classical result can be considered as a special case of the quantum mechanical version in which the transmission coefficient for each subband is unity:

$$G = (2e^2/h) \sum |t_m|^2 \qquad (1.3)$$

This raises the question of the influence of the contact/channel junction on the transmission coefficient and this is taken up by Kirczenow [1988] and by Van Der Marel and Haanappel [1989]. Both groups find oscillatory fine structure in the transmission coefficient curve and explain it by reference to ' end effects '. We can be more quantitative; the observed peaks from our network result coincide exactly with those obtained from a 1D calculation of the transmission coefficient assuming a barrier height equal to the confinement energy of the channel and a length equal to the channel length.

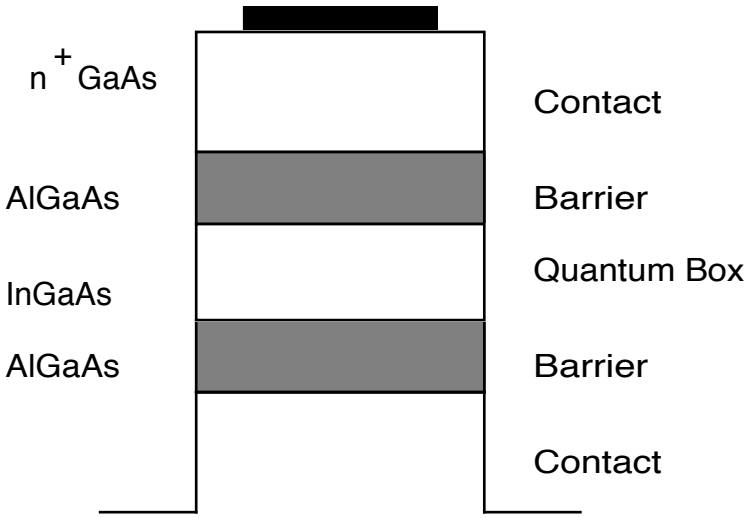
Berggren et al [1986] describe the effect of a magnetic field on ballistic conductance. The effect is dimensionally dependent. In a 1D channel a flux in the plain of the 2DEG but transverse to the subbands raises energy levels and the magneto - conductance is stepped, as successive subbands are pushed through the Fermi energy. The effect is less pronounced for a flux parallel to the subbands. In 2D, by contrast, only a parallel field has this effect. Either orientation has another effect, to lift the spin degeneracy thus producing extra quantized levels in the conductance. Both these magnetic phenomena are demonstrated by Wharam et al [1988].

We return to Kirczenow's results and to the thorny subject of the Landauer formulae in more detail in chapter 4.

The addition of barriers to the channel could result in mixing of the lateral modes and in any case the resonant tunnelling is modified by the lateral confinement. This is discussed by Bryant [1988] in relation to a vertical device grown by Reed et al [1988]. The device is shown in Fig 7, energy separation of the lateral modes being much less than for the two resonances in the vertical direction. Fine structure was observed on the resonant tunnelling curve and Bryant offers two explanations for this, under the general title of Quantum Box Resonant Tunnelling (QBRT). Diagonal Tunnelling (no inter - subband scattering) can produce one fine - structure peak for every contact subband occupied on each vertical resonance curve. If off - diagonal tunnelling is taken into account a *set* of peaks occurs for each occupied contact subband with possible contributions from all lateral states inside the quantum box.

Smith et al [1989] have made measurements of the conductance for a lateral QBRT device where the box is square so that quantization in both directions of the box are the same and the energy levels are degenerate.

Fig 7 Quantum Box Device



The class III devices exploit interference effects. A wavepacket entering the ring from the left will split into two coherent parts, one travelling round each half of the ring. If the path length to the output wire on the right is the same for both parts they will interfere constructively and the device is on. If however the relative path length is altered they interfere destructively and the device is off. This change of path length can be affected by means of a magnetic flux threading the central hole, creating a vector potential along the paths. This changes the phase of each wavepacket by equal and opposite amounts proportional to the integral of the potential along the path. This effect was first proposed by Aharonov and Bohm [1959]. Recently numerical modelling of this effect has been carried out by Finch [1989].

Very little attention has so far been devoted to class IV devices. Clearly a potential step is involved at the junction between two channels of different widths since for any particular mode the lateral confinement energy is different. For a tapered junction the potential difference becomes graded over a finite distance. Frohne and

Datta [1988] predict very little reflection for electrons injected into a wide contact from a narrow channel, in agreement with Kriman and Ruden's [1985] study with an infinitely wide contact . Barker and Laughton [1989] expect to be able to exploit analogy with microwave waveguide theory to model these devices .

Quantum dot arrays (class V) have been fabricated by several groups, via lateral processing of quantum well material. Temkin et al [1987] used milling techniques on InGaAs/InP and Cibert et al [1986] used ion implantation , both groups reporting increased efficiency of the photo - luminescence. The underlying reason for this phenomenon is the change in density of states with dimensionality; for 2D it is flat, for 1D it rises as a step function and decays as $E^{-1/2}$, whilst for 0D (quantum dots) it is a delta function. The absorption is hence spiked for the quantum dot and the exciton intensity is increased. The alternative class V device forms a network of 1D or quasi - 1D channels.

2 TRANSMISSION COEFFICIENT METHODS

2.1 Introduction

In this chapter we examine existing transmission coefficient methods in one and two dimensions. In section 2.2 we look at 1D models. These may be useful in themselves e.g. for class I devices in which the lateral energy spacing is effectively a continuum and the potential profile is constant in that direction, or in the special case of a narrow channel with no potential and therefore no interaction between confinement effects in the two directions. The 1D models are also required as an integral part of some of the 2D methods. In section 2.3 we detail two ways of modelling structures in 2D and note their advantages and disadvantages.

2.2 One Dimensional Models

2.21 Transfer Matrix

This work is primarily concerned with systems in which both ends are open and all states are current carrying (although the network method has limited application to bound states). For this reason models that are geared to bandstructure calculations, such as Tight - Binding [Rucker et al 1986] or envelope function [Bastard 1981] are not of interest here. The continued fraction method developed by Vigneron and Lambin [1980] is an excellent way of obtaining the transmission curve but is not amenable to calculating the wavefunction. We therefore devote this section to a description of the transfer matrix method and two derivatives of it.

The idea of the transfer matrix, or T-matrix, is described by Cohen - Tannoudji, Die and Laloe. The concept is very simple. The wavefunction on one side of a region of space is related to the wavefunction on the other side by decomposing the wavefunction into component parts.

The T-matrix appears quite generally in the form:

$$\begin{bmatrix} \mathbf{a}_I \\ \\ \mathbf{b}_I \end{bmatrix} = T \begin{bmatrix} \mathbf{a}_{III} \\ \\ \mathbf{b}_{III} \end{bmatrix} \qquad (2.1)$$

Where the wavefunction in region I is given by the Column vector [$\mathbf{a}_I$ $\mathbf{b}_I$] and the **a**,**b** may themselves be column vectors.

Here we are concerned with 1D so we may write the wavefunction as $\Psi_x = A_x + B_x$ and the **a**,**b** are complex numbers. In order to make use of this concept for arbitrary potentials, the region must be discretised; this is done most conveniently by representing it as a series of square barriers of varying height. An infinitesimal slit is imagined to exist between successive barriers and each transfer matrix relates the wavefunction on either side of one of these barriers. This is not necessary, the T-matrix could be defined across each interface and the slits done away with, but it saves time and is not intrinsically less accurate. The elements of the T-matrix are obtained from the matching conditions at an interface at x=X:

$$\Psi(X^-) = \Psi(X^+) \qquad (2.2)$$

$$(1/m(X^-))\, d\Psi(X^-)/dx = (1/m(X^+))\, d\Psi(X^+)/dx \qquad (2.3)$$

and the form of the wavefunction is given by the superposition mentioned above:

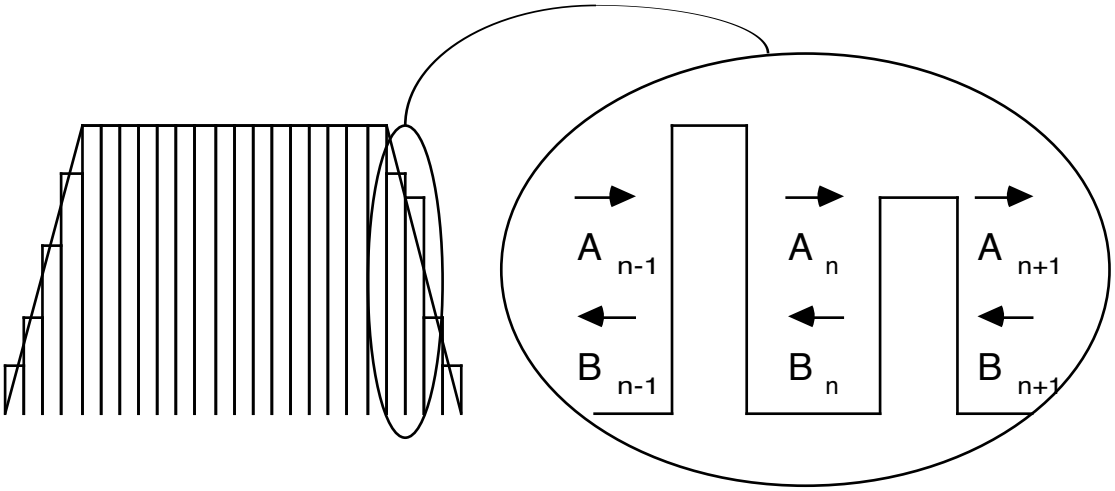
$$\Psi_x = a_x e^{ikx} + b_x e^{-ikx} \qquad (2.4)$$

and

$$k = [2m_x(E-V_x)]^{1/2}/h \qquad (2.5)$$

The use of the T-matrix approach to propagate through the system is explained with reference to Fig 8.

Fig 8 Use of the T-Matrix over an Arbitrary Potential



Any two matrices can be combined to relate the wavefunction across a pair of barriers:

$$\begin{pmatrix} A_n \\ B_n \end{pmatrix} = \mathbf{T}_{n,n-1} \begin{pmatrix} A_{n-1} \\ B_{n-1} \end{pmatrix} \qquad (2.6)$$

$$\begin{pmatrix} A_{n+1} \\ B_{n+1} \end{pmatrix} = \mathbf{T}_{n+1,n} \begin{pmatrix} A_n \\ B_n \end{pmatrix} \qquad (2.7)$$

and hence:

$$\begin{pmatrix} A_{n+1} \\ B_{n+1} \end{pmatrix} = \mathrm{T}_{n+1,n-1} \begin{pmatrix} A_{n-1} \\ B_{n-1} \end{pmatrix} \qquad (2.8)$$

The total system is then described by the T-matrix $\mathbf{T}_{N,0}$:

$$\begin{pmatrix} A_N \\ B_N \end{pmatrix} = \prod \mathbf{T}_{n,n-1} \begin{pmatrix} A_0 \\ B_0 \end{pmatrix} = \mathbf{T}_{N,0} \begin{pmatrix} A_0 \\ B_0 \end{pmatrix} \qquad (2.9)$$

Some authors use a different definition for the T-matrix, relating the a's and b's of equation (2.4). In this case an iteration matrix must be included to take care of the $e^{\pm ikx}$ phase factors so that the T-matrix for each barrier is not explicitly spatially dependent (of course the T-matrix depends on the potential, which is spatially dependent). This is discussed by Barker [1986].

The transmission coefficient is defined by:

$$|t|^2 = |A_N|^2/|A_0|^2 \qquad (2.10)$$

Using the boundary conditions $B_N = 0$, $A_N = 1$ we have the two equations:

$$1 = T_{11} A_0 + T_{12} B_0 \qquad (2.11)$$

$$0 = T_{21} A_0 + T_{22} B_0 \qquad (2.12)$$

From which:

$$1 = (\mathrm{Det}\, \mathbf{T}/T_{22})\ A_o \qquad (2.13)$$

For zero bias the determinant is equal to the current and is normalised to 1. Thus the transmission coefficient is:

$$|t|^2 = 1/|T_{22}|^2 \qquad (2.14)$$

For non - zero bias Collins [1985] has shown that the ratio of currents is given by:

$$|t_c|^2 = (1/|T_{22}|^2)\,(k_{III}/k_I) \qquad (2.15)$$

Unfortunately the Transfer matrix method suffers from numerical instability in its straightforward form. This is associated with the unbounded hyperbolic functions which arise when k in equation (2.5) is imaginary. Three routes have been taken to cope with this problem. One is simply to ignore solutions where the exponent is real and the magnitude is very large or very small [Mailhoit and Smith 1986, Brand and Hughes 1987]. These solutions correspond to highly localized states and so this method is not very useful if it is localized states that are being sought ! In any case we have evidence to suggest that the problem persists even for extended states (see section 2.22)

A more satisfactory approach is to effect a change of variables, eliminating the hyperbolic functions in terms of the bound sinusoidal functions. This method, due to Peres [1983] is discussed in the next sub-section. In section 2.23 we look at an alternative strategy due to Ko and Inkson [1988], based on replacing the transfer matrix with the scattering matrix.

2.22 Group Theoretical Transfer Matrix

2.221 Construction of the SO(2,1) Transfer Matrix

In fact the original transfer matrix displays the properties of a group, in particular it is pseudo - unitary and obeys the metric equation:

$$\sigma\,\mathbf{T}^{\dagger}\,\sigma = \mathbf{T}^{-1} \qquad (2.16)$$

where the metric σ is one of the Dirac matrices and $\mathbf{T}^{\dagger}$ is the Hermitian conjugate of the T-matrix. We make use of this property later. Here though we wish to make use of a different group whose commutation relations allow a change of variables via association with a higher - dimensional representation. The technique follows that expounded by Peres [1983] and a summary of the steps needed is given

below.

The 1D Schroedinger equation is first written in matrix form. To this end the wavefunction is written in terms of a combination of forward and backward propagating waves:

$$\psi = f + g \tag{2.17}$$

with,

$$f = Fe^{ikx} \ , \ g = Ge^{-ikx} \tag{2.18}$$

We can always choose to calculate the wavefunction at the points either side of the potential elements representing the continuous potential (Fig 8) and the propagation constant is therefore given unambiguously by :

$$k = (2mE)^{1/2}/h \tag{2.19}$$

In order to make use of this decomposition of the wavefunction we write f and g in terms of the elements connected by the transfer matrix :

$$f = [\psi + \psi'/ik]/2 \tag{2.20}$$

$$g = [\psi - \psi'/ik]/2 \tag{2.21}$$

a dash denoting differentiation with respect to x.

The derivative of the wavefunction can now be written in terms of the wavefunction itself. First differentiating f and g :

$$f' = ik[f - (f + g)V/2E] \tag{2.22}$$

$$g' = ik[g - (f + g)V/2E] \tag{2.23}$$

Using the substitutions :

$$U = V/E \tag{2.24}$$

$$t = -kx \tag{2.25}$$

we get,

$$i\,\Psi' = \begin{pmatrix} 1 - U/2 & -U/2 \\ U/2 & U/2 - 1 \end{pmatrix} \Psi \tag{2.26}$$

with,

$$\Psi = \begin{pmatrix} \mathbf{f} \\ g \end{pmatrix} \tag{2.27}$$

The important step is to notice that the matrix equation (2.26) can be written in terms of the following representation of the SU(1,1) group :

$$\mathbf{s}_X = \tfrac{1}{2}\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix},\ \mathbf{s}_Y = \tfrac{1}{2}\begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix},\ \mathbf{s}_T = \tfrac{1}{2}\begin{pmatrix} i & 0 \\ 0 & i \end{pmatrix} \tag{2.28}$$

Equation (2.26) then becomes :

$$\Psi' = [-U\mathbf{s}_Y - (2-U)\mathbf{s}_T]\,\Psi \tag{2.29}$$

By replacing the two-dimensional, complex representation of the SU(1,1) group with a three-dimensional, real representation of the SO(2,1) group - having the same commutation relations - and the two-component vector Ψ with a real, three-component vector **R** we can replace (2.29) by :

$$\mathbf{R}' = [-U\mathbf{S}_Y - (2 - U)\mathbf{S}_T]\,\mathbf{R} \tag{2.30}$$

with ,

$$\mathbf{S}_Y = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & -1 & 0 \end{pmatrix}, \qquad \mathbf{S}_T = \begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \tag{2.31}$$

In component form this is :

$$\begin{pmatrix} X' \\ Y' \\ T' \end{pmatrix} = \begin{pmatrix} 0 & 2-U & 0 \\ U-2 & 0 & U \\ 0 & U & 0 \end{pmatrix} \begin{pmatrix} X \\ Y \\ T \end{pmatrix} \tag{2.32}$$

Since the SO(2,1) group preserves the metric :

$$T^2 - X^2 - Y^2 = Z^2 \tag{2.33}$$

(Z constant) polar coordinates may be introduced as follows :

$$X = (T^2 - Z^2)^{1/2} \cos\phi \tag{2.34}$$

$$Y = (T^2 - Z^2)^{1/2} \sin\phi \tag{2.35}$$

This differs from the system used by Peres [1983] but the present author sees no advantage in re - introducing the hyperbolic functions which we are trying to avoid.

It will be seen from (2.32) that we have :

$$X' = (2 - U)Y \tag{2.36}$$

$$Y' = (U - 2)X + UT \tag{2.37}$$

$$T' = UY \tag{2.38}$$

It is then a simple matter to determine the equations of motion required, namely :

$$T' = U(T^2 - Z^2)^{1/2} \text{ SIN } \phi \tag{2.39}$$

$$\phi' = U\text{COS } \phi \quad T/(T^2 - Z^2) \quad - 2 + U \tag{2.40}$$

Returning to the Peres model we still need to relate the rather abstract vector **R** to the wavefunction, ψ. This is done as follows :

$$X = f^*g + fg^* = [\psi^*\psi - \psi^*\psi]/2 \tag{2.41}$$

$$Y = i[f^*g - fg^*] = [\psi^*\psi + \psi^*\psi]/2 \tag{2.42}$$

$$T = f^*f + g^*g = [\psi^*\psi + \psi^*\psi]/2 \tag{2.43}$$

By inspection we see that :

$$|\psi|^2 = X + T \tag{2.44}$$

In order to include the phase information we need to introduce a second angle, α. On substituting the general form of ψ :

$$\psi = f + g \;=\; Fe^{i\theta 1} + Ge^{i\theta 2} \tag{2.45}$$

into (2.41-2.43) we obtain :

$$\theta_1 = (\alpha+\phi)/2 \tag{2.46}$$

$$\theta_2 = (\alpha-\phi)/2 \tag{2.47}$$

The quantity $F^2 - G^2$ is proportional to the current, Z^2 and from this we can determine :

$$F = [(T + Z)/2]^{1/2} \tag{2.48}$$

$$G = [(T - Z)/2]^{1/2} \tag{2.49}$$

The equation of motion for α is found to be, on substituting (2.45 - 2.49) into (2.32) :

$$\alpha' = -(U /(T^2 - Z^2)^{1/2}) \ \cos\phi \tag{2.50}$$

2.222 Boundary Conditions and Generalisation to Bound States

The boundary conditions for a tunnelling experiment (i.e a current carrying state with Z=1) are usually specified as, on the right - hand side :

$$f = 1 \tag{2.51}$$

$$g=0 \tag{2.52}$$

With the wavefunction given by :

$$\psi = [(T + 1)/2]^{1/2} e^{i(\alpha + \phi)/2} + [(T - 1)/2]^{1/2} e^{i(\alpha - \phi)/2} \tag{2.53}$$

it is easily seen that the values :

$$\alpha = 0 \tag{2.54}$$

$$\phi = 0 \tag{2.55}$$

$$T = 1 \tag{2.56}$$

satisfy this requirement. The transmission coefficient is easily obtainable from (2.53):

$$|t|^2 = (T_{RHS} + 1)/(T_{LHS} + 1) \qquad (2.57)$$

Both the transmission coefficient and the modulus of the wavefunction may be calculated solely from the real equations of motion for ϕ and T.

Peres states that a different group must be used for the calculation of bound states. We disagree, finding that boundary conditions pertinent to bound (i.e. non-current- carrying) states may also be used within this formalism.

In the case of bound states the quantity $F^2 - G^2$ must be zero and the metric (2.33) preserved by the variables X,Y,T,Z is re-scaled. In the limit $Z \rightarrow 0$, the hyperboloid becomes a cone (Fig 9), the surface of which is the domain of the bound states. With Z=0 and taking α=0 as before the wavefunction becomes :

$$\psi = (2T)^{1/2} \, COS(\phi/2) \qquad (2.58)$$

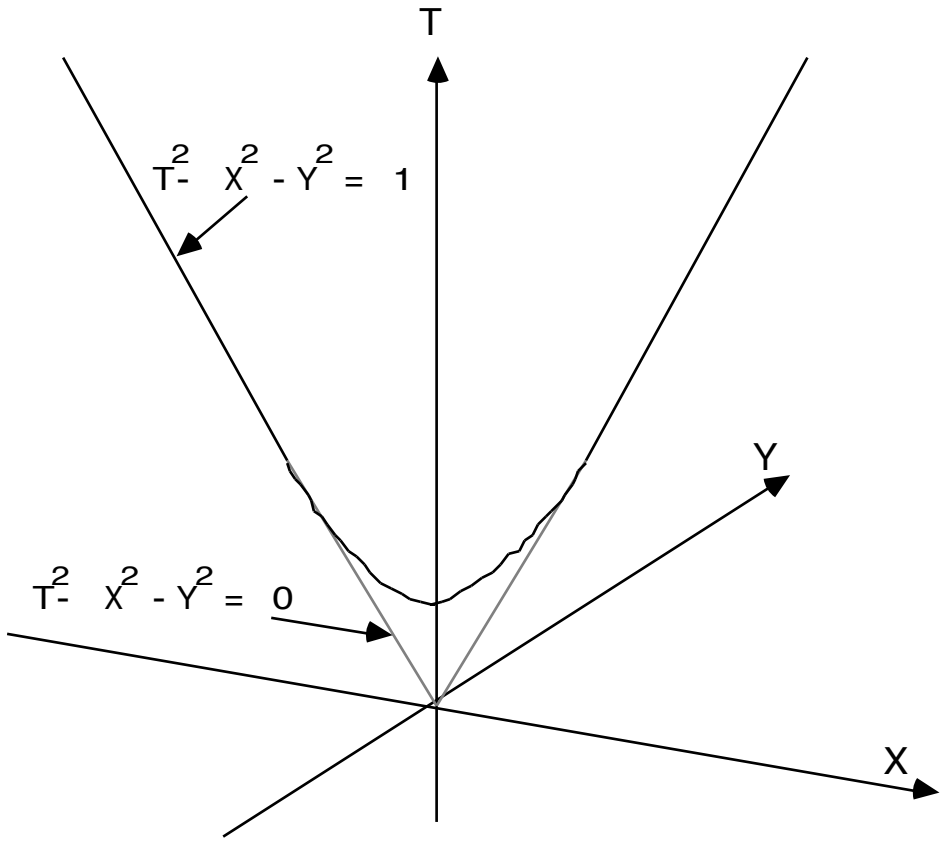
For a bound state the wavefunction vanishes at the boundaries so we must have in this case that $\phi = m\pi$ (m an integer). The value of T at the boundary is arbitrary but for X and Y to be real it must satisfy :

$$T^2 \geq Z^2 \qquad (2.59)$$

This last point is useful in the implementation of the scheme since it allows us to make full use of the range of numbers allowed on the computer by starting with a very low value of T.

To determine a bound state we need only integrate the expressions (2.39) and (2.40) and search for the boundary condition $\phi = m\pi$. Particular states may be located by counting the nodes in ϕ in the usual manner. Thus both free and bound states may be calculated under the same formalism involving only real variables simply by a change of boundary conditions and the realization that the zero of energy may be located at any convenient point - in this case the bottom of any potential well.

Fig 9 Hyperboloid/Cone surface for X,Y,T representation of wavefunction



2.223 Variable Mass

The Schroedinger equation with variable mass is :

$$(-h^2/2)\, d/dx\, ((1/m_x)d\psi_x/dx) + V_x\psi_x = E\psi_x \qquad (2.60)$$

where the quantities to be continuous are ψ and $(1/m_x)d\psi_x/dx$. We therefore require a new decomposition and we replace (2.20, 2.21) with :

$$f = [\psi + \psi'/ikm_r]/2 \tag{2.61}$$

$$g = [\psi - \psi'/ikm_r]/2 \tag{2.62}$$

with m_r the ratio of the mass to some constant mass, m_1 (in this case the V=0 mass):

$$m_r = m_x/m_1 \tag{2.63}$$

By the same argument as led to (2.19) the propagation constant remains a function of energy only and is given by :

$$k = (2m_1 E)^{1/2}/h \tag{2.64}$$

Proceeding as before we arrive at the matrix equation :

$$\begin{pmatrix} X' \\ Y' \\ T' \end{pmatrix} = \begin{pmatrix} 0 & m_r + 1 - U & 0 \\ U - 1 - m_r & 0 & m_r - 1 + U \\ 0 & m_r + U - 1 & 0 \end{pmatrix} \begin{pmatrix} X \\ Y \\ T \end{pmatrix} \tag{2.65}$$

Equations (2.34, 2.35) remain un-altered as do (2.41-2.43). The new equations of motion are :

$$T' = (m_r\text{-}1+U)(T^2 - Z^2)^{1/2} \text{ SIN } \phi \tag{2.66}$$

$$\phi' = (m_r\text{-}1+U)\text{COS } \phi \;\; T/(T^2 - Z^2) \;\; + U\text{-}m_r\text{-}1 \tag{2.67}$$

$$\alpha' = -((1\text{-}m_r\text{-}U)\ /(T^2 - Z^2)^{1/2}) \;\; \text{COS } \phi \tag{2.68}$$

2.224 Discussion

The accuracy of the method is demonstrated in Fig 10 where it is compared with

the analytic result. The failure of the usual transfer matrix (UTM) is shown by Figs 11 and 12. In both cases the result for the UTM looks plausible but is in error both near resonance (Fig 11) and even for very thin barriers (Fig12) where the phase is not properly maintained. An application of the method to a much larger system consisting of a disordered superlattice is illustrated in Fig 13. The barrier and well widths of the superlattice are both 50Å and the height of the barriers is 250meV. The energy is 79.5meV, one of the resonances.

A major disadvantage with this method is that the number of points required in the discretisation depends on the ratio of the energy to the potential and must therefore be high for propagation through barriers at low energies. This is explained as follows. We are essentially saying that from the matrix equation:

$$d\Psi/dt = \Omega\Psi \qquad (2.69)$$

the value of Ψ at successive grid points is given by:

$$\Psi_n = (\mathbf{1} + \Omega dt)\Psi_{n-1} \qquad (2.70)$$

Which requires, for a valid approximation to the integration, that $\Omega dt << \mathbf{1}$. From (2.24) and (2.26) this implies that the spacial step dt is restricted by the inequality:

$$dt << E/V \qquad (2.71)$$

Note that the form of (2.70) reveals why the identity member of the group does not appear in the transfer matrix itself.

2.23 Scattering Matrix Propagation

This solution to the instability problem of the transfer matrix was first demonstrated by Ko and Inkson [1988]. Two S-matrices can be ' added ' by eliminating the wavefunction terms appearing in both. In 1D the form of the S-matrix connecting the inputs and outputs to the region between x=m and x=m-1 is:

$$\begin{pmatrix} A_{m-1} \\ \\ B_{m} \end{pmatrix} = \begin{pmatrix} {}^{m}S_{11} & {}^{m}S_{12} \\ \\ {}^{m}S_{21} & {}^{m}S_{22} \end{pmatrix} \begin{pmatrix} B_{m-1} \\ \\ A_{m} \end{pmatrix} \qquad (2.72)$$

Where A's are the components of the wavefunction travelling into the region and B's travel out of it. Similarly for the region between m-1 and m-2 we have:

$$\begin{pmatrix} A_{m-2} \\ \\ B_{m-1} \end{pmatrix} = \begin{pmatrix} {}^{m-1}S_{11} & {}^{m-1}S_{12} \\ \\ {}^{m-1}S_{21} & {}^{m-1}S_{22} \end{pmatrix} \begin{pmatrix} B_{m-2} \\ \\ A_{m-1} \end{pmatrix} \qquad (2.73)$$

The required composite S-matrix is:

$$\begin{pmatrix} A_{m-2} \\ \\ B_{m} \end{pmatrix} = \begin{pmatrix} S_{11} & S_{12} \\ \\ S_{21} & S_{22} \end{pmatrix} \begin{pmatrix} B_{m-2} \\ \\ A_{m} \end{pmatrix} \qquad (2.74)$$

and the elements are determined by eliminating A_{m-1} and B_{m-1} from (2.73) and (2.74):

$$S_{11} = {}^{m-1}S_{11} + {}^{m-1}S_{12}\,{}^{m}S_{11}\,[\,1 - {}^{m-1}S_{22}\,{}^{m}S_{11}\,]^{-1}\,[{}^{m-1}S_{21}] \qquad (2.75)$$

$$S_{12} = {}^{m-1}S_{12}\,{}^{m}S_{11}\,[1 - {}^{m-1}S_{22}\,{}^{m}S_{11}]^{-1}\,[{}^{m-1}S_{22}]\,{}^{m}S_{12} + {}^{m-1}S_{12}\,{}^{m}S_{12} \qquad (2.76)$$

$$S_{21} = {}^{m}S_{21}\,[\,1 - {}^{m-1}S_{22}\,{}^{m}S_{11}\,]^{-1}\,[{}^{m-1}S_{21}] \qquad (2.77)$$

$$S_{22} = {}^{m}S_{22} + {}^{m}S_{21}\,[\,1 - {}^{m-1}S_{22}\,{}^{m}S_{11}\,]^{-1}\,[{}^{m-1}S_{22}]\,{}^{m}S_{12} \qquad (2.78)$$

Thus the total S-matrix for the system can be built up by ' adding ' the matrix for successive regions. We find that in 1D this scheme is indeed stable numerically for both the transmission coefficient and for the wavefunction at all points, the latter being found by converting the total S-matrix to each point to a T-matrix and supplying two boundary conditions at one end.

It is not clear why this scheme overcomes the problem with the T-matrix. Ko and Inkson present an alternative formula for the wavefunction components at each point m in terms of the matrix elements for the S-matrices for regions x=0 to x=m and x=m to x=0 and claim the stability is due to the need to use only the inputs to the entire region as boundary conditions. This they contrast with the propagation of errors using the T-matrix as components at each point are derived from the components at the previous point. This of course is a false description of the operation of the T-matrix; it is possible to construct the total T-matrix to any point and use only the initial boundary conditions. Moreover the use of the two S-matrices to allow one boundary condition from each end is not necessary in 1D . In 2D we find that the use of one boundary condition from each end is not a sufficient condition to ensure stability. We return to a more detailed discussion of this in chapter 4.

2.3 Two-Dimensional Models

2.31 Un-Coupled Matching-States Method

Recently this method has been employed by Kirczenow [1988, 1989] to study injection of electrons into a confined channel. The method rests on the ability to separate variables, with the transverse part of the solution being a known analytic function. We briefly outline the method and discuss its limitations.

The system is divided into three regions, two infinitely wide contacts separated by a narrow channel. Within each region the potential is a function of the transverse

direction (y) only. In the left-hand contact the hamiltonian is:

$$H_{co} = \nabla^2 + V(y) \tag{2.79}$$

and the wavefunction is a superposition of the incident wave in the positive x-direction and the reflected waves in the negative x-direction, each term being multiplied by a single plane wave in the transverse direction:

$$\Psi_K(r) = e^{ikx} e^{i\kappa y} + \sum_{\kappa'} a_{\kappa'} e^{-ik'x} e^{i\kappa' y} \tag{2.80}$$

In the channel again a separation of variables is assumed, the transverse part being eigenfunctions of the potential V(y). Thus the wavefunction here is written as:

$$\Psi_K(r) = \sum_n [\, a_n^+ e^{ik_n x} + a_n^- e^{-ik_n x} \,] \phi_n(y) \tag{2.81}$$

In the right-hand contact we have only the outgoing waves:

$$\Psi_K(r) = \sum a_{\kappa'} e^{ik'x} e^{i\kappa' y} \tag{2.82}$$

The relationship between the wavevectors is:

$$K^2 = k^2 + \kappa^2 \tag{2.83}$$

where K is related to the total energy, a constant in the system.

The procedure is to match the wavefunction and its derivative at the two discontinuities to produce equations for the $a_n^{\pm}$. The conductance is not derived from the transmission coefficient but rather the total current in the channel is divided by the bias voltage across it. The current is given by:

$$J \propto \int dy \, [\, \Psi_K^* \, (\partial \Psi_K / \partial x) - \Psi_K \, (\partial \Psi_K^* / \partial x) \,] \tag{2.84}$$

This scheme has reproduced the quantized resistance in narrow channels but is of

limited application. The solution requires the use of the overlap integrals between the transverse states in the channel and contacts which are only obtainable analytically if the states themselves are simple analytic functions. A barrier may be introduced into the channel, increasing the number of constant potential regions to five, but a continuously varying potential would demand that the summation in equation (2.73) occurred at every grid line and the error introduced into the current continuity from curtailing the infinite sum would tend to multiply. Laughton [1989] has suggested that even using finite width contacts is enough to severely increase difficulties with current continuity.

2.32 Asymptotic Time-Dependent Solution

This method has been used by Finch [1989] to investigate the conductance and magneto-conductance of ring structures. The transmission coefficient is simply defined as the ratio of the magnitude of the part of the wavepacket reaching the right-hand contact to the magnitude of the initial wavepacket after a sufficiently large time has elapsed.

The method is based on the integration of the time evolution operator equation:

$$\Psi(r,t+\Delta t) = e^{-i\Delta tH/h}\,\Psi(r,t) \qquad (2.85)$$

The exponential operator may be understood by a series expansion of the exponential. Using the analysis of Potter [1973] a first order Taylor-expansion was found to be un-stable and the Crank-Nickolsen scheme was adopted. By eliminating $\Psi(t)$ from the two expressions connecting it to $\Psi(t+\Delta t)$ and to $\Psi(t-\Delta t)$ we get:

$$e^{i\Delta tH/h}\,\Psi(t+\Delta t) = e^{-i\Delta tH/h}\,\Psi(t-\Delta t) \qquad (2.86)$$

This is stable under a first order expansion which leads to the equation:

$$\Psi(t+\Delta t) = [(1 - i\Delta \mathbf{t}\mathbf{H}/2h)/(1 + i\Delta \mathbf{t}\mathbf{H}/2h)] \Psi(t) \qquad (2.87)$$

This form involving first order terms in both directions is known as the Cayley Expansion. As well as the stability test the unitarity is also considered. The unitarity condition on a matrix or operator is that the Hermitian conjugate of the operator is equal to the inverse. The conjugate of the time evolution operator is:

$$\mathbf{L}^{\dagger} = e^{i\Delta \mathbf{t}\mathbf{H}^{\dagger}/h} \qquad (2.88)$$

and so the equation $\mathbf{L}^{\dagger}\mathbf{L} = \mathbf{1}$ is satisfied. The Crank-Nickolson scheme also satisfies this requirement.

Discretisation in space of the scheme results in a tri-diagonal matrix equation for the wavefunction across the system, constructed using the equation:

$$a(r)\Psi(t+\Delta t,r+\Delta r) + b(r)\Psi(t+\Delta t,r) + c(r)\Psi(t+\Delta t,r-\Delta r) = d(r,t) \qquad (2.89)$$

where a,b and c are derived from the potential V(r) and d includes the wavefunction at time t.

The advantage of a tri-diagonal matrix equation is that the solution time is proportional to N rather than N^3 for an N X N matrix. In order to make use of this time saving Finch uses a partition of the 2D Hamiltonian resulting in two coupled 1D equations of motion.

The Hamiltonian is written as:

$$\mathbf{H} = \mathbf{H}_x + \mathbf{H}_y \qquad (2.90)$$

and the Cayley expansion (2.87) becomes:

$$\Psi(t+\Delta t) = [\,(1 - \mathbf{H}_x\tau - \mathbf{H}_y\tau)/(1 + \mathbf{H}_x\tau + \mathbf{H}_y\tau)\,]\,\Psi(t) \qquad (2.91)$$

with

$$\tau = i\Delta t/2h \qquad (2.92)$$

Equation (2.91) can be factorised, leaving out terms of $O\mathbf{H}^3$ or higher into:

$$\Psi(t+\Delta t) = \frac{(1 - H_x\tau)(1 - H_y\tau)}{(1 + H_x\tau)(1 + H_y\tau)}\,\Psi(t) \qquad (2.93)$$

thus allowing the expression of (2.91) as a product of two equations having the same form as (2.87):

$$\Psi(t+\Delta t/2) = [\,(1 - \mathbf{H}_y\tau)/(1 + \mathbf{H}_x\tau)\,]\,\Psi(t) \qquad (2.94)$$

$$\Psi(t+\Delta t) = [\,(1 - \mathbf{H}_x\tau)/(1 + \mathbf{H}_y\tau)\,]\,\Psi(t+\Delta t/2) \qquad (2.95)$$

This construction is known as Alternating Direction Implicit (ADI). The solution for the wavefunction now involves two tri-diagonal matrix equations to be solved successively.

An error is introduced by the ADI construction since the effective operator is now the product:

$$\mathbf{L}' = e^{-i\mathbf{H}_x\Delta t/h}\, e^{-i\mathbf{H}_y\Delta t/h} \qquad (2.96)$$

whereas the true operator is:

$$\mathbf{L} = e^{-i(\mathbf{H}_x + \mathbf{H}_y)\Delta t/h} \qquad (2.97)$$

This error becomes important for high magnetic fields when the non-

commutivity of the two parts of the Hamiltonian destroys unitarity. The other major disadvantage of this method is the time required to reach the asymptotic result; the saving on the tri-diagonal matrix form is offset by the 10^4 time iterations needed. Thus a grid 590 X 256 takes approximately 10000s cpu time. The same system can be modelled by the network method using a grid of only 96 by 48 which requires only about 8000s cpu time for the wavefunction and only about 150s cpu time per transmission coefficient.

3 NETWORK METHOD

3.1 Introduction

In this chapter we introduce the main subject of this thesis, the new network method for calculating the wavefunction and transmission coefficient in 2D systems. In section 3.2 we review earlier work on networks and explain the difference between this and our approach. Section 3.3 contains the basic details of the new method with some practical points on numerical stability made in section 3.34. In section 3.35 we indicate the limitations imposed by the bandstructure of the network.

3.2 Early Network Models

The network approach was first adopted in the form of straight, one-dimensional lines connecting nodes by free-propagating plane waves. This was applied to molecules by Ruedenberg and Scherr [1953] and it was found that the eigenvalues of the network agreed well with the values derived from the traditional tight-binding (combination of atomic orbitals) method. Coulson [1953] drew the same conclusion by applying the method to metals. We now review the main points of Coulson's work.

The continuity equations at the nodes are pertinent for any network model where scattering at the nodes is elastic. Referring to Fig 14 for notation (we restrict ourselves to 2D for clarity, it making no difference to the points we wish to make) we have the continuity of the wavefunction itself:

$$\phi_1 = \phi_2 = \phi_3 = \phi_4 \qquad (3.1)$$

and of the current:

$$\sum_i \nabla\phi_i = 0 \qquad (3.2)$$

Along each line the wavefunction obeys the 1D, free-space schroedinger equation:

$$d^2\phi/dr^2 + q^2\phi = 0 \qquad (3.3)$$

where q is the free-space wavevector proportional to $E^{1/2}$.

At this point Coulson also introduces the Bloch condition:

$$\Psi(\mathbf{r}) = e^{i\mathbf{k}.\mathbf{r}} U(\mathbf{r}) \qquad (3.4)$$

where $\mathbf{k} = \mathbf{k}(k_x , k_y)$

This last restraint implies a periodic lattice and this is only the case for an infinite lattice or one that is wrapped round to form a closed surface and periodic boundary conditions may be used. With this in mind we proceed to calculate the bandstructure of this network. Along each line the wavefunction takes the general form:

$$\phi = COS(qr) + a_{x(y)} SIN(qr) \qquad (3.5)$$

and the form of the Bloch function is:

$$U(r) = e^{-ikr} [COS(kr) + a SIN(kr)] \qquad (3.6)$$

By using these conditions to eliminate the constants $a_{x,(y)}$ a bandstructure equation is obtained:

$$2COS(q\Delta) = COS(k_x \Delta) + COS(k_y \Delta) \qquad (3.7)$$

This then is the bandstructure of the periodic lattice itself and is only obtainable if the periodic Bloch condition is assumed. It seems to us un-surprising that the result obtained agrees so well with the tight-binding model; both imply not only a fixed functional form along each line but also the exact same coefficients for this form along each ' line ' (for the tight-binding case the ' line ' means the overlap considered between neighbours). The lines in the x and y directions in this model are not coupled, again implied by the use of the Bloch condition.

In the present scheme it is the bandstructure of the network itself which is of

interest. In the next section we present a new network model formulated under a different concept; the network is used to represent a system whose features are defined on a scale larger than the period of the network and it is the properties of the system that are of interest, the bandstructure of the network acting as a constraint on the legitimate energy domain of the model.

Alexander [1983] uses the same continuity equations for a superconducting network and arrives at the connectivity equation:

$$\Psi_{n,m} \sum_{n',m'} \frac{\mathrm{COT}\theta_{n,m,n',m'}}{\xi_{n,m,n'm'}} = \sum_{n',m'} \frac{\Psi_{n',m'}}{\xi \mathrm{SIN}\theta_{n,m,n',m'}} \qquad (3.8)$$

Here $\theta_{n,m,n'm'}$ involves the length between nodes at n,m and n',m' and an 'intrinsic correlation length' $\xi_{n,m,n',m'}$. The equation (3.8) reduces to the one we derive (3.17) if the vector and scaler potentials are zero and θ takes the value $\Delta^2 k^2/2$, Δ being the distance between nodes along the network lines. Alexander uses this equation to investigate some simple networks including a transmission line with side branches. The model is not the same as the one we propose however. Firstly the value of θ (related to Δk) is allowed to vary on different lines. Secondly, as with Coulson it is the bandstructure (i.e. the bound states) that is sought and therefore the solution for each network is for Alexander the particular version of (3.8). In order to get the transmission coefficient it is necessary to solve not merely for the amplitude of the wavefunction (i.e solving (3.8)) but for the components of it going in either direction along each line. We show in appendix 2 that whilst in principle a 2D network can be reduced to an equivalent 1D transmission line a systematic method is required. Moreover reducing to 1D is not actually very useful since for a 2D system there are boundary conditions, and therefore solutions required, on all the lines emerging from the system.

Montroll [1970] finds the bandstructure for networks that are periodic with the potential and for empty (zero potential) square networks where the boundary conditions are periodic with the whole system. In the latter case he assumes some particular form for the wavefunction through the whole system.

3.3 New Network Model

3.31 Introduction

We must of necessity adopt the same continuity conditions as specified in the previous section. However since we are rejecting the notion of periodicity of the potential being commensurate with that of the network we do not include the Bloch function. Instead we adopt an S-matrix (scattering matrix) to describe mixing of the x and y direction wavefunctions at the nodes. This mixing function is intended to restore the general 2D nature of the solution and does not include the effect of the particular potential profile (this is included along the lines, between nodes). Consequently the S-matrix is the same for all nodes.

3.32 Mixing S-matrix and Local Connectivity

We intend to represent the two-dimensional wavefunction by constraining it to exist on a one-dimensional network of interconnected lines. Since physically the electron has only one energy we must be able to represent it on the network by wavefunctions which are linked between nodes by propagation using a single wavevector. In order to understand that this is possible it is helpful to realise that although the wavefunctions on the four lines meeting at a node have to be coincident at that point the derivatives do not and therefore it is not equivalent to merely propagating one-dimensional solutions across the system in two directions; indeed to reproduce the correct two-dimensional result the network lines must be genuinely coupled, emulating a two-dimensional space.

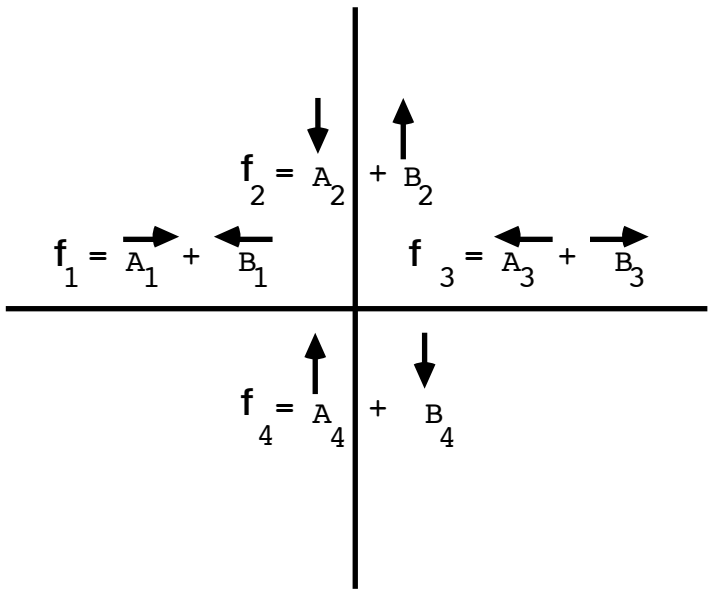
The next step then consists in choosing the correct S-matrix to describe the nodes. Consider Fig 14; the four lines meet at the infinitesimal point where the 1D wavefunctions must converge i.e. $\phi_1 = \phi_2 = \phi_3 = \phi_4 = \Psi$, Ψ being the 2D wavefunction. We must also have current continuity. In terms of the formulation $\phi = A + B$ on each line this is expressed by:

$$\Sigma |A_j|^2 = \Sigma |B_j|^2 \qquad (3.9)$$

A's representing the ingoing waves here. In terms of the S-matrix it requires that the latter is unitary. As far as the mixing function of the S-matrix is concerned we require a homogenous network to genuinely represent the 2D space; in this instance homogenous means that the node appears the same from whichever line it is approached from. An S-matrix satisfying these conditions is the following:

$$\begin{pmatrix} B_1 \\ B_2 \\ B_3 \\ B_4 \end{pmatrix} = \frac{1}{2} \begin{pmatrix} -1 & 1 & 1 & 1 \\ 1 & -1 & 1 & 1 \\ 1 & 1 & -1 & 1 \\ 1 & 1 & 1 & -1 \end{pmatrix} \begin{pmatrix} A_1 \\ A_2 \\ A_3 \\ A_4 \end{pmatrix} \qquad (3.10)$$

Fig 14 Wavefunctions at a node



After discovering the above matrix we learned that Avron, Raveh and Zur [1988], who review Alexander's work, also present a general form for the S-

matrix:

$$\mathbf{S} = -\mathbf{1} + (\nu/2)\,\mathrm{COS}(\theta)e^{i\theta}\mathbf{I} \qquad (3.11)$$

Where ν is the number of vertices at the node and $\mathbf{I}$ is the square ν by ν matrix whose elements are all 1. Our Matrix takes this form with θ taking the value 0, 2π etc. This value for θ implies, according to Avron et al, that the scattering potential at the node is zero and is thus in accord with our description of the node.

We return now to the question of the energy. The network representation requires the use of half the physical energy. We explain this by comparison with the usual finite difference form. We need to construct an equivalent to the finite difference relationship relating the wavefunction at a node to those at its four nearest neighbours (see Abramowitz and Stegun [1984]):

$$\Psi_{n,m} = \frac{\sum \Psi_{n',m'}}{4 - \Delta^2 k^2 + (2m^* \Delta^2 V_{n,m} / h^2)} \qquad (3.12)$$

where Δ is the distance between nodes along the grid lines and k is proportional to the root of the energy. In dealing with the energy we adopt the trick used by Peres [1983] and imagine infinitesimal holes in the potential at the nodes so that the energy and wavevector are related at the node by the free-space expression.

Let us Taylor expand the neighbouring wavefunctions in terms of $\Psi_{n,m}$, discarding terms of third order and higher:

$$\Psi_{n \pm 1, m} = \Psi_{n,m} \pm \Delta \frac{d\Psi_{n,m}}{dy} + \frac{\Delta^2}{2!} \frac{d^2\Psi_{n,m}}{dy^2} \qquad (3.13)$$

$$\Psi_{n, m \pm 1} = \Psi_{n,m} \pm \Delta \frac{d\Psi_{n,m}}{dx} + \frac{\Delta^2}{2!} \frac{d^2\Psi_{n,m}}{dx^2} \qquad (3.14)$$

The wavefunction obeys the 1D Schroedinger equation which we may write as, in the x-direction:

$$\frac{d^2\Psi}{dx^2} + k^2\Psi - \frac{2m^*V\Psi}{h^2} = 0 \qquad (3.15)$$

Using (3.15) we can re-write (3.14) as:

$$\Psi_{n, m \pm 1} = \Psi_{n,m} \left[1 + \frac{\Delta^2}{2!} \left(\frac{2m^*V_{n,m}}{h^2} - k^2 \right) \right] \pm \Delta \frac{d\Psi_{n,m}}{dx} \qquad (3.16)$$

Adding the four equations of type (3.16) we get:

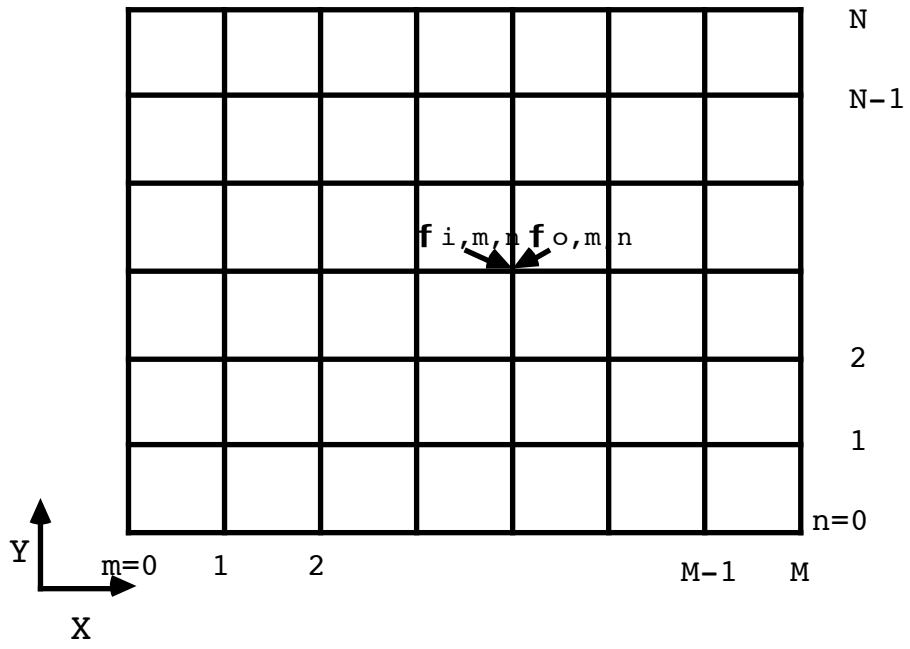
$$\sum \Psi_{n',m'} = 4\Psi_{n,m} \left[1 + \frac{\Delta^2}{2!} \left(\frac{2m^*V_{n,m}}{h^2} - k^2 \right) \right] \qquad (3.17)$$

or

$$\Psi_{n,m} = \frac{\sum \Psi_{n',m'}}{4 - 2\Delta^2 k^2 + (4m^*\Delta^2 V_{n,m} / h^2)} \qquad (3.18)$$

Comparing (3.18) with (3.12) we see an extra factor of two in the energy terms. Since the finite difference and network schemes must agree it is necessary to use only half of the true energy in the network model.

Fig 15 Complete Network Showing X and Y Indices



We now use the S-matrix to derive the relationship between the nodal wavefunctions. The grid to be considered is shown in Fig 15. Note that the wavefunction is chosen to be zero along the boundaries parallel to the x-direction. The labels on the coefficients are re-defined to concur with a sense of input and output to a line of nodes parallel to the y-direction (we later introduce a T-matrix propagating in the positive x-direction).

To explain the basic connectivity we expand part of the grid and reduce the complexity of the indices as shown in Fig 16. This nearest neighbour coupling may be used to build up the entire grid. In the notation of Fig 16 the S-matrix connects column vectors as follows:

$$\begin{pmatrix} B_i \\ B_r \\ B_o \\ B_l \end{pmatrix} = \mathbf{S} \begin{pmatrix} A_i \\ A_r \\ A_o \\ A_l \end{pmatrix} \qquad (3.19)$$

Where **S** is the S-matrix from (3.9). We first derive the input/output relationships for the node. From (3.19):

$$B_o = 0.5 [A_i + A_r + A_l - A_o] \qquad (3.20)$$

$$B_i = 0.5 [A_r + A_o + A_l - A_i] \qquad (3.21)$$

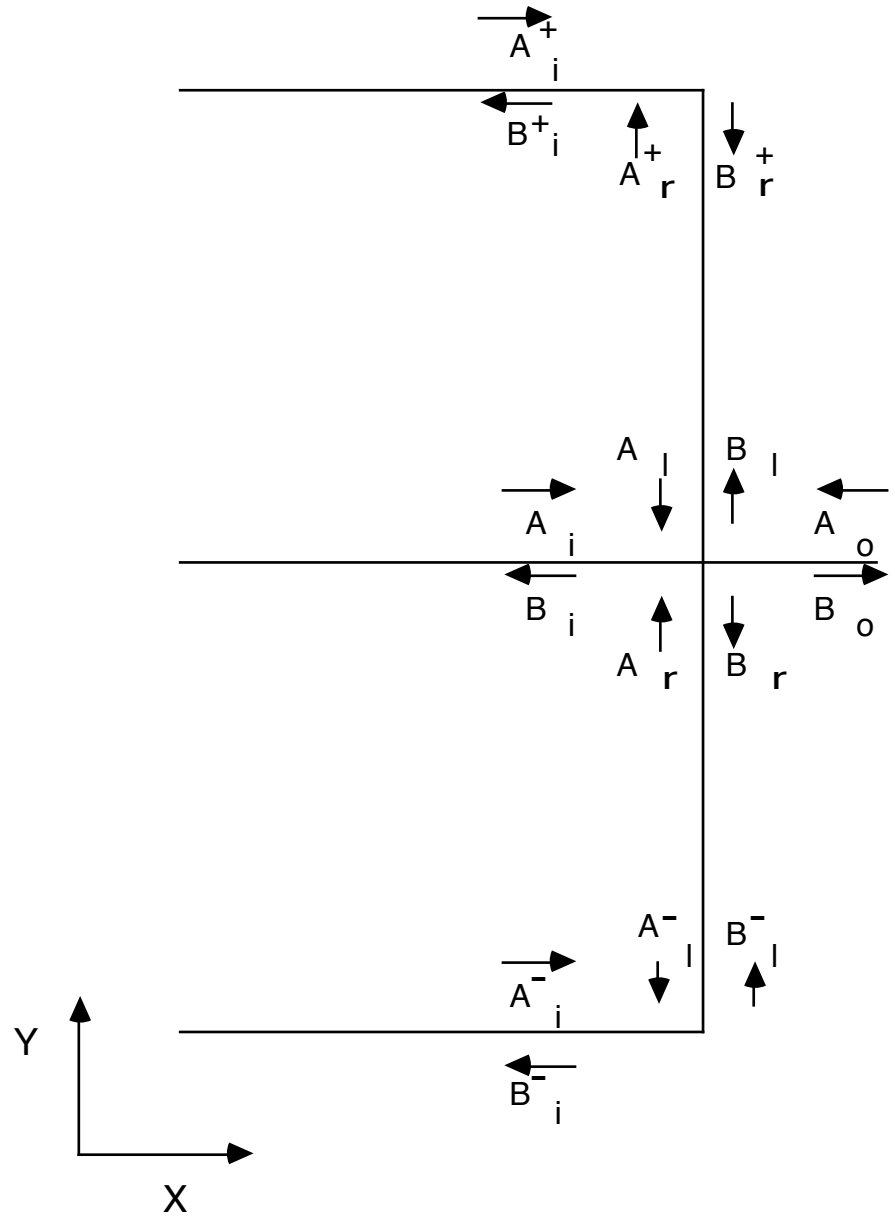
From (3.21):

$$A_o = 2B_i + A_i - A_r - A_l \qquad (3.22)$$

and ,

$$B_o = A_r + A_l - B_i \qquad (3.23)$$

Fig 16 Section of Network Showing Nearest - Neighbour Coupling



In matrix form the input/output relationship is then

$$\begin{pmatrix} A_o \\ B_o \end{pmatrix} = \begin{pmatrix} 1 & 2 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} A_i \\ B_i \end{pmatrix} + \begin{pmatrix} -1 & -1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} A_r \\ A_l \end{pmatrix} \qquad (3.24)$$

The next stage is to eliminate A_r , A_l . To do this we employ the T-matrices along the lines connecting nodes at n to those at n+1 and n-1. It is very easy to get the order of the vectors wrong during the combinatorial process and we consider it prudent to explicitly define our matrices in our notation. Because we write A's as always going into a node the T-matrix takes the form:

$$\begin{pmatrix} B_l \\ A_l \end{pmatrix} = \mathbf{T} \begin{pmatrix} A_o \\ B_o \end{pmatrix} \qquad (3.25)$$

i.e. the first vector is in the opposite order to that usually written. Taking the line between y=n and y=n+1 we have:

$$\begin{pmatrix} B_l \\ A_l \end{pmatrix} = \begin{pmatrix} {}^{+}\mathbf{t}_{11} & {}^{+}\mathbf{t}_{12} \\ {}^{+}\mathbf{t}_{21} & {}^{+}\mathbf{t}_{22} \end{pmatrix} \begin{pmatrix} A_r^{+} \\ B_r^{+} \end{pmatrix} \qquad (3.26)$$

and,

$$A_l = {}^{+}\mathbf{t}_{21}A_r^{+} + {}^{+}\mathbf{t}_{22}B_r^{+} = ({}^{+}\mathbf{t}_{21} - {}^{+}\mathbf{t}_{22})A_r^{+} + {}^{+}\mathbf{t}_{22}\Psi^{+} \qquad (3.27)$$

$$B_l = ({}^{+}\mathbf{t}_{11} - {}^{+}\mathbf{t}_{12})A_r^{+} + {}^{+}\mathbf{t}_{12}\Psi^{+} \qquad (3.28)$$

We get A_r^+ from (3.28) and substitute it into (3.27) to get:

$$A_l = \frac{\Omega^+ \Psi + ({}^+t_{22} - \Omega^+ {}^+t_{12})\Psi^+}{1 + \Omega^+} \qquad (3.29)$$

with Ω^+ given by:

$$\Omega^+ = \frac{{}^+t_{21} - {}^+t_{22}}{{}^+t_{11} - {}^+t_{12}} \qquad (3.30)$$

Similarly for the line connecting y=n to y=n-1:

$$A_r = \frac{\Omega^- \Psi + ({}^-t_{22} - \Omega^- {}^-t_{12})\Psi^-}{1 + \Omega^-} \qquad (3.31)$$

with $^-$t's and Ω^- defined in the same way as above. We can now replace the A_r and A_l in (3.24):

$$A_o = A_i + 2B_i - \left[\frac{\Omega^-(A_i + B_i) + ({}^-t_{22} - \Omega^- {}^-t_{12})(A_i^- + B_i^-)}{1 + \Omega^-}\right]$$

$$- \left[\frac{\Omega^+(A_i + B_i) + ({}^+t_{22} - \Omega^+ {}^+t_{12})(A_i^+ + B_i^+)}{1 + \Omega^+}\right] \qquad (3.32)$$

Rearranging this we get;

$$A_o = \left(1 - \frac{\Omega^-}{1+\Omega^-} - \frac{\Omega^+}{1+\Omega^+}\right)A_i + \left(2 - \frac{\Omega^-}{1+\Omega^-} - \frac{\Omega^+}{1+\Omega^+}\right)B_i$$

$$- \Gamma^- A_i^- - \Gamma^- B_i^- - \Gamma^+ A_i^+ - \Gamma^+ B_i^+ \qquad (3.33)$$

where $\Gamma^{-(+)}$ is given by:

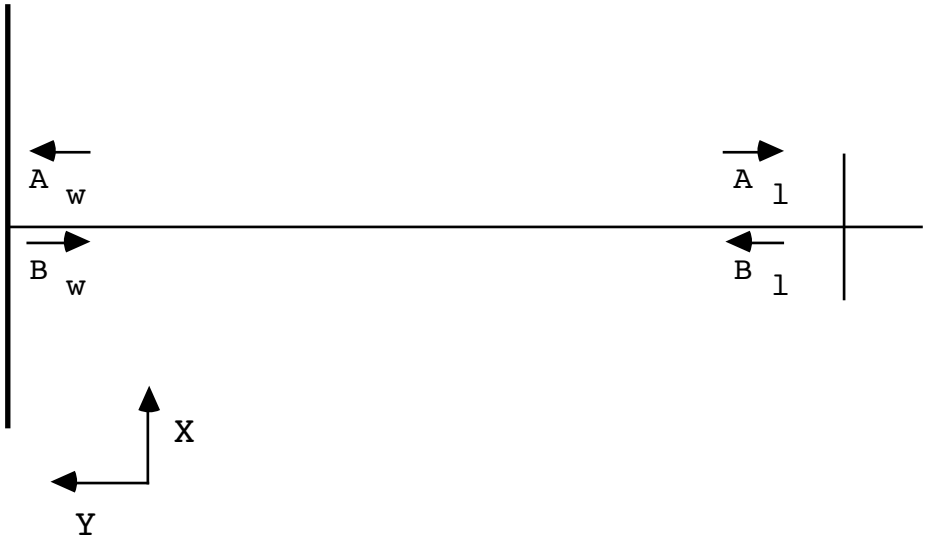
$$\Gamma^- = ({}^-t_{22} - \Omega^- {}^-t_{12})/(1 + \Omega^-) \qquad (3.34)$$

A similar expression for B_o is obtained and we have the result:

$$\begin{pmatrix} A_o \\ B_o \end{pmatrix} = \begin{pmatrix} 1 - \frac{\Omega^-}{1+\Omega^-} - \frac{\Omega^+}{1+\Omega^+} & 2 - \frac{\Omega^-}{1+\Omega^-} - \frac{\Omega^+}{1+\Omega^+} \\ \frac{\Omega^-}{1+\Omega^-} + \frac{\Omega^+}{1+\Omega^+} & -1 + \frac{\Omega^-}{1+\Omega^-} + \frac{\Omega^+}{1+\Omega^+} \end{pmatrix} \begin{pmatrix} A_i \\ B_i \end{pmatrix}$$

$$+ \Gamma^- \begin{pmatrix} -1 & -1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} A_i^- \\ B_i^- \end{pmatrix} + \Gamma^+ \begin{pmatrix} -1 & -1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} A_i^+ \\ B_i^+ \end{pmatrix} \qquad (3.35)$$

Fig 17 Termination of one of the Y-Direction Lines



It can be easily shown that (3.35) applies to nodes next to the boundary. From Fig 17 we may write:

$$\begin{pmatrix} B_l \\ A_l \end{pmatrix} = \begin{pmatrix} {}^{w}t_{11} & {}^{w}t_{12} \\ {}^{w}t_{21} & {}^{w}t_{22} \end{pmatrix} \begin{pmatrix} A_w \\ B_w \end{pmatrix} \qquad (3.36)$$

We get a similar equation to (3.29) but since in this case Ψ^+ is the wavefunction on the boundary and is therefore zero we use (3.35) with $A_i^+ = B_i^+ = 0$. At the other end of the line $A_i^- = B_i^- = 0$.

The 1D transfer matrix has been dealt with in chapter 2. We choose to use here the S-matrix technique of section 2.23. Although it is not as efficient as the Peres method it uses the same parametrization as the multi-dimensional transfer matrix introduced in the next section and therefore slots into the scheme without difficulty. It is not clear to the author how the Peres decomposition is converted to the A,B decomposition used in the network model.

What we have achieved so far is to express the output coefficients of all the nodes of a line in terms of input coefficients. In the next section we shall use this to construct a multi-dimensional transfer matrix which can be iterated through the system.

3.33 Multi-Dimensional Transfer Matrix

It is necessary now to extend the spacial extent of the connectivity into something which can be iterated to cover the whole system. Fig 18 shows the element of the network that we now analyse. For clarity we leave out the index indicating the position in the y-direction; the form of the equations are of course independent of position.

We write the equation (3.35) in symbolic form, using the notation of Fig 18, as:

$$\begin{pmatrix} A_o \\ B_o \end{pmatrix} = \mathbf{T}_s \begin{pmatrix} A_i \\ B_i \end{pmatrix} + \mathbf{T}_l^{+} \begin{pmatrix} A_i^{+} \\ B_i^{+} \end{pmatrix} + \mathbf{T}_l^{-} \begin{pmatrix} A_i^{-} \\ B_i^{-} \end{pmatrix} \qquad (3.37)$$

and define the following new T-matrices:

$$\begin{pmatrix} A_i \\ B_i \end{pmatrix} = \mathbf{T}_r \begin{pmatrix} B_o \\ A_o \end{pmatrix}_{m-1} \qquad (3.38)$$

$$\begin{pmatrix} A_i^- \\ B_i^- \end{pmatrix} = \mathbf{T}_r^- \begin{pmatrix} B_o^- \\ A_o^- \end{pmatrix} \tag{3.39}$$

$$\begin{pmatrix} A_i^+ \\ B_i^+ \end{pmatrix} = \mathbf{T}_r^+ \begin{pmatrix} B_o^+ \\ A_o^+ \end{pmatrix} \tag{3.40}$$

Substituting the above into (3.35) we get the equation linking the lines x=m and x=m-1:

$$\begin{pmatrix} A_o \\ B_o \end{pmatrix} = \mathbf{T}_{sr} \begin{pmatrix} B_o \\ A_o \end{pmatrix}_{m-1} + \mathbf{T}_{lr}^+ \begin{pmatrix} B_o^+ \\ A_o^+ \end{pmatrix} + \mathbf{T}_{lr}^- \begin{pmatrix} B_o^- \\ A_o^- \end{pmatrix} \tag{3.41}$$

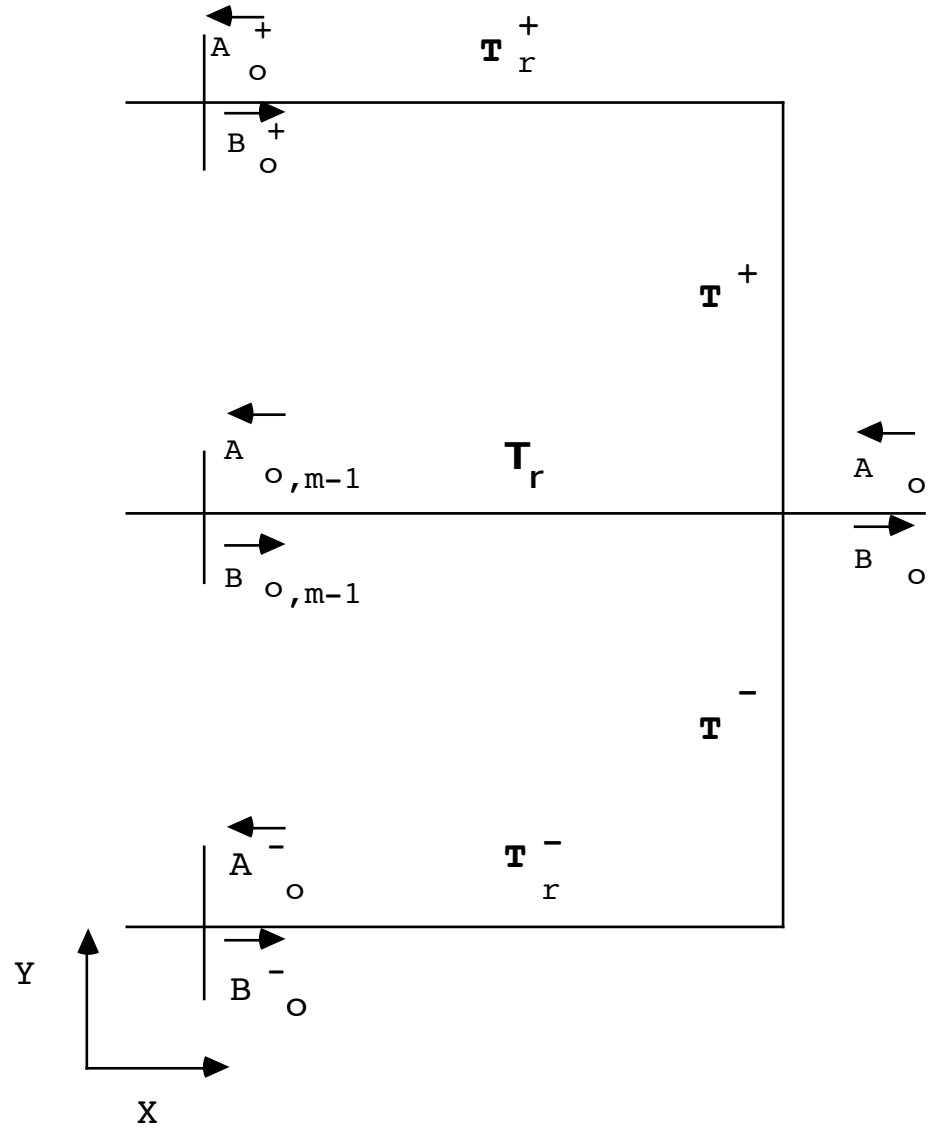
Where,

$$\mathbf{T}_{sr} = \mathbf{T}_s \mathbf{T}_r \tag{3.42}$$

$$\mathbf{T}_{lr}^+ = \mathbf{T}_l^+ \mathbf{T}_r^+ \tag{3.43}$$

$$\mathbf{T}_{lr}^- = \mathbf{T}_l^- \mathbf{T}_r^- \tag{3.44}$$

Fig 18 One - Dimensional Transfer Matrices Connecting Line n to Line n-1



A T-matrix describing the whole line x=m may be set up from equation (3.41). By inspection it will be seen that such a T-matrix will be banded with three sub-diagonals and three super-diagonals. We illustrate the form for a grid having four lines parallel to the x-direction (non-zero elements are denoted by a X):

$$\begin{pmatrix} A_1 \\ B_1 \\ A_2 \\ B_2 \\ A_3 \\ B_3 \\ A_4 \\ B_4 \end{pmatrix}_m = \begin{pmatrix} X & X & X & X & 0 & 0 & 0 & 0 \\ X & X & X & X & 0 & 0 & 0 & 0 \\ X & X & X & X & X & X & 0 & 0 \\ X & X & X & X & X & X & 0 & 0 \\ 0 & 0 & X & X & X & X & X & X \\ 0 & 0 & X & X & X & X & X & X \\ 0 & 0 & 0 & 0 & X & X & X & X \\ 0 & 0 & 0 & 0 & X & X & X & X \end{pmatrix} \begin{pmatrix} A_1 \\ B_1 \\ A_2 \\ B_2 \\ A_3 \\ B_3 \\ A_4 \\ B_4 \end{pmatrix}_{m-1} \tag{3.45}$$

These can be multiplied together in the usual way to propagate the wavefunction along the system.

3.34 Stability and S-matrix Combinatorics

As in the 1D case we find that the T-matrix is numerically unstable under propagation. Stability may be tested by checking the metric equation. For the pseudo-unitary T-matrix the metric equation:

$$\mathbf{m}\mathbf{T}^{\dagger}\mathbf{m} = \mathbf{T}^{-1} \tag{3.46}$$

has the metric:

$$\mathbf{m} = \mathbf{U}\,\sigma_3\,\mathbf{U}^{\dagger} \tag{3.47}$$

Here σ_3 is one of the Dirac matrices and is diagonal with unity elements of alternating sign (see Arfken [1970]). The matrix **U** is the operator required to transform the T-matrix from the form in equation (3.45) to a form where the components of the wavefunction going in the same direction are grouped together (it is this new form that appears in (3.46)). We explain this below.

Using this test the T-matrix is shown to diverge from the constraining metric equation quite rapidly. For example the product of the left hand and right hand parts of (3.46) should give the unit matrix. However for a 2 by 100 network (giving rise to a 4 by 4 T-matrix) the elements that should be zero are of order 1 at the end of the propagation.

To circumvent this problem we turn to the propagation of the S-matrix as an alternative. Interestingly we find the instability cure is more restricted in the 2D case than in the 1D case. We identify three different approaches to the use of the S-matrix but first we show how it is constructed.

The combinatorics detailed in section 2.23 are used to build up an S-matrix spanning many network lines, but the S-matrices for each segment are derived from the T-matrix. First we have to transform the T-matrix into a new form as indicated above.

We introduce a unitary matrix which re-orders the column vectors in (3.45):

$$\mathbf{U} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \qquad (3.48)$$

for the case of the 6 X 6 T-matrix. The general formula for the N X N case is:

$$U(I + 1, 2I + 1) = 1$$

$$U(I + 1 + N/2, 2I + 2) = 1 \quad , \quad I = 0,1,2 \ldots N/2 - 1 \qquad (3.49)$$

Thus applying this transformation to (3.45) we get:

$$\mathbf{U}\,\Psi_m = \mathbf{U}\,\mathbf{T}\,\mathbf{U}^{\dagger}\,\mathbf{U}\,\Psi_{m-1} \qquad (3.50)$$

where $\Psi_{m(m-1)}$ are the column vectors in (3.45), or in terms of the re-ordered column vectors:

$$\begin{bmatrix} \mathbf{a} \\ \mathbf{b} \end{bmatrix}_m = \begin{bmatrix} \mathbf{T'}_{11} & \mathbf{T'}_{12} \\ \mathbf{T'}_{21} & \mathbf{T'}_{22} \end{bmatrix} \begin{bmatrix} \mathbf{a} \\ \mathbf{b} \end{bmatrix}_{m-1} \tag{3.51}$$

where

$$\mathbf{T'} = \mathbf{U}\,\mathbf{T}\,\mathbf{U}^{\dagger} \tag{3.52}$$

and each **a** and **b** are column vectors containing the components into and out of the region respectively. From the new T-matrix **T'** we can construct the S-matrix for the region:

$$\begin{bmatrix} \mathbf{a}_{m-1} \\ \mathbf{b}_m \end{bmatrix} = \begin{bmatrix} {}^m\mathbf{S}_{11} & {}^m\mathbf{S}_{12} \\ {}^m\mathbf{S}_{21} & {}^m\mathbf{S}_{22} \end{bmatrix} \begin{bmatrix} \mathbf{b}_{m-1} \\ \mathbf{a}_m \end{bmatrix} \tag{3.53}$$

using the following sub-matrix conversion formulae (dropping the dash from the T's), obtained by comparing (3.53) with (3.51):

$$\mathbf{S}_{11} = -[\mathbf{T}_{11}]^{-1}\,\mathbf{T}_{12} \tag{3.54}$$

$$\mathbf{S}_{12} = [\mathbf{T}_{11}]^{-1} \tag{3.55}$$

$$\mathbf{S}_{21} = \mathbf{T}_{22} - \mathbf{T}_{21}\,[\mathbf{T}_{11}]^{-1}\,\mathbf{T}_{12} \tag{3.56}$$

$$\mathbf{S}_{22} = \mathbf{T}_{21}\,[\mathbf{T}_{11}]^{-1} \qquad (3.57)$$

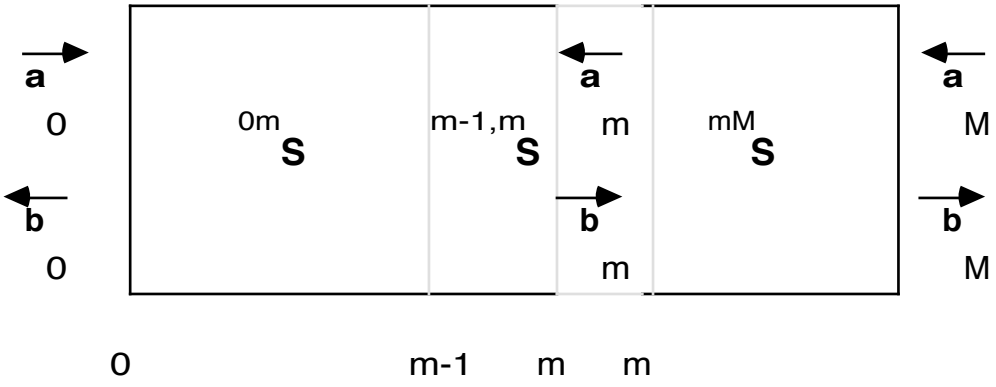
To construct the S-matrix over an extended region then the procedure is as follows. First find the T-matrix for the region x=m to x=m+1. Carry out the transformation (3.52) and convert to the S-matrix using (3.54-3.57). Do the same for the region x=m+1 to x=m+2 and combine the two S-matrices using the formulae of section 2.23 with the single component A's and B's replaced by the column vectors **a** and **b** and the S_{ij}'s replaced by the sub-matrices $\mathbf{S}_{ij}$'s. The next region x=m+2 to x=m+3 is then included in the same way and so on. Under this scheme the metric restraint is conserved i.e. :

$$\mathbf{S}^{\dagger}\mathbf{S} = \mathbf{1} \qquad (3.58)$$

and current continuity is maintained. Maintaining current continuity does not however guarantee the correct wavefunction or transmission coefficient (this is illustrated in appendix 1 where a 2 by 2 network is analysed for these two quantities using both the T-matrix and S-matrix propagation).

Of the three ways of using the S-matrix that we identify all three produce a satisfactory result for the transmission coefficient but only one can give us the correct wavefunction for large networks. According to Ko and Inkson [1988] the stability of the S-matrix propagation technique derives from the necessity to define only the ingoing (or outgoing) components of the wavefunction at each end as boundary conditions. Since all three of our alternatives posses this property we disagree with that proposition. In fact the stability would appear to be associated with the suitability for the chosen object for inversion. We shall explain this after quoting the three different results, with reference to Fig 19 for the notation. (The details of the calculations are in appendix 3).

Fig 19 S-matrices used in equations (3.59-3.61)



The three equations of interest here are:

$$\begin{bmatrix} \mathbf{b}_0 \\ \mathbf{b}_M \end{bmatrix} = \begin{bmatrix} \mathbf{S}_{11} & \mathbf{S}_{12} \\ \mathbf{S}_{21} & \mathbf{S}_{22} \end{bmatrix} \begin{bmatrix} \mathbf{a}_0 \\ \mathbf{a}_M \end{bmatrix} \qquad (3.59)$$

$$\begin{bmatrix} \mathbf{a}_m \\ \mathbf{b}_M \end{bmatrix} = \begin{bmatrix} {}^{mM}\mathbf{S}_{11} & {}^{mM}\mathbf{S}_{12} \\ {}^{mM}\mathbf{S}_{21} & {}^{mM}\mathbf{S}_{22} \end{bmatrix} \begin{bmatrix} \mathbf{b}_m \\ \mathbf{a}_M \end{bmatrix} \qquad (3.60)$$

$$\begin{bmatrix} \mathbf{b}_0 \\ \mathbf{b}_m \end{bmatrix} = \begin{bmatrix} {}^{0m}\mathbf{S}_{11} & {}^{0m}\mathbf{S}_{12} \\ {}^{0m}\mathbf{S}_{21} & {}^{0m}\mathbf{S}_{22} \end{bmatrix} \begin{bmatrix} \mathbf{a}_0 \\ \mathbf{a}_m \end{bmatrix} \qquad (3.61)$$

The boundary conditions used for all three routes are the inputs to the system $\mathbf{a}_0$ and $\mathbf{a}_M$. As usual with tunnelling problems we specify $\mathbf{a}_M = \mathbf{0}$. Thus from (3.59):

$$\mathbf{b}_0 = \mathbf{S}_{11}\mathbf{a}_0 \tag{3.62}$$

$$\mathbf{b}_M = \mathbf{S}_{21}\mathbf{a}_0 \tag{3.63}$$

The first alternative is equivalent to converting the S-matrix to a T-matrix, using the reverse procedure of the equations (3.54-3.57) and the expression for the wavefunction is:

$$\Psi_m = ({}^{0m}\mathbf{S}_{22} + \mathbf{1})({}^{0m}\mathbf{S}_{12})^{-1}\mathbf{b}_0 + [{}^{0m}\mathbf{S}_{21} - (\mathbf{1} + {}^{0m}\mathbf{S}_{22})({}^{0m}\mathbf{S}_{12})^{-1}{}^{0m}\mathbf{S}_{11}]\mathbf{a}_0 \tag{3.64}$$

where Ψ_m is the column vector representing the wavefunction along line x=m. This is the version that may be used to accurately determine the wavefunction in 1D but fails in 2D.

The second way to get at the wavefunction is to use only the ${}^{mM}\mathbf{S}$ matrix. In this case the expression for the wavefunction is simpler but equally unsuccessful:

$$\Psi_m = ({}^{mM}\mathbf{S}_{11} + \mathbf{1})({}^{mM}\mathbf{S}_{21})^{-1}\mathbf{b}_M \tag{3.65}$$

The third and reliable alternative is to use a combination of both the ${}^{0m}\mathbf{S}$ and ${}^{mM}\mathbf{S}$ matrices:

$$\Psi_m = ({}^{mM}\mathbf{S}_{11} + \mathbf{1})(\mathbf{1} - {}^{0m}\mathbf{S}_{22}{}^{mM}\mathbf{S}_{11})^{-1}\mathbf{S}_{21}\,\mathbf{a}_0 \tag{3.66}$$

Clearly it is not a question of any of the sub-matrices themselves being unstable since all the elements are required in the iteration process which maintains

unitarity. Also it is noted that even fixing the $\mathbf{b}_M$ and $\mathbf{b}_0$ elements so that they form an exact eigenfunction does not improve the accuracy of (3.64) or (3.65).The difference between the first two expressions (3.64, 3.65) and the third (3.66) is the form of the inverse required. The form in the third expression is the same as in the propagation equations of section 2.23:

$$(\mathbf{1} - \mathbf{S}_{22}\mathbf{S}_{11})^{-1} \qquad (3.67)$$

Unfortunately the only reliable route is also the slowest. The ${}^{mM}\mathbf{S}$ matrix has to be calculated from scratch for each value of m. A method of ' subtracting ' the m-1 to m matrix from the previous ${}^{m-1M}\mathbf{S}$ matrix has been found but produces overflow errors for large networks. The details are given in appendix 3. We illustrate in Figs 20, 21 the results of the two unstable S-matrix routes (3.64, 3.65) when applied to an open-ended, empty (no potential) channel. In Fig 22 we show the stable result (3.66). From the choice of the energy at the first transverse eigenvalue we know that Fig 22 shows the expected result (this is explained in section 4.32). The transmission coefficient for each case is unity. This is calculated by taking the ratio of the total current leaving the system on the right to that entering on the left:

$$T = |t|^2 = \sum |B_{M,n}|^2 \; / \; \sum |A_{0,n}|^2 \qquad (3.68)$$

where the summation is over all the lines n parallel to the x-direction. The boundary conditions are that $\mathbf{a}_M$ is zero and $\mathbf{a}_0$ is an exact eigenfunction of the transverse (y) direction (since the wavefunction is single-valued at the nodes this implies a plane wave of constant amplitude in the x direction). Thus for transverse state λ_n the elements of $\mathbf{a}_o$ are given by:

$$\mathrm{Re}[\mathbf{a}_0(n)] = \mathrm{SIN}(n.\lambda_n.\pi/N) \qquad (3.69a)$$

$$\mathrm{Im}[\mathbf{a}_0(n)] = 0 \qquad (3.69b)$$

the phase being arbitrary since the negative going wave is still to be determined.

3.35 Bandstructure of Network

As we have indicated the bandstructure of the network itself is not the desired objective here although it is crucial to reflecting the 2D nature of the system as we show in chapter 4. The effect is contradictory however; the bandstructure also acts as a perturbation on the result and too much influence will render the model invalid. As such it is important to be aware of the effects of this perturbation and what limitations it imposes on the applicability of the method.

Since our potential is not periodic with the network we cannot apply the bandstructure expression derived by Coulson (3.7) (if we did, with $k_x = k_y$ we would have the free-electron result i.e. q = k). we can though make some empirical observations. The band edges will be at $k\Delta = 0$ and $k\Delta = \pi$. The other natural period in the system is the width of the system itself, since we have demanded that the wavefunction vanishes at these boundaries. This provides the position of the band-centre i.e. the energy of maximum transmission. This occurs at $k.\Delta.N = \pi$ for a network N lines wide. Since $\Delta = L_y/N$ this quantity is independent of Δ for a system of a given width. The wavevector $k = \pi/L_y$ of course corresponds to the first transverse bound state energy. The value of Δ does have an effect on the shape of the transmission curve as we illustrate for a network on a zero-potential system 100Å wide using two values of N (Fig 23). The value of M (the number of lines in the x-direction) is irrelevant since we are only looking at the transmission from the boundary at x=0 to x=Δ. The black curve shows the transmission curve for N=3. Clearly with such a course grid the bandstructure is an un-acceptable effect. Compare this with the green curve where N is 50. Here we see the curve is flattened out around the band-centre ($k = \pi/L_y$) and compares well with the 1D analytic solution (red curve). The latter is obtained by assuming that the constriction of the channel, width Ly, acts as an effective potential barrier of height $E_{y,0}$ and length Δ, where $E_{y,0}$ is the energy corresponding to $k = \pi/L_y$. The equation is given by [see e.g. Winter, 1979]:

$$T = \frac{1}{1 + \frac{E_{y,0}^2 \, SIN^2(k\Delta)}{4E(E - E_{y,0})}} \qquad E > E_{y,0}$$

$$T = \frac{1}{1 + \frac{E_{y,0}^2 \, SINH^2(k\Delta)}{4E(E_{y,0} - E)}} \qquad E < E_{y,0} \qquad (3.70)$$

The conclusion to be drawn here is that for a sufficiently fine network the bandstructure has little influence around an operating point k = π/Ly although only at the operating point can the effect be totally ignored.

4 APPLICATIONS

4.1 Introduction

In this section we apply the network method for the transmission curve and the wavefunction to various structures and discuss the results with comparison to two other theoretical models. We start with a simple channel and progress to more unusual waveguide type structures. We explain the differences between our network model and the other theoretical models with particular reference to the boundary conditions.

4.2 Bound States

Of primary interest are applications to travelling-wave problems but the bound state solution allows us to demonstrate the correctness of our energy scaling argument (section 3.32) and the genuinely coupled nature of the network.

In Figs 25a,b we show one of the degenerate eigenfunctions for the eigenvalue E = 1110meV for an empty (zero potential) quantum well 50Å by 50Å. In terms of the decomposition $\Psi = \phi_{x,m}\,\phi_{y,n}$ the solution is the product of the second 1D state in the x direction (m=2) with the first 1D state in the y direction (n=1). By the analytic formula for the eigenvalue:

$$E = E_x + E_y = (h^2\pi^2/2m^*)[\ (m/l_x)^2\ +\ (n/l_y)^2\ ] \qquad (4.1)$$

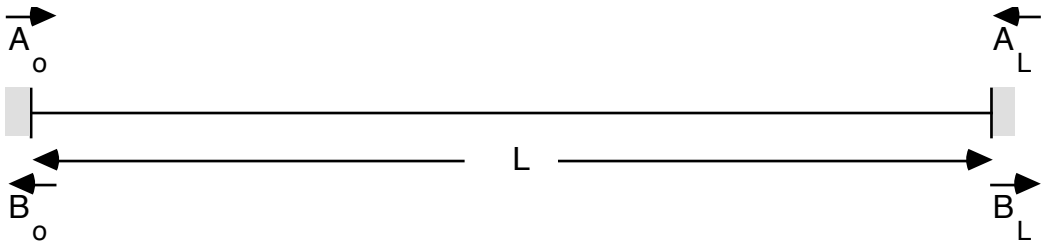
the value of 1110 meV is distributed unevenly between the x and y states. Thus in using half the eigenvalue in our simulation, i.e. the average of E_x and E_y , we are not merely propagating the separate 1D forms and combining them; there is no decomposition into x and y direction states. The boundary conditions used for Figs 25a,b are the T-matrix ones i.e. $\mathbf{a}_0 = -\mathbf{b}_0$. This produces the correct result but is unstable for large grids (see section 3.34).

For stability the boundary conditions given must be the inputs at either end of the system. In 1D this presents no problem. With reference to Fig 24 we know that the

value of B_0 must be such that the sum $\Psi_0 = A_0 + B_0$ is zero and A_L is simply related to A_0 by a phase factor. Thus:

$$A_L = -e^{-ikL}A_0 \qquad (4.2)$$

Fig 21 Boundary Components of the Wavefunction for a 1D Line



For a coupled network it is not obvious how to determine the values of the A_L's. Empirically we have found that using equation (4.2) as if the lines were decoupled, with the value of k derived from half the true eigenvalue, gives a solution which vanishes along the boundaries but has an incorrect phase. Thus in Figs 26a,b we see that the peaks are shifted to the right and the fall to zero at the ends is too steep. Of course expression (4.2) is in any case no use for calculating bound states for a region containing a changing potential.

4.3 Simple Channel

4.31 Theory of Conductance Quantisation

The unusual feature of the quantum channel, i.e. a channel whose width is narrow enough for transverse quantisation to be significant, is the steplike nature of the conductance. The theoretical development of this concept starts with Landauer's [1957, 1970] proposal connecting the conductance to the transmission coefficient. Landauer's original formula was strictly one-dimensional and based on a marriage between classical diffusion and quantum mechanical probability of occupancy of states to give:

$$g \propto T/R \tag{4.3}$$

where T is the modulus squared of the transmission and R similarly for the reflection. The expression (4.3) was generalised by Fisher and Lee [1981] to give:

$$g \propto Tr(\mathbf{T}^{\dagger}\mathbf{T}) \tag{4.4}$$

for the multi-channel case where Tr denotes the trace and the quantity in the brackets is the product of the transmission matrix with its hermition conjugate. This itself is a generalisation of an alternative single channel result due to Economou and Soukoulis [1981]:

$$g \propto T \tag{4.5}$$

Equation (4.4) is essentially defining the transmission coefficient for a multi channel system. For our purposes therefore, since we have already defined the 2D transmission coefficient explicitly in terms of current ratios (3.68), we may write (4.4) as $g \propto T$, where T is the 2D transmission coefficient derived from the multi channel network. Thus we consider the 2D transmission coefficient for a single transverse mode as emanating from a single channel system and consistency with (4.4) dictates that for many channels i.e. more than one transverse mode the conductance is given by:

$$g \propto \sum T_n \tag{4.6}$$

As van Wees et al [1988] point out this is a special case of the Fisher and Lee multi channel formula (4.4) where the T-matrix is diagonal, i.e. there is no mode mixing. (Remember this refers to the different transverse modes; of course there is mixing between the channels that form the 2D network)

Alternatively we may calculate the total transmission coefficient in one go by using a superposition of modes as the initial condition and use (4.5). Thus for occupancy of the first two transverse modes we would have:

$$\mathbf{b}_M = \mathbf{S}_{21}\,(\mathbf{a}_{O(1)}+\mathbf{a}_{O(2)}) \tag{4.7}$$

for the output vector on the right hand side from which the transmission coefficient is obtained using (3.68).

In summary then we find that there are two routes to the transmission coefficient. In both cases we sum over currents on all lines and define the transmission coefficient as the ratio of output current on the right hand side to the input current on the left hand side. Using (4.7) and (4.5) the system is considered as a single channel whilst (4.6) assumes several non-mixing channels. In both cases the total transmission and the conduction are synonymous. In appendix 4 we show that although the output wavefunction $\mathbf{b}_0$ is not affected by the choice of route we do not in general expect identical results for the transmission and therefore conductance.

It will be noticed immediately that (4.5) implies a finite conductance and therefore a non-zero resistance. We return to this point below. The expression (4.5) also provides the quantization of the conductance, demonstrated experimentally by Thornton et al [1986] van Wees et al [1988] and Wharam et al [1988], as we now explain.

Inside the channel the time-independent solution of the wavefunction in the transverse direction must be an eigenfunction of the confining potential. Since this has an energy $E_{y,n}$ for the nth mode the energy of the electron is:

$$E = E_{y,n} + E_x \qquad (4.8)$$

Thus for a total energy that is less than $E_{y,n}$ the x-direction energy is negative, the wavevector k_x imaginary and the wavefunction either grows or decays. In other words at energies below the transverse energy for the mode the electron energy is below the top of an effective barrier of height $E_{y,n}$ and undergoes reflection accordingly. Where more than one transverse mode is occupied the energy of the effective barrier height is different for each mode and hence each mode starts to contribute to the transmission at different energies. This is the explanation of the stepped structure of the transmission and hence of the conductance.

We disagree therefore with the suggestion of Imry [1986], Buttiker [1986] and Landauer [1957,1970] that this behaviour is due to a ' contact resistance '. This is a misleading term since the stepped effect is present even with a perfectly

transmitting interface (as we show below) but disappears when the channel length is reduced or channel width increased, thereby reducing the effective barrier height. This is shown in Figs 31 and 32. The curves are for the first six modes, using (4.7) and (4.5). The dimensions of the channel are 1Å long by 50Å wide for Fig 31 and 50Å long by 2500Å wide for Fig 32. The more modes that are included the more the steps in the curve are smoothed out. In other words the reflection undergone is a function of the whole channel/interface system and not just the interface.

4.32 Results of the Three Different Models

We present results for the transmission curve for a channel 50Å wide by 500Å long in Figs 27 and 28. Fig 27 is for the lowest mode only, Fig 28 has contributions from the first two modes. In Fig 29 we show the first two modes for a channel 50Å by 50Å. In all the plots there are three curves; the network result, Kirczenow's result [1988] and an analytic result which we explain below. The network result in Figs 28 and 29 is obtained using (4.6). In Fig (30) we compare results for (4.6) with those for (4.7), for the parameters of Fig 28. Clearly (4.6) and (4.7) give almost identical results for these initial conditions.

It can be seen that all three models reproduce the step-like structure with the steps at roughly the energies of the lateral eigenvalues, indicated by arrows on the energy axes. However from Figs 27 and 28 particularly we can see several differences in the fine structure between the models. Most strikingly there is an oscillation in the network result that persists to higher energies than in the other two cases. Also the amplitude of oscillations in the fine structure differs from model to model and the position of the peaks is different in the Kirczenow case from the other two. (Since the Kirczenow results are traced from plots with no divisions on the energy axes a slight misalignment must be considered possible but not sufficient to alter the comparisons qualitatively.

Let us consider first the position of the peaks. Kirczenow notes that the peak separation for his curve can be related to the de Broglie wavelength of the electron. The oscillations are therefore, he concludes, associated with resonant energy levels that occur when an integral number of half wavelengths fit into the length of the channel, adjusted for end effects. To account for the end effects Kirczenow introduces a fitting parameter γ. Thus his resonant condition is [1988]:

$$n\lambda_F = 2\gamma L_x \qquad (4.9)$$

For the network we find a similar structure but, for a single mode, the positions of the peaks coincide exactly with those of the analytic model. The latter is simply the 1D transmission coefficient for a single barrier whose length is the length of the channel and whose height is given by the eigenvalue for the transverse mode being considered. The expression is (see section 3.35):

$$T = \frac{1}{1 + \dfrac{E_{y,1}^2 SIN^2(k_2 L_x)}{4E(E - E_{y,1})}} \qquad E > E_{y,1}$$

$$T = \frac{1}{1 + \dfrac{E_{y,1}^2 SINH^2(k_2 L_x)}{4E(E_{y,1} - E)}} \qquad E < E_{y,1} \qquad (4.10)$$

Of course the resonant condition is governed here by the same kL_x product as in (4.9), with $\gamma = 1$, but the physical picture is different. Using (4.10) we require not only the idea of the longitudinal ' organ pipe ' resonance but also the effective barrier, explicitly employed to predict the position and separation of the peaks.

It is not surprising that the Kirczenow model gives slightly different results since the system being modelled differs from that of the network. Kirczenow includes the interfaces between the contact regions and the channel, which is an additional source of possible reflection. His wavefunction in the contacts is a superposition of x-direction states all modified by a single plane wave in the y direction. This is matched to the wavefunction in the channel *over the width of the channel only*.

By contrast the network model considers the channel *in isolation* from the contacts or, in what amounts to the same thing, attached to the contacts by perfectly

transmitting interfaces. For a long thin barrier, as in Figs 27 and 28, it is not surprising then that the channel itself dominates the transmission and the network and Kirczenow results are in close but not exact agreement.

The analytic model is somewhere between the other two. Since it is a 1D model it implies a plane wave in the transverse direction although this is true in all regions, not just in the contacts. No contact/channel interface is therefore included. In the x direction there is only one k value at any point in space, not a superposition. In this respect it is again more like the network model. We demonstrate later (Fig 34) that the network model allows no mode mixing for the simple channel and therefore it is not surprising that it resembles the analytic model quite closely.

From Fig 29 we see that the comparisons made above differ little for a shorter channel of the same width. The network and analytic results are for a 50Å by 50Å channel; Kirczenow quotes only the aspect ratio and in fact all the models scale in this way (the curves for channels of different dimensions but the same aspect ratio will coincide if the energy scale is adjusted to lie between the same transverse eigenvalues of the channel, e.g. between $E_{y,1}$ and $E_{y,3}$ as in Figs 27-29. This is simply because $E_{y,n}$ is proportional to n^2/L_y^2. This is demonstrated for the network by Figs 31 and 32). The relatively good agreement between the Kirczenow and network results suggests that the effective barrier of the channel dominates the transmission even at an aspect ratio of 1:1. This view is supported by the much reduced structure in all the curves.

Kirczenow's treatment of the interface is not the only one possible. In reality the wavefunction just outside the interface will not be a plane wave in the y direction because of scattering, as pointed out by van der Marel and Haanappel [1989]. In considering current flow only over the width of the channel and assuming a plane wave any effects of the interfaces are greatly reduced and it is quite consistent that the transmission tends to unity as the energy becomes larger than the effective barrier height. For this reason we disagree with the implication of Kirczenow and van der Marel and others, contained in (4.4), that the transmission is independent of the form of the wavefunction in the contacts. Not only is the wavefunction constrained in the construction of a *finite* transfer matrix (in our case by demanding it to be zero on the transverse boundaries) but also because the formulation of adding currents:

$$T = |B_{M,1}|^2 + |B_{M,2}|^2 + \ldots |B_{M,N}|^2 / |A_{0,1}|^2 + |A_{0,2}|^2 + \ldots |A_{0,N}|^2 \quad (4.11)$$

is not independent of the column vector $\mathbf{a}_0$, the initial condition.

We must now address the persistent oscillation in the network transmission curve. We suggest three possible explanations and rule out two of them. Perhaps the most obvious possibility is that it is an effect of the discretisation i.e. the bandstructure of the network itself is modifying the result. From section 3.35 we know that this is possible, however we offer two pieces of evidence against it. Firstly, by altering the separation between grid points we can alter the bandstructure of the network. This however makes no change to the transmission curve, as shown in Fig 33 for grids of 30 by 30 (black) and 50 by 50 (green) for a 100Å by 100Å channel. Secondly we use a quantitative analysis of the effect of mixing a network state with the physical state of the system to demonstrate that this is not occurring.

We wish to observe the effect of mixing a network state , wavevector k_2, with the physical state, wavevector k_1. (Remember that for the network $k_x=k_y=k$) We choose k_1 to be high enough to expect unity transmission from the other two models and hence we can write the wavefunction as :

$$\Psi = a_1 e^{ik_1 x} \phi_1 + a_2 e^{ik_2 x} \phi_2 \quad (4.12)$$

Assuming the perturbing state has a smaller weighting i.e. $a_1 \gg a_2$ and both real, we can write:

$$|\Psi|^2 = a_1^2 \phi_1^2 + 2a_1 a_2 \mathrm{COS}(\Delta k x)\phi_1\phi_2 \quad (4.13)$$

where Δk is k_2-k_1. The nearest network state is the $N\text{-}1^{th}$ Fourier component of the system (the N^{th} component is exactly zero) as seen from Fig 34, the Fourier spectrum for a 100Å by 100Å channel with the first mode occupied:

$$\phi_1 = \mathrm{SIN}(n\pi/N)$$

$$\phi_2 = \mathrm{SIN}((N-1)n\pi/N) \quad (4.14)$$

Thus $\phi_1 \approx \phi_2$ and $\phi_1\phi_2 \approx |\phi_1|^2$ and hence expression (4.13) contains two terms; the first the zero reflection result and the second a standing wave perturbation with a wavelength:

$$\lambda = 2\pi/\Delta k \qquad (4.15)$$

From (4.14) we can see that $k_2 = (N-1)\pi/L_y$. Since therefore $k_1 \ll k_2$ the wavelength λ should change significantly with any change in N. Fig 35 showing contour plots of the modulus of the wavefunction at 400 meV for networks of 20 by 20 (black) and 50 by 50 (green) demonstrates that this is not the case (expected change in $\lambda \approx 60\%$).

The second possibility is that mixing is occurring between two or more physical states but this is disproved by the Fourier spectrum, Fig 34. The third hypothesis, which we can neither confirm nor deny, is that the persistent oscillation is due to the channel/contact interface. There are however reasons for thinking this unlikely. Our model can be thought of as not excluding the interface but rather assuming a special case, i.e. one that is perfectly transmitting. Thus if we were to include a non-perfect interface in a revised model we would still presumably have this oscillation present; the oscillation is consistent with an effective potential that increases with energy (see equation (4.10)) and it is difficult to imagine the interface compensating for this by providing a negative potential with a similar energy dependence.

In order to try and discover more about the transmission curve we have compared the network results with those of an analytic model for the wavefunction and transmission. This model differs in respect of (4.10) in that the S-matrix boundary conditions (i.e. we specify the inputs to the system) are used in order not to introduce a systematic phase difference in the wavefunction between the network and analytic models. We assume the separation $\Psi=\phi_x\phi_y$ as before and consider only energies above the effective barrier $E_{y,1}$. The x-direction wavefunction in the three regions is:

region I $\qquad \phi_x = e^{ik_1x} + be^{-ik_1x}$

$$\text{region II} \quad \phi_x = ce^{ik_2x} + de^{-ik_2x} \tag{4.16}$$

$$\text{region III} \quad \phi_x = fe^{ik_1x}$$

where region I is to the left of the channel, region II is inside the channel and region III is to the right of the channel. The wavevectors are given by:

$$k_1 = [2m^*E]^{1/2}/h$$

$$k_2 = [2m^*(E-E_{y,1})]^{1/2}/h \tag{4.17}$$

Matching at the interfaces x=0 and x=L_x provides expressions for c and d:

$$c = 2k_1(k_1+k_2)/[\,(k_1+k_2)^2 - (k_1-k_2)^2e^{i2k_2L_x}\,]$$

$$d = 2k_1(k_2-k_1)/[\,(k_1+k_2)2e^{-i2k_2L_x} - (k_1-k_2)^2\,] \tag{4.18}$$

From (4.16) we can see that the wavefunction in region II (the region of interest) can only be non-oscillatory if there is no reflected part i.e. d and b are both zero. From the matching conditions this means 1 + b = c + d and therefore c must be unity. Re-writing c as:

$$c = \frac{2k_1(k_1+k_2)}{[\,(k_1+k_2)^2 - (k_1-k_2)^2\mathrm{COS}(2k_2L_x) - i\,(k_1-k_2)^2\mathrm{SIN}(2k_2L_x)\,]} \tag{4.19}$$

we see by inspection that the imaginary part disappears on the resonance condition $2k_2L_x=\pi,2\pi$ etc but the expression only goes to unity if $k_1 \approx k_2$ i.e. at high energies $E_{y,1} \ll E$. Thus the magnitude of the oscillation in the wavefunction for this model is essentially a function of energy and not of transmission. The transmission coefficient is simply $T = |f|^2$ where f is:

$$f = (k_2/k_1)[\, c\, e^{i(k_2-k_1)L} - d\, e^{-i(k_1-k_2)L}\,] \qquad (4.20)$$

We have chosen a channel of 100Å by 100Å since the fine structure is less complicated than for a channel with an aspect ratio such as in Figs 27 and 28. In Fig 36 we show the network (black) and analytic (4.20 green and 4.10 red) transmission curves for the first transverse mode. In Figs 37-41 we show contour plots of the modulus of the wavefunction for the network model (black) and (4.18, green) for various energies. In order to compare like with like it is the total energy that is used in each case i.e. for the analytic model it is $E_x + E_y$, *not just* E_x. The energies are chosen at salient points of the network transmission curve (Fig 36); at the first and second resonances (E=111meV, E=280meV), at the first un-expectedly low trough (E=400meV), at four times the first resonance (E=444meV) and at an energy where the oscillation in the analytic curve (4.10) has completely disappeared (E=700meV).

Clearly from Fig 36 the analytic 1D result depends markedly on the boundary conditions. In Fig 42 we show that this is even more important in the 2D network result. In Fig 42 the T-matrix b.c.'s are used i.e. referring to Fig 20 (ch 3) $\mathbf{a}_M$ and $\mathbf{b}_M$ are specified. This result also depends very much on the size of the network, in contrast to the S-matrix result.

The first thing to note from the contour plots is that the amplitude of the oscillation in the network wavefunction result maintains a constant value as the energy increases. This at least is consistent with the persistent oscillation in the network transmission curve. The only exception to this is at E=111meV (Fig 37). This energy corresponds to the k-value actually used in the calculation fitting exactly the transverse eigenstate. In this case the network lines are effectively decoupled in the x and y directions, i.e. both Ψ and $\nabla\Psi$ are continuous across a node in either direction, hence there is no mixing mechanism to introduce the effective potential, $E_{y,1}$, and thus T=1 with no modulation of Ψ. The analytic model of course explicitly includes the effective barrier and so modulation is present at this energy.

At E=444meV, Fig 40, although the k-value again exactly fits an eigenvalue (n=2) the initial condition is still for n=1 and so the decoupling does not take place. From all the energies we see that the wavelength of the wavefunction modulation is the same for both models. For the analytic model the wavelength depends on the kL_y product and this is therefore further evidence that the factor of two in the energy

used in the network is correct.

There is however a phase difference between the results of the two models, clearly seen in Fig 38 (E=280meV). By E=400meV, Fig 39, there is a significant difference in magnitude of modulation between the two models and this is even more marked at E=700meV (Fig 41).

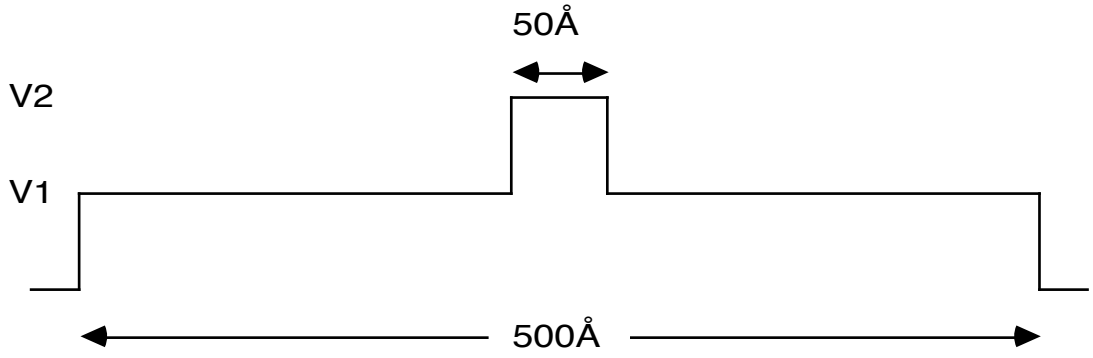
This investigation does not tell us the origin of the oscillation in the transmission curve but does confirm that the wavefunction is consistent with that oscillation.

4.4 Single and Double Barriers inside a Channel

In Fig 43 we show the conductance curve for a single barrier inside a channel for the three models, network (4.6), Kirczenow [1989] and analytic (a single transfer matrix is calculated for each region of flat potential in Fig 44 and these are multiplied together to create the T-matrix for the whole system). The channel is 500Å long by 50Å wide and the barrier is 50Å long, across the whole width of the channel, centered half way along. The barrier height is 222meV, this being the first transverse eigenvalue of the channel. In the case of the analytic model the potential profile used is shown in Fig44. The height of the barrier v_2 depends on the transverse mode,n; v_2 is 222meV above v1 ($E_{y,n}$).

It will be seen from Fig 43 that there is again broad agreement between all three models for the course structure, although to a lesser extent than for the empty channel. The steps in the network curve appear at lower energies than for the Kirczenow and analytic models. If we compare the change in the analytic result due to adding the barrier (Fig 45) with the change in the network result (Fig 46) we see that the former is affected to a greater extent, the sharpness of the step-structure being smoothed out. The same thing happens using Kirczenow's model. Presumably this effect is associated with the increase from two to four interfaces at which reflection can occur. Why it is less pronounced in the network result we do not know.

Fig 44 Equivalent 1D Potential Profile for the System of Fig 43



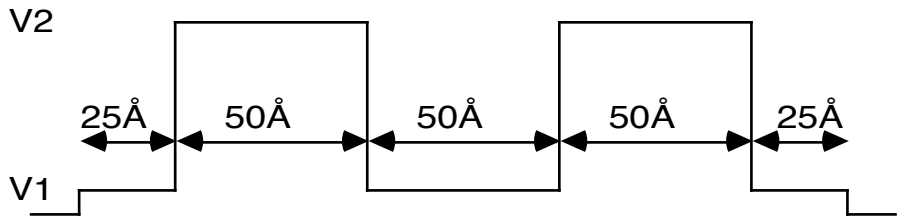
In Fig 47a we show the wavefunction at 325(18^2) meV for the network model. This energy is marked on the energy axis of Fig 43, as is the effective barrier for the first mode at 444meV. A contour version of Fig 47a is shown in Fig 47b. The transmission at this energy is 0.7 for the network model, considerably less for the other two models. It will be seen that the wavefunction appears perfectly plausible and suggests no possible explanation for the higher transmission. We can only suggest that it is either a genuine dimensional effect not reproduced by the Kirczenow model or it is due to the use of the S-matrix boundary conditions.

Finally for the single barrier we compare results for the network using (4.6) and (4.7). From Fig 48 we see that as for the empty channel the results are very similar.

We turn now to the double barrier inside a channel. In Fig 49 we show the transmission curve for the network model with the first two modes included. A channel aspect ratio of 1:1 (200Å by 200Å) is chosen to reduce the fine structure of the empty channel thus allowing barrier effects to be clearly identified. The energy range for Fig 49 is from zero to the third transverse eigenvalue of the channel and the first and second eigenvalues are marked, as is the total resonant well energy. The barrier and well dimensions (well length and barrier length both 50Å, barrier height 250meV) ensure that the density of transverse states is greater than that of the longitudinal resonant well states. We can see straight away that the step structure has been almost completely destroyed. According to Bryant [1989] there should be one peak for each occupied contact subband (in this case two) for each

resonant well state (in this case one). Fig 49 is in qualitative agreement with this. In Fig 50 we show the first mode only (black curve) on an extended energy scale, up to 300meV (36 meV above the effective barrier for the first mode). We also show an analytic 1D result obtained from the transfer matrix as for the single barrier (red curve). The potential for this model is shown in Fig 51.

Fig 51 Equivalent 1D Potential Profile for the System of Fig 50



Again this is in agreement with Bryant's supposition, although it is clear from Fig 50 that for the black curve, grid size 48 by 48, the peaks are not very close to the analytic ones. The green and blue curves are the network results using grids of 24 by 24 and 16 by 16 respectively. There appears to be a trend towards the analytic result. This seems reasonable since the smaller the grid, the less mixing is allowed between the lines in the x and y directions and the more the x -direction lines are carrying an un-interrupted 1D potential. Note also the decreasing width of the resonances with grid size. Our restraint $k.\Delta x \ll \pi$ (band edge of the network, section 3.35) prevents any smaller grids being tried.

In Fig 52 we compare the double barrier result for the first mode (black curve) with the empty channel result (green curve). The persistent oscillation in the empty channel result has sufficient magnitude to influence the double barrier result.

Returning to Fig 50 and assuming the black curve to be the most representative of the 2D result we see that the resonances occur at lower energies than in the 1D model. In Fig 53a we plot the modulus of the wavefunction at the first network resonance, 50meV, with a contour plot shown in Fig 53b. Clearly it resembles the lowest symmetric quasi-eigenstate of the well. However from the contour plot it can be seen that the peak of the wavefunction is not centered halfway between the barriers, the edges of which are shown in green, as would be the case for a 1D model,

but over to one side. The analytic result is for the total energy (see section 4.32) i.e. $E_{y,1}$ is included. Therefore the lowest quasi-eigenvalue of the 2D well is simply 95meV, read from the red curve. This is approximately twice the first network resonance (50meV). In Fig 54a we show the modulus of the wavefunction at the second resonant energy, 170meV. Once again the quasi-state is in the lowest transverse mode but in approximately the second mode in the x direction. This time we see from the contour plot, Fig54b, that the two peaks are both centered *inside* the barriers. From the analytic curve this quasi-state has an eigenvalue of 285meV. This is rather less than twice the network value. However it will be seen that whilst for the first peak there is no change in energy between the 24 by 24 and 48 by 48 results, this is not the case for the second. We have plotted a portion of the curves for 64 by 64 and 96 by 96 grids around the second peak (red and green portions respectively). The peaks for these are $\approx$ 165meV and $\approx$ 162.5 meV, thus it is reasonable to conclude that the limit of the network result provides a peak at no less than 160meV. This is still more than half of the analytic value (185/2 = 142.5meV).

We can see no reason for this discrepancy in the resonant energies between the two models. It is tempting to link it to the energy scaling introduced in the expression for the Schroedinger equation at the nodes but this has created no such similar discrepancies in results for the empty channel. Neither can it be simply due to the destruction of the phase relationship associated with the 1D resonances (i.e. the wavelength fitting the wavelength of the quasi well state); if it was there would be no strong resonance at all in the network result. However the resolution of this problem would seem to lie in interpretation; there cannot be a resonance below the minimum eigenvalue of the 2D well.

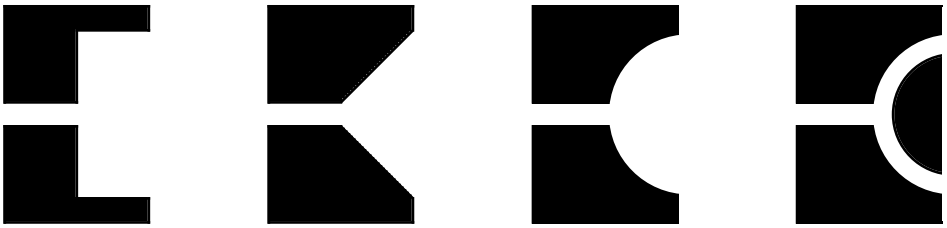
4.5 Quantum Waveguides

In this section we present results for the wavefunction and transmission for a selection of novel waveguide geometries. These are shown in Figs 55a-55f. They are the abrupt junction throttle, the tapered throttle, the half circle, the half ring, the full circle and the full ring. All results in this section are for the network.

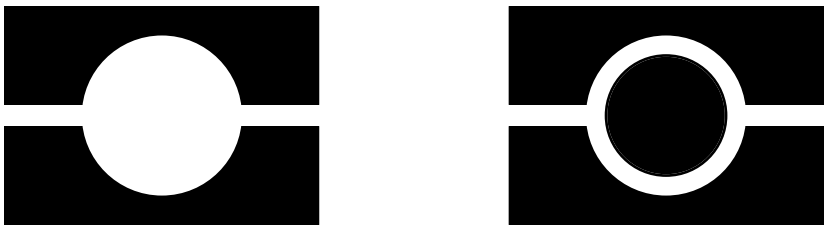
The dimensions are those used by Finch [1989] for his circle and ring models; maximum channel/ring width 4000Å with narrow channels one sixth of this (≈ 660Å). Grid size is 48 by 48 for aspect ratio 1:1 and 96 by 48 for aspect ratio 2:1. The first transverse mode only is used as the initial condition in all cases.

Fig 55 Novel Geometry Waveguides

a) Abrupt Junction b) Tapered Junction c) Half Circle d) Half Ring



e) Full Circle f) Full Ring



Starting with the abrupt junction throttle we show the transmission curve in Fig 56, over an energy range that includes the first three transverse eigenvalues for the narrow channel. The first transverse eigenvalue of the wide channel, $E_{y,1}(w)$ is at 0.03meV, $1/36^{th}$ of $E_{y,1}(n)$. We see a cutoff at around the $E_{y,1}(n)$ value of 1.27meV. We see also the small amplitude oscillation persisting to high energies. Unexpectedly there are also two sharp dips in the curve. However we would not expect to see a simple step up to unity transmission since although the energy is well above $E_{y,1}(w)$ the input into the wide channel is not in the first mode; indeed it is not in an eigenmode at all. The form of the wavefunction emerging from the narrow channel is a strongly localized portion of a sinusoid, implying a Fourier spectrum involving high frequencies. Note that because the ratio of channel widths is 6:1 the emerging waveform is never commensurate with one of the wide channel eigenmodes. Perhaps the surprising thing after all is not the existence of the two dips but that the

transmission is generally high for energies above $E_{y,1}(n)$. These dips are not predicted by Frohne and Datta [1988] or by Kriman and Ruden [1985].

In Figs 57a,b we show the modulus of the wavefunction at 2.54meV, twice $E_{y,1}(n)$. The transmission is 0.78. From the contour plot we can see strong scattering into the 7^{th} and 14^{th} modes at the interface but a strong relaxation towards the end of the wide channel with a high percentage of the 5^{th} mode present. Figs 58a,b show the wavefunction at the same energy for a channel width ratio of 2:1 (i.e. the narrow channel is $\approx$ 2000Å wide). This time we are looking from the wide channel in the 3D plot. We can see that the relaxation is more advanced, a high proportion of the first mode being present at the end of the wide channel. The high frequency spikes are still present at the interface. If we reduce the physical scale to 100Å maximum width (Figs 59a,b at the equivalent energy $2E_{y,1}(n)$) the spikes disappear and the wavefunction is almost entirely in the first mode at the end of the channel. The transmission coefficients for Figs 58 and 59 are 0.9 and 0.94. The wavefunction at a higher energy, 11.27meV, is shown in Figs 60a,b, for the original geometry, for which the transmission is 0.98. Comparing Figs 60b and 57b we see that at the higher energy the localized form of the wavefunction penetrates deeper into the wide channel.

In Figs 61a,b and 62a,b we show the wavefunction at the 8th and 12th transverse eigenvalues of the wide channel. These energies are marked on the transmission curve Fig 56 and coincide with the dips. These plots offer no obvious explanation for the latter.

A tapered junction is now studied. From the transmission curve, Fig 63, we see immediately that there are no large dips above the cutoff. From the wavefunction plot at 2.54meV, Figs 64a,b we see a higher proportion of the first mode present than for the abrupt junction. Also scattering into higher modes at the interface is less prominent. The transmission is also higher at 0.96. The higher energy plot, 11.27meV, is shown in Figs 65a,b, for which the transmission is the same as for the abrupt junction at 0.98.

The half circle and half ring devices are treated in Figs 66 - 69. Comparing the half circle, Figs 66a,b, at the low energy (2.54meV) with the abrupt and tapered junctions we see slightly spikier structure than for the tapered version with the transmission about the same, both being greater than for the abrupt case. The

wavefunction relaxes mainly into the second mode by the end of the half circle. This is all consistent with the potential changing at a rate less than that of the abrupt case but greater than for the taper. There are no real differences observed between any of the three types of junction at the higher energy (11.27meV, Figs 67a,b for the half circle).

For the half-ring at the lower energy, Figs 68a,b, there is a high transmission (0.91) with the wavefunction remaining mainly in the first mode round the arms of the ring, with some spikes on the outside edge of the ring near the interface as for previous devices. At the higher energy, Figs 69a,b, the scattering is much less clean with a spike on the inside edge of the ring opposite the junction and some reflection back into the input channel in the second mode (identifiable from the contour plot). By the end of the arms of the ring the wavefunction has relaxed mainly into the third mode. The energy of 11.27meV is just below the third transverse eigenvalue of the narrow channel of 11.43meV.

We end with the full circle and full ring devices. The low energy result for the full circle is shown in Figs 70a,b. The insert in Fig 70b is Finch's [1989] result for the time dependent model after 10ps, the wavepacket having been allowed to diffuse from a localized form in the input channel on the left. In both the network and time-dependent results we can see concentric circle structure forming. This structure is not perfect but then the energy is not an exact eigenvalue of the structure and we do not have a circle, we have a circle with two gaps in it. Both results also display low probability in the central region of the circle and a concentration in front of the output channel interface. The first eigenvalue of the true circle with these dimensions is at 0.008meV and so 2.27meV is well into the high density of states energy range. At the higher energy, Figs 71a,b, structure is more complicated but we still see a low concentration in the central region and a high concentration in front of the output channel.

For the full ring at both energies, Figs 72a,b and 73a,b - 75a,b there are again similarities with Finch's results for the time-dependent model. As in the half ring the wavefunction remains mainly in the first mode round the ring and into the output channel at the lower energy. For the higher energy we show three plots for which the barrier potential (from which the channels are carved out) is 5000meV (Figs73a,b), 2000meV (Figs74a,b) and 1000meV (Figs 75a,b). The result most resembling Finch's is the one with the 2000meV barriers, for which the

wavefunction is scattered mainly into the first and third modes in the ring, relaxing back into the first mode in the output channel. The transmission for the three potentials is 0.67, 0.62 and 0.20 respectively for 5000meV, 2000meV and 1000meV. It is curious that the change in potential above 1000meV should make such a difference to an electron with an energy of only 11.27meV. Finch's figure for the transmission is 0.45. Of course the results are for different systems, one time-dependent the other time independent.

4.6 Summary

The use of the network method has been demonstrated in section 4.2 with application to the bound state problem. We have found that using T-matrix boundary conditions reproduces the correct eigenfunction thus vindicating the energy scaling argument and the truly coupled nature of the network.

In section 4.3 we have introduced the topic of conductance quantization in a quantum channel and shown that both 1D analytic and 2D models produce this result. Furthermore we have shown that the position of the peaks in the fine structure coincide for the network and 1D analytic models. We have argued that the fine structure present on the transmission curves depends crucially on the boundary conditions used and on the treatment of the channel/contact interfaces. We have also suggested that the course structure (the quantisation) should not be thought of as an interface effect. The logical conclusion of these points is that there is no one 'correct' transmission curve and it is more useful to adopt the conductance expression that depends on the transmission curve, and thus on the boundary conditions, rather than on the invariant transmission (or scattering) matrix which is abstracted from the particular physical conditions pertaining.

We have been unable to eliminate or explain a persistent oscillation in the transmission curve for a channel, using the network method, of about 5% although we have eliminated mode mixing and discretisation error and we think interface complication unlikely to be the cause.

For the single barrier inside a channel (section 4.4) we see similar agreement between the three models although adding the barrier affects the analytic and Kirczenow results to a greater extent than the network, the sub-barrier peaks being

nearer to unity for the latter than for the other two models. When we add a double barrier to the channel we find that the clear quantization structure is almost completely destroyed. Comparing the network result using various grid sizes, with the analytic result we find that the resonant peaks do *not* coincide. There does appear however to be a trend towards the analytic result as the grid size is reduced. The wavefunction at the network resonances is *not* the same as we would get from a 1D result; the peaks are shifted i.e. the wavefunction at resonance does not coincide with the quasi-eigenfunctions of the 2D well.

The waveguide section (4.5) provides some interesting results. The transmission curve for an abrupt junction throttle has sharp dips in it which we cannot explain. These disappear for a tapered junction. In the abrupt, half circle and tapered guides we see a tendency for the wavefunction to relax to the original first mode in the wider channel, having been scattered into higher modes at the interface. This is most advanced in the tapered junction at low energy (energy not much higher than the first transverse eigenvalue).

For the half ring we find the scattering dependent on energy with the first mode persisting in the ring for low energy and the third mode being prominent for an energy close to the third transverse eigenvalue. For the full circle we can draw broad comparisons with Finch's time-dependent result for the wavefunction at low energy. This is also true for the full ring at the low energy. The scattering for the full ring at the higher energy depends to a surprisingly great extent on the value of the potential used to define the channels. For potentials of 1000meV and 5000meV the results are less close to Finch's than for a potential of 2000meV.

5 FURTHER WORK AND CONCLUSION

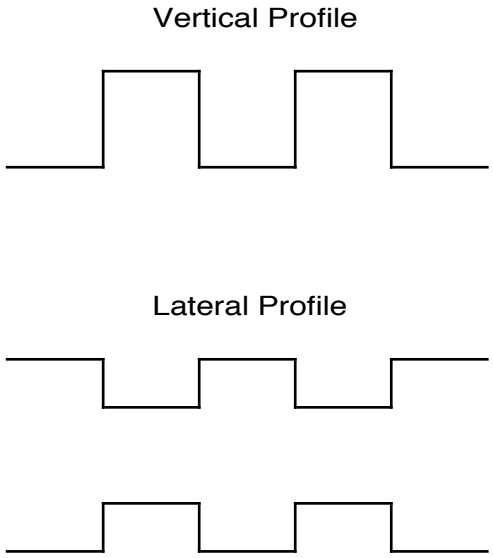
5.1 Further work

There are several topics requiring further attention within the scope of the work already carried out. Most pressingly the question of the persistent oscillation in the transmission curve for the channel needs to be resolved since this could well affect the results for more complex structures such as the double barrier where structure may be expected on the same scale as the magnitude of the oscillation. There is also the unusual result for the abrupt junction waveguide - the large dips in the transmission curve above the cut-off. Calculations for other initial conditions might provide a clue as to the origin of these. Perhaps the most revealing area to explore would be the single and double barrier systems. We still have to answer the question " Under what conditions can these systems be represented by the 1D model and what exactly is the relationship between the transverse and longitudinal quantization ? ".

An interesting possibility is the resonant tunnelling structure with incomplete barriers. From Fig 76 the constrictions on the lateral device should have the same effect as the barriers in the 1D profile, since in both cases longitudinal energy is reduced as transverse energy is increased.

In order to address the above problems and to tackle new ones the technical side of the model needs enhancing. At the moment the maximum grid possible is 200 by 100; this is the limit imposed by the maximum virtual machine size on the IBM 3090 mainframe on which these calculations are performed. This limit is in the process of being eliminated by a hardware/software update, although increasing the grid size is by itself not sufficient because of the prohibitive cost in time (proportional to N^3 for a grid of width N). However there are approaches which could be made to increase speed of computation. Firstly the programme is not as efficient as it could be. This is especially true when calculating the wavefunction, as we point out at the end of chapter 3. Secondly the Fortran code could be re-written to be more suitable to vectorization. Thirdly we have not investigated precisely the limits to how closely the potential profile must be followed by the discretisation. Finally we could also utilize conformal mapping in order to make more efficient use of the grid by ' fitting ' the potential to it.

Fig 76 Equivalent Resonant Tunnelling Potential Profiles



Other devices that we have not had time to investigate include tuning stubs, the superlattice, quantum dot arrays (or quasi-one-dimensional channels) and Y-junctions. Also disordered systems would be very interesting to study; by calculating the wavefunction we would be able to see the most likely paths through a disordered region of potential, for instance under the gate of an FET.

A major extension to the model would be to include the time dependent wavefunction. This would be very simple to do. We calculate any number of time independent wavefunctions using the present model and construct a wavepacket with each component modified by the appropriate time dependent factor:

$$\Psi(r,t) = \sum \phi_m(r,E_m)\, e^{-iE_m t/h} \qquad (5.1)$$

The time required is mainly to create the initial condition, i.e. the time independent wavefunctions, the time propagation being much quicker (multiplying by the time factor and adding both being simply proportional to the size of the matrix). To compare with Finch's model for a specific example, the 4000Å ring, the network takes about four-fifths of the time of Finch's 10ps time dependent model to calculate one time independent wavefunction. Thus to calculate a time dependent model using a wavepacket with ten components would take about eight times as long plus a

short time extra for the time propagation.

The choice of wavepacket would be interesting. We could follow Finch's [1989] scheme and use wavefunctions with the same transverse component but with different energies. Alternatively we could use wavefunctions all having different transverse modes. Most generally we could use a combination of these two.

Another extension to the model would be the inclusion of a magnetic field. The Schroedinger equation with a magnetic field present is:

$$[\ (\underline{ih\nabla - e\mathbf{A}})^2 + V\ ]\Psi = E\Psi \qquad (5.2)$$
$$2m$$

where **A** is the vector potential introduced by the magnetic field. Expanding the operator we get:

$$-(h^2/2m)\underline{d}^2 + \gamma_x \underline{d} + U_x + V\ ]\Psi = E\Psi \qquad (5.3)$$
$$dx^2 \qquad dx$$

where:

$$\gamma_x = -iehA_x/m\ , \quad Ux = e^2A_x{}^2/2m \qquad (5.4)$$

and similarly for the y direction lines. The vector potential **A** can be calculated from either of the relationships with the magnetic field:

$$\mathbf{B} = \nabla \times \mathbf{A} \qquad (5.5)$$

$$\int \mathbf{A}dl = \int \mathbf{B}ds \qquad (5.6)$$

Clearly the new equation (5.3) will alter the form of the wavefunction as well as adding a scalar potential U_x, and therefore the transmission along the lines would need re-examining although the continuity conditions remain unaltered.

Extension to three dimensions (or more) is conceptually simple although technically it might become complicated. From Avron et al's formula for the

scattering matrix (3.10) the 3D S-matrix describing the six port node is:

$$\mathbf{S}_{3D} = \frac{1}{3}\begin{pmatrix} -2 & 1 & 1 & 1 & 1 & 1 \\ 1 & -2 & 1 & 1 & 1 & 1 \\ 1 & 1 & -2 & 1 & 1 & 1 \\ 1 & 1 & 1 & -2 & 1 & 1 \\ 1 & 1 & 1 & 1 & -2 & 1 \\ 1 & 1 & 1 & 1 & 1 & -2 \end{pmatrix} \qquad (5.7)$$

5.2 Conclusion

It is fair to begin this section by acknowledging that we have asked more questions than we have answered.

We have introduced the family of lateral devices via a brief history of the progress of microelectronics and have noted that new phenomena are available for study by this route. These phenomena are time independent - conductance quantization - and time dependent - interference effects. Also an existing phenomenon - resonant tunnelling - is modified by the dimensionality of these devices.

We have reviewed existing transmission methods in both 1D and 2D and pointed out both attractive and negative features of these. In particular the most promising of the time independent models, the 2D uncoupled matching states method due to Kirczenow [1988], suffers from restrictions on the variation of potential and an associated problem with current conservation.

We have therefore introduced a new method of calculating the time independent wavefunction and transmission coefficient based on a representation of the 2D system by a 1D network. This model requires a single energy; there is *no* decomposition into x and y states. Rather a single k vector propagates 1D solutions along the lines of the network which are uniquely coupled at the nodes by a scattering matrix. Because

of this there is *no* constraint on the potential profile. The bandstructure of the network both ensures the 2D nature of the model and acts to perturb that model. This perturbation is acceptable if the $k\Delta$ product is kept away from the band edges.

The numerical stabilizing effect of using the S-matrix propagation has been shown to be different in 2D from the 1D case. In 1D either converting to a T-matrix and using T-matrix boundary conditions at each point in the discretisation or using S-matrices from either end and using S-matrix boundary conditions gives a stable solution. In 2D only the latter method is stable and even this is dependent on the particular route taken.

The mixing action of the S-matrices automatically creates the effective potential induced by the transverse quantization of a quantum channel and thus this model produces the transmission/conductance quantization recently seen experimentally. We have shown that this quantization is in broad agreement with both Kirczenow's 2D model and a 1D analytic model. However we have also shown that choice of boundary conditions can radically alter results in both 1D and 2D. We have suggested that the conductance quantization should not be thought of as a contact resistance effect since its detail depends on the whole channel. In support of this we have shown that the network result, which doesn't include interface effects, and the Kirczenow result, which does, are quite close even for a channel with an aspect ratio of 1:1.

For the more complicated double barrier in a channel our results demonstrate that the longitudinal tunnelling is affected in a non trivial way by the transverse quantization; the resonant peaks in the network result are not simply shifted in energy by the value of the effective potential induced by the transverse constriction. Moreover the network wavefunction on resonance resembles one of the quasi states of the 2D well, but is shifted in phase.

When we study waveguide devices involving a change of channel width we find a tendency in all cases for the wavefunction to relax back into the initial condition (in this case the first transverse mode; we do not know if this tendency would be to the second mode if that was the initial condition). This tendency is strongest for a tapered junction at low energy. The degree of scattering into higher modes at the interface is reduced both by reducing the channel widths ratio and by reducing the absolute size of the system. We find also that there are large dips in the transmission curve for an abrupt junction waveguide. This is contrary to the predictions of Frohne and Datta [1988] and Kriman and Ruden [1985].

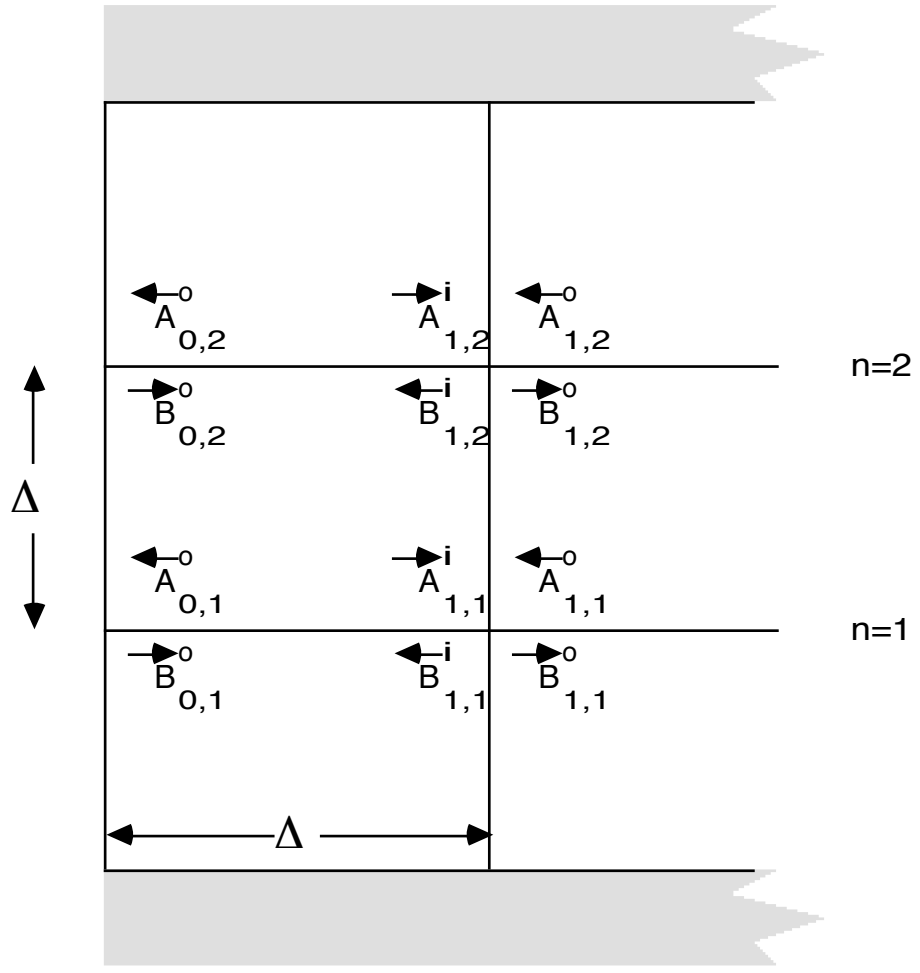
When a ring is introduced we find that although the relaxation tendency still operates over the whole device, the wavefunction scatters to a large degree into the third mode in the ring itself if the energy is near the third transverse eigenvalue. If the energy is much lower, nearer the first eigenvalue then the first mode is favoured in the ring. This effect is at its most pronounced half way round the ring. For the full circle device we find qualitative agreement with Finch's time dependent result for the modulus of the wavefunction.

APPENDIX

Appendix 1 Analysis of 2 by 2 Network

We show here that for the 2 by 2 network with zero potential the wavefunction is continuous and effectively de-coupled i.e. 1D and the current continuity is maintained but that this does not mean that the transmission coefficient is also identical to the 1D result. We also demonstrate the difference in the transmission coefficient if the T-matrix is converted to an S-matrix.

Fig A1 Two by Two Network



For an empty system (zero potential) we know that the components of the wavefunction at m=1 are connected to those at m=0 via the phase factor α:

$$A^{i}_{11} = \alpha B^{o}_{0,1} \qquad \text{(A1.1)}$$

$$B^{i}_{11} = \alpha^{-1} A^{o}_{0,1} \qquad \text{(A1.2)}$$

where

$$\alpha = e^{ik\Delta} \qquad \text{(A1.3)}$$

We take the general equation for the output components at m in terms of the input components:

$$\begin{pmatrix} A_o \\ B_o \end{pmatrix} = \begin{pmatrix} 1 - \frac{\Omega^-}{1+\Omega^-} - \frac{\Omega^+}{1+\Omega^+} & 2 - \frac{\Omega^-}{1+\Omega^-} - \frac{\Omega^+}{1+\Omega^+} \\ \frac{\Omega^-}{1+\Omega^-} + \frac{\Omega^+}{1+\Omega^+} & -1 + \frac{\Omega^-}{1+\Omega^-} + \frac{\Omega^+}{1+\Omega^+} \end{pmatrix} \begin{pmatrix} A_i \\ B_i \end{pmatrix}$$

$$+ \Gamma^- \begin{pmatrix} -1 & -1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} A^-_i \\ B^-_i \end{pmatrix} + \Gamma^+ \begin{pmatrix} -1 & -1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} A^+_i \\ B^+_i \end{pmatrix} \qquad \text{(A1.4)}$$

and substitute in the phase factor α (since the $k\Delta$ product is the same on all lines only one phase factor appears) to get the appropriate form for the outputs at m=1,n=1.

$$\begin{pmatrix} A^{o}_{1,1} \\ \\ B^{o}_{1,1} \end{pmatrix} = \begin{pmatrix} 1+\dfrac{2\alpha^2}{1-\alpha^2} & 2+\dfrac{2\alpha^2}{1-\alpha^2} \\ \dfrac{-2\alpha^2}{1-\alpha^2} & -1-\dfrac{2\alpha^2}{1-\alpha^2} \end{pmatrix} \begin{pmatrix} A^{i}_{1,1} \\ \\ B^{i}_{1,1} \end{pmatrix}$$

$$+ \frac{\alpha}{1-\alpha^2} \begin{pmatrix} -1 & -1 \\ & \\ 1 & 1 \end{pmatrix} \begin{pmatrix} A^{i}_{1,2} \\ \\ B^{i}_{1,2} \end{pmatrix} \qquad \text{(A1.5)}$$

Substituting (A1.1) and (A1.2) into (A1.5) we get:

$$A^{o}_{1,1} = (1+\alpha^2)/(1-\alpha^2)]\alpha B^{o}_{0,1} + [2\,\alpha^{-1}/(1-\alpha^2)]\,A^{o}_{0,1}$$

$$+ [\alpha^2/(1-\alpha^2)]\,B^{o}_{0,2} - [1/(1-\alpha^2)]\,A^{o}_{0,2} \qquad \text{(A1.6)}$$

$$B^{o}_{1,1} = -[2\,\alpha^3/(1-\alpha^2)]\,B^{o}_{0,1} - (1+\alpha^2)/(1-\alpha^2)]\alpha^{-1}\,A^{o}_{0,1}$$

$$+ [\alpha^2/(1-\alpha^2)]\,B^{o}_{0,2} + [1/(1-\alpha^2)]\,A^{o}_{0,2} \qquad \text{(A1.7)}$$

Adding (A1.6) to (A1.7) we get:

$$\Psi^{o}_{1,1} = \alpha B^{o}_{0,1} + \alpha^{-1}\,A^{o}_{0,1} \qquad \text{(A1.8)}$$

Thus the wavefunction appears to be de-coupled at this point. We now demonstrate the current continuity. Because of the symetry of the system we can expect the same current on both the n=1 and n=2 lines. At m=0 the current is given the arbitrary value of 1 on each line, that is:

$$A^{o}_{0,1} = 0 \qquad \text{(A1.9)}$$

$$B^{o}_{0,1} = 1 \qquad \text{(A1.10)}$$

$$I = |B^{o}_{0,1}|^2 - |A^{o}_{0,1}|^2 \qquad (A1.11)$$

The task is to show that the current at the output of node 1,1 is the same as at node 0,1 for all energies (we are excluding the pre-factor to the current expression involving k here of course). From (A1.6) and (A1.7) we have:

$$I = |(\alpha - 2\alpha^2)/(1 - \alpha^2)|^2 - |(1 + \alpha + \alpha^2)/(1 - \alpha^2)|^2 \qquad (A1.12)$$

a factor of $\alpha.\alpha^*$ having been eliminated from each term. After the necessary manipulation this is re-written as the real expression:

$$I = [\ 16 - 32COS^2(k\Delta) + 16COS^4(k\Delta)\ ] / 16SIN^4(k\Delta) \qquad (A1.13)$$

which is identically equal to 1. The transmission coefficient, again making use of the symmetry, is given by:

$$T = |B^{o}_{1,1}|^2/|B^{o}_{0,1}|^2 \qquad (A1.14)$$

and $B^{o}_{0,1} = 1$. From the first term in (A1.12) the transmission turns out to be:

$$T = \frac{20 - 16COS(k\Delta) - 20\ COS^2(k\Delta) + 16COS^3(k\Delta)}{16SIN^4(k\Delta)} \qquad (A1.15)$$

We now derive the associated expression by using the S-matrix. The T-matrix for the 2 by 2 network is:

$$\begin{pmatrix} A^{o}_{1,1} \\ A^{o}_{1,2} \\ B^{o}_{1,1} \\ B^{o}_{1,2} \end{pmatrix} = \begin{pmatrix} T_{11} & T_{12} & T_{13} & T_{14} \\ T_{21} & T_{22} & T_{23} & T_{24} \\ T_{31} & T_{32} & T_{33} & T_{34} \\ T_{41} & T_{42} & T_{43} & T_{44} \end{pmatrix} \begin{pmatrix} A^{o}_{0,1} \\ A^{o}_{0,2} \\ B^{o}_{0,1} \\ B^{o}_{0,2} \end{pmatrix} \qquad (A1.16)$$

and we require the S-matrix:

$$\begin{pmatrix} A^{o}_{0,1} \\ A^{o}_{0,2} \\ B^{o}_{1,1} \\ B^{o}_{1,2} \end{pmatrix} = \begin{pmatrix} S_{11} & S_{12} & S_{13} & S_{14} \\ S_{21} & S_{22} & S_{23} & S_{24} \\ S_{31} & S_{32} & S_{33} & S_{34} \\ S_{41} & S_{42} & S_{43} & S_{44} \end{pmatrix} \begin{pmatrix} B^{o}_{0,1} \\ B^{o}_{0,2} \\ A^{o}_{1,1} \\ B^{o}_{1,2} \end{pmatrix} \qquad \text{(A1.17)}$$

Using the boundary conditions $A^{o}_{1,1}$ and $A^{o}_{1,2}$ equal to zero and $B^{o}_{0,1}$ and $B^{o}_{0,2}$ equal to 1 we need S_{31} and S_{32} to determine $B^{o}_{1,1}$. Using the conversion formulae for the sub-matrices:

$$\mathbf{S}_{21} = \mathbf{T}_{22} - \mathbf{T}_{21}(\mathbf{T}_{11})^{-1}\mathbf{T}_{12} \qquad \text{(A1.18)}$$

we form the relationship between the S and T elements:

$$\begin{pmatrix} S_{31} & S_{32} \\ S_{41} & S_{42} \end{pmatrix} = \begin{pmatrix} T_{33} & T_{34} \\ T_{43} & T_{44} \end{pmatrix} - \begin{pmatrix} T_{31} & T_{32} \\ T_{41} & T_{42} \end{pmatrix} \begin{pmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{pmatrix}^{-1} \begin{pmatrix} T_{13} & T_{14} \\ T_{23} & T_{24} \end{pmatrix} \qquad \text{(A1.19)}$$

Expanding the matrices we get:

$$S_{31} = T_{33} - T_{31}[T_{22}T_{13} - T_{12}T_{23}]/[T_{22}T_{11} - T_{12}T_{21}]$$

$$- T_{32}[T_{11}T_{23} - T_{21}T_{13}]/[T_{22}T_{11} - T_{12}T_{21}] \qquad \text{(A1.20)}$$

$$S_{32} = T_{34} - T_{31}[T_{22}T_{14} - T_{12}T_{24}]/[T_{22}T_{11} - T_{12}T_{21}]$$

$$- T_{32}[T_{11}T_{24} - T_{21}T_{14}]/[T_{22}T_{11} - T_{12}T_{21}] \qquad \text{(A1.21)}$$

To determine the T elements we compare the matrix equation (A1.16) with (A1.6) and (A1.7). This leads to the two expressions:

$$S_{31} = [2 - 3\alpha^2 - \alpha^3 + 3\alpha^4]\alpha/[(1 - \alpha^2)(4 - \alpha^2)] \quad \text{(A1.22)}$$

$$S_{32} = [\alpha + \alpha^2 - 3\alpha^3 + \alpha^5]\alpha/[(1 - \alpha^2)(4 - \alpha^2)] \quad \text{(A1.23)}$$

From (A1.17) and the boundary conditions $B^0_{1,1}$ is simply $S_{31} + S_{32}$ and the transmission coefficient is then:

$$T = I\,[2 + \alpha - 2\alpha^2 - 4\alpha^3 + 3\alpha^4 + \alpha^5]/\,[(1 - \alpha^2)(4 - \alpha^2)]\,I^2 \quad \text{(A1.24)}$$

After some extensive algebraic work this can be re-written as:

$$T = N/D \quad \text{(A1.25)}$$

with:

$$N = 2426 + 276\text{COS}(k\Delta) - 3756\text{COS}(2k\Delta) - 646\text{COS}(3k\Delta) + 1768\text{COS}(4k\Delta) + 510\text{COS}(5k\Delta) - 494\text{COS}(6k\Delta) - 156\text{COS}(7k\Delta) + 56\text{COS}(8k\Delta) + 16\text{COS}(9k\Delta) \quad \text{(A1.26)}$$

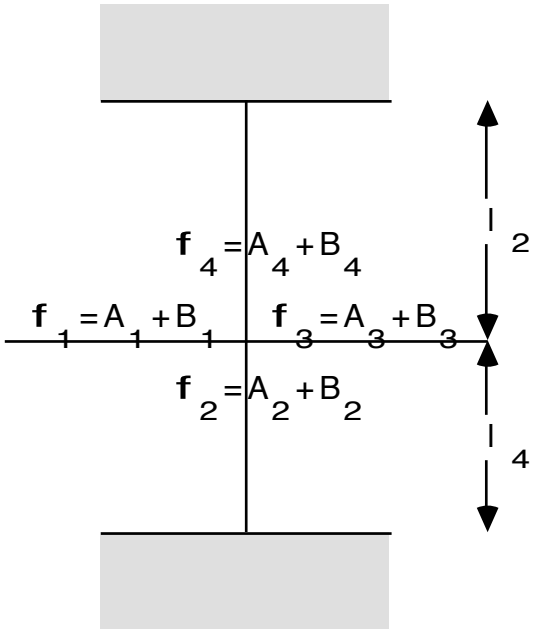
$$D = 16\text{SIN}^4(k\Delta)[\,625 - 800\text{COS}^2(k\Delta) + 256\text{COS}^4(k\Delta)\,] \quad \text{(A1.27)}$$

In Fig A2 we compare the transmission coefficients using (A1.15) and (A1.25). It can be seen that the S-matrix version lies below the value of 1 except near $k\Delta=0$ and $k\Delta=\pi$. The growth in these areas is probably due to not taking the limit correctly when the SIN term in D approaches zero. Both versions only give unity transmission at $k\Delta.N = \pi$ (for the 2 by 2 network N is 3).

Appendix 2 Dimensional Reduction of S-Matrices

We demonstrate here in some detail the reduction of a four-port to a two-port and also sketch the process whereby a 2D network may in principle be reduced to a 1D transmission line.

Fig A3 Wavefunctions Around a Four-Port



Referring to Fig A3 we require the S-matrix $\mathbf{S}^2$ (superscript here denotes a two-port, not a power) relating the outputs and inputs on lines 1 and 3, eliminating lines 2 and 4:

$$\begin{pmatrix} B_1 \\ B_3 \end{pmatrix} = \mathbf{S}^2 \begin{pmatrix} A_1 \\ A_3 \end{pmatrix} \qquad (A2.1)$$

From the four-port S-matrix:

$$\begin{pmatrix} B_1 \\ B_2 \\ B_3 \\ B_4 \end{pmatrix} = \frac{1}{2} \begin{pmatrix} -1 & 1 & 1 & 1 \\ 1 & -1 & 1 & 1 \\ 1 & 1 & -1 & 1 \\ 1 & 1 & 1 & -1 \end{pmatrix} \begin{pmatrix} A_1 \\ A_2 \\ A_3 \\ A_4 \end{pmatrix} \qquad (A2.2)$$

we need the following sub-matrices:

$$\begin{pmatrix} B_1 \\ B_3 \end{pmatrix} = \mathbf{S}^{13} \begin{pmatrix} A_1 \\ A_3 \end{pmatrix} + \mathbf{S}'^{13} \begin{pmatrix} A_2 \\ A_4 \end{pmatrix} \qquad (A2.3)$$

$$\begin{pmatrix} B_2 \\ B_4 \end{pmatrix} = \mathbf{S}^{24} \begin{pmatrix} A_2 \\ A_4 \end{pmatrix} + \mathbf{S}'^{24} \begin{pmatrix} A_1 \\ A_3 \end{pmatrix} \qquad (A2.4)$$

where:

$$\mathbf{S}^{13} = \mathbf{S}^{24} = \begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix} \qquad (A2.5)$$

$$\mathbf{S}'^{13} = \mathbf{S}'^{24} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \qquad (A2.6)$$

The problem is to eliminate A_2 and A_4 from (A2.3) but none of the submatrices have inverses. With the wavefunctions at the ends of lines 2 and 4 equal to zero we can write:

$$\begin{pmatrix} B_2 \\ B_4 \end{pmatrix} = \mathbf{S}'^{2} \begin{pmatrix} A_2 \\ A_4 \end{pmatrix} \qquad \text{(A2.7)}$$

$$\mathbf{S}'^{2} = \begin{pmatrix} -\alpha & 0 \\ 0 & -\beta \end{pmatrix} \qquad \text{(A2.8)}$$

and the phase factors are given by:

$$\alpha = e^{2ik_2 l_2} \qquad \text{(A2.9)}$$

$$\beta = e^{-2ik_4 l_4} \qquad \text{(A2.10)}$$

Substituting (A2.7) into (A2.4) we get:

$$\begin{pmatrix} A_2 \\ A_4 \end{pmatrix} = (\mathbf{S}'^{2} - \mathbf{S}^{24})^{-1} \mathbf{S}'^{24} \begin{pmatrix} A_1 \\ A_3 \end{pmatrix} \qquad \text{(A2.11)}$$

and substituting this into (A2.3):

$$\begin{pmatrix} B_1 \\ B_3 \end{pmatrix} = (\mathbf{S}^{13} + \mathbf{S}'^{13}\mathbf{G}\mathbf{S}'^{24}) \begin{pmatrix} A_1 \\ A_3 \end{pmatrix} \qquad \text{(A2.12)}$$

where G is:

$$\mathbf{G} = (\mathbf{S}'^{2} - \mathbf{S}^{24})^{-1} = \begin{pmatrix} \frac{1}{2} - \alpha & -\frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} - \beta \end{pmatrix}^{-1} \qquad \text{(A2.13)}$$

Evaluating the inverse we get:

$$\mathbf{G} = \begin{pmatrix} \frac{1}{2} - \beta & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} - \alpha \end{pmatrix} \frac{1}{\alpha\beta - \left(\frac{\alpha+\beta}{2}\right)} \qquad \text{(A2. 14)}$$

Inserting this into (A2.12) we get the required S-matrix:

$$\mathbf{S}^2 = \frac{1}{4}\begin{pmatrix} U-2 & U+2 \\ U+2 & U-2 \end{pmatrix} \qquad \text{(A2. 15)}$$

where:

$$U = [\,4 - 2(\alpha+\beta)\,]/[\,2\alpha\beta - \alpha - \beta\,] \qquad \text{(A2.16)}$$

For full transmission we require U = 2 and from (A2.16) this implies the product of α and β is equal to one. In other words the side arms 2 and 4 are invisible to the two-port as long as the product kl is the same in both arms. Obviously this is true only for the specific configuration we have chosen, where the wavefunction vanishes at the ends of lines 2 and 4. Never-the-less it is interesting that an integral number of half-wavelengths is not required in the side-arms to eliminate reflection at the two-port.

Using (A2.15) a chain with side-branches may be reduced to a 1D transmission line. We now show, in principal, how a more extensive 2D network may be similarly treated. For simplicity we now demand that all the lines on the network are the same length , Δ, so that only one kl product appears. From the symetry of the four-port S-matrix (A2.2) any two arms may be eliminated using (A2.15) so long as they terminate at a $\Psi = 0$ boundary. Thus a 6 by 3 network may be immediately reduced to the form shown in Fig A4. In the inset of Fig A4 the next reduction required is shown, four-port to three-port. This is described by the S-matrix:

$$\begin{pmatrix} B_a \\ B_b \\ B_c \end{pmatrix} = \frac{1}{2}\begin{pmatrix} \gamma - 1 & \gamma + 1 & \gamma + 1 \\ \gamma + 1 & \gamma - 1 & \gamma + 1 \\ \gamma + 1 & \gamma + 1 & \gamma - 1 \end{pmatrix}\begin{pmatrix} A_a \\ A_b \\ A_c \end{pmatrix} \qquad \text{(A2.17)}$$

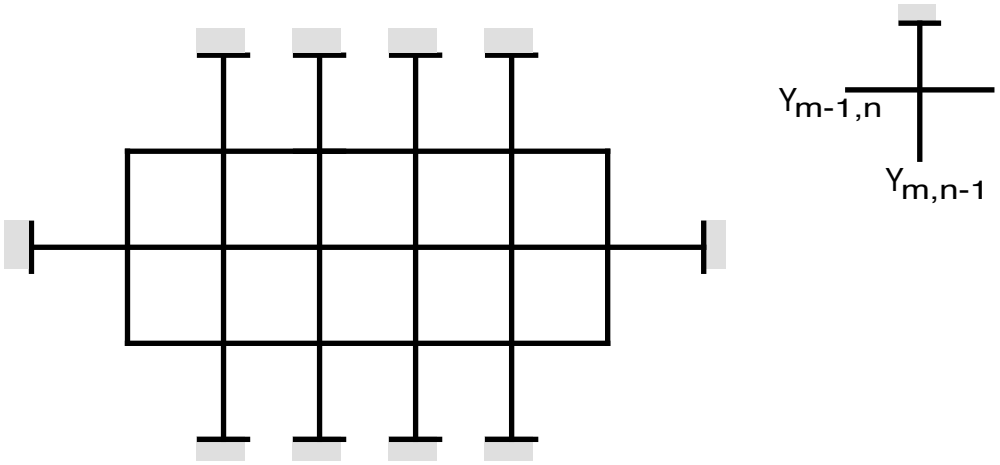
with

$$\gamma = 1/(1-2\omega^2) \qquad \text{(A2.18)}$$

$$\omega = e^{ik\Delta} \qquad \text{(A2.19)}$$

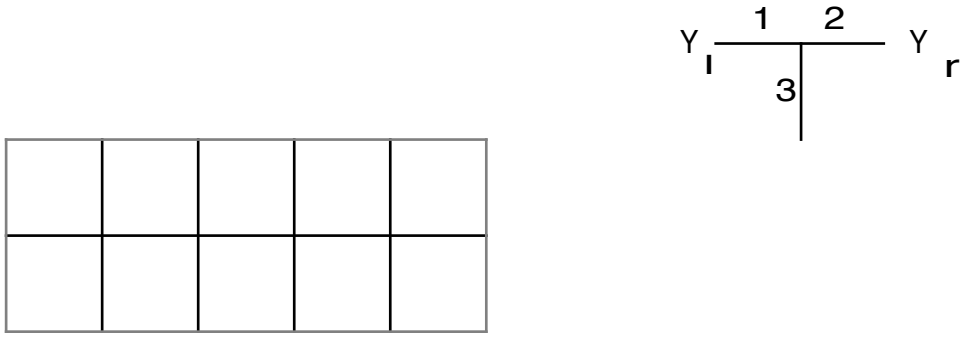
and the a,b,c are any three of the four lines off the node.

Fig A4 Six by three network with corner four-ports reduced to two-ports



In Fig A5 we see the result of applying (A2.17) and in the inset is the next reduction required, either (but not both) of the lines labelled 1 or 2 being eliminated. Note that from here on the ends of lines to be eliminated do not terminate on the $\Psi = 0$ boundary and so we have un-known quantities entering into the equations. Thus at the end we will have a 1D representation for the network but it will contain components of the wavefunction at nodes throughout the network.

Fig A5 Network with all branches to boundary eliminated



To remove the left line we need:

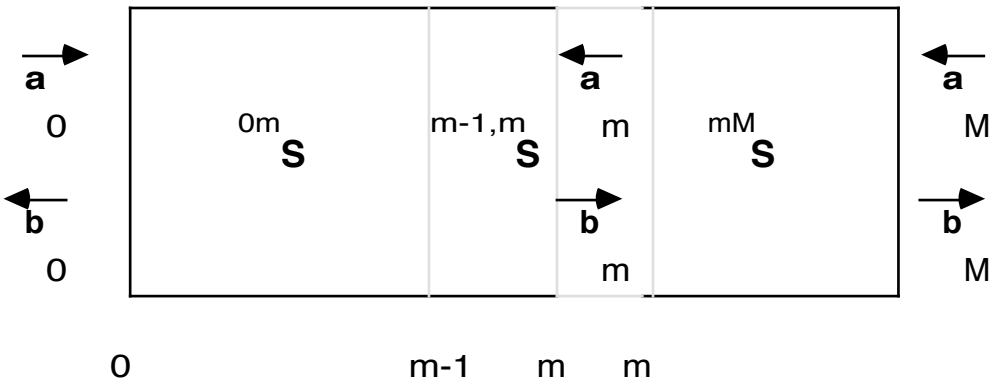
$$\begin{pmatrix} B_2 \\ B_3 \end{pmatrix} = \frac{1}{2}\begin{pmatrix} \gamma - 1 & \gamma + 1 \\ \gamma + 1 & \gamma - 1 \end{pmatrix}\begin{pmatrix} A_2 \\ A_3 \end{pmatrix} + \frac{1}{2}(\gamma + 1)\omega\begin{pmatrix} B_l \\ B_l \end{pmatrix} \qquad (A2.20)$$

with a similar equation involving B_1, B_2, A_1, A_2, and B_r to eliminate the right line. Thus all the dotted lines in Fig A5 can be removed. A modified version of (A2.5) with two extra terms involving components of the wavefunction at the ends of the two discarded lines will finally achieve the 1D transmission line. Solving this would give us the transmission through the 2D network but we would have to determine components of the wavefunction appearing in the equation but not on the 1D line. It is to overcome this problem that the multi-dimensional transfer matrix is introduced in chapter 3.

Appendix 3 Three Different Routes to the Wavefunction

The quantities appearing in the equations below are shown in Fig A6.

Fig A6 S-matrices used in equations (A3.1-A3.3)



The three equations of interest here are:

$$\begin{bmatrix} \mathbf{b}_0 \\ \mathbf{b}_M \end{bmatrix} = \begin{bmatrix} \mathbf{S}_{11} & \mathbf{S}_{12} \\ \mathbf{S}_{21} & \mathbf{S}_{22} \end{bmatrix} \begin{bmatrix} \mathbf{a}_0 \\ \mathbf{a}_M \end{bmatrix} \qquad \text{(A3. 1)}$$

$$\begin{bmatrix} \mathbf{a}_m \\ \mathbf{b}_M \end{bmatrix} = \begin{bmatrix} {}^{mM}\mathbf{S}_{11} & {}^{mM}\mathbf{S}_{12} \\ {}^{mM}\mathbf{S}_{21} & {}^{mM}\mathbf{S}_{22} \end{bmatrix} \begin{bmatrix} \mathbf{b}_m \\ \mathbf{a}_M \end{bmatrix} \qquad \text{(A3. 2)}$$

$$\begin{bmatrix} \mathbf{b}_0 \\ \\ \mathbf{b}_m \end{bmatrix} = \begin{bmatrix} {}^{0m}\mathbf{S}_{11} & {}^{0m}\mathbf{S}_{12} \\ \\ {}^{0m}\mathbf{S}_{21} & {}^{0m}\mathbf{S}_{22} \end{bmatrix} \begin{bmatrix} \mathbf{a}_0 \\ \\ \mathbf{a}_m \end{bmatrix} \qquad \text{(A3.3)}$$

First let us construct the wavefunction by using only the ${}^{0,m}\mathbf{S}$ matrix. From (A3.3) we have:

$$\mathbf{b}_m = {}^{0,m}\mathbf{S}_{21}\mathbf{a}_0 + {}^{0,m}\mathbf{S}_{22}\mathbf{a}_m \qquad \text{(A3.4)}$$

and,

$$\mathbf{b}_0 = {}^{0,m}\mathbf{S}_{11}\mathbf{a}_0 + {}^{0,m}\mathbf{S}_{12}\mathbf{a}_m \qquad \text{(A3.5)}$$

Thus:

$$\mathbf{a}_m = ({}^{0,m}\mathbf{S}_{12})^{-1} [\, \mathbf{b}_o - {}^{0,m}\mathbf{S}_{11}\mathbf{a}_0 \,] \qquad \text{(A3.6)}$$

Substituting (A3.6) into (A3.4) we get:

$$\mathbf{b}_m = {}^{0,m}\mathbf{S}_{21}\mathbf{a}_0 + {}^{0,m}\mathbf{S}_{22} ({}^{0,m}\mathbf{S}_{12})^{-1} [\, \mathbf{b}_o - {}^{0,m}\mathbf{S}_{11}\mathbf{a}_0 \,] \qquad \text{(A3.7)}$$

Adding (A2.6) to (A2.7) we get the wavefunction:

$$\Psi_m = ({}^{0,m}\mathbf{S}_{22} + \mathbf{1})({}^{0,m}\mathbf{S}_{12})^{-1}\mathbf{b}_0$$

$$+ [\, {}^{0,m}\mathbf{S}_{21} - ({}^{0,m}\mathbf{S}_{22} + \mathbf{1})({}^{0,m}\mathbf{S}_{12})^{-1}{}^{0,m}\mathbf{S}_{11} \,]\mathbf{a}_0 \qquad \text{(A3.8)}$$

Alternatively we may use (A3.2) to determine the wavefunction in terms of the

$^{m,M}\mathbf{S}$ matrix. From (A3.2):

$$\mathbf{a}_m = {}^{m,M}\mathbf{S}_{11}\mathbf{b}_m \qquad \text{(A3.9)}$$

and

$$\mathbf{b}_m = ({}^{m,M}\mathbf{S}_{21})^{-1}\mathbf{b}_M \qquad \text{(A3.10)}$$

($\mathbf{a}_M = \mathbf{0}$ is a boundary condition).

Substituting (A3.10) into (A3.9):

$$\mathbf{a}_m = {}^{m,M}\mathbf{S}_{11}\ ({}^{m,M}\mathbf{S}_{21})^{-1}\mathbf{b}_M \qquad \text{(A3.11)}$$

and adding (A2.10) to (A2.11) we have:

$$\Psi_m = ({}^{m,M}\mathbf{S}_{11} + \mathbf{1}\ )\ ({}^{m,M}\mathbf{S}_{21})^{-1}\mathbf{b}_M \qquad \text{(A3.12)}$$

The third route involves both the matrices. Substituting (A2.9) into (A2.4) we get:

$$\mathbf{b}_m = [\mathbf{1} - {}^{0,m}\mathbf{S}_{22}{}^{m,M}\mathbf{S}_{11}\ ]^{-1}\ {}^{0,m}\mathbf{S}_{21}\mathbf{a}_0 \qquad \text{(A3.13)}$$

and adding this to (A2.9) we have:

$$\Psi_m = (\ \mathbf{1} + {}^{m,M}\mathbf{S}_{11}\ )[\mathbf{1} - {}^{0,m}\mathbf{S}_{22}{}^{m,M}\mathbf{S}_{11}\ ]^{-1}\ {}^{0,m}\mathbf{S}_{21}\mathbf{a}_0 \qquad \text{(A3.14)}$$

In order to avoid re-calculating the $^{m,M}\mathbf{S}$ matrix from scratch at each different value of m we need to be able to ' subtract ' the matrix $^{m-1,m}\mathbf{S}$ from the old value, i.e. the opposite of adding $^{m-1,m}\mathbf{S}$ to the old value of $^{0,m}\mathbf{S}$.

We quote the result for the submatrix conversion $\mathbf{S}' = \mathbf{S}\ \Theta^2\mathbf{S}$:

$$\mathbf{S'}_{11} =$$

$$\left\{\mathbf{1} - [(^2\mathbf{S}_{22})^{-1}\ {}^2\mathbf{S}_{21}[\mathbf{1} - (^2\mathbf{S}_{11})^{-1}\mathbf{S}_{11}]^{-1}\ (^2\mathbf{S}_{11})^{-1}]^2\ \mathbf{S}_{12}\right\}^{-1}$$

$$\text{x} \quad (^2\mathbf{S}_{22})^{-1} \qquad \text{(A3.15)}$$

$$\mathbf{S'}_{12} =$$

$$-\left\{\mathbf{1} - [(^2\mathbf{S}_{22})^{-1}\ {}^2\mathbf{S}_{21}[\mathbf{1} - (^2\mathbf{S}_{11})^{-1}\mathbf{S}_{11}]^{-1}\ (^2\mathbf{S}_{11})^{-1}]\,{}^2\mathbf{S}_{12}\right\}^{-1}$$

$$\text{X}\ [(^2\mathbf{S}_{22})^{-1}\ {}^2\mathbf{S}_{21}[\mathbf{1} - (^2\mathbf{S}_{11})^{-1}\mathbf{S}_{11}]^{-1}\ (^2\mathbf{S}_{11})^{-1}]\mathbf{S}_{12} \qquad \text{(A3.16)}$$

$$\mathbf{S'}_{21} = -\ \mathbf{S}_{21}[\mathbf{1} - (^2\mathbf{S}_{11})^{-1}\mathbf{S}_{11}]^{-1}\ (^2\mathbf{S}_{11})^{-1}\ {}^2\mathbf{S}_{12}\,\mathbf{S'}_{11} \qquad \text{(A3.17)}$$

$$\mathbf{S'}_{22} = \mathbf{S}_{22}$$

$$+\ \mathbf{S}_{21}[\mathbf{1} - (^2\mathbf{S}_{11})^{-1}\mathbf{S}_{11}]^{-1}\ (^2\mathbf{S}_{11})^{-1}\ [\mathbf{S}_{12} - {}^2\mathbf{S}_{12}\mathbf{S'}_{12}]$$

(A3.18)

Appendix 4 Composite (multi-mode) Initial Condition Versus Separate Calculations

We demonstrate here that although the wavefunction result is not affected by combining two initial conditions and doing a single calculation, as opposed to summing the results from two separate calculations, the transmission coefficient will in

general be altered.

Taking the 2D S-matrix equation for the system where x=0 is on the LHS and x=L is on the RHS:

$$\begin{pmatrix} \mathbf{b}_0 \\ \mathbf{b}_L \end{pmatrix} = \begin{pmatrix} \mathbf{S}_{11} & \mathbf{S}_{12} \\ \mathbf{S}_{21} & \mathbf{S}_{22} \end{pmatrix} \begin{pmatrix} \mathbf{a}_0 \\ \mathbf{a}_L \end{pmatrix} \qquad \text{(A4. 1)}$$

where the **a**, **b**'s are column vectors and the boundary conditions are $\mathbf{a}_L = 0$, $\mathbf{a}_0 \neq 0$, we have:

$$\mathbf{b}_L = \mathbf{S}_{21}\mathbf{a}_0 \qquad \text{(A4.2)}$$

If the initial condition consists of two component modes then (A4.2) becomes:

$$\mathbf{b}_L = \mathbf{S}_{21}[\mathbf{a}_{O(1)} + \mathbf{a}_{O(2)})] \qquad \text{(A4.3)}$$

Clearly this is the same result for the output wave at x=L as we get if the calculation is done separately for each mode:

$$\mathbf{b}_L = \mathbf{S}_{21}\mathbf{a}_{0(1)} + \mathbf{S}_{21}\mathbf{a}_{O(2)} \qquad \text{(A4.4)}$$

The transmission coefficient is given by:

$$T = \Sigma\, |B_{L,n}|^2 \;/\; \Sigma\, |A_{0,n}|^2 \qquad \text{(A4.5)}$$

where the $A_{0,n}$ and $B_{L,n}$ are the elements of the column vectors $\mathbf{a}_0$ and $\mathbf{b}_L$. For the two modes taken separately we get the expression:

$$T = T_1 + T_2$$

$$= \Sigma\, |B_{L,n(1)}|^2 \;/\; \Sigma\, |A_{0,n(1)}|^2 \;+\; \Sigma\, |B_{L,n(2)}|^2 \;/\; \Sigma\, |A_{0,n(2)}|^2 \qquad \text{(A4.6)}$$

Whilst combining the initial condition produces the expression:

$$T = \sum |B_{L,n(1)} + B_{L,n(2)}|^2 \; / \; |A_{0,n(1)} + A_{0,n(2)}|^2 \qquad (A4.7)$$

From the last two expressions we expect, in general, to get different results for the transmission coefficients from the two routes. The apparent coincidence of the results we have obtained in chapter 4 must derive from the choice of real sinusoids as the initial conditions.

References

Abramowitz M, Stegun I A (eds) " Pocketbook of Mathematical Functions " (Verlag Harri Deutsch ,1984), p379

Aharonov Y, Bohm D Phys Rev Lett **115** p485 (1959)

Alexander S Phys Rev B **27 3** p1541 (1983)

Avron J E, Raveh A, Zur B Rev Mod Phys **60** 4 p873 (1988)

Barker J R, Laughton M J private communication

Barker J R Proceedings in Physics 13 " The Physics and Fabrication of Microstructures and Microdevices " (1986)

Bastard G Phys Rev B **24** 10 p5693 (1981)

Beaumont S P, Wilkinson C D W et al Solid State Science 83 " Physics and Technology of Submicron Devices " (1988)

Berggren K-F, Newson D J, Semicond Sci Technol **1** p327 (1986)

Bohm D "Quantum Mechanics" (1951)

Borland R E Proc R Soc A **274** p529 (1963)

Brand S, Hughes D T Semicond Sci Technol **2** p607 (1987)

Bryant G W Phys Rev B **39** 5 p3145 (1989)

Buttiker M a.Phys Rev B **33** p3020 (1986)

b.Phys Rev Lett **57** p1761 (1986)

Capasso F Physica B **129** p92 (1985)

Cibert J, Petroff P M, Dolan G J, Pearton S J, Gossard A C, English J H Appl Phys Lett **49** p1275 (1986)

Cohen-Tannoudji B, Die, Laloe F " Quantum Mechanics " Vol 1 (Wiley)

Collins S PhD Thesis (1986)

Collins S, Lowe D, Barker J R J Phys C **18** p637 (1985)

Coulson C A Proc Phys Soc Lon **67** p608 (1954)

Dingle R Festkoerperprobleme XV (Advances in Solid State Physics) Edit H J Queiser (Pergamon Viewig, 1975)

Economou E N, Soukoulis C M Phys Rev Lett **46** p618 (1980)

Esaki L IEEE J Quant Elect **22** 9 p1611 (1986)

Esaki L, Chang L L Proc 12th Int Conf Low Temp Phys p551 (1970)

Esaki L, Tsu R Appl Phys Lett **22** 11 p562 (1973)

Finch M PhD Thesis, Glasgow University (1989)

Fisher D S, Lee P A Phys Rev B **23** p6851 (1981)

Frohne R, Datta S J Appl Phys **64** 8 p4086 (1988)

Grondin R O, Porod W, Ho J, Ferry D K Sup Micro **1** 2 p183 (1985)

Grondin R O, Wang J C Sup Micro **2** 3 p197 (1986)

Hill A J, Ladbrooke P H GEC J Research **4** 1 p1 (1886)

Imry Y " Directions in Condensed Matter Physics, Memorial Volume in Honour of Shrankeng Ma " (World scientific, Singapore, 1986) p101

Kirczenow G a. Solid State Comms **68** 8 p715 (1988)

b. J Phys Cond Matt **1** p305 (1989)

Ko D Y K, Inkson J C Semicond Sci Technol **3** p791 (1988)

Kriman A M, Ruden P P Phys Rev B **32** p8013 (1985)

Landauer R a.IBM J Res Develop **5** p183 (1961)

b.Phil Mag **21** p863 (1970)

c.Z Phys B **68** p217 (1987)

Laughton M J Private communication

Lebwohl P A, Tsu R J Appl Phys **41** 6 p2664 (1970)

Lee P A, Ramakrishnan T V Rev Mod Phys **57** 2 p287 (1985)

Levine B F, Choi K K , Bethea C G, Walker J Appl Phys Lett **50** 16 p1092 (1987)

Marsh A C, Inkson J C J Phys C **17** p6561 (1984)

Mailhoit C, Smith D L, Phys Rev B **33** p8360 (1986)

Mead C, Conway L "Introduction To VLSI Systems" Addison-Wesley, Reading, Massachusetts (1973)

Miller R C, Kleinman D A, Gossard A C Phys Rev B **29** 12 p7085 (1984)

Montroll E W J Math Phys **11** 2 p635 (1970)

Okumura H, Misawa S, Yoshida S, Gonda S Appl Phys Lett **46** 4 p377 (1985)

Pendry J B J Phys C **20** p733 (1987)

Peres A J Math Phys **24** 5 p1110 (1983)

Potter D " Computational Physics " Wiley (1973)

Reed M A, Randall J N, Aggarwal R J, Matyi R J, Moore T M, Wetsel A E Phys Rev Lett **60** p535 (1988)

Ricco B, Olivo P Sup Micro **2** 1 p79 (1986)

Rucker H, Hanke M, Bechstedt F, Enderlein R Sup Micro **2** 5 p477 (1986)

Ruedenberg K, Scherr C W J Chem Phys **21** p1565 (1953)

Schulman J N Phys Rev B **31** 4 p2056 (1985)

Sharvin Y V JETP Lett **1** p152 (1965)

Smith C G, Pepper M, Ahmed H, Frost J E F, Hasko D G, Peacock D C, Ritchie D A, Jones G A C J Phys C **21** L893 (1988)

Stone A D, Szafer A IBM J Res Develop **32** 3 p384 (1988)

Summers C J, Brennan K F, Rogers H D, Wagner B K Sup Micro **3** 2 p147 (1987)

Temkin H, Dolan G J, Panish M B, Chu S N G Appl Phys Lett **50** p413 (1987)

Thornton T J, Pepper M, Ahmed H, Andrews D, Davies G J Phys Rev Lett **56** p1198 (1986)

Van Wees B J, Van Houten H et al Phys Rev Lett **60** 9 p848 (1988)

Van Der Marel D, Haanappel E G Phys Rev B **39** 11 p7811 (1989)

Vigneron J P, Lambin P J Phys A **13** p1135 (1980)

Wilkinson C D W, Beaumont S P Proceedings in Physics 13 " The Physics and Fabrication of Microstructures and Microdevices " (1986)

Welch D F, Wicks G W, Eastman L F J Appl Phys **55** 8 p3176 (1984)

Wharam D A, Thornton T J et al J Phys C **21** 8 L209 (1988)

Williams C K, Glisson T H, Littlejohn M A, Hauser J R IEEE EDL **4** 6 p161 (1973)

Winter, " Quantum Physics " (Wadsworth, California, 1979) p51