Silicon Dangling Bonds

Non-equilibrium Dynamics and Applications

by

Marco Taucer

A thesis submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Physics

University of Alberta

Abstract

Dangling Bonds (DBs) on the silicon surface exist when a silicon atom lacks a bonding partner, resulting in a localized orbital which is not involved in any chemical bonds. On the hydrogen-terminated Si(100) surface, such DBs introduce a mid-gap state. DBs can be created on this surface by the selective desorption of a surface hydrogen atom using a Scanning Tunneling Microscope (STM). This thesis deals with characterization and fabrication of DBs in STM, as well as some of their potential applications. We discuss the unusual appearance of the silicon DB in STM, which can be understood in detail by considering the non-equilibrium charging effects that take place during imaging. We also show that the single-electron tunneling events that lead to non-equilibrium charging of the DB are directly observable in STM experiments. The tip-sample tunnel junction serves as a single electron-sensitive charge sensor, which measures the fluctuating charge of a single silicon DB as electrons tunnel on and off of the DB. Corresponding single-electron transfer rates are extracted, and these agree with the previously proposed model of non-equilibrium charging. Progress in DB fabrication is also discussed. Image analysis and desorption algorithms permit creation of DBs at pre-determined locations, leading to the creation of DB patterns of various sizes, from several DBs to thousands. Finally, a potential application of DBs, Quantum-dot Cellular Automata (QCA), is discussed. QCA is an emerging technology which promises tremendous advantages over today's Complementary Metal-

Oxide-Semiconductor (CMOS) technology, if it can be realized at the atomic or molecular scale. Silicon DBs are a promising platform for QCA devices. Here, we focus on the issue of quantum correlations in QCA circuits, an issue which has not been important in prototype QCA demonstrations, but which may play an increasingly central role as QCA is brought to the atomic scale. Through computational simulations, we find that the inclusion of intercellular correlations qualitatively alters the ground state and thermal steady state of the QCA circuit.

Preface

This thesis is an original work by Marco Taucer. Some of the work presented in this thesis is also presented in recent publications. In particular, parts of Chapter 3 were published in Livadaru *et al.* (2011),[1] parts of Chapter 4 were published in Taucer *et al.* (2014),[2] and parts of Chapter 7 were published in Taucer *et al.* (2015).[3] All figures are original unless otherwise indicated in the figure caption.

To Elmer V. Smith,
my grandfather.

Acknowledgements

First, I am thankful for a family that has helped me in countless ways throughout my life. My parents have given me the best education anyone could hope for. From the kitchen table to enrolment in my first university courses, their guidance opened the door for me to walk this path. To my partner, Fiorella, I am deeply grateful too, for unfailing support through thick and thin in these past years.

I owe many thanks to my collaborators, past and present, for stimulating discussions, for having patience with me, for spurring me into new territory. I have found my collaboration with Prof. Konrad Walus and Dr. Faizal Karim particularly eye-opening. Many ideas in this thesis stem from discussions and correspondence with them.

Also to the wonderful people I have worked with, Dr. Josh Mutus, Dr. Shoma Sinha, Peter Legg, Cristian Vesa, Chad Diedrichs, Dr. Stas Dogel, Dr. Jason Pitters, Martin Cloutier, Mark Salomons, Dr. Radovan Urban, Dr. Paul Piva, Dr. Bruno Martins, Dr. Hatem Labidi, Dr. Moe Rashidi, Dr. Hedieh Kalachi, Dr. Mohammad Koleini, Roshan Achal, Taleana Huff, Erika Lloyd, and John Wood. There are many reasons to thank each of these people. Whether troubleshooting in the lab, collaborating on articles, developing code, or working out tricky concepts on the whiteboard, I have learned so much from my interactions with these colleagues, and it has made my PhD a rich and energizing experience. I was fortunate enough to benefit from the guidance of several of these people as well, which has been a great encouragement to me. I owe a special thanks to Dr. Lucian Livadaru, who patiently answered an endless string of questions when I started in this program, until, at some

point, I started to feel that I was a little less lost, and perhaps even able to make a contribution. Many of the ideas he introduced me to have become quite central to this thesis.

I would like to thank Prof. Michael Woodside and Prof. John Beamish, members of my supervisory committee, who have given valuable guidance and advice. Sarah Derr has been wonderfully helpful since the first day of my PhD, and she has saved me on more than one occasion from accidentally dropping out of the program by missing a deadline through sheer absent-mindedness.

It has been a great pleasure to have the opportunity, during part of my PhD, to work with a tremendous team at Quantum Silicon. I feel honoured to have been part of the early days of this effort to bring new technological capabilities to market. Ken Gordon has been a great leader at QSi, always encouraging bold ambitions and giving the feeling that great things can be achieved.

Finally, there is my supervisor and mentor, Prof. Robert Wolkow. I am grateful to Prof. Wolkow for sharing his passionate curiosity, for imparting an intangible physical intuition about atoms and molecules, for the patience he has shown me and which at times I have needed, for supporting and encouraging me in innumerable ways, for showing by example what it is to be intellectually honest, to distinguish understanding from rote, for bringing together an exceptional group of researchers, for creating a peaceful workplace where we can get lost in atoms. There is not enough room in all the pages of this thesis to express the gratitude I feel for all of this and much more. Fewer words will have to do... Bob, thank you.

Contents

List of Tables

List of Figures

A noiseless patient spider,
I mark'd where on a little promontory it stood isolated,
Mark'd how to explore the vacant vast surrounding,
It launch'd forth filament, filament, filament, out of itself,
Ever unreeling them, ever tirelessly speeding them.

— *Walt Whitman*, from *A Noiseless Patient Spider*

1 Introduction

This thesis deals primarily with Dangling Bonds (DBs) on the hydrogen-terminated silicon (100) surface. The experimental tool used is the Scanning Tunneling Microscope (STM). The first aim of this thesis is to make sense of the varied behaviours that can be observed at the DB in STM. Since the ultimate aim of this effort is to create atom-scale technologies, the later part of this thesis will discuss the process of creating DBs on a larger scale with atomic precision, as well as one of the potential applications of this atom-scale technology, Quantum-dot Cellular Automata (QCA). This first chapter aims to build up some basic concepts that underly the physics of DBs and STM. While the tone is pedagogical, it is not meant to be a complete overview of required concepts. Rather, the intention is to provide a heuristic description to illustrate a way of thinking about the basics. The concepts specific to STM and silicon will be further developed in Chapter 2.

1.1 Quantum Mechanical Concepts

In this section, I will present some introductory thoughts on quantum mechanics. The goal will be to introduce concepts that will be used throughout the thesis, like quantum states, superpositions, and time evolution. Rather than attempting to cover all concepts in a rigorous way, I will attempt to convey a way of thinking about quantum mechanical things which will hopefully be useful whether or not the reader is already familiar with the intricacies of quantum theory.

Many introductory texts describe quantum theory as the theory of very small things — atoms and molecules — yet there is nothing in the quantum

theory that dictates this limitation. Personally, I find it useful to think of quantum theory simply as a theory, and to ask what the world looks like in that theory, at all scales. Of course, the quantum description of the world will seem to have very little to do with our common experience. Nonetheless, I feel that the exercise helps develop an intuition about quantum mechanics. Furthermore, there is an interpretation of quantum mechanics which postulates that the strange reality that results from applying quantum theory at all scales is real and, despite appearances, is not at odds with our experience.

1.1.1 Superpositions

Superpositions are the feature of quantum mechanics that most strikingly differentiates it from classical mechanics. The existence of superpositions broadens the class of states that is open for discussion in any given system — and by a state, we may simply mean "a state of affairs." Our classical intuition tends to consider a certain set of states for a particular system. For example, "my coffee cup is on the left side of my desk" is a state that we can denote $|\mathrm{L}\rangle$. And we can denote the state "my coffee cup is on the right side of my desk" by $|\mathrm{R}\rangle$. $|\mathrm{L}\rangle$ and $|\mathrm{R}\rangle$ represent two perfectly reasonable states of affairs. Quantum mechanics departs from intuition by asserting that for any states that are allowed within a system, any superposition of them is also a perfectly valid state. This means that $|\mathrm{both}\rangle \equiv \left(|\mathrm{L}\rangle + |\mathrm{R}\rangle\right)/\sqrt{2}$ is also a perfectly valid state. So in quantum mechanics, there is a state that is made up of equal parts $|\mathrm{L}\rangle$ and $|\mathrm{R}\rangle$, where the coffee cup is both on the left *and* on the right side of the desk. It is not in the middle of the desk, and it is not moving from one side of the desk to the other, nor is it split into two pieces. The whole coffee cup is in a state which has a component "here," and also a component "there," all at once.

The parts of a superposition do not need to be equal, however, so any state $(a|\mathrm{L}\rangle + b|\mathrm{R}\rangle)$ is also valid — a superposition can have more or less of this or that state. In addition, there can be a phase relationship between the components that make up the superposition, so we can have a state $\left(a|\mathrm{L}\rangle + e^{i\phi}b|\mathrm{R}\rangle\right)$, which is different from $\left(a|\mathrm{L}\rangle + e^{i\theta}b|\mathrm{R}\rangle\right)$, if ϕ and θ are different. The most

general state we can make with one coffee cup is $(\alpha|\mathrm{L}\rangle + \beta|\mathrm{R}\rangle)$, where α and β are complex numbers. Actually, this state is too general, since the overall phase does not matter. That is, $(\alpha|\mathrm{L}\rangle + \beta|\mathrm{R}\rangle)$ and $e^{i\phi}\,(\alpha|\mathrm{L}\rangle + \beta|\mathrm{R}\rangle)$ represent exactly the same physical state. What is important is the *relative* phase of the different components: in this case, the phase of $|\mathrm{R}\rangle$ relative to $|\mathrm{L}\rangle$ is the only meaningful phase. Finally, and least importantly, states are normalized so that $|\alpha|^2 + |\beta|^2 = 1$.

So a system that in classical mechanics has only two states, in quantum mechanics has a continuous range of states. Such a two-state classical system is called a bit, and of course the quantum analog is the famous qubit. The qubit states are continuous in two senses: the weight of the two states can vary continuously from all $|\mathrm{L}\rangle$ to all $|\mathrm{R}\rangle$, and the relative phase between them can vary from $\phi = 0$ to $\phi = 2\pi$. Yet of all the states of the general form $(\alpha|\mathrm{L}\rangle + \beta|\mathrm{R}\rangle)$, only the pure $|\mathrm{L}\rangle$ and pure $|\mathrm{R}\rangle$ states are classically intuitive. They are the only ones we ever observe in the macroscopic world.

From the perspective of today's understanding of physics, quantum mechanics is fundamental, and our everyday experience is puzzling. Why are superpositions of things like coffee cups never observed if they are perfectly valid states? The bridge from quantum concepts and dynamics to classical ones is made with the aid of a variety of *ad hoc* postulates and tricks.[4] One example is the Copenhagen interpretation, which postulates wavefunction "collapse" at the moment of measurement. For a wide variety of situations, this view succeeds in giving sensible results from quantum mechanical calculations. However, the addition of a "measurement" step in the conceptualization of quantum dynamics is arbitrary and inelegant. Furthermore, it raises the question of what exactly a measurement is, a question which remains unanswered. Philosophical attempts to sidestep this problem posit that reality is created by observation, or that quantum mechanics is merely a tool used to predict the outcome of classical experiments, and should not be considered to represent reality. These streams of thought are often called phenomenalist, or operationalist.

Among those who do not accept an operationalist view of science — that it

is a mere tool to predict the outcome of experiments, rather than a window into the *reality* of nature — the question of how the classical world comes about in quantum mechanics remains controversial, subtle, and unresolved. Some fundamental insights may come from considering the interactions of systems with their environments from a quantum mechanical perspective. From the invention of quantum mechanics to the present, the overwhelming majority of work was done on isolated quantum systems, with environmental interactions treated in a cursory way, or treated in analogy with classical physics as a source of friction or energy dissipation. However, recent work has shown the tremendous importance of the system-environment interaction in understanding quantum mechanics and its relation to our classical experience. These insights have not yet been taken up in textbooks, classrooms, or the broad perspective of the field. The seeds of these insights can be traced back to the earliest work in the field, but their elaboration took place starting in the 1980s and generally falls under the rubric of the "program of decoherence"*.

1.1.2 Wavefunctions

The wavefunction describes the state of a quantum system and it does not necessarily have anything to do with waves. The states described above, such as $|\mathrm{L}\rangle$, $|\mathrm{R}\rangle$, or $(\alpha|\mathrm{L}\rangle + \beta|\mathrm{R}\rangle)$, are all wavefunctions. A wavefunction is defined by specifying the magnitude and phase of each of its components. The components are expressed in terms of the basis states, which are usually chosen to be orthogonal (so that if you are in one basis state, $|\mathrm{L}\rangle$, you are certainly not in another, $|\mathrm{R}\rangle$), and they should span the space of all possible states of the system. They are usually also normalized, thereby constituting a *complete orthonormal basis*.

The space of all possible states, including superpositions, is a Hilbert space, and so far we have talked about a very simple one where only two states can span the entire Hilbert space: a single particle is either here or there, or in a superposition of here and there. But we can imagine a much larger Hilbert

*The recent work of people like W. Zurek, E. Joos, D. Zeh, and others[4–6] is in some ways a continuation of ideas that go back as far as Schrödinger and von Neumann.[7]

space, where a particle (an electron or a coffee cup) can be at *any* position. Then we need to specify an amplitude (magnitude and phase) for each possible state — that is, each possible position if we are in the position basis. If we use ψ to denote that amplitude, then we need a unique ψ for each position in space. The function $\psi(\mathbf{r})$ specifies the amplitude at each point in space. This function, $\psi(\mathbf{r})$, is also a wavefunction, and it describes a quantum state when the particle can be at any position in space. In fact, this function is the reason that the wavefunction is called the wavefunction. Wavefunctions of electrons in space, expressed as $\psi(\mathbf{r})$, often have a wavy (oscillatory) character, and also evolve in time in a rippling, wave-like way.

1.1.3 Time Evolution

The change of a system in time presents another interesting difference between classical and quantum mechanics — or between our experience of the world and the description provided by quantum theory. It is possible to say, in both quantum and classical mechanics, that the time evolution of a system is determined by the system's state. That is, time evolution is deterministic. But something subtly different is meant in each case. In classical mechanics, in order to predict the evolution of a system, we need to know the system's coordinates in *phase space*, which means that we need to know not only the positions of the particles, but also the rates of change of those positions. In quantum mechanics, we only need to know the quantum state at any given instant to know the future evolution of the system. Put another way, in classical mechanics, a picture of the system does not determine what happens next; we need two pictures separated by an infinitesimal time interval. In quantum mechanics, a single picture of the quantum state (including the phase of each component) suffices to determine its future course.

In paticular, the rate of change of a quantum state is determined by the Schrödinger equation,

$$i\hbar\frac{\partial}{\partial t}|\psi\rangle = H|\psi\rangle, \tag{1.1}$$

where H is the Hamiltonian operator, and $\hbar$ is the reduced Planck constant,

which can be written in the position basis as

$$H = -\frac{\hbar^2}{2m}\nabla^2 + V(\mathbf{r}), \tag{1.2}$$

where m is the particle mass, and $V(\mathbf{r})$ is the potential energy of the particle at each position. $V(\mathbf{r})$ is an operator which is diagonal in the position basis. Eigenvalues of the Hamiltonian satisfy

$$\left[-\frac{\hbar^2}{2m}\nabla^2 + V(\mathbf{r})\right]\psi_n(\mathbf{r}) = E_n\psi_n(\mathbf{r}), \tag{1.3}$$

which is the time-independent Schrödinger equation, often just called the Schrödinger equation. The states $\psi_n(\mathbf{r})$ and energies E_n are the eigenstates and eigenvalues of the Hamiltonian.

1.2 Covalent Bonds, Ionic Bonds, and Qubits

We can illustrate how this works with a toy model. A double well provides a good picture to work with. We can imagine any confining potential for each of two identical wells — it could be the confining potential of an atom, or of a quantum dot, for example — and we will consider only one particular bound state (it could be the ground state of a harmonic oscillator, for instance), whose energy is E_0. We then imagine bringing the two wells closer together until there is a small overlap between the wavefunction centered on the first well, and the wavefunction centered on the second. This overlap results in a "hopping" amplitude for an electron localized in one well to hop to the other.

The Hamiltonian for this double well system is

$$\hat{H} = E_0 - t\left(|L\rangle\langle R| + |R\rangle\langle L|\right) \equiv \begin{pmatrix} E_0 & -t \\ -t & E_0 \end{pmatrix}, \tag{1.4}$$

where t is the hopping amplitude (with units of energy) between the left and right well, and the $\{|\mathrm{L}\rangle, |\mathrm{R}\rangle\}$ basis states denote the eigenfunctions localized in the left and right wells, respectively. The ground state is

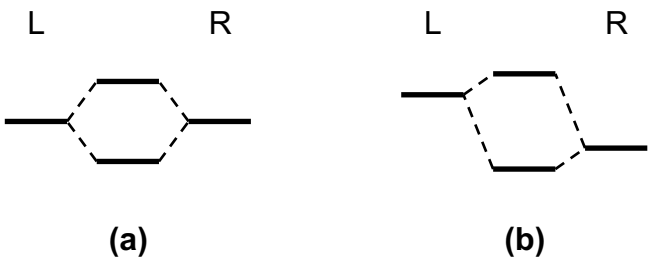


Figure 1.1: The ground and excited states of (a) ionic and (b) covalent bonds are superpositions of the localized states on the left and right sites.

$$|+\rangle = \frac{1}{\sqrt{2}}\big(|\mathrm{L}\rangle + |\mathrm{R}\rangle\big) \tag{1.5}$$

with energy $-t$, and the excited state is

$$|-\rangle = \frac{1}{\sqrt{2}}\big(|\mathrm{L}\rangle - |\mathrm{R}\rangle\big) \tag{1.6}$$

with energy $+t$. This can be thought of as a covalent bond, where energy levels, with energy E_0, combine to form a symmetric wavefunction — the bonding orbital — and an antisymmetric wavefunction — the anti-bonding orbital. The difference in energy between these two eigenfunctions is exactly $2t$, which leads to the familiar picture of the covalent bond, shown in Figure 1.1a.

As a thought experiment, we can imagine preparing the wavefunction of the electron in the state on the left, $|\psi(\tau = 0)\rangle = |\mathrm{L}\rangle$, where τ denotes time (to avoid confusion with the hopping parameter t). We can then ask how the wavefunction will evolve coherently in time:

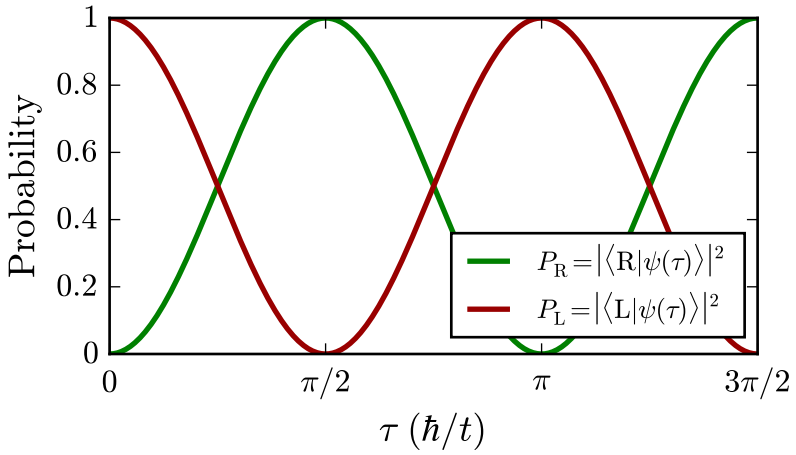


Figure 1.2: Coherent evolution of a double-well system (analogous to a covalent bond). Given an initial state localized on the left, the electron oscillates from left to right and back.

$$
\begin{aligned}
|\psi(\tau)\rangle &= e^{-iH\tau/\hbar}|\mathrm{L}\rangle \\
&= e^{-iH\tau/\hbar}\frac{|+\rangle + |-\rangle}{\sqrt{2}} \\
&= \frac{e^{-iE_g\tau/\hbar}|+\rangle + e^{-iE_e\tau/\hbar}|-\rangle}{\sqrt{2}} \\
&= \frac{\left(e^{-iE_g\tau/\hbar} + e^{-iE_e\tau/\hbar}\right)|L\rangle + \left(e^{-iE_g\tau/\hbar} - e^{-iE_e\tau/\hbar}\right)|R\rangle}{2} \\
&= \cos\left(\frac{t\tau}{\hbar}\right)|\mathrm{L}\rangle + i\sin\left(\frac{t\tau}{\hbar}\right)|\mathrm{R}\rangle. \qquad (1.7)
\end{aligned}
$$

Figure 1.2 shows the probability of finding the electron in states $|\mathrm{L}\rangle$ and $|\mathrm{R}\rangle$ as a function of time. We see that an electron that starts in the left well will coherently evolve until it has completely shifted to the right well, and will then evolve back to the left well returning to its original state, and so on. The period of this oscillation is $h/2t$ (where h is the Planck constant). This can be thought of as coherent tunneling.

We can think of a two level system more generally, without the assumption of resonance between the two unperturbed on-site energies. This represents a bond between orbitals of unequal energy, for instance, like an ionic bond, as

shown in Figure 1.1b. In that case, the energies E_L and E_R may be different, such that $E_R - E_L \equiv \Delta$. The Hamiltonian for this system then becomes

$$H = \begin{pmatrix} -\Delta/2 & -t \\ -t & \Delta/2 \end{pmatrix}, \tag{1.8}$$

Where we have assumed the average energy, $(E_L + E_R)/2$ to be zero — if this is not the case, a constant energy can be added trivially. If we express the Hamiltonian in units of $\Delta/2$, then the only important parameter left in the Hamiltonian is the ratio $2t/\Delta$. The problem turns out to be greatly simplified by considering a right angle triangle whose side lengths are $2t$ and Δ, as shown in Figure 1.3. The Hamiltonian then becomes

$$\frac{H}{\Delta/2} = \begin{pmatrix} -1 & -\tan\theta \\ -\tan\theta & 1 \end{pmatrix}, \tag{1.9}$$

where $\theta \equiv \tan^{-1}(2t/\Delta)$.

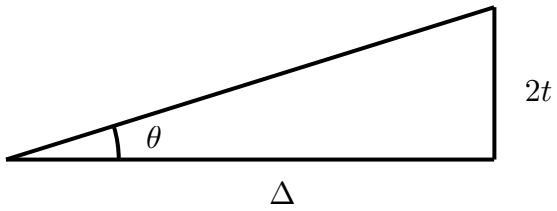


Figure 1.3: The hybridization of two non-resonant energy levels can be understood by thinking of a right angle triangle.

The eigenstates and eigenvalues of the Hamiltonian can then be expressed as

$$\begin{aligned} E_g &= -\tfrac{\Delta}{2}\sec\theta \quad ; \quad |g\rangle = \begin{pmatrix} \cos(\theta/2) \\ \sin(\theta/2) \end{pmatrix} \\ E_e &= \tfrac{\Delta}{2}\sec\theta \quad ; \quad |e\rangle = \begin{pmatrix} \sin(\theta/2) \\ -\cos(\theta/2) \end{pmatrix}. \end{aligned} \tag{1.10}$$

Note that the difference in energy between the ground state and the excited state is $\Delta\sec\theta = \sqrt{\Delta^2 + (2t)^2}$, which is the hypotenuse of the triangle shown

in Figure 1.3.

The ratio of $2t$ to Δ (or we might say the angle θ) determines the degree to which hybridization is important. When $2t$ is much smaller than Δ, very little hybridization occurs: the ground and excited state energies are very close to the on-site energies at each location, and the corresponding wavefunctions are dominated by one or the other localized state. Only when $2t$ becomes comparable to the splitting, Δ, is there a significant degree of hybridization, as the energies and eigenstates depart from those of the localized states. In the opposite limit, where $2t \gg \Delta$, we return to the case of the covalent bond.

1.3 The Tight Binding Model and Bloch Functions

This idea of creating an abstraction where we define discrete sites, each with its own energy, and with a probability for hopping from site to site, is widely used and is known as the tight-binding model. It is readily used to construct "toy models," which often provide insight into the important factors in a problem. In many cases, the tight-binding model, if appropriately parametrized, can actually provide a good description of real systems. It is also a starting point for many more complicated methods, such as the Hubbard model and its many variations. Here, we will employ it as a useful way to illustrate some general topics in solid state physics.

1.3.1 A Simple 1D Chain

A very simple tight-binding Hamiltonian, where sites along a 1D chain have a constant hopping t between nearest neighbours, can be written

$$\hat{H} = -t \sum_{n=0}^{N-2} \left(|n\rangle\langle n+1| + |n+1\rangle\langle n| \right), \tag{1.11}$$

where n has been used to label the sites, which go from $n = 0$ to $n = N - 1$. Each term in the summation couples site n to site $n + 1$, and the summation runs up to $N - 2$ since there is no site to the right of the right-most site. This is depicted in Figure 1.4

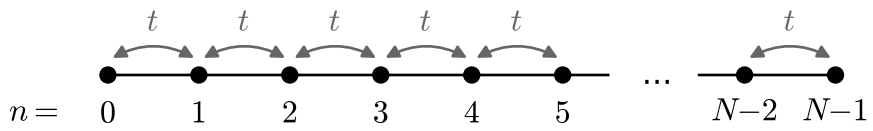


Figure 1.4: The simple 1D chain consists of a series of equally spaced sites with a hopping constant, t, coupling neighbouring sites.

For $N = 1$, the system is trivial: there is only one site, and therefore no hopping, and the Hamiltonian is equal to zero. For $N = 2$, we recover exactly the double well Hamiltonian outlined above, whose two eigenstates are the bonding and anti-bonding orbitals. For $N = 3$, there are three eigenstates of the Hamiltonian, each with a distinct energy.

Figure 1.5 shows the orbitals and energy levels for values of N from 1 to 5. For each value of N there are precisely N energy eigenstates, which span a range of energies which increases with N. The ground state of an isolated site has been taken to be a Gaussian wavefunction. This is chosen simply for the sake of illustration, but could for example represent the case of a locally quadratic confining potential at each site. Certain patterns in the wavefunctions are apparent. The lowest energy eigenstate consistently has all orbitals in phase (chosen to be positive — shown as blue — in this diagram), while the highest energy orbital has the phase of the orbitals alternate. In general, as the energy increases, the wavefunction crosses zero more (precisely $n-1$ times in fact, for the n^{th} energy eigenstate). These turn out to be general features.

As the number of orbitals in the one dimensional chain is increased, the features identified above persist. Figure 1.6 shows energy level diagrams for chains of increasing length, from $N = 1$ to $N = 50$. The energy eigenstates span an increasingly wide range of energies, but asymptotically approach a total width of $4t$. The total number of eigenstates, however is always N, meaning that the total density of energy eigenstates continues to increase as the number of orbitals in the chain increases. Clearly, the discrete states that make up the energy level structure of a tight-binding chain effectively become

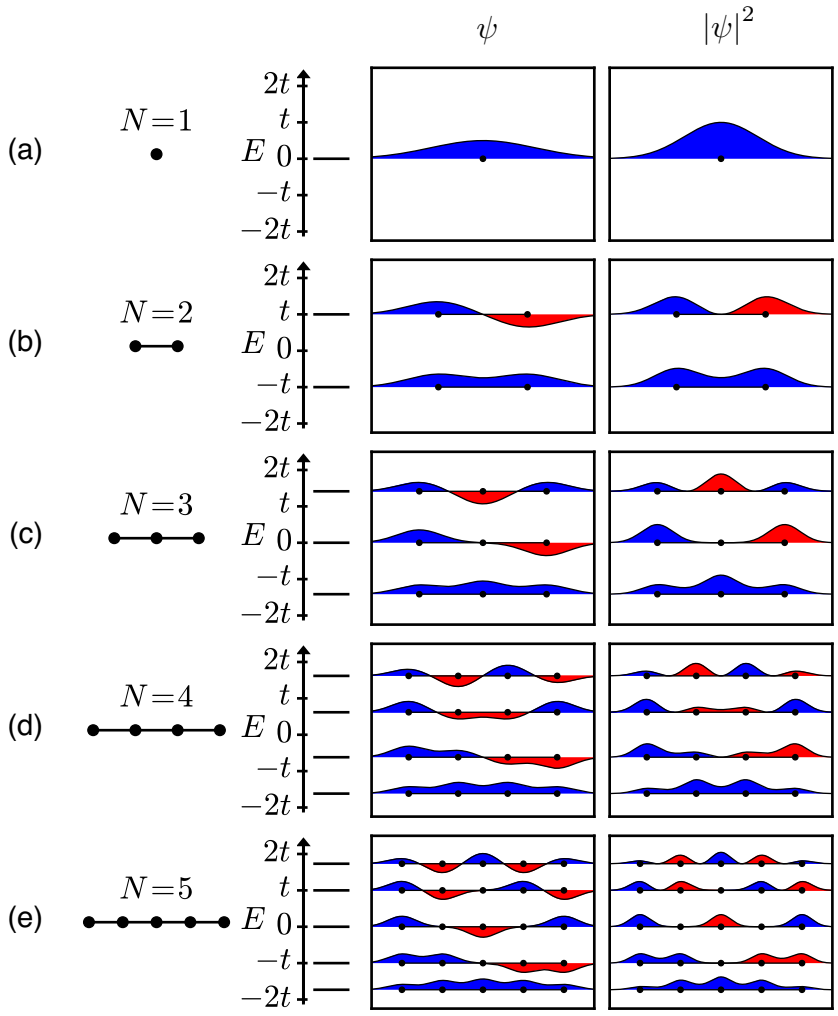


Figure 1.5: (a-e) Illustration of the energy levels and wavefunctions of a one-dimensional chains of orbitals within the tight binding model. Starting from the left, each subfigure shows a schematic representation of the tight-binding chain, followed by the energy level diagram, followed by a graphical representation of the wavefunctions and modulus square of the wavefunctions, with offsets to bring them in line with the energy level diagram. The absolute y-scale of the wavefunctions is arbitrary, although the relative scale is consistent within each panel. Color represents phase, with blue indicating positve values, and red indicating negative values of the wavefunction (there are no imaginary components).

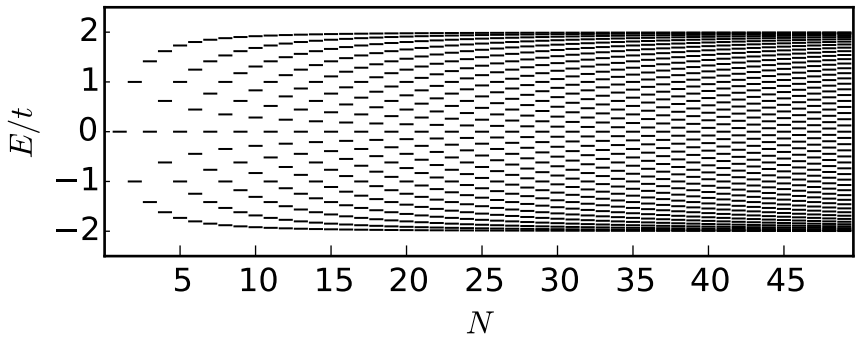


Figure 1.6: Diagram showing the formation of a band as orbitals are sequentially added to a one-dimensional chain within the tight-binding model. For each number of orbitals in the chain, N, the energy levels are plotted as horizonal lines.

a continuum at some point.

The example just described illustrates the formation of a band. In a solid, bands form when the orbitals of its constituent atoms overlap, allowing transfer of electrons from the orbital of one atom to the orbitals of its neighbours. This results in a spreading of the initially discrete states of the constituent atoms into bands which span a range of energies. In principle, bands are composed of a large number of discrete states, but since the number of atoms, N, is typically very large for real solids, bands are considered to be continuous. Nonetheless, it is clear from Figure 1.6 that the energy eigenstates are not distributed equally across the band, but are instead more dense at some energies than at others (in this case, there are more eigenstates at the extremes of the band, near $\pm 2t$, than in the middle, near 0). For that reasons, one of the ways of looking at the band structure of a solid is through the solid's Density of States (DOS).

As stated above, it is also possible to consider the wavefunctions of tight-binding chains in terms of the phase of the wavefunctions that constitute it. For short chains, the energy of the wavefunction was higher for wavefunctions that changed sign more frequently. This points to a general feature of wavefunctions: energies are highest for wavefunctions whose phase oscillates more rapidly. This fact can be seen by inspection of the Schrödinger equation (Equation 1.3) and is illustrated in Figure 1.5. This suggests that wavefunc-

tions might be categorized (or labelled) according to their "wavyness". This can be made formal by labeling wavefunctions by their wavevector, k. We will now switch to analytical calculations to show this.

We can define a state $|k_j\rangle$ as a linear combination of all the localized states, $|n\rangle$,

$$|k_j\rangle \equiv \frac{1}{\sqrt{N}} \sum_{n=0}^{N-1} e^{ik_j x_n} |n\rangle \quad ; \quad x_n = ns, \tag{1.12}$$

where x_n is the position of the n^{th} site, s is the spacing between adjacent sites, and k_j is a wavevector defined by

$$k_j = \frac{2\pi j}{Ns} \quad ; \quad j = -N/2, -N/2+1, \ldots, N/2-1, \tag{1.13}$$

where we have assumed that N is even (trivial adjustments need to be made if it is not). The states $|k_j\rangle$ then form their own orthonormal basis. The localized states can be expressed in terms of the wavevector components,

$$|n\rangle \equiv \frac{1}{\sqrt{N}} \sum_{j=-N/2}^{N/2-1} e^{-ik_j x_n} |k_j\rangle. \tag{1.14}$$

The $|k\rangle$ states are the discrete Fourier transform of the the localized states, $|n\rangle$. Substituting the above expression for $|n\rangle$ into Equation 1.11 diagonalizes the Hamiltonian, yielding

$$H = \sum_{k_j} \left[-2t\cos(k_j s)\right] |k_j\rangle\langle k_j|, \tag{1.15}$$

which allows us to express energy as a function of the wavevector:

$$E(k) = -2t\cos(ks). \tag{1.16}$$

Figure 1.7a shows $E(k)$, which is the band structure for this simple model — it plots the distribution of energies in reciprocal space. The discrete points show the energies and wavevectors of the discrete eigenstates in the case of $N = 50$. As N is increased, more wavevectors are allowed and the points along the curve $E(k)$ become more closely spaced. As N tends to infinity, the

discrete points completely fill the continuous curve represented by the solid line.

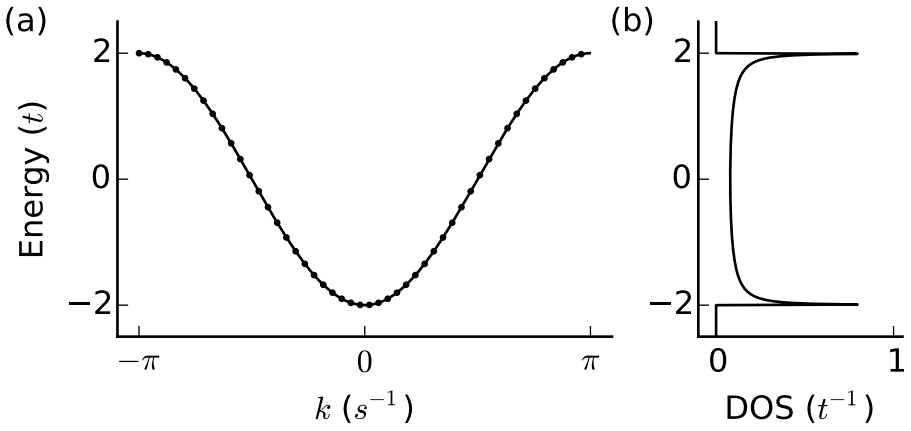


Figure 1.7: (a) Band diagram in reciprocal space for a simple 1D tight-binding chain. The solid line represents the energies of k states for an infinite line, and the points represent the energies of discrete k states for the case of $N = 50$. (b) Density of States (DOS) for the infinite chain, normalized by the number of atoms, N.

We can note that in the case of finite N, the allowed wavevectors are evenly spaced along the k axis, and their density is

$$\frac{dN}{dk} = \frac{Ns}{2\pi}. \tag{1.17}$$

The density of states in energy, in the limit of large N, can then be expressed as

$$DOS(E) = \frac{dN}{dE} = \frac{dN}{dk}\frac{dk}{dE} = \frac{N}{2\pi\sqrt{(2t)^2 - E^2}}. \tag{1.18}$$

The density of states is shown in Figure 1.7b. The divergences at $E = \pm 2t$ are characteristic of the one-dimensional character of the present model, and they are known as van Hove singularities. In other dimensionalities, different types of singularities are observed.

We can also note that if each orbital can accommodate two electrons (one

for each spin), we would then have to double the density of states calculated above. If each orbital also comes with a single electron, for instance if we imagine stringing together a chain of neutral hydrogen atoms, then the band that is formed will be exactly half filled. At zero temperature, electrons will exclusively occupy the energy levels in the lower half of the band. As temperature increases, the border between occupied and unoccupied energy levels is blurred. The occupation is described by Fermi statistics,

$$f(E) = \frac{1}{1 + e^{(E-\mu)/k_B T}}, \tag{1.19}$$

where μ is the chemical potential. In this thesis chemical potential, μ, and Fermi energy, E_F, will be used interchangeably. The border between occupied and unoccupied energy levels at low temperature would be located at $E = 0$, in the case of the band structure described above, for a half-occupied band. This would describe a metallic electronic structure, since there is a non-zero density of states at the Fermi level.

1.3.2 Broadening of a Discrete Level in Contact with a Continuum

It is possible to ask other questions about the formation of a band as well. Figure 1.5 gives some sense of how a band is formed as atoms are sequentially added. But we can ask exactly how an individual atom goes from having its own discrete state to becoming intricately connected with the extended energy levels of an entire crystal (in this case, a simple one-dimensional chain). Figure 1.8 shows precisely this process. One orbital is brought in from the left, in Figure 1.8a, until it is at the standard spacing for the one-dimensional chain, in Figure 1.8d. In this diagram, the wavefunctions are represented using the thickness and darkness of lines, which are both proportional to $|\psi_\alpha(n)|^2$, where n labels the sites as usual, and α is an index which identifies the energy eigenstate. A vertical slice represents the energy levels present at each site, essentially the *local* density of states (LDOS), while the horizontal slices can be interpreted as the modulus squared of the wave functions themselves.

The chain is composed of 20 equally spaced sites, at positions 0 through 19,

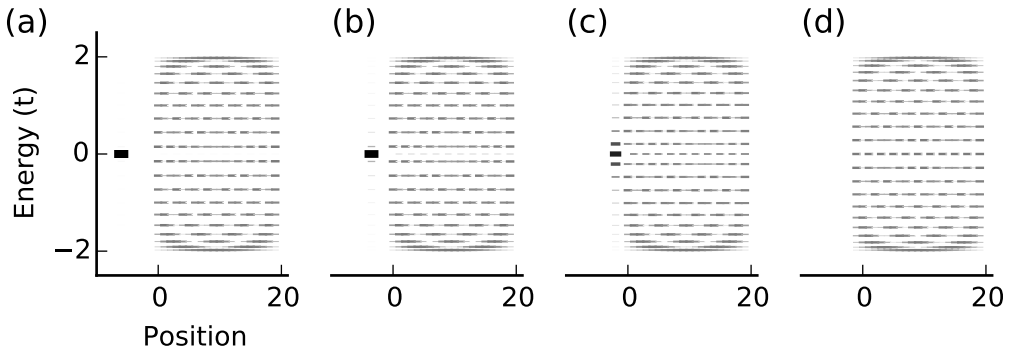


Figure 1.8: A 1-dimensional chain of sites in the tight-binding model, composed of 20 equally spaced sites at positions 0 through 19, with an additional site brought in from the left. The positions of the additional site as it approaches from the left are (a) $-6s$, (b) $-3.5s$, (c) $-2s$, and (d) $-1s$, where s is the standard spacing between nearest neighbours. This last case, therefore corresponds to a chain of 21 regularly spaced sites.

with an additional site brought in from the left. The positions of the additional site as it approaches from the left are $-6s$, $-3.5s$, $-2s$, and $-s$, respectively for Figures 1.8a to d. Here, we have assumed an exponential dependence of the hopping parameter on separation from the nearest neighbour, of the form $t(s) = t_o e^{1-s}$, which decreases by a factor of e for each unit of separation.

When the additional site is far from the chain, at $x = -6s$ it has a single energy level, as shown in Figure 1.8. In Figure 1.8b, the additional site begins to hybridize with the energy levels of the band which are closest in energy. We see that in addition to the very prominent energy density at $E = 0$, the additional site begins to show some density corresponding to the nearest extended energy eigenstates. Additionally, the eigenstate at $E = 0$, which is dominantly localized on the additional site, begins to have a delocalized component which extends across the 1D chain. At the closer separation in Figure 1.8c, the delocalized part of the wave function is even stronger, and the energy levels at the site are spread out over a significant part of the band, though clearly with most density still in the neighbourhood of $E = 0$. Finally,

in Figure 1.8d, the additional site becomes a part of the periodic structure of the 1D chain. The energy level structure at the additional site has broadened from a single level to include varying amplitudes across all energies of the band, from $-2t$ to $+2t$. At the same time, an extended wavefunction with energy $E = 0$ has also appeared.[†]

This broadening of a discrete energy level in contact with an extended reservoir is a well-studied phenomenon in the context of quantum transport. The coupling of a discrete level with a continuum can be described by a transfer rate, Γ, or equivalently, by an energy defined as $\gamma = \hbar\Gamma$. The density of states of the discrete level can be described in terms of this coupling energy using the relation[8]

$$\mathrm{DOS}(E) = \frac{1}{2\pi} \frac{\gamma}{(E - E_0)^2 + (\gamma/2)^2}, \tag{1.20}$$

where E_0 is the energy of the unbroadened energy level. This equation is plotted in Figure 1.9 for various values of γ. It says that as we bring the discrete state into closer contact with an extended reservoir (increasing the rate Γ), the state is gradually broadened according to a Lorentzian distribution centered on E_0. The integral of the DOS is always equal to precisely unity, regardless of the degree of broadening. In that sense, we can think of the initially discrete state as having been "smeared out".

While it may be useful to use terms like "smearing" or "broadening," it is worth keeping in mind that what is really happening is that the discrete state hybridizes with the extended levels of the reservoir, preferentially with those spatially closest and closest in energy. When the previously localized orbital is coupled to the bulk, it is no longer an energy eigenstate of the system. The true energy eigenstates are altered so that what used to be extended states confined to the reservoir gain a small component on the additional site, while the discrete state itself is altered to acquire a component which is delocalized inside the reservoir.

[†]While it is tempting to think of the initially localized eigenstate as having transformed into the extended state at $E = 0$ in Figure 1.8d, we should instead think of it as having merged into the entire band. This point would be made clearer by the addition of yet another atom to the chain, after which there would in fact be *no* extended eigenstate at $E = 0$. In general, simple tight-binding chains have an eigenstate at $E = 0$ only when the number of sites is odd.

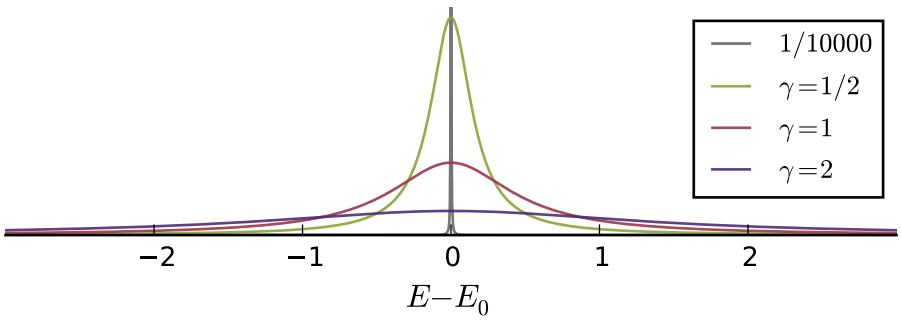


Figure 1.9: DOS of a broadened energy level in contact with a reservoir for various values of the coupling energy, γ. The integral of each each curve is precisely unity.

1.3.3 The Dimerized 1D Chain

We can also construct different types of band structures using this same tight-binding model, by modifying the hopping constants. For instance, we can define a new Hamiltonian, in which atoms are paired by a large tunnel-coupling, t_0, and pairs are in turn coupled through a weaker tunnelling constant, t_1,

$$\begin{aligned} H = & \sum_{p=0}^{N/2-1} -t_0\Big(|2p\rangle\langle 2p+1| + |2p+1\rangle\langle 2p|\Big) \\ & + \sum_{p=0}^{N/2-1} -t_1\Big(|2p+1\rangle\langle 2p+2)| + |2p+2\rangle\langle 2p+1|\Big), \end{aligned} \tag{1.21}$$

where p labels the pairs, and the number of sites, N, has been assumed to be an even number. In effect, this amounts to a chain of sites, with alternating hopping constants, however it is useful to think of the Hamiltonian as describing coupled pairs, as shown in Figure 1.10. The first summation describes the Hamiltonian for each pair, while the second summation describes the interactions between neighbouring pairs. This Hamiltonian turns out to be analytically diagonalizable. We will solve the Hamiltonian by a series of changes of basis.

We saw above that it is straightforward to solve the Hamiltonian for a

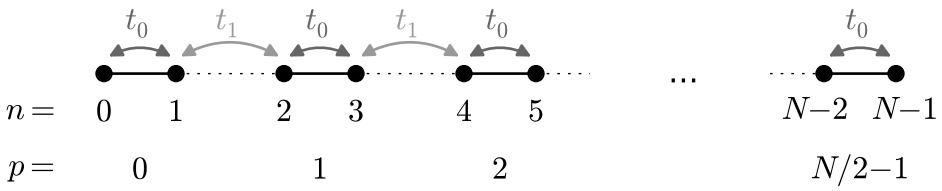


Figure 1.10: The dimerized 1D chain consists of pairs of sites, coupled by a strong hopping constant, t_0, which are in turn tunnel coupled by a weak hopping constant, t_1. Pairs are sequentially labeled by p and sites are sequentially labeled by n, such that $n = 2p$ and $n = 2p + 1$ for the two sites in pair p.

tunnel-coupled pair of sites with resonant energies. This suggests that we might start by solving the Hamiltonian for each pair of atoms, and then working with the resulting wavefunctions. Each term in the first summation in Equation 1.21 corresponds to the Hamiltonian for a particular pair, and we can note that this Hamiltonian is essentially identical to the one in Equation 1.4. Each pair has a lower energy state, $|+_p\rangle$, with energy $-t_0$, and a higher energy state, $|-_p\rangle$, with energy $+t_0$. These are the symmetric and antisymmetric combinations of the two localized states of the pair, $|p\rangle$ and $|p+1\rangle$.

The states $|+_p\rangle$ and $|-_p\rangle$ define a new orthonormal basis. In that basis, the Hamiltonian can be written as

$$\begin{aligned} H = & \sum_{p=0}^{N/2-1} \left\{ -t_0 |+_p\rangle\langle+_p| - \frac{t_1}{2}\Big(|+_p\rangle\langle+_{p+1}| + |+_{p+1}\rangle\langle+_p| \Big) \right\} \\ + & \sum_{p=0}^{N/2-1} \left\{ +t_0 |-_p\rangle\langle-_p| + \frac{t_1}{2}\Big(|-_p\rangle\langle-_{p+1}| + |-_{p+1}\rangle\langle-_p| \Big) \right\} \\ + & \sum_{p=0}^{N/2-1} \left\{ -\frac{t_1}{2}\Big(|+_p\rangle\langle-_{p+1}| + |-_{p+1}\rangle\langle+_p| \Big) \right. \\ & \qquad \left. + \frac{t_1}{2}\Big(|-_p\rangle\langle+_{p+1}| + |+_{p+1}\rangle\langle-_p| \Big) \right\}. \end{aligned} \tag{1.22}$$

This form is particularly telling. The first and second summations resemble the Hamiltonian in Equation 1.11 — that is, they appear to describe a one di-

mensional chain of orbitals (the bonding or antibonding orbitals of the pairs), which are tunnel coupled with their neighbours through a hopping constant of $\pm t_1/2$. Were it not for the third summation, the Hamiltonian would immediately be solved by taking the Fourier transforms of the bonding orbitals, and also of the anti bonding orbitals. The the third summation, however, complicates things by mixing the bonding orbitals with the anti bonding orbitals. In light of this, we can label the three summations as three distinct parts of the Hamiltonian, H_+, H_-, and $H_{+/-}$, respectively.

We proceed by diagonalizing the first two parts, H_+ and H_-. This requires a change to the basis,

$$\begin{aligned} |k_j\rangle &= \sqrt{\frac{2}{N}} \sum_{p=0}^{N/2-1} e^{ik_j x_p} |+_p\rangle \quad ; \quad k_j = \frac{2\pi j}{Ns} \\ |q_j\rangle &= \sqrt{\frac{2}{N}} \sum_{p=0}^{N/2-1} e^{iq_j x_p} |-_p\rangle \quad ; \quad q_j = \frac{2\pi j}{Ns}. \end{aligned} \tag{1.23}$$

Note that "k" and "q" are used to label states in the lower and upper bands respectively, but that the wavevectors k_j and q_j are equal. The separation, s, refers to the distance from the center of one pair to the center of the next, so that $s = s_0 + s_1$, with s_0 the separation between the two sites of a pair, and s_1 the separation between the two closest atoms of neighbouring pairs. As promised, the first two parts of the Hamiltonian are diagonalized,

$$H_+ = \sum_{j=0}^{N/2-1} \{-t_0 - t_1 \cos(k_j s)\} |k_j\rangle\langle k_j| \tag{1.24}$$

and

$$H_- = \sum_{j=0}^{N/2-1} \{+t_0 + t_1 \cos(q_j s)\} |q_j\rangle\langle q_j|, \tag{1.25}$$

and the third part of the Hamiltonian, which "mixes" these two bands, becomes

$$H_{+/-} = \sum_{j=0}^{N/2-1} \{-it_1 \sin(k_j s)\} |k_j\rangle\langle q_j| + \sum_{j=0}^{N/2-1} \{+it_1 \sin(k_j s)\} |q_j\rangle\langle k_j|. \tag{1.26}$$

While this third part of the Hamiltonian means that we still have not quite diagonalized the Hamiltonian, we see that the mixing between the two bands is of a very particular form. A state in one band with a particular value of j — that is, with a particular wavevector — mixes only with the state in the other band with precisely the same wavevector, and no others. Furthermore, the basic properties of the band structure are already revealed at this point: a lower-energy band is formed out of the symmetric bonding orbitals, $|+\rangle$, and a higher energy band, with opposite curvature, is formed out of the antisymmetric anti-bonding orbitals, $|-\rangle$. The rest of the process of diagonalizing only adds a small correction to this picture.

The final step in diagonalizing the Hamiltonian is then to combine the terms for each value of j, from the three parts of the Hamiltonian (H_+, H_-, and $H_{+/-}$), into a single two-state Hamiltonian called H_j,

$$\begin{aligned} H_j = & \quad [-t_0 - t_1\cos(k_j s)]\,|k_j\rangle\langle k_j| \quad + \quad [+t_0 + t_1\cos(k_j s)]\,|q_j\rangle\langle q_j| \\ & + \quad [-it_1\sin(k_j s)]\,|k_j\rangle\langle q_j| \quad + \quad [+it_1\sin(k_j s)]\,|q_j\rangle\langle k_j|. \end{aligned} \tag{1.27}$$

The total Hamiltonian is then simply written as a series of uncoupled 2×2 Hamiltonians, $H = \sum_j H_j$. In matrix form, we can write

$$H_j = \begin{pmatrix} -\Delta_j/2 & -it_j \\ +it_j & +\Delta_j/2 \end{pmatrix} \quad ; \quad \begin{aligned} \Delta_j &= 2t_0 + 2t_1\cos(k_j s) \\ t_j &= t_1\sin(k_j s), \end{aligned} \tag{1.28}$$

in the basis $\{|k_j\rangle, |q_j\rangle\}$. These 2×2 matrices are then diagonalized in the exact same manner as Equation 1.8. For each wave vector, k_j, we find two energy eigenstates, one in the lower band and one in the upper band, which we will label $|k_j^+\rangle$ and $|k_j^-\rangle$, respectively.

The eigenstates and eigenvalues from the lower band are then given by

$$|k_j^+\rangle = \cos\left(\frac{\theta_j}{2}\right)|k_j\rangle - i\sin\left(\frac{\theta_j}{2}\right)|q_j\rangle \quad ; \quad E_j^+ = \frac{-\Delta_j}{2}\sec\theta_j \tag{1.29}$$

while the eigenstates and eigenvalues for the upper band are given by

$$|k_j^-\rangle = \sin\left(\frac{\theta_j}{2}\right)|k_j\rangle + i\cos\left(\frac{\theta_j}{2}\right)|q_j\rangle \quad ; \quad E_j^- = \frac{+\Delta_j}{2}\sec\theta_j, \tag{1.30}$$

where θ_j is defined as

$$\theta_j \equiv \tan^{-1}\left(\frac{2t_j}{\Delta_j}\right). \tag{1.31}$$

Figure 1.11a shows the band structure for this type of 1-dimensional chain of pairs of atoms. The bonding and anti-bonding orbitals of the pairs have given rise to two bands with opposite curvature. The width of each band is $2t_1$, and the two bands are centered at $\pm t_0$.

It is informative to consider the eigenfunctions of the Hamiltonian in terms of the eigenstates of the individual pairs, $|+_p\rangle$ and $|-_p\rangle$. In that basis, we can express the eigenstates belonging to the +-band as

$$|k_j^+\rangle = \sqrt{\frac{2}{N}}\sum_{p=0}^{N/2-1} e^{ik_j x_p}\left[\cos\left(\frac{\theta_j}{2}\right)|+_p\rangle - i\sin\left(\frac{\theta_j}{2}\right)|-_p\rangle\right]. \tag{1.32}$$

This illustrates a very general feature of crystal wavefunctions: they are composed of a part which is identical from unit cell to unit cell, multiplied by a phase which rotates as a function of the unit cell position. In this case, the factor $e^{ik_j x_p}$ is a phase which is different for each pair (and the pairs *are* the unit cells), and the part in the square brackets is exactly the same from one unit cell to the next (the amplitudes of the bonding and antibonding orbital in each unit cell are a function of j, but they are independent of p).

Wavefunctions that have this property are called Bloch functions. In this case, we have written the wavefunction in the form

$$\psi_{\mathbf{k}}(\mathbf{r}) = \sum_{\mathbf{R}} e^{i\mathbf{k}\mathbf{R}}\psi_{0\mathbf{k}}(\mathbf{r}-\mathbf{R}), \tag{1.33}$$

where $\mathbf{R}$ is a Bravais vector — a vector which points from one unit cell to another, and $\psi_{0\mathbf{k}}$ is a wavefunction localized in the unit cell. Note that $\psi_{\mathbf{k}}(\mathbf{r}) = \psi_{\mathbf{k}+\mathbf{K}}(\mathbf{r})$, where K is a reciprocal space lattice vector, defined such that $\mathbf{K}\cdot\mathbf{R}$ is a multiple of 2π. The phase factor for each unit cell is unchanged by the

addition of a reciprocal lattice vector. Likewise, the wavefunction that applies to each unit cell, $\psi_{0\mathbf{k}}$, is also unchanged by a change in the reciprocal lattice vector from $\mathbf{k}$ to $\mathbf{k}+\mathbf{K}$. This can be seen by inspection of Equation 1.32, where a reciprocal lattice vector would be any integer multiple of $2\pi/s$.

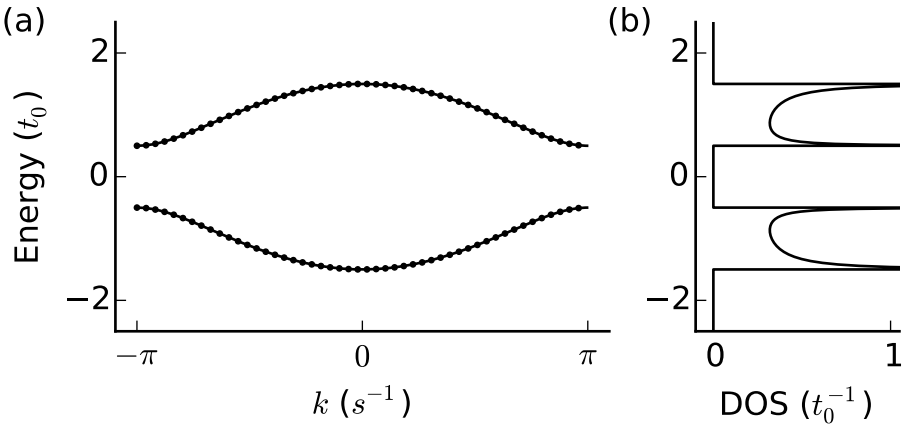


Figure 1.11: (a) Band structure for the dimerized 1D tight-binding model. The continuous line corresponds to the limit of infinite N, and the points correspond to $N = 100$, or 50 pairs. (b) Density of States (DOS) of the dimerized tight-binding chain in the limit of large N, normalized by N.

More often, Bloch functions are expressed as[9]

$$\psi_{\mathbf{k}}(\mathbf{r}) = e^{i\mathbf{kr}} u_{\mathbf{k}}(\mathbf{r}), \tag{1.34}$$

where $u_{\mathbf{k}}(\mathbf{r})$ is a periodic function, $u_{\mathbf{k}}(\mathbf{r}) = u_{\mathbf{k}}(\mathbf{r}+\mathbf{R})$. The equality $\psi_{\mathbf{k}}(\mathbf{r}) = \psi_{\mathbf{k}+\mathbf{K}}(\mathbf{r})$ is less obvious in this case, since the phase factor $e^{i\mathbf{kr}}$ is not equivalent for $\mathbf{k} \to \mathbf{k}+\mathbf{K}$. Indeed the wavefunction appears to be dramatically "curvier" rotating its phase one additional time *per unit cell*. The equality comes from the fact that the added curviness in this phase factor can be compensated by a change in the periodic function, $u_{\mathbf{k}}(\mathbf{r})$. In practice, this issue is usually unimportant since we generally consider wavevectors in the first Brillouin zone only — the natural unit cell of reciprocal space.

It is also possible to calculate the DOS, as was done for the simple 1D chain. While one can express the DOS in terms of energy, it can also be immediately expressed in terms of the wavevector, k, as

$$\mathrm{DOS}(k) = \frac{N\sqrt{(t_0 + t_1 \cos ks)^2 + (t_1 \sin ks)^2}}{2\pi t_0 t_1 \sin ks}. \tag{1.35}$$

The DOS is plotted in Figure 1.11b. If we again consider the scenario in which each localized orbital comes with a single electron, then the available energy levels will again be half-filled. In the case of this paired 1D chain, this results in a lower band which is completely occupied, and an upper band which is completely unoccupied. This type of band structure describes a semiconductor or an insulator, depending on the size of the energy gap which separates occupied from unoccupied states.

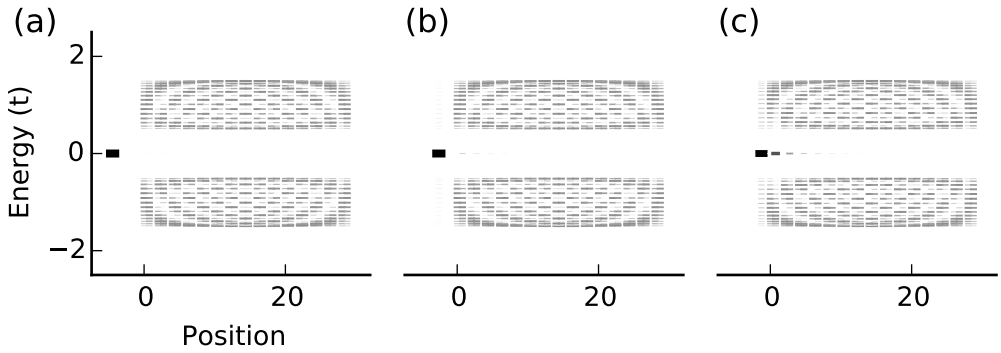


Figure 1.12: A 1-dimensional dimerized chain of sites in the tight-binding model, composed of 30 equally spaced pairs of sites, with an additional site brought in from the left. The positions of the additional site as it approaches from the left are (a) $-4s_1$, (b) $-2s_1$, and (c) $-s_1$, where s_1 is the spacing between nearest neighbours of different but adjacent pairs. In the last case, therefore, the additional site is in the next lattice site to the left of the 30-pair chain.

Figure 1.12 shows the process of bringing one additional orbital to a dimerized chain of 60 orbitals (30 paris), with the separation of the additional atom from the rest of the chain as $4s_1$, $2s_1$, and s_1, for (a), (b), and (c) respec-

tively, the last one therefore representing the normal spacing between pairs. As the additional orbital is brought into tunnel coupling with the rest of the chain, the discrete level gains a decaying tail in the gap of the dimerized chain. The extended states, meanwhile, gain a small amplitude at the additional site. However, unlike the case of the simple 1D chain, shown in Figure 1.8, when the orbital comes to the adjacent lattice site, in the last panel of both figures, the localized level does not delocaliize across the entire crystal. Instead, it retains its localized character, developing an exponentially decaying tail into the band gap.

The localized state developed at the unpaired atom in Figure 1.12c can be thought of as a model for a dangling bond. Because it is missing its normal bonding partner, its unpaired orbital fails to hybridize with the extended states of the crystal, and sits instead in the midgap. In this sense, it bears a very close resemblance to a dangling bond on the hydrogen-terminated silicon surface, where the lack of a capping hydrogen atom leaves the unpaired orbital as a deep-level defect with a localized wavefunction.

1.4 Tunneling

1.4.1 Tunneling Through a Barrier

Tunneling is a quantum phenomenon whereby a particle can traverse a potential energy barrier which is higher than the particle's total energy. In classical mechanics, this is impossible: a particle incident upon a barrier will be reflected at the point where the potential energy equals the particle's total energy. This point is called a classical turning point. Figure 1.13 shows such a barrier. While the classical trajectory bounces off of the barrier at the left turning point, x_a, a quantum wavefunction has a probability to be transmitted through the barrier. Even in the classically forbidden region from x_a to x_b, the wavefunction can have a non-zero amplitude, which allows the electron to emerge in the second classically allowed region to the right of x_b.

Usually, in introductory treatments of tunneling, this description is made formal by considering a rectangular barrier of width $|x_b - x_a|$ and of height V_0.

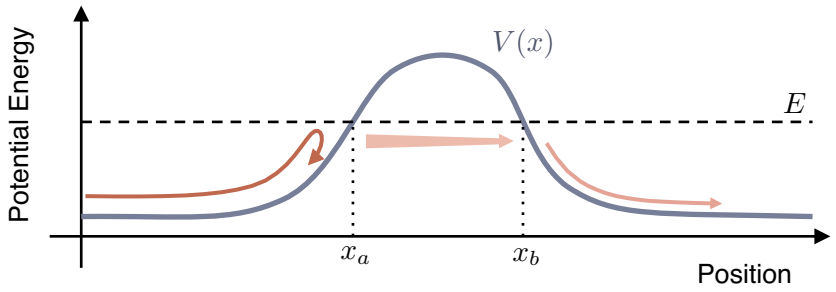


Figure 1.13: A quantum mechanical wavepacket incident from the left on a potential barrier, which it would not be able to cross classically, nonetheless has an amplitude to emerge on the other side. The energy of the patricle is E, and x_a and x_b label the classical turning points.

The Hamiltonian is

$$H = \frac{-\hbar^2}{2m}\frac{d^2}{dx^2} + V(x) \quad ; \quad V(x) = \begin{cases} V_0 & \text{for } x_a < x < x_b \\ \\ 0 & \text{otherwise.} \end{cases} \tag{1.36}$$

The Schrödinger equation can be solved separately in three regions, to the left of the barrier, inside the barrier, and to the right of the barrier, as described, for instance, in Cohen-Tanoudji.[10] In the classically allowed regions, the solutions are just the solutions of the free electron, proportional to $e^{\pm ikx}$. Inside the barrier, the Schrödinger equation dictates an exponential function of the form $e^{\pm\kappa x}$. The wavevector, k, and the exponential constant, κ, are given by

$$k = \sqrt{\frac{2mE}{\hbar^2}} \quad \text{and} \quad \kappa = \sqrt{\frac{2m(V_0 - E)}{\hbar^2}}. \tag{1.37}$$

The solutions found in each of the the three distinct regions are

$$\left.\begin{aligned}\psi_L^+ &= A_L^+ e^{ikx} \\ \psi_L^- &= A_L^- e^{-ikx}\end{aligned}\right\} \quad \text{for} \quad x < x_a$$

$$\left.\begin{aligned}\psi_{\text{bar}}^+ &= A_{\text{bar}}^+ e^{\kappa x} \\ \psi_{\text{bar}}^- &= A_{\text{bar}}^- e^{-\kappa x}\end{aligned}\right\} \quad \text{for} \quad x_a < x < x_b \tag{1.38}$$

$$\left.\begin{aligned}\psi_R^+ &= A_R^+ e^{ikx} \\ \psi_R^- &= A_R^- e^{-ikx}\end{aligned}\right\} \quad \text{for} \quad x < x_a$$

Typically, A_L^+ and A_L^- are interpreted as the amplitudes of incoming and outgoing waves to the left of the barrier, and likewise for A_R^+ and A_R^-. By setting A_R^- to zero, we consider the case in which no particle is "approaching from the right." A_L^+ is then interpreted as the amplitude of the incident particle (from the left), A_L^- is interpreted as the reflected amplitude, and A_R^+ is interpreted as the transmitted amplitude. The amplitudes in each region are then adjusted such that the wavefunction and its derivative are continuous at the edges of the barrier, x_a and x_b. In this way, a valid solution of the Schrödinger equation can be constructed from an incoming wave from the left, which decays through the barrier and has a small component to the right of the barrier. The transmission probability is then described by

$$T = \frac{|A_R^+|^2}{|A_L^+|^2} = \frac{4E(V_0 - E)}{4E(V_0 - E) + V_0^2 \sinh^2(\kappa |x_b - x_a|)} \tag{1.39}$$

This simple result demonstrates the basic phenomenon of tunneling: it shows a non-zero probability for a particle to pass through the barrier. But the interpretation of the result is rather peculiar in the context of most quantum calculations. We have described a wavefunction which satisfies the Schrödinger equation — that is we have found a stationary state. If the particle is prepared in this state, the time evolution of the system will leave the state unchanged aside from an overall phase. And yet we interpret this stationary state in a dynamic way, as a particle incident from the left and subsequently reflecting partly and transmitting partly. This is a strange way to think about a stationary state. Furthermore, in the context of STM, electrons occupy the extended

states of the sample and tip, and they are in some sense *always* incident on the barrier. If the above expression gives the probability to traverse the barrier, we then need to know how often each electron tries to traverse it. In some scenarios, an attempt frequency is used to convert a transmission probability into a transmission rate, for instance by using the frequency of a Bohr orbital in the context of atoms.[11] However, in many cases it is not entirely clear that such an "attempt frequency" is applicable. It would be useful to develop an approach to tunneling which aims to calculate dynamics from the start — something we will do presently, first for tunneling between discrete states, then for tunneling from a discrete state to a continuum.

1.4.2 Tunneling Between Discrete States

Sections 1.2 and 1.3 dealt with Hamiltonians in which localized sites were coupled by hopping constants, which describe tunneling from site to site. Here, we will describe the process by which realistic Hamiltonians can be approximated by simple tight-binding models. This will shed light on the origin and meaning of the hopping constants. We will also comment on some general features of tunneling between discrete states.

First, we consider overlap between eigenstates of two separate atoms, which we will denote ψ_L, an eigenstate of the left atom, and ψ_R, an eigenstate of the right atom,

$$H_\alpha|\psi_\alpha\rangle = E_\alpha|\psi_\alpha\rangle \quad ; \quad \alpha \in \{L, R\} \ . \tag{1.40}$$

When the two atoms are well separated, these two wavefunctions are good eigenstates of the total Hamiltonian,

$$H = \frac{p^2}{2m} + V_L + V_R, \tag{1.41}$$

where V_L and V_R are the confining potentials for the left and right atoms respectively. As the atoms are brought closer, however, they cease to be perfect eigenstates, and a probability emerges for electrons to be transferred from one orbital to the other.

In order to see how these eigenfunctions are modified because of their

interaction, we consider a two state basis spanned by these states only. We will comment on this assumption later. In this basis, the diagonal elements of the Hamiltonian are

$$\begin{aligned} H_{LL} &= \langle\psi_L|H|\psi_L\rangle &= E_L + \langle\psi_L|V_R|\psi_L\rangle \\ H_{RR} &= \langle\psi_R|H|\psi_R\rangle &= E_R + \langle\psi_R|V_L|\psi_R\rangle. \end{aligned} \tag{1.42}$$

This can be interpreted as meaning that the normal energies of the diagonal terms are corrected by an amount equal to the interaction of the left orbital with the confining potential of the right atom, and vice versa. The more important off-diagonal terms are

$$\begin{aligned} H_{LR} = H_{RL}^* &= \langle\psi_L|H|\psi_R\rangle \\ &= E_L\langle\psi_L|\psi_R\rangle + \langle\psi_L|V_R|\psi_R\rangle. \end{aligned} \tag{1.43}$$

The first term is proportional to the overlap between the two wavefunctions, $S \equiv \langle\psi_L|\psi_R\rangle$. The second term involves one of the confining potentials as well. However, if we express it as an integral in space, we see that the integrand is small except where the overlap is non-negligible. As long as the two sites are not too close together, this region will be far from the center of either confining potential, where the confining potential nearly vanishes, so that the second term can often be neglected. Roughly speaking, then, we can say that the matrix element which connects these two states is proportional to the overlap, S.

Equation 1.38 described the exponential decay of wavefunctions in classically forbidden regions. It says that a wavefunction must decay with an exponential constant determined by the barrier height. In principle, exponential increase is also possible, but is excluded on physical grounds. Using that result, we can expect that orbital overlap, and therefore the off-diagonal

matrix element, will decay with separation, s, as

$$S \propto e^{-\kappa s} \quad ; \quad \kappa \equiv \sqrt{\frac{2mE_i}{\hbar^2}}, \tag{1.44}$$

where E_i is the confining potential.

It is worth noting that if the two orbitals described above belong to two realistic atoms, then there are other eigenstates of each atom that need to be taken into account. Section 1.2 discussed the case where a hopping constant connected two non-resonant orbitals. When the hopping constant became comparable with the energy splitting between the orbitals (that is, when $2t$ became comparable with Δ), significant hybridization occurred. By the same token, whenever orbital overlap with other eigenstates becomes large enough that the resulting hopping constant is on the order of the energy difference, those additional orbitals need to be taken into account. This is the limitation of the assumption made above. The restriction to two states is only valid when the hopping constants are much smaller than the spacing of energy levels around those eigenstates for each atom.

The tight-binding model is a good approximation for confining potentials with relatively widely spaced energy levels, and where overlap between orbitals of neighbouring sites is small in comparison. This allows us to use the heuristic results from Section 1.2 to consider the transfer of a single electron between discrete states. For resonant states, the electron is transferred back and forth sinusoidally between the two orbitals, as depicted earlier in Figure 1.2. The probability of being in the initial state, as a function of time, is

$$P_i = |\psi_i(\tau)|^2 = \cos^2\left(\frac{t\tau}{h}\right), \tag{1.45}$$

which of course assumes coherent time-evolution. Such coherent oscillations have been proposed as the basis for charge qubits made from silicon DBs.[12,13]

This conceptualization of tunneling is a step in the right direction compared with the previous treatment that considered an incident wave, but it still presents difficulties of its own. In this picture, tunneling is not a one-way process as we normally think of it in the context of STM. The electron cycles

indefinitely. The quantity $\Gamma = 2t/h$ has units of frequency and represents the rate at which the electron cycles back and forth. It is difficult to use this in considering STM, however, since it would seem to indicate that any electron transferred to the sample from the tip would then return to the tip over an equal time. Clearly this does not happen. The cycling rate that describes tunneling between discrete states may give an idea of a natural timescale for tunneling, but it should not be thought of as a "transfer rate." In order to effectively describe tunneling as a one-way process, we need to consider transfer of an electron not from discrete state to discrete state, but instead from a discrete state to a continuum, or from continuum to continuum.

1.4.3 Tunneling from Disctrete State to Continuum

Time-dependent perturbation theory provides a means to address problems of dynamics whenever a small perturbation is added to a known and solved potential. In this case, we can imagine tunneling from a discrete state of a known potential to a continuum of states, such as the continuum formed by the band structure of a metal or semiconductor. We are interested in the transfer of electrons from an initial state, ψ_i, to the states of the continuum, ϕ_n.

The general idea is to imagine a perturbation turned on at time $t = 0$ (in this section, we will use t to represent time as usual, not to be confused with the hopping t used earlier in this chapter). In this case, the perturbation will be the overlap between the discrete orbital and the eigenstates of the continuum. We might imagine the moment at $t = 0$, where the overlap is turned on, as representing a moment when the discrete orbital is physically brought close to the continuum states. More accurately, this is a tool to consider what happens starting from an initial state where an electron occupies the discrete state, ψ_i. How are electrons subsequently transferred to states of the continuum, ϕ_n?

We start by assuming that a varying potential is added to a known Hamiltonian,

$$H = H_0 + V(t), \tag{1.46}$$

where the known Hamiltonian is already solved,

$$H_0|n\rangle = E_n|n\rangle. \tag{1.47}$$

The time-dependent potential used in this context has the simplest possible possible time-dependence,[14]

$$V(t) = \begin{cases} 0\ , & \text{for } t < 0 \\ V\ , & \text{for } t \geq 0, \end{cases} \tag{1.48}$$

so that aside from turning on, the potential actually has no time dependence at all. We consider an initial state, $|\psi_i\rangle$, which is an eigenstate of the unperturbed Hamiltonian, H_0, with energy E_i, and consider transitions to the contiunuum states, $|\phi_n\rangle$, also eigenstates of the unperturbed Hamiltonian, with associated energies, E_n.

What is the significance of the added potential V in the context of tunneling? When V is zero, ψ_i and ϕ_n are exact eigenstates of the Hamiltonian. This scenario was depicted earlier, in Figure 1.8a, where the discrete state is widely separated from the continuum states. When they are brought closer, it becomes possible for the discrete state to hybridize with the extended states, ϕ_n, as depicted, for instance, in Figures 1.8b and c. In Section 1.3.2, we discussed the fact that this hybridization leads to a "broadening" of the discrete state, however it also has the effect of enabling an initially localized electron to be transferred to the extended (continuum) states. V is an addition to the Hamiltonian which accounts for this small degree of hybridization, and thereby enables the transfer of electrons, as we will show presently.

The time evolution operator is $U(t) = e^{-iHt/\hbar}$ so that the wavefunction at time, t, is given by

$$|\psi(t)\rangle = e^{-iHt/\hbar}|\psi_i\rangle. \tag{1.49}$$

The probability of a transition to state $|\phi_n\rangle$ is then

$$P_{i\to n}(t) = \left|\langle\phi_n|\psi(t)\rangle\right|^2, \tag{1.50}$$

which, to first order in perturbation theory, is given by,[14]

$$P_{i\to n}(t) = \frac{4|V_{in}|^2}{(E_n - E_i)^2} \sin^2\left[\frac{(E_n - E_i)t}{2\hbar}\right], \tag{1.51}$$

where V_{in} is the matrix element of the perturbing potential, $\langle\psi_i|V|\phi_n\rangle$. For small values of t, this expression reduces to $|V_{in}|^2 t^2/\hbar^2$. Comparing this to the analogous behaviour described in the case of tunneling between discrete states, we can see that V_{in} here plays the role that the hopping constant played in that case. That is, V_{in}, can be thought of as a hopping constant connecting ψ_i to ϕ_n.

Equation 1.51 presents all the same problems as did the equivalent expression in the case of tunneling between discrete states (which is to be expected since we have not yet used the fact that the states ϕ_n form a continuum). Specifically, it is periodic with time, so that it does not seem to describe a one-way transition from the discrete state to the continuum. Furthermore, the short-time behaviour is quadratic with time, whereas a process characterized by a tunneling rate is expected to linearly transition on short timescales.

Sensible results are found by considering transitions not just to a single level, n, but to the set of all levels of the continuum, which we will denote $\{n\}$. The probability of transitioning to this set of states, is then

$$P_{i\to\{n\}} = \sum_n P_{i\to n} \approx \int_{-\infty}^{+\infty} dE_n \rho(E_n) P_{i\to n}(E_n), \tag{1.52}$$

where $\rho(E_n)$ is the density of states of the continuum per unit energy, and we have assumed that the transition probability, $P_{i\to n}$, can be written as a function of the final energy only. This amounts to assuming that the modulus squared of the matrix element V_{in} can be written as a function only of E_n, or alternatively, that it can be replaced by a coarse-grained average of its values for energies near E_n. We can denote this coarse-grained value of the modulus

squared as $|V_{i\{n\}}|^2$. This allows us to write

$$\begin{aligned} P_{i\to\{n\}} &= \int_{-\infty}^{+\infty} dE_n \rho(E_n) \frac{4|V_{in}|^2}{(E_n - E_i)^2} \sin^2\left[\frac{(E_n - E_i)t}{2\hbar}\right] \\ &= \frac{2|V_{i\{n\}}|^2 \rho_n(E_i) t}{\hbar} \int_{-\infty}^{\infty} dx \, \mathrm{sinc}^2(x), \end{aligned} \tag{1.53}$$

using the substitution $x = (E_n - E_i)t/\hbar$. The presence of $(E_n - E_i)^2$ in the denominator of the first line indicates that the integrand is sharply peaked at energies near the energy of the initial state (but the divergence is avoided by the presence of the sin function). Since the integrand is sharply peaked, we can assume that the density of states of the continuum varies slowly over the important range of energies, allowing the density of states to be replaced with its value at the initial energy, $\rho_n(E_i)$.

The integral in the second line of Equation 1.53 is equal to π, which gives the following result for the probability of a transition to the final states, $\{n\}$,

$$P_{i\to\{n\}} = \left(\frac{2\pi}{\hbar}|V_{i\{n\}}|^2 \rho_n(E_i)\right) t. \tag{1.54}$$

This result is known as Fermi's Golden Rule, and the rate associated with the transition, $\Gamma_{i\to\{n\}}$, is simply the quantity in brackets. We also note that even though we have considered transitions to a continuum of states, it has turned out that only transitions to states close in energy play an important role. In that sense, tunneling resonant: the electron can tunnel from a state with energy E_i to states of the continuum with the same, or very nearly the same, energy.

We saw earlier that the added potential, V, was related to the possibility to transition from the discrete state to the continuum of states. We now see that the transition rate from the discrete state to the continuum depends on the modulus squared of V_{in}. What determines the magnitude of these matrix elements? In the previous section, on tunneling between discrete states, we discussed the origin of the hopping constant, t, which connected two states, ψ_L and ψ_R. The close analogy allows us to think of V_{in} as a hopping constant connecting state ψ_i to state ϕ_n, and following the same reasoning as described

above for the case of tunneling between discrete states, we can say that the value of V_{in} is determined primarily by the overlap between the wavefunctions ψ_i and ϕ_n.

This gives a sensible rate for an electron in a discrete state to tunnel to the energy levels of a continuum of states. The linear transition is a result of the existence of a continuum of final states; while the transition to any one of these final states is actually quadratic in time for small times, the total transition probability turns out to be linear after we take a sum over all the possible final states.

The description of tunneling from a continuum to a continuum is essentially no different from this one. We simply consider the initial state, ψ_i to be one of the continuum states of one electrode, and the states ϕ_n again represent the continuum states of the other electrode. The total transition rate from one electrode to the other then requires an additional sum (or integral) over the states of the "initial" electrode.

This description of tunneling is the basis of the first successful theory of tunneling in STM, put forward by Tersoff and Hamann,[15] which will be described in more detail in the next chapter.

2 Scanning Tunneling Microscopy and the Silicon Surface

2.1 Scanning Tunneling Microscopy

The Scanning Tunneling Microscope (STM) belongs to the family of Scanned Probe Microscopy (SPM). Other members of the SPM family are Atomic Force Microscopy (AFM) and its many variations, which all measure the forces of interaction between the tip and sample, Scanning Near-field Optical Microscopy (SNOM), and many others, most of which came after the invention of STM. What all SPM techniques have in common is that a probe is scanned over a sample in order acquire spatially-resolved information about the interaction between the probe and the sample. The spatial resolution of the technique is determined by the nature of this interaction and the sharpness of the probe. The advantage of STM is that it relies on the rapidly exponentially decaying tunneling current through vacuum in order to probe samples. This is the feature of STM that allows it to achieve atomic resolution routinely.

The STM was invented in 1982 by Binnig and Rohrer,[16] who had first demonstrated the ability to form a controllable tunneling gap through vacuum,[17] a necessary prerequisite for the STM. Figure 2.1 shows the first reported atomically resolved STM image, which the authors described as "our shining example of an STM graph." It shows two unit cells of the 7×7 reconstruction of the Si(111) surface.

In the years that followed its invention, the field of STM exploded, as the technique was applied to a variety of metallic and semiconducting sur-

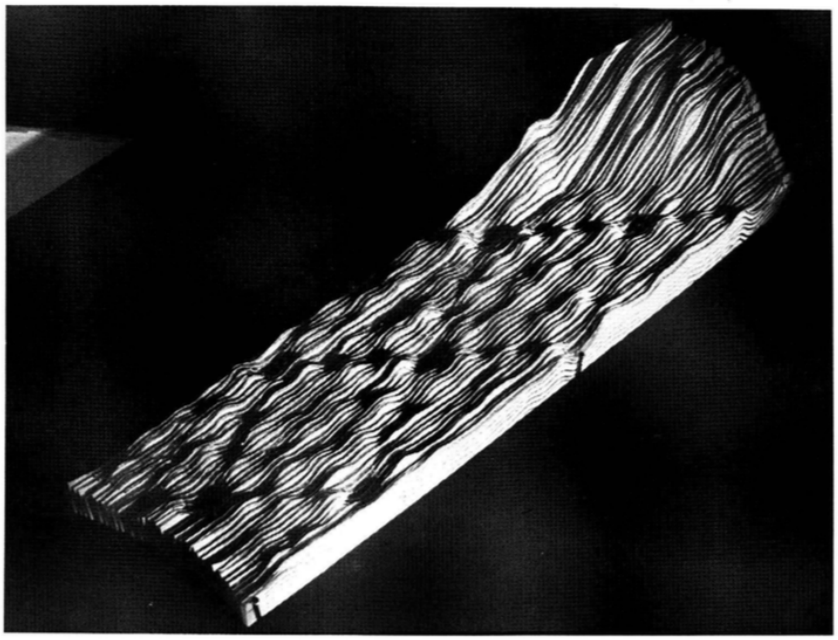


Figure 2.1: First reported STM topographical image showing two unit cells of the Si(111) 7×7 surface. Reprinted figure with permission from Binnig *et al.*, *Physical Review Letters* **50**, 120, 1983.[18] Copyright (1983) by the American Physical Society.

faces, providing real-space images of atomic surface reconstructions for the first time.[18–23] The field soon grew further to include the study of surface chemistry[24,25] and atomic scale manipulations.[26] In 1986, Binnig and Rohrer were awarded the Nobel prize for their design of the STM.

2.1.1 Basic Principles of STM

The simplest description of the STM is a sharp metallic tip scanning a conducting surface, as shown in Figure 2.2. A bias, V_S, is applied between the tip and sample, and a tunneling current, I_T, is measured when the tip is within $\lesssim 1$ nm. By convention, the bias of the sample relative to the tip is used, no matter which of the two is grounded in reality. Sometimes, sample bias is simply called "bias" in the context of STM, and denoted V. Tunneling current is maintained constant by the use of a feedback loop, which causes the tip to move toward the sample whenever current is too low, and move away from

the sample whenever current is too high. Typical sample biases can range from tens of mV for metallic samples, to $\sim$ 2 V for semiconducting samples. Typical tunneling currents may vary from 10 pA to 1 nA.

Images are acquired in STM by raster scanning a tip over a region of a surface. As the tip moves along a line over the surface, the feedback loop maintains a constant tunneling current, and therefore a (roughly) constant tip-sample separation, as shown in Figure 2.2. In practice, the measured topography is convolved with information about the sample DOS, as will be discussed in the next section. As the tip moves along a line, for example in the x-direction, the tip height, z, is measured. The tip is then moved a small distance in the perpendicular direction, y in this example, and another line along x is measured. This is repeated until a (typically square) image is acquired. The first reported STM image from Binnig and Rohrer illustrates this nicely, since the individual line profiles making up the topographic image are directly used to show the topography, as seen in Figure 2.1. This type of depiction of topographic maps is no longer in common use. Instead, it is customary to show STM images as a colormap of topography. In black and white images, brightness corresponds to protrusions and darkness to depressions.

The schematic in Figure 2.2 shows three piezoelectric elements, labeled x, y, and z. Such materials experience a change in their physical length in response to an applied bias. The position of the tip in each dimension can be controlled by applying different biases to each of these piezo elements. The position of the tip, relative to its centered position, where all piezo elements are unbiased, is then proportional to the applied bias on each piezo. Typically, biases of up to hundreds of volts can be applied to the piezo elements, and their total range is typically one or several microns. It is often necessary to move the tip much further than this range, for instance to move to a different area of the sample, or to approach the sample surface from a distance on the order of 1 mm. Such coarse positioning is accomplished by a separate mechanism, often stick-slip motors, which can move through a range on the order of centimetres in steps on the order of 100 nm.

All STMs are extremely sensitive to mechanical noise, since tip position

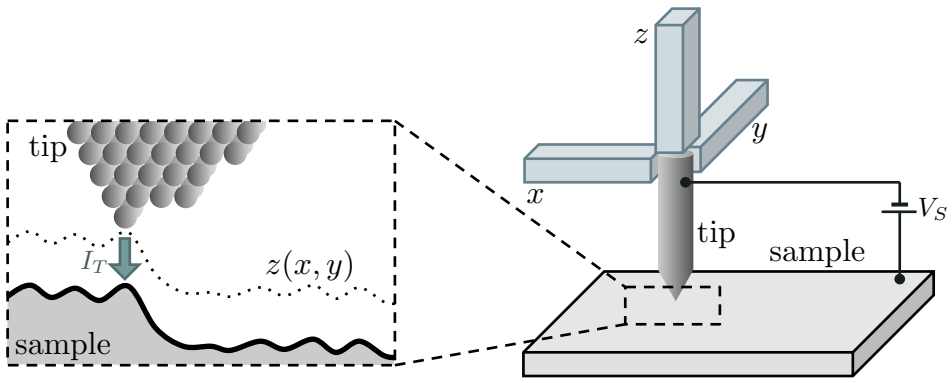


Figure 2.2: Schematic representation of an STM. The left panel zoom-in shows that an atomically sharp tip traces the atomic corrugations of the sample, $z(x, y)$, by maintaining a constant tunneling current, I_T. The right panel shows that the STM tip scans the surface using piezoelectric elements, labeled x, y, and z, to control its position on the atomic scale, with a potential V_S applied between the tip and sample.

needs to be maintained within a fraction of a typical bond length. Any significant vibration of the tip relative to the sample results in an uncontrolled crash of the sharp apex into the sample, often putting an end to the day's research. Care needs to be taken to isolate the STM from all mechanical noise, starting from the room in which the STM is located, ideally on the lowest floor of a building and even with its own foundation, and proceeding through various other means of isolation, possibly including air-legs or active isolators to float the entire microscope, springs to hang the scanner and sample, and eddy-current damping. And because very small currents are always being measured in STM, equal care needs to be taken to eliminate all sources of electronic noise.

While it is possible to do STM in air or other environments, the vast majority of work, and many of the most exciting results, have come from STM done in Ultra-High Vacuum (UHV). UHV corresponds to pressures below $\sim 10^{-9}$ Torr. Such pressures are required to ensure that the atomic structure of surfaces can be studied in the absence of interfering atoms and molecules

from the environment. Even at pressures as low as 10^{-8} Torr, a typical surface atom is struck by a molecule from the environment on the order of once per minute. This leads to a fairly high chance of sample contamination as well as an increased chance of tip changes, as molecules can react with the tip. It also leads to the possibility of artifacts in measurement, due, for instance, to physisorbed molecules. Ideally, experiments are performed at pressures around 10^{-10} Torr or below.

2.1.2 Theory of STM

The basic theory of STM was developed by Tersoff and Hamann,[15] and is based on Bardeen's earlier approach to tunneling which he developed to describe tunneling between superconductors.[27] We will take this approach as the basis for the presentation in this section. It applies Fermi's Golden Rule to transitions from tip to sample, or vice versa. Transitions from a state μ of the tip (scanned probe) to the state ν of the sample are considered to take place at the rate

$$\Gamma_{\mu\nu} = \frac{2\pi}{\hbar^2}|M_{\mu\nu}|^2, \tag{2.1}$$

where $M_{\mu\nu}$ is usually called the "matrix element" connecting state μ to state ν — a remarkably uninformative nomenclature. Comparing this expression to Equation 1.54, we see that the matrix element $M_{\mu\nu}$ is the matrix element of the perturbing potential, V, which was added to the known Hamiltonian, H_0, in order to account for hybridization between states of the tip and sample, and enabling transfer of electrons. The "matrix element" is therefore a small matrix element in the Hamiltonian, related to the overlap of the wavefunctions ψ_μ and ψ_ν.

Bardeen showed that the matrix element could be expressed as

$$M_{\mu\nu} = \frac{\hbar^2}{2m}\int d\mathbf{S}\cdot\left(\psi_\mu^*\nabla\psi_\nu - \psi_\nu\nabla\psi_\mu^*\right), \tag{2.2}$$

where the surface integral is over any surface within the tunneling barrier region separating the tip from the sample. This form provides a means of directly calculating the value of $M_{\mu\nu}$ whenever the wavefunctions of the tip

and sample are explicitly known. However, some general features of the matrix element can be worked out using the argument outlined in Section 1.4.3. In particular, we expect wavefunctions in the vacuum barrier to decay with an exponential constant, $\kappa = \sqrt{2mE_i/\hbar^2}$. We can then expect, since the matrix element $M_{\mu\nu}$ is a function of wavefunction overlap, that

$$|M_{\mu\nu}| \propto e^{-\kappa s}, \tag{2.3}$$

where s is the separation between tip and sample. Since the biases used in STM are usually small in comparison to the work functions, we expect M to be a relatively slowly varying function of energy. Aside from the dependence on energy, the exponential dependence on distance makes it clear that the STM is extremely sensitive to surface topography.

The total current is given by a sum over all the transition rates from all the states of the probe to all the states of the sample. It also needs to take into account the occupations of these states. That is, electrons can only tunnel from a state μ to a state ν if μ is occupied and ν is unoccupied. The probability of μ being occupied is given by the Fermi function, $f(E_\mu; E_F^{\text{tip}})$, where E_F^{tip} is the Fermi level of the tip*. Likewise, the probability of the state ν being *un*occupied is $[1 - f(E_\nu; E_F^{\text{sam}})]$. This leads to the following expression for tunneling from tip to sample:

$$I = \frac{2\pi e}{\hbar} \sum_{\mu,\nu} f(E_\mu; E_F^{\text{tip}}) \big[1 - f(E_\nu; E_F^{\text{sam}})\big] |M_{\mu\nu}|^2 \delta(E_\mu - E_\nu), \tag{2.4}$$

where the delta function accounts for the fact that tunneling is elastic.

The sums over states of the tip and sample can be converted to an integral over energies, making use of the density of states per unit energy in both tip and sample. The resulting equation for current is

$$I = \frac{2\pi e}{\hbar} \int dE f(E; E_F^{\text{tip}}) \big[1 - f(E; E_F^{\text{sam}})\big] |M_{\mu\nu}(E)|^2 \rho_{\text{tip}}(E) \rho_{\text{sam}}(E), \tag{2.5}$$

*In the case of semiconductors and insulators, the Fermi level may not be well defined. When the term "Fermi level" is used in such cases, it is to be understood as signifying a chemical potential, since Fermi levels and chemical potentials are used interchangeably throughout this thesis.

or, assuming low temperature, so that the Fermi functions can be replaced by step functions, one can write more simply,

$$I = \frac{2\pi e}{\hbar} \int_{E_F^{\rm sam}}^{E_F^{\rm tip}} dE |M_{\mu\nu}(E)|^2 \rho_{\rm tip}(E) \rho_{\rm sam}(E), \tag{2.6}$$

where the matrix elements connecting states of the tip and sample are now assumed to be a function only of energy.

In general, since we want to study the properties of the sample, we try (or hope) to use a tip with a nearly flat density of states near its Fermi level (this makes metal tips more suitable than semiconducting tips). In an ideal case, the density of states of the tip is constant over the range of energies of interest. In that case, we can further simplify the expression for the current to

$$I = \frac{2\pi e}{\hbar} \rho_{\rm tip} \int_{E_F^{\rm sam}}^{E_F^{\rm sam}+eV} dE |M_{\mu\nu}(E)|^2 \rho_{\rm sam}(E). \tag{2.7}$$

This shows that the tunneling current can be considered as coming from an integral from Fermi level to Fermi level, of the density of states, weighted by the matrix element. It also has the consequence that the sample density of states is immediately accessible, since the derivative with respect to sample bias is

$$\begin{aligned} \frac{dI}{dV} &= \frac{2\pi e}{\hbar} \rho_{\rm tip} |M_{\mu\nu}(E_F^{\rm sam} + eV)|^2 \rho_{\rm sam}(E_F^{\rm sam} + eV) \\ &\propto \rho_{\rm sam}(E_F^{\rm sam} + eV). \end{aligned} \tag{2.8}$$

Sometimes the density of states is taken instead to be proportional to the differential conductivity, dI/dV, normalized by the total conductivity, I/V. In general, this normalization is only a small correction, but creates some cancellation of terms, and accounts to some extent for the energy-dependence of $M_{\mu\nu}$.

One assumption we have made in this line of argument, which may be questioned at points in this thesis, is the use of Fermi-Dirac statistics to describe the occupation of energy levels of the tip and sample. For mid-gap states and

at low temperatures, it is possible for states in a semiconducting sample to be out of equilibrium with respect to their own Fermi level because of the injection or extraction of electrons from or to the STM tip. We will explore this issue more in several places throughout this thesis.

2.1.3 STM of Semiconductors

STM of semiconductors introduces some peculiarities that require special consideration. Figure 2.3 shows a band diagram depicting a tip tunnel-coupled to a semiconductor sample. One notices that unlike in the case of STM of metals, the potential difference between tip and sample does not only drop across the vacuum barrier, but continues to have an effect deep into the sample. This is because semiconductors have a much smaller density of free electrons to screen charge, compared to metals. While in a metal, electrons quickly rush to (or from) the surface to exclude any field from penetrating, semiconductors may have few or no free carriers to do this. The result is that most of the potential difference typically drops across the vacuum barrier, but there is still a residual potential drop in the near-surface region of the semiconductor.

The potential drop near the surface of the semiconductor is called Tip-Induced Band Bending (TIBB), since the spatially changing electrostatic potential shifts the sample's energy levels up or down. The degree of TIBB depends on the mismatch between the vacuum level, E_{vac}, in the metal and the vacuum level deep inside the sample. The depth of the TIBB in the sample depends on the screening length in the sample, which itself is determined by doping density and carrier type — screening of upward band bending may be different from screening of downward band bending in a given sample.[28]

The defining feature of the semiconducting sample is its bandgap at the Fermi level — an energy window with zero density of states. Considering Equation 2.7, it is clear that in a range of energies corresponding to the bandgap, we can expect zero current. Equations aside, this is clear since in that range of energies, there are no sample states that can accept a tunneling electron from the tip, nor supply one to it. This makes the bandgap a dangerous place for STM. Since tip height in STM is adjusted to ensure constant tunneling cur-

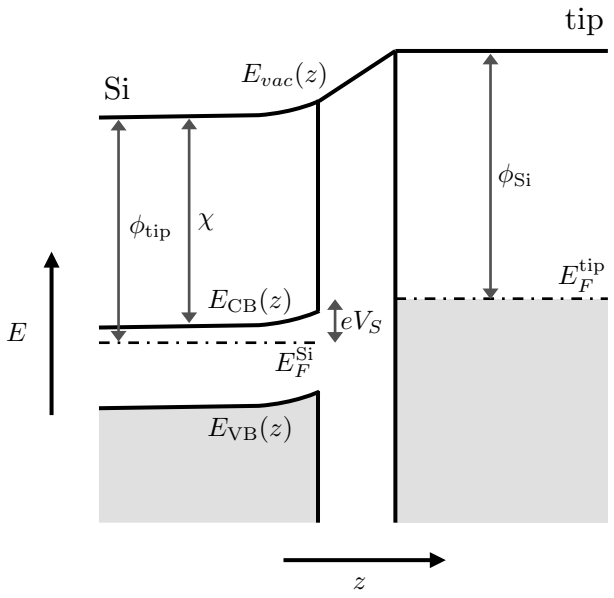


Figure 2.3: Band diagram illustrating STM of a semiconductor, and indicating many of the important variables.

rent, the inability of the sample to provide any tunneling current will result in an uncontrolled approach of the STM tip towards the sample, as it seeks to maintain the desired non-zero current. Since this current never turns on, the approach of the tip can continue until the tip has been forced into the sample, ruining its once sharp apex in the process. This is the dreaded "tip crash."

It is important to note that vacuum levels may be misaligned even at 0 V. The condition $V_S = 0$ V states that the Fermi levels in the tip and samples are aligned, $E_F^{\mathrm{tip}} = E_F^{\mathrm{Si}}$. However, different samples can have different work functions, ϕ, and if $\phi_{\mathrm{tip}} \neq \phi_{\mathrm{Si}}$ then there can be a misalignment in vacuum levels between tip and sample even at 0 V. The difference between the work functions at two surfaces is called the Contact Potential Difference (CPD).

2.1.4 The Origin of the Contact Potential Difference

The origin of the differing work functions in different materials is their differing ability to bind electrons. We can imagine two materials, A and B, both neutral so that the total number of electrons is equal to the total number of

protons in each, and initially not in contact with one another. For a variety of reasons (orbital structure, lattice configuration, *etc.*), material A might bind its electrons more efficiently than does B, so that they lie deeper below the vacuum level; the work function of A will be greater than that of B. As long as both materials remain neutral (and assuming charge does not re-arrange itself within each material, for instance preferring certain facets of the material's surface), the vacuum level is equal everywhere, within both materials and outside them as well.

Note that with both materials neutral, the Fermi levels are necessarily misaligned, with A's being lower. Now if the two materials are brought into contact, since A bound its electrons more deeply, it may have unoccupied levels which are lower in energy than some occupied levels of B. Electrons will be transferred from B to A. The increase in the electron density in A will shift its bands upward, along with its Fermi level and its vacuum level. Likewise, these same levels in B will be shifted downward, until the Fermi levels of A and B are aligned. At the end of this process, the potential difference between the two materials is zero (since their Fermi levels are aligned), but there is now a misalignment in their vacuum levels. The magnitude of the misalignment in the vacuum levels is precisely the CPD. In order to re-align the vacuum levels between the two materials, a potential difference must be applied, equal in magnitude to the CPD — a state of affairs referred to as "the flat band condition."

It is strange, but true, that two tunnel coupled materials at 0 V may experience a fairly large electric field across the vacuum barrier, and despite this, no net current flows from one material to the other. Tunneling current is determined by differences in Fermi levels.

2.1.5 Quantum Effects

There are other effects, peculiar to STM of semiconductors, which come from quantum effects. In general, we consider bands to "bend" rigidly. That is, the energy level structure of the material stays constant, aside from a rigid shift upward or downward depending on the electrostatic potential at each location.

This is a powerful and very broadly applicable way to think of semiconductors. However, one might question just how far one can take this. In the case of TIBB, bands may bend by an amount on the order of one eV over a distance of a few nanometers — approaching the atomic scale. One might well suspect that the assumption of rigid shifting of energy levels might break down, and indeed it does in some cases.

Figure 2.4 illustrates two quantum effects that may need to be taken into account in some instances. Bands are shown as bending abruptly downward at the surface.

The first quantum effect, in the CB, is the quantization of levels. If bands shifted rigidly, the lowered CB levels near the surface would have a large density of states. However, the lowered CB instead acts as a potential well, able to support only discrete quantized bound states. The shape of the confining potential is similar to a half-lens, roughly circular in the plane of the surface, under the tip, and extending only a shallow depth into the semiconductor.

A second quantum effect is the evanescent tails of wavefunctions from the VB, reaching toward the surface. This effect is a simple consequence of the ability of wavefunctions to extend into classically forbidden regions, as described above. The larger the barrier separating the wavefunction from the surface, the smaller its evanescent tail at the surface, as depicted by the two wavefunctions shown in the VB in Figure 2.4. Of course, the symmetrical effects are equally possible when the bias is opposite, where the evanescent tails extend from the CB, and likewise, near-surface states in the VB may be quantized.

2.2 Silicon

2.2.1 Bulk Silicon

Bulk silicon has a diamond structure in which every atom is tetrahedrally bonded to four nearest-neighbour silicon atoms, as shown in Figure 2.5. We have labelled the top four corners of the unit cell with the cardinal directions. Of course, they should instead be described using the unit vectors of the unit

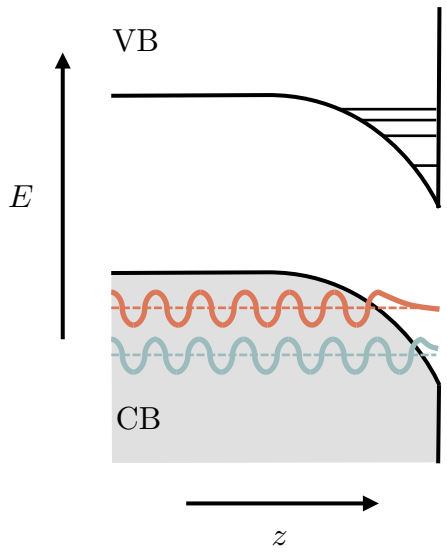


Figure 2.4: Band diagram of a semiconductor experiencing downward band bending near its surface. Two quantum effects are visible: discretization of confined states in the near surface region of the CB, and evanescent tails of the extended states of the VB decaying toward the surface. This figure is based on a figure from.[29]

cell, but since it can be easy to lose track of the meanings of directions like $[\bar{1}10]$, I will use the cardinal directions for convenience in parts of this section.

The diamond structure is Face-Centred Cubic (FCC) with a basis. In Figure 2.5, the silicon atoms belonging to each of the two FCC lattices are coloured in slightly different shades of grey. The tetrahedral bonds connect each atom with the four neighbouring atoms of the other sublattice. The light grey lattice is given by the dark grey one translated by a vector equal to one quarter the distance across the unit cell diagonally, from the "bottom West" corner to the opposite, "top East" corner. The bond length is then $a\sqrt{3}/4$, where a is the lattice constant (sidelength of the unit cell). This works out to a bond length of 2.352 Å.

A crystal consists of a large number of unit cells repeated periodically, as shown in Figure 2.6. Bond lengths are equal for every Si-Si bond throughout the crystal (except in the vicinity of defects or surfaces). At first sight, it might

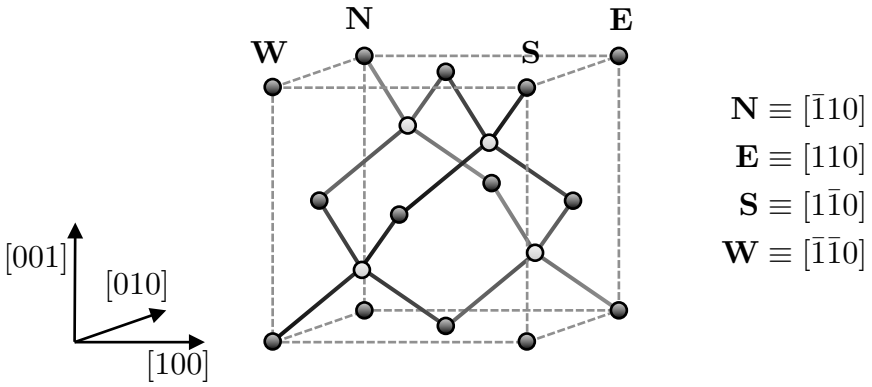


Figure 2.5: The conventional unit cell of silicon. The crystal structure is FCC with a basis. Silicon atoms from each sublattice are colored in different shades of grey. The cardinal directions are used as shorthands for the indicated lattice directions. The lattice constant at room temperature is 5.431 Å.

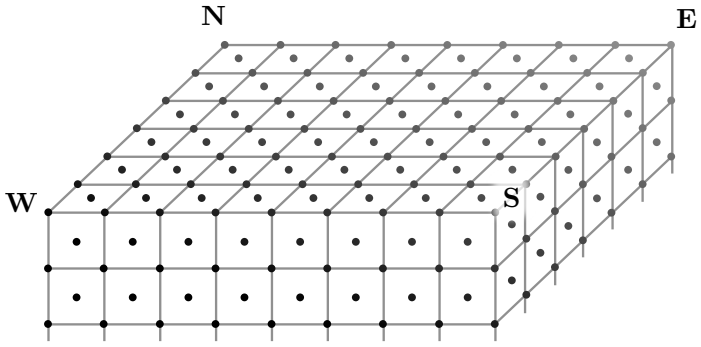


Figure 2.6: The silicon lattice made up of many unit cells. Only surface atoms are shown, and the three visible surfaces are the (100), (010), and (001) surfaces.

seem that this arrangement of atoms should lead to a single band without gaps, like in the case of the simple one dimensional chain described using the tight-binding model above, where atoms were equally spaced. This turns out to not be the case. In three dimensions (actually, for dimensions greater than one), it is possible to have a pairing structure even with a single bond length throughout the crystal, because of the spatial arrangement of atoms. In this case, the smallest possible unit cell contains two atoms, not one. In order to

conceptualize a tight-binding approach, one would first have to consider the eigenstates of the unit cell, as was shown for the dimerized one-dimensional chain in Section 1.3.3. Like in that case, this leads to a splitting in the resulting bands for silicon, which opens a gap between occupied and unoccupied states, making silicon a semiconductor.

In order to calculate more realistic band structures, one must include all of the orbitals of the individual atoms, which can give rise to several bands, as well as hybridization between the bands originating from different orbitals. In practice, band structures are calculated with more sophisticated methods such as the Orthogonalized Plane Wave (OPW) method[30] or the pseudopotential method.[31] Figure 2.7 shows the band structure of silicon as calculated by F. Herman in 1955 using the OPW method.[32] The silicon band structure has an indirect band gap of 1.12 eV at room temperature, which goes to 1.17 eV near 0 K.

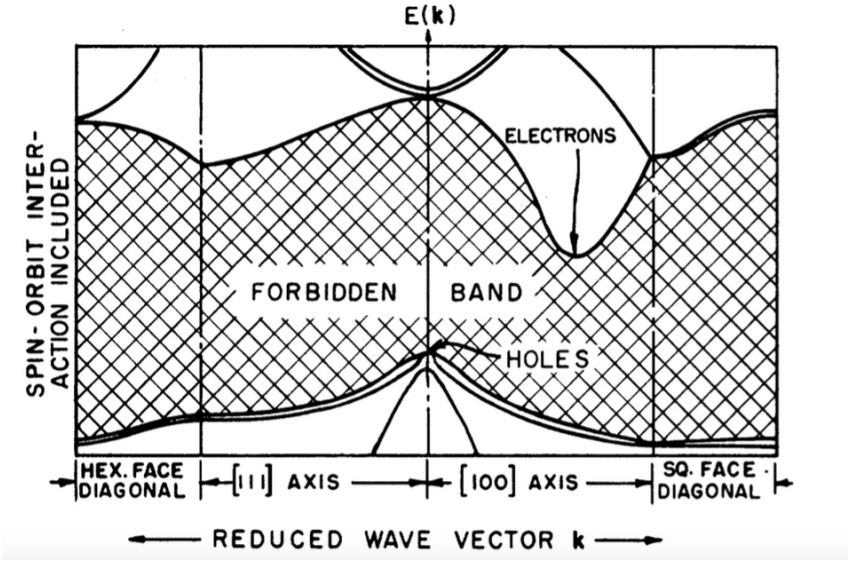


Figure 2.7: Band structure of silicon as a function of wavevector. Reprinted figure with permission from F. Herman, *Procedings of the IRE* **43**, 1703, 1955.[32] Copyright (1955) IEEE.

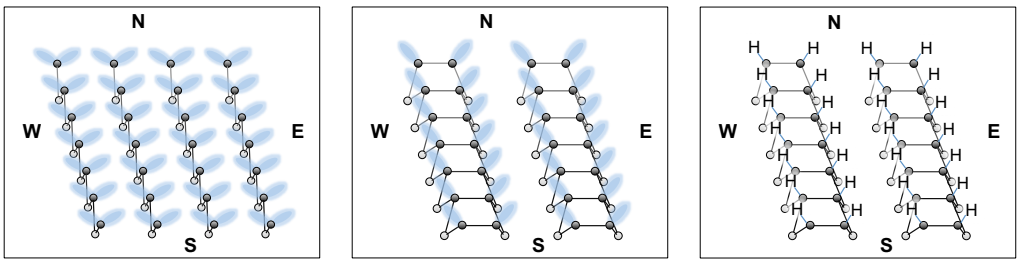
2.2.2 The Silicon (100) 2×1 Surface and Dangling Bonds

The three surfaces of the crystal shown in Figure 2.6 are the (100), (010), and (001), which are all identical by symmetry. Throughout this thesis, we will generally refer to it as the (100) surface, although it is sometimes also called the (001) surface. This surface is of tremendous technological relevance. It is the standard surface used in CMOS fabrication, which makes the semiconductor devices in nearly all high-tech products. Silicon, and this surface in particular, is the basis of the semiconductor industry.

In reality, unreconstructed surfaces like this one are often unstable, as this particular surface certainly is. We can see why this is by considering more closely the bonding structure for the top layers of atoms of the (001) surface. Figure 2.8a shows the unreconstructed surface atoms as well as the next layer of atoms below. Each surface atom is bonded to two silicon atoms below, one to the South and one to the North, leading to a zigzag of bonds running from South-to-North (along the $[\bar{1}10]$ direction). This zigzag of bonds is already visible in the unit cell in Figure 2.5. One can also infer from the unit cell that the distance between neighbouring atoms on the unreconstructed (100) surface is $a/\sqrt{2} = 3.84$ Å.

The instability of the surface comes from the fact that the surface atoms are missing their bonding partners above. As a result, two unbonded sp^3 orbitals extend upward and to the West, and upward and to the East, in keeping with the tetrahedral bonding tendency of silicon. The unbonded orbitals are called Dangling Bonds (DBs), and are shown as blue lobes in Figure 2.8. Each dangling bond on this surface can lower its energy by forming a chemical bond. Clearly there is no shortage of potential bonding partners on this surface, and while the nearest dangling bond on the surface is further away than the normal nearest neighbour distance, the lattice can be strained to bring pairs of surface atoms closer together, allowing new bonds to form. The surface reconstructs to form rows of pairs of surface atoms, as shown in Figure 2.8b. The resulting reconstructed surface is the clean Si(100)-2 × 1 surface, and it has precisely one dangling bond per surface atom.

The dangling bonds that occur at a surface can be removed by reacting the

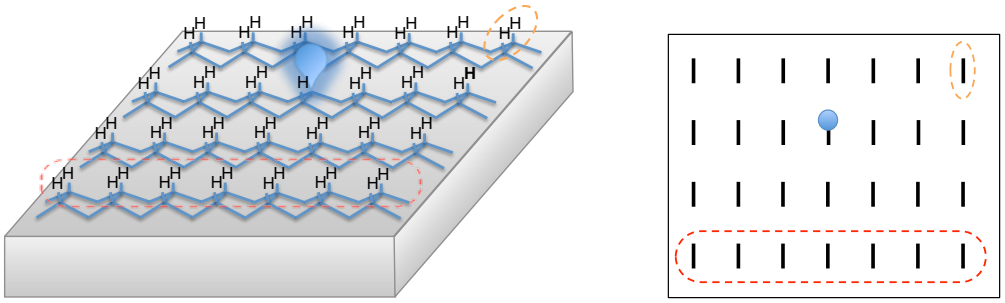


(a) (b) (c)

Figure 2.8: (a) The unreconstructed Si(100) surface. (b) The reconstructed Si(100)-2 × 1 surface. (c) The hydrogen-terminated surface, H-Si(100)-2 × 1.

surface with a molecule or atom. Hydrogen radicals (that is, H, as opposed to H_2) are ideal capping elements, since they are extremely reactive, and only have one unpaired valence electron, and thus are satisfied with only one chemical bond. In practice, Hydrogen-termination of the (100) silicon surface is relatively straightforwardly accomplished by exposure to hydrogen radicals at a temperature of around 330°C. Each dangling bond reacts with a hydrogen atom, so that the resulting surface consists of a 2 × 1 reconstructed surface with a single hydrogen atom capping each surface silicon atom, written H-Si(100) 2 × 1 and depicted in Figure 2.8c. The pairs of atoms on this surface are known as dimers, and the rows formed by these pairs are known as dimer rows. We can schematically represent this surface with diagrams like the one shown in Figure 2.9b.

We can also note that if we were to repeat the same argument as above, but considering the surface atoms to be from the lighter-coloured sublattice, for instance by considering the atoms a single atomic layer lower, then we would have to take into account that their bonding structure, while also tetrahedral, is rotated by 90° relative to the darker-coloured sublattice. It follows from this that the surface reconstruction would then have to be rotated by 90°. Indeed, atomically resolved images of this surface show that the dimer rows on adjacent terraces (separated by a single atomic step) are always perpendicular, as shown, for instance, in Figure 2.10.



(a) (b)

Figure 2.9: (a) The H-Si(100)-2 × 1 surface, and (b) its schematic representation. The blue lobe/circle represents a dangling bond, the orange dashed line encircles a single dimer, and the red dashed line encircles a dimer row.

Isolated DBs will exist on this surface wherever a single hydrogen atom is missing from an otherwise hydrogen-terminated surface, as shown in Figure 2.9a. The presence of an isolated DB on the H-Si(100) 2×1 surface introduces a localized state with an associated mid-gap energy level, much like the unpaired atom described at the end of Section 1.3.3. Like in that case, the localization of the DB orbital is a result of the fact that the hydrogen-terminated silicon crystal offers no resonant energy levels with which the DB energy level can hybridize. This opens great opportunities for engineering the electronic structure and wavefunctions of DB structures and devices; these can be fabricated with atomic control of positions, and, in the window of energies corresponding to the silicon bandgap, the substrate is in some sense invisible — it acts as a "solid state vacuum."

2.3 The Dangling Bond Orbital

We can describe the wavefunction of an isolated DB using a Slater-Type Orbital (STO), which has the two-lobe form characteristic of p-orbitals,

$$\psi_{\mathrm{DB}}(r,\theta) = Nze^{-\zeta(z)r}, \tag{2.9}$$

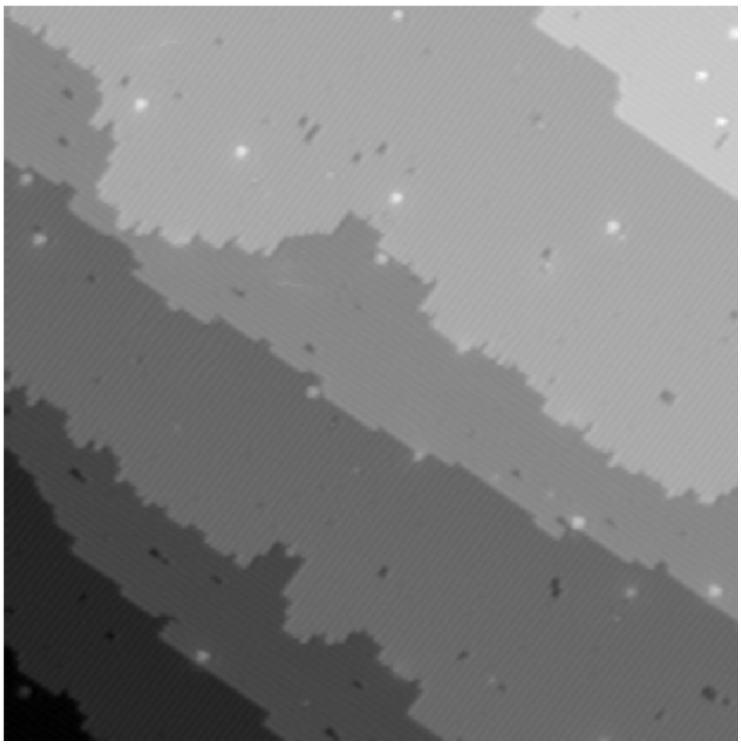


Figure 2.10: 80 nm × 80 nm image of the H-Si(100)-2 × 1 surface, showing seven atomic terraces separated by single atomic steps. Bright localized protrusions are DBs. $I_T = 80$ pA, $V_S = -2.0$ V.

where ζ is an exponential decay constant and N is a normalization constant. This orbital is different from normal STOs in the fact that the decay constant, ζ, has a spatial dependence, reflecting the fact that the orbital occurs at a surface and sees two very different potential landscapes toward vacuum and toward bulk. In particular, we generally expect the decay constant of a wavefunction in a barrier to have the value $\sqrt{2m(V-E)/\hbar^2}$, as described in Section 1.4, where V is the energy of the barrier and E is the energy of the orbital. The quantity $V - E$ can be called the ionization potential, W_i. Since the DB is at the interface between silicon and vacuum, it sees a barrier height

equal to the vacuum level (on the order of 5 eV above the DB energy level) on the vacuum side, and equal to the CB edge (on the order of 0.5 eV above the DB energy level) on the bulk side. We can therefore assume a functional form for the ionization potential which transitions from the bulk value at large negative values of z to the value in vacuum for large positive values of z,

$$W_{\mathrm{i}}(z) = W_{\mathrm{bulk}} + \frac{1}{2}(W_{\mathrm{vac}} - W_{\mathrm{bulk}})[\tanh(z/w) + 1], \tag{2.10}$$

which transitions from W_{bulk} to W_{vac} over a width w centred at the surface. The decay constant is then simply

$$\zeta(z) = \sqrt{\frac{2mW_{\mathrm{i}}(z)}{\hbar^2}}, \tag{2.11}$$

which can easily be expressed in polar coordinates and incorporated it into Equation 2.9.

The orbital described by Equation 2.9 can be used to describe both the singly occupied (neutral) orbital as well as the doubly occupied (negative) orbital, with the difference that the higher energy of the doubly occupied orbital reduces the ionization potential, $W_{\mathrm{i}}(z)$, everywhere, which causes the DB orbital to decay less rapidly. Figure 2.11 shows the neutral and negative DB orbitals.

This description of the DB orbital is useful because it can be described with a simple analytical formula. It also accurately describes the exponential tails of the wavefunction toward the vacuum and toward the bulk as a function of the DB energy level, as well as the relative weights of the lobes in bulk and in vacuum. However, some assumptions are made which we know to be untrue. The STO orbital used makes the assumption of a p-orbital, when in fact we expect that the orbital will have a significant s-component. More to the point, a realistic calculation would show corrugation of the wavefunction with a structure related to the lattice. The STO, at best, describes an envelope for the realistic wavefunction. Finally, the STO has an axial symmetry with respect to rotations about the z-axis, which reflects the assumed symmetry in the potential energy landscape in the vicinity of the DB. In fact, we know

that no such symmetry exists because of the bonding configuration of the host atoms and the reconstruction of the surface, both of which break the rotational symmetry. Taking this into account, the only expected symmetry is a reflection symmetry through the plane to which the direction of the dimer rows is normal. Keeping all this in mind, the STO is a useful computational tool which captures many of the important features.

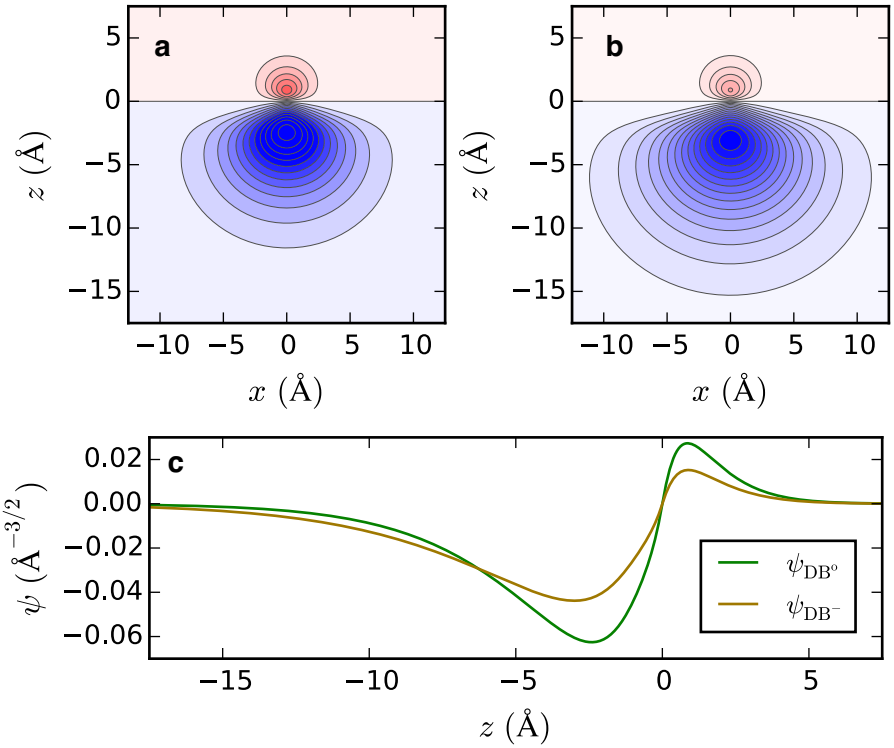


Figure 2.11: (a) neutral and (b) negative DB orbitals, showing isodensity surfaces of the DB calculated according to Equation 2.9, both plotted with the same color scale. The negative DB is expanded relative to the neutral one. (c) Cross sections of the neutral and negative orbitals along the z-axis. The energy of the neutral DB level is set at 0.35 eV above the VB edge, and the charging energy, U, is set at 0.3 eV. While this is an underestimate for the charging energy, it is used in order to be able to plot both orbitals on the same color scale.

2.4 DB Charge States

2.4.1 Charge States of the DB

As mentioned above, Dangling Bonds (DBs) introduce a single surface state whose energy level is within the bandgap. DBs are interesting entities from a number of different perspectives, many of which derive from the fact that they can be variably occupied. Being a single mid-gap energy level, they offer the possibility of being in one of three occupations: completely unoccupied, singly occupied (with either spin state), and doubly occupied (with one electron of each spin state). Pauli exclusion forbids more.

The silicon atom which hosts a DB has three of its sp^3 orbitals involved in chemical bonds, along with the associated valence electrons[†]. The remaining sp^3 orbital constitutes the DB. If the fourth valence electron remains in that orbital, then the silicon atom's nuclear charge will be compensated, and the DB will be neutral overall. If a second electron occupies the orbital, there will be a diffuse negative charge associated with the doubly occupied orbital. It is also possible to render the orbital completely unoccupied, in which case there is no electron cloud to compensate the nuclear charge so that the silicon atom behaves as a point-like positive charge, and we say that the DB is in a positively charged state. We refer to these three scenarios as the negative, neutral, and positive charge states, denoted DB^-, DB^o, and DB^+.

So far we have talked about a DB energy level. This is somewhat misleading, since in reality the DB energy level is not fixed. Firstly, it can be shifted up or down by an electrostatic potential. Secondly, it can shift depending on its own occupation: the energy for the first electron to occupy the dangling bond is lower than the energy for an additional electron to occupy it. This is primarily because when the second electron is introduced an additional energy price is paid as a result of the repulsion between the two electrons. This is sometimes called the charging energy, U.

[†]In reality, the electrons which participate in the bond become delocalized across the whole crystal, but since all electrons involved in bonding are also delocalized, the overall charge density is usually not changed (for covalent bonds).

Density Functional Theory (DFT) has been used to determine the energy levels of the DB charge states, and places the neutral state, DB^o, at 0.35 eV and the negative state, DB^-, at 0.85 eV above the Valence Band (VB)[12,33] (Figure 4.1). Other studies have calculated the energy level of the positive charge state, DB^+. We will not discuss the energy level associated with the positive charge state here, since it is not clear that it even has meaning. This issue will be discussed further in the next section.

2.4.2 Charge States vs. Energy Levels

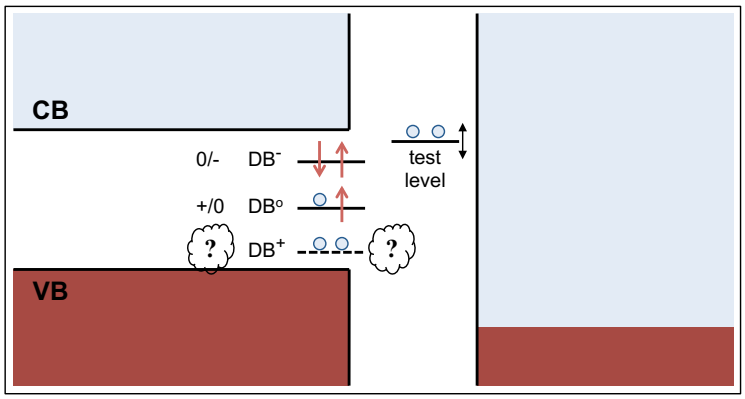
The energy level for the first and second electrons to occupy the DB are sometimes called transition levels, labeled $+/0$ and $0/-$ respectively. In this thesis, I will sometimes use the simple terms neutral level and negative level instead, as is common in the literature. The nomenclature of transition levels, however, has an advantage in that it avoids a confusion. When presented with a neutral and a negative energy level, one is tempted to ask what the energy of the positive charge state is. In my opinion, the only justified answer to this question is that there is no energy level associated with the positive charge state. The energy levels are associated with transitions between charge states, and although there are three charge states, there are only two transition levels, and therefore only two energy levels. This appears to be a point of disagreement in the community, and since this view is not universally accepted,[1,34] I will elaborate these thoughts presently.

When we talk about energy levels or draw energy level diagrams in the context of solid state physics, we are talking about electronic energy levels. We make the approximation that the Schrödinger equation for electrons can be solved with the degrees of freedom of the nuclei treated as parameters (the Born-Oppenheimer aproximation). From that point on, whether we are discussing the hydrogen atom or a many-body wavefunction, the energy levels that we calculate, measure, or discuss, are electronic energy levels. In short, there are two energy levels of the DB because there is the energy of a single electron, and then there is the energy of two electrons. The question of the *electronic* energy of *no electrons* is ill-posed.

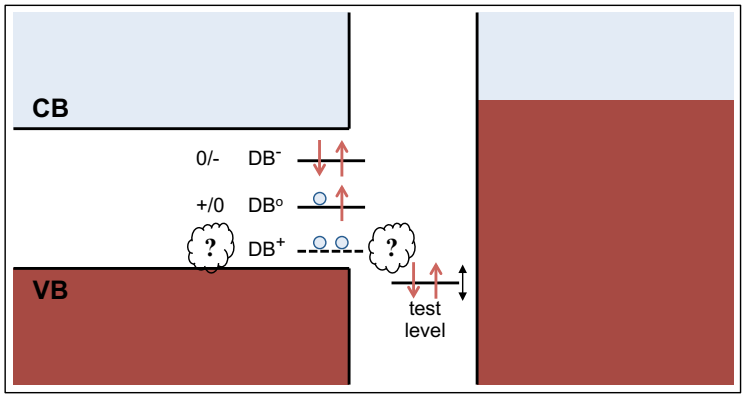
We might ask if considering the notion of holes rather than electrons gives some meaning to the notion of an energy level associated with the positive charge state. After all, when a DB is neutral it has one bound hole. When it becomes positive, it acquires an additional bound hole. Could the energy level of the positive charge state be the binding energy of the second hole? We often draw energy level diagrams for a DB as shown in Figure 2.12. Again, since we draw a singly occupied level, and a doubly occupied level, it seems natural to draw the unoccupied level. It may even seem *necessary* in order to preserve the symmetry between electrons and holes. Again, this is a fallacy. The third energy level, shown with question marks in Figure 2.12, in fact *breaks* the symmetry. This is not apparent because of the conventions of nomenclature and labelling that we use.

Consider the symmetry between electrons and holes. If the positive DB level were real, we would expect an equivalence like: "The DB^- level is to electrons as the DB^+level is to holes." We can test this with a thought experiment. Imagine a discrete, unoccupied, energy level near the surface, which we will call the "test level." We will imagine that it empties quickly into some empty reservoir, so that it always stays unoccupied. Since it is near the surface, tunneling between the DB and the test level is possible whenever a resonance condition is met. Let us further imagine that we can tune its energy. Starting from the top of the band gap, as we lower its energy, we first come into resonance with the DB^- level. Since the test level is unoccupied, an electron can be transferred from the DB to the test level. As we lower the test level further, we then come into resonance with the DB^o level. Again, since the test level is empty, the remaining electron that occupies the DB can be transferred to the test level. As we continue to lower the test level, it eventually comes into resonance with the hypothetical DB^+ level. Since both the DB level and the test level are already empty, nothing happens.

Now consider the symmetric case. We consider a test level that is filled, and connected to a filled reservoir, and we will imagine raising its energy level starting from the bottom of the bandgap. The first thing that happens is that the test level comes into resonance with the imagined DB^+ level. If the



(a)



(b)

Figure 2.12: Band diagrams depicting a thought experiment. (a) An empty test level, in contact with an empty reservoir and tunnel-coupled to a DB, is swept from an energy at the top of the band gap to an energy at the bottom, crossing the DB energy levels on its way. (b) A filled test level is swept in the opposite direction, from the bottom of the gap to the top.

situation were symmetric, then one hole would be transferred from the DB to the test level. In other words, one electron should be transferred from the tip to the DB. But we know that this cannot happen, since an energy of DB^o is required for one electron to occupy the DB. So we immediately see that the scheme that includes a DB^+ level does not exhibit any symmetry between electrons and holes. To continue the thought experiment, we will find that one hole is transferred from the DB to the test level when the test level comes into resonance with the DB^o level (equivalently, an electron is transferred *to* the DB). Subsequently, another hole is transferred from the DB to the tip when the test level comes into resonance with the DB^- level (equivalently, another electron is transferred *to* the DB).

Note that whether the test level was filled or empty, nothing happened when the test level came into resonance with the hypothetical DB^+ level. Clearly, in this picture, there is no symmetry between the DB^+ and the DB^- level.

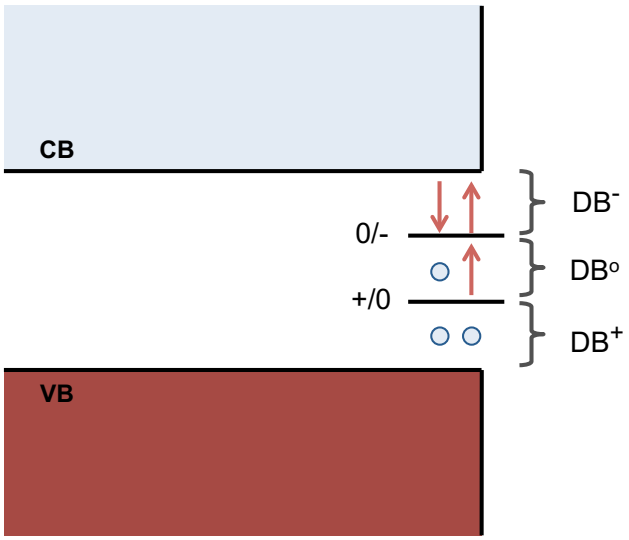


Figure 2.13: Band diagram indicating the correct way to think of the transition levels of the DB. The energies referred to as the neutral and negative energy levels are more accurately referred to as the $+/0$ and $0/-$ transition levels respectively. The position of the Fermi level with respect to these transition levels determines the charge state of the DB.

The symmetry is restored by removing the DB^+ level altogether. This is justified, since even in a thought experiment, it has no effect whatsoever on the dynamics. In fact, to the best of my knowledge, there is no conceivable experiment that could show a signature of a third energy level associated with the positive charge state. At best, it seems to be a theoretical extrapolation that does not represent a real energy level. In my opinion, the best way to think of these energy levels is as thresholds or transition levels. When the sample Fermi level is below both transition levels, the DB is positive, when it is between the two, the DB is neutral, and when it is above both, the DB is negative, as depicted in Figure 2.13 (assuming equilibrium statistics apply).

One final note might be necessary as a caution. Although we have drawn the singly occupied DB energy level and the doubly occupied level, that is, the $+/0$ and the $0/-$ transition levels, one above the other, we should not think of this as an excited state. These two levels are really a single eigenstate, which shifts depending on its occupation. If one electron already occupies the DB, then the next electron does not see an energy level at the $+/0$ transition, but instead only sees an energy level at the higher $0/-$ transition level.

2.4.3 More Transition Levels

The previous section described why one should think of two transition levels associated with a single DB, rather than three levels, associated with the three charge states. Here, I will describe a perspective that gives *six* transition levels for a single DB. While these have not been observed experimentally, this subtlety of the system may be important to understand future observations.

The arguments made so far have all relied on the Born-Oppenheimer approximation — that the lattice is effectively still on timescales associated with electronic processes. This is undoubtedly true for most electronic processes, however it will be shown later in this thesis that some electronic processes, which determine the charge state of the DB, can be extremely slow, on the order of milliseconds. This gives the lattice ample time to respond to the changing electronic configuration. This does not imply that the Born-Oppenheimer approximation is invalid; the electrons almost certainly occupy eigenstates of

the *electronic* wavefunction, with the lattice coordinates treated simply as parameters. Nonetheless, if each charge state gives rise to a unique configuration of the lattice in the vicinity of the DB, then the transition levels of the DB will be solutions of different Schrödinger equations depending on the charge state of the DB.

The result of this is that our previous picture of two transition levels, assuming an absolutely fixed lattice, becomes altered, so that the two transition levels are multiplied by the number of lattice configurations, as shown in the following diagram:

$$\begin{array}{ccc} & & (0/- \;\; ; \;\; \mathrm{Si}^{-}) \\ & & (+/0 \;\; ; \;\; \mathrm{Si}^{-}) \\ \\ (0/-) & \rightarrow & (0/- \;\; ; \;\; \mathrm{Si}^{o}) \\ (+/0) & & (+/0 \;\; ; \;\; \mathrm{Si}^{o}) \\ \\ & & (0/- \;\; ; \;\; \mathrm{Si}^{+}) \\ & & (+/0 \;\; ; \;\; \mathrm{Si}^{+}) \end{array} \quad ,$$

where $\mathrm{Si}^{+,0,-}$ labels the configuration of the silicon lattice. This alters somewhat the picture of changes of charge state of the DB. Electrons can tunnel into or out of the localized levels determined by the present configuration of the lattice, thereby changing the charge state. After this tunneling event, the lattice is temporarily in an energetically unfavourable configuration. Unless the original charge state is rapidly restored[‡], the lattice will subsequently relax to a new configuration, altering the electronic energy levels nearby, including the transition levels.

In practice, this distinction may not be significant in most experiments, but such lattice deformations, coupled to the DB charge state, are consistently reported in DFT calculations of DBs.[1,34,35] Indeed, the relaxation of the lattice after a change in charge state has been suggested as a mechanism to drastically increase the rate of inelastic electronic recombination by coupling the electronic transition to lattice phonons.[36]

[‡]Lattice relaxation times are on the order of picoseconds.

2.4.4 Excited States

Earlier, we made an important distinction between transition levels and excited states. It has been suggested that the DB may in fact have a localized excited state, separate from the previously discussed transition levels. For the delta-doped Si(111) surface, a rather unusual surface where each surface atom is capped *from below* by a Boron acceptor, there is strong evidence for a bound excited state at isolated DBs.[37] This excited state was identified by Nguyen *et al.* by a strong peak in dI/dV spectroscopy, and could be mapped in dI/dV imaging, as shown in Figure 2.14. The authors noted that the excited state was only accessible if one avoided significant (near one) occupation of the $0/-$ transition level.

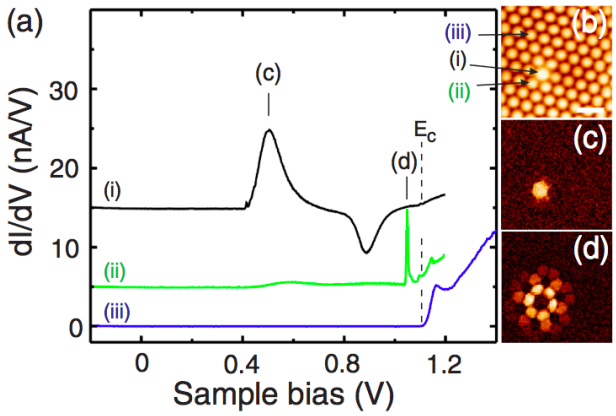


Figure 2.14: (a) dI/dV spectroscopy at three sites near a DB on the Si(111)-$(\sqrt{3} \times \sqrt{3})R30°$ surface. (b) Topographical map showing the DB and labeling the three sites for spectroscopy. (c) and (d) show dI/dV maps at bias voltages corresponding to the ground and first excited states, respectively. Reprinted figure with permission from Nguyen *et al.*, *Physical Review Letters* **105**, 226404, 2010.[37] Copyright (2010) by the American Physical Society.

In the case of the H-Si(100) surface, there is less evidence for such a localized excited state. At present, the main argument for the existence of this excited state is its inclusion in a model to describe the unusual topography

of multi-DB structures in STM.[34] The excited state is used to explain certain protrusions in topography which do not correspond to the locations of any DBs. These features in STM images of multi-DB structures will be discussed in Chapter 6.

3 Non-Equilibrium Imaging of DBs in Empty States

In this chapter I will describe STM imaging of DBs on highly doped n-type silicon surfaces, making use of the theoretical description given by Livadaru *et al.* (2011).[1]

3.1 STM of Dangling Bonds

Typical empty and filled state images of a DB are shown in Figure 3.1a and b. As a very general observation, we can say that DBs are dark in empty states and bright in filled states. This is already well understood as the result of a negative charge state for the DB. A negatively charged DB has a doubly occupied energy level, and because of Pauli exclusion, it cannot accept any more electrons from the STM tip. Since empty state imaging involves injection of electrons from tip to sample, one expects that a fully occupied sample energy level would not contribute anything to the observed tunneling current. In this sense, the DB can be expected to be invisible to the STM in empty states, and hence dark. By the same token, the doubly occupied DB should present a good source of electrons to the STM tip in filled states, where the tip extracts electrons from the surface. In this case, one expects the DB to contribute significantly to the current in STM, and should therefore be visible. Indeed, in filled states (at high biases — we will discuss other scenarios below), the DB appears as a very bright protrusion.

This first interpretation, based solely on the occupation of the DB, seems to agree well with the experiment at a glance. Clearly, however, there are

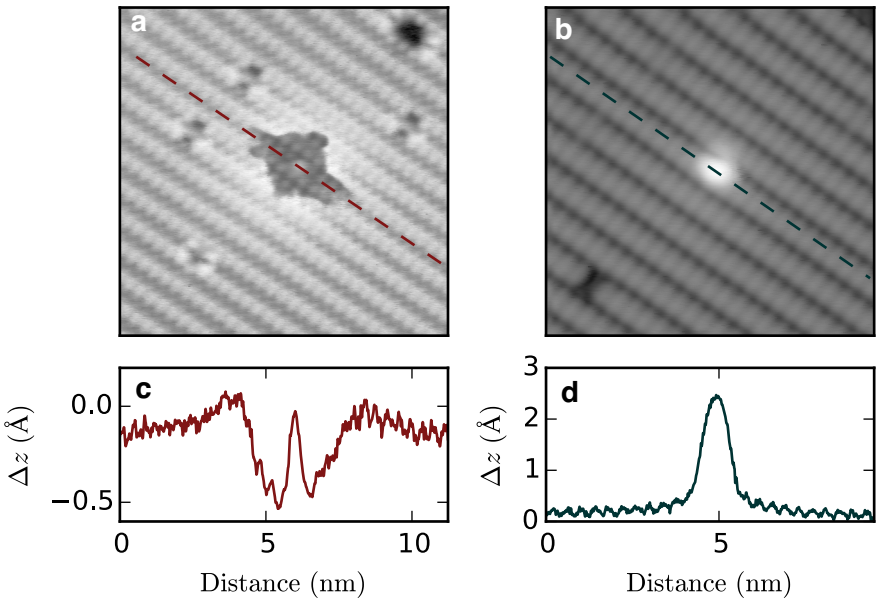


Figure 3.1: (a) 10×10 nm^2 empty state image of a DB. $I_T = 40$ pA and $V_S = 2.0$ V. (b) 8×8 nm^2 filled state image of a DB. $I_T = 40$ pA and $V_S = -2.0$ V. (c-d) Cross-sections of the DB along the lines shown in (a) and (b), respectively. Height, Δz, is given relative to the mean height in each image.

further effects at play. In particular, it is clear from Figure 3.1a that the DB has an extended effect on the silicon in its vicinity, as far as a few nanometers away. Based on the argument of occupation, one can understand why the DB orbital might not be seen in empty state imaging, but it is not clear from this idea alone that the DB should have any effect on the current from the tip to the bulk silicon levels. Evidently, in the region of the halo, not only is there no extra current through the DB, but there is a *suppression* of current from the STM tip to the extended states of the silicon sample. Indeed, these alterations of the normal tunneling current in the vicinity of the DB are seen in the cross-sections both in empty and filled states, as shown in Figure 3.1c and d. To understand the influence of the DB on the tip-sample current at some lateral separation, we need to consider the band bending that occurs because of the localized charge of the DB.

3.1.1 The Dangling Bond as a Gate

In STM of Dangling Bonds, one of the most important effects of the DB is its electrostatic effect on the surrounding silicon. The DB's charge shifts nearby energy levels in the silicon conduction band and valence band. Figures 3.2a-c show the band diagram as a function of the lateral position, x, along the surface, crossing the dangling bond for the positive, neutral, and negative charge states respectively. The completely unoccupied DB has no bound electrons to compensate the nuclear charge, so that it acts as a point-like positive charge. The resulting electrostatic effect is a Coulomb potential modified by the dielectric of silicon and screening from free carriers, bending bands downward near the DB. The singly occupied DB is to a first approximation neutral, and therefore has little or no effect on the surrounding energy levels. The doubly occupied DB has an overall negative charge coming from the excess electron density, which is spread out over the DB orbital. Far from the orbital, the negatively charged DB's electrostatic effect can again be approximated by a modified Coulomb potential, bending nearby energy levels upward.

Figures 3.2d-f show the corresponding band diagrams as a function of z, the direction normal to the surface, including the presence of an STM tip above the

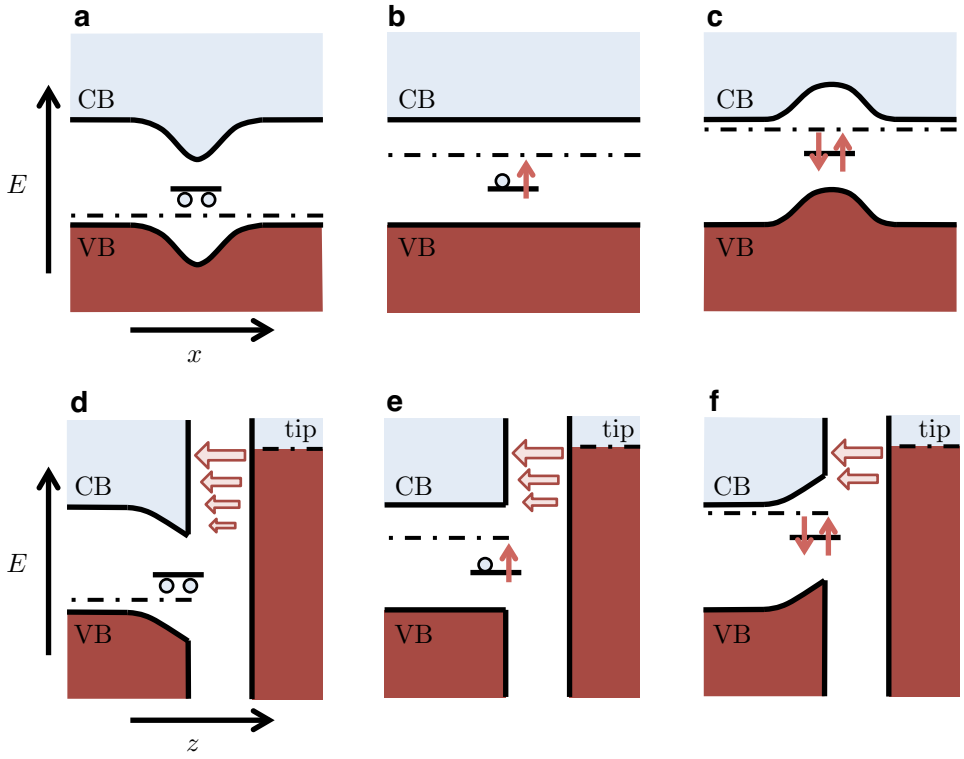


Figure 3.2: Band diagrams showing the band bending induced by a charged DB. (a-c) Band bending as a function of lateral position, crossing the DB, for the positive, neutral, and negative DB charge states respectively. (d-f) Band bending as a function of the surface-normal coordinate, z, showing the STM tip above the surface, for the positive, neutral, and negative charge states, respectively. TIBB is ignored in these diagrams.

surface. These diagrams depict the effect of the DB on the tunneling current from the tip to the sample, and ignore the tunneling current which passes through the DB itself. It is clear that for empty state imaging, the tunneling current is enhanced by a nearby positively charged DB (Figure 3.2d), since the downward band bending in the sample opens the possibility to tunnel into more states of the CB. Likewise, tunneling current is reduced near a negatively charged DB (Figure 3.2f). Because of the approximately Coulombic nature of the electrostatic effects in both cases, we expect that either the enhancement of current due to a positively charged DB, or the suppression of current in the case of the negatively charged DB, should become more pronounced closer to the DB, decaying as lateral separation increases.

These considerations provide a good explanation for the extended effect of the DB shown in Figure 3.1. The DB acts as a gate, increasing or decreasing the current between the tip and the sample, even at considerable lateral separation. A negatively charged DB explains a suppression of current near the DB in empty state imaging, as seen in the halo region. Within the halo region, also, the silicon appears to slope downward toward the DB, consistent with Coulombic band bending. Furthermore, reversing the argument for filled states, one would expect that a negatively charged DB would lead to an enhancement of current. Indeed, the cross-section shown in Figure 3.1d shows that the silicon appears to slope upward toward the DB in filled states, consistent with an enhancement of current from a localized negative charge.

3.1.2 Bias Dependent Imaging of DBs

Initial descriptions of DBs on the silicon surface used the above arguments to understand the topography of DBs in STM. The halo surrounding the DB in empty states was correctly attributed to a Coulomb-like band bending near the DB. This description is appealing in its simplicity: on the n-type samples studied, one expects negatively charged DBs, and this interpretation says that STM images are a simple map of the silicon surface with negatively charged DBs. However, carefully considering high quality images of DBs shows that there are fundamental elements missing from this simple description. The

reality is that the STM is intimately involved in the charge dynamics at the surface, and has a large, sometimes dominant, effect on the charge state of these mid-gap states.

Figure 3.3 shows empty state images of the DB at biases from 2.0 V to 1.2 V. At $V_S = 2.0$ V (Figure 3.3a) the DB halo is about 2 nm in diameter, with a mirror symmetry about an axis perpendicular to the dimer rows, and with a very sharp edge. The first problem with the description of the halo given so far, is that the sharpness of the edge of the halo is not consistent with band bending from a localized negative charge, which is expected to decay smoothly as the distance to the DB increases. Instead, the suppression of current near the DB gradually weakens as lateral separation increases, up to the edge of the halo, at which point it suddenly and drastically stops — even reverses, becoming an enhancement of current. The edge of the halo can be extremely sharp, even much less than the apparent size of an atom, particularly for high bias images.

As the bias is lowered, the halo expands continuously and its edge becomes less well defined until it extends outside of the image frame in Figure 3.3i*. The bias dependence of the size of the halo poses another problem for the simple picture of Coulombic band bending. In that picture, the halo is not expected to have an edge at all, so it clearly cannot account for this systematic change in the shape of the halo edge.

In all the images in Figure 3.3, the silicon slopes downward toward the DB in the region inside the halo. Outside of this region, there is a brightening of the silicon as the silicon appears to slope upward toward the DB, most apparent at high biases. We sometimes refer to this as a "volcano effect," since the silicon rises initially and then drops suddenly into a "crater" (the halo). This effect is also visible in the cross section in Figure 3.1c. The silicon surface initially appears to rise as one approaches the DB from either side, and then

*Note that at the lowest biases in Figure 3.3, with $V_S \lesssim 1.6$ V, there is a region within the halo, nearest to the DB, which is noticeably darker than the rest of the halo. This is likely a simple DOS effect, resulting from the tip DOS, the silicon DOS, or their convolution. As the tip approaches a negatively charged DB in empty states, the effective bias decreases because of the DB-induced band bending. The decreasing effective bias can eventually exclude sample or tip states from the tunneling process. This "halo within a halo" is not observed for all tips.

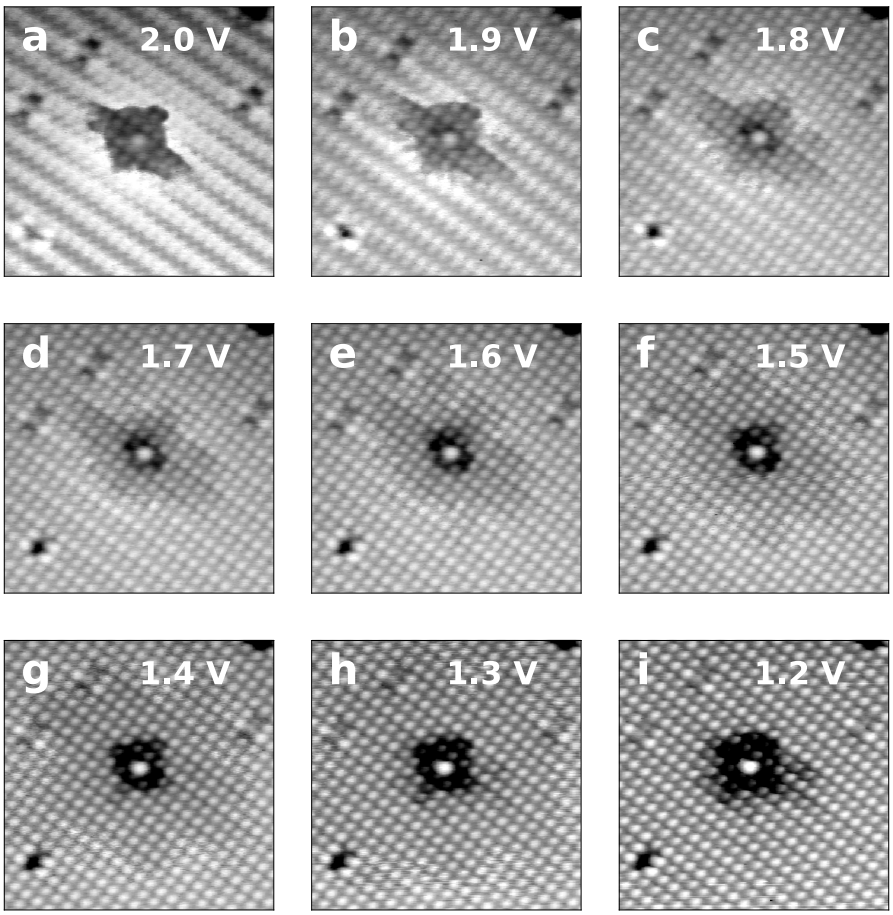


Figure 3.3: 8 nm × 8 nm images of a single DB at biases from $V_S = +2.0\ V$ to $+1.2\ V$, and tunneling current of $I_T = 80$ pA.

drops suddenly into the halo. The enhancement of current outside of the halo in empty states again poses a challenge to the simple interpretation described above, in which DBs are negative, and the STM passively images them. In fact, from the description of the DB's electrostatic effect given above, and of its resulting gating effect on the silicon around it, one can see the that enhancement of current outside of the halo is indicative of a positive charge state for the DB. A positively charged DB is surprising on an n-type sample, where the Fermi level is normally at the top of the band gap, ensuring that all gap states are fully occupied.

Finally, the bias-dependent filled state imaging of DBs poses another prob-

lem to such simplified interpretations of STM of DBs. Usually in filled states, the DB appears as a bright protrusion, with the silicon around it sloping upward, as in Figure 3.1b. In fact, at low biases, a halo can also be observed in filled states, as shown by Labidi *et al.*.[38] While this halo is not shown here, it is similar to the empty state halo, including the so-called volcano effect, with the difference that there is often no protrusion at the atomic site in the center of the halo. Such filled state halos again indicate unusual charge states for the DB, which appears negative outside the halo, but positive inside the halo (the opposite of the case we just described for empty states).

3.1.3 Effects of the Tip on the Sample

The previous section described several problems with the interpretation of STM images of DBs as simple maps of the silicon surface in the presence of a statically charged defect. One of the difficulties presented by the topography in the vicinity of a DB is that the charge state required to explain the apparent sloping of the silicon is different in different regions of the silicon. In particular, the topography appears to indicate a negative charge state inside the DB halo, as expected for n-type silicon, but it also indicates a *positive* charge state in the region outside the halo, which is surprising on an n-type sample. The fact that the charge state of the DB appears to depend sensitively on the position of the STM means that we need to consider the effect of the STM tip on the sample more closely.

The effect of the tip on the sample can be divided into two distinct components:

(1) Tip-Induced Band Bending (TIBB), and

(2) Injection (in empty state imaging) or extraction (in filled state imaging) of electrons to or from the sample.

These two effects can be considered separately, and in fact, they pull in opposite directions. TIBB, in empty state imaging, tends to bend bands upward, along with any localized energy levels, and in some cases even brings these levels above the sample Fermi level. In that sense, its tendency is to *empty*

nearby energy levels. Another way of saying this is to note that in empty state imaging, the tip bias is negative relative to the sample. This means that the tip will have an excess of electrons, relative to the sample, and these electrons will tend to be more concentrated near its apex. The field effect of this sharp, negatively charged tip at the surface tends to repel nearby electrons in the sample. This "field" description is exactly equivalent to the idea of TIBB. On the other hand, the second effect, injection of electrons from the tip (in empty state imaging), directly places electrons into the energy levels of the sample at the location of the tip. So the first effect tends to empty nearby energy levels, and the second effect tends to fill them.

Usually in STM, it is safe to describe the occupation of levels in the sample according to Fermi-Dirac statistics set by the sample Fermi level. In other words, it is safe to neglect the second effect, that of injection or extraction of electrons. This amounts to assuming that dynamics within the sample are much faster than the transfer rates between the tip and sample. When that condition is satisfied, any transfer which changes the occupation of energy levels in the sample is quickly followed by a reconfiguration within the sample to restore its equilibrium. This seems reasonable, given that the atoms of the sample are tightly (chemically) bonded to other atoms of the sample. One expects transfer of electrons to be fast within such a tight-knit lattice of atoms. On the other hand, the tip is separated by a vacuum barrier, and wavefunctions from the tip and sample usually have relatively small overlap only in the region of their exponential tails.

Localized mid-gap states are the exception to the rule. They are separated by a large energy from any nearby energy levels of the sample, and usually the only resonant states are other mid-gap states, which are located at distant defects, too far to be tunnel-coupled. This isolation, both in energy and in space, can lead to a very slow approach to equilibrium within the sample. As a result, the dynamics of electron transfer between tip and sample can out-compete the dynamics within the sample, and the occupation of such mid-gap levels is no longer dictated by equilibrium statistics. Instead, the competition between filling and emptying rates determines the occupation of

mid-gap states. The DB is a perfect example of such a mid-gap state, and exhibits a wide range of unusual behaviours as a result.

Far from the tip apex, the DB is negatively charged, since the sample is degenerately doped n-type. As the tip approaches the DB, however, the DB transition levels are shifted upward, eventually bringing the $0/-$ transition level above the sample Fermi level, and causing the equilibrium charge state of the DB to be the neutral charge state. As the tip gets closer to the DB, the $+/0$ transition level can be shifted above the sample Fermi level, causing the equilibrium charge state to be the positive charge state. This brings both transition levels above the sample Fermi level. So the DB under the influence of the tip *field* — or equivalently, under the influence of TIBB — is positively charged in equilibrium. TIBB alone does not bring the sample out of equilibrium. Rather, it shifts the energy levels of the sample, which changes their alignment with respect to the sample Fermi level. The occupation of the sample's energy levels does change as a result of TIBB, but can continue to obey the sample's equilibrium statistics.

This description of TIBB with equilibrium occupation in the sample already describes a situation in which the DB charge state can change as a function of the tip position. But the change in the DB charge state under the influence of TIBB appears to be the opposite of what is observed in the empty state image and cross-section of Figure 3.1a and c. That is, the picture of TIBB describes a transition from a negative charge state to a positive charge state as the tip approaches the DB. The observed topography, however is consistent with an abrupt transition in the opposite direction, from positive to negative. TIBB explains why the signature of a positive DB is seen on an n-type sample where the DB is expected to be negative, but understanding the halo requires the second effect described above — that of the injection of electrons from the tip to the sample, which creates the possibility for the DB to have an occupation different from what its Fermi-Dirac statistics dictate.

3.2 Occupation and Nonequilibrium Current

Deviation from equilibrium occurs when injection of electrons from the tip into the DB becomes possible. In that case, the tip begins to fill the DB at the rate Γ_{F}, while the sample attempts to restore equilibrium by emptying the DB at a rate Γ_{E}. In general, both of these rates depend upon the charge state of the DB. Obviously, Γ_{E} must be zero for a DB which is already completely unoccupied, and Γ_{F} must be zero for a DB which is doubly occupied. This leaves two filling rates and two emptying rates. For now, we will assume that there is a single filling rate which applies to the transitions $(+ \rightarrow 0)$ and $(0 \rightarrow -)$, and a single emptying rate which applies to the transitions $(- \rightarrow 0)$ and $(0 \rightarrow +)$. I will show later that experiments justify this simplification, since it turns out (surprisingly, perhaps) that the two emptying rates are similar as are the two filling rates, in the cases where we can measure them.

If the DB orbital can accommodate only *one* electron, then it is easy to show that the overall rate for electrons to traverse through the DB is

$$\Gamma_{1e^-} = \frac{\Gamma_{\mathrm{E}}\Gamma_{\mathrm{F}}}{\Gamma_{\mathrm{E}} + \Gamma_{\mathrm{F}}}, \tag{3.1}$$

and the non-equilibrium occupation of the orbital is

$$f^*_{1e^-} = \frac{\Gamma_{\mathrm{F}}}{\Gamma_{\mathrm{E}} + \Gamma_{\mathrm{F}}}. \tag{3.2}$$

The overall rate, Γ_{1e^-}, approaches the slower of the two rates when one rate dominates, and is equal to one half of either rate when the two rates are equal. The occupation increases slowly from 0 to 1 as the Γ_{F} overtakes Γ_{E}.

In reality, the orbital can accommodate two electrons, which changes the overall rate to

$$\Gamma_{2e^-} = \frac{\Gamma_{\mathrm{E}}^2\Gamma_{\mathrm{F}} + \Gamma_{\mathrm{E}}\Gamma_{\mathrm{F}}^2}{\Gamma_{\mathrm{E}}^2 + \Gamma_{\mathrm{E}}\Gamma_{\mathrm{F}} + \Gamma_{\mathrm{F}}^2}, \tag{3.3}$$

which also approaches the slower of the two rates whenever one rate dominates, but is equal to two thirds of either rate whenever the two rates are equal. Before defining an occupation, we define the probabilities for the neutral and

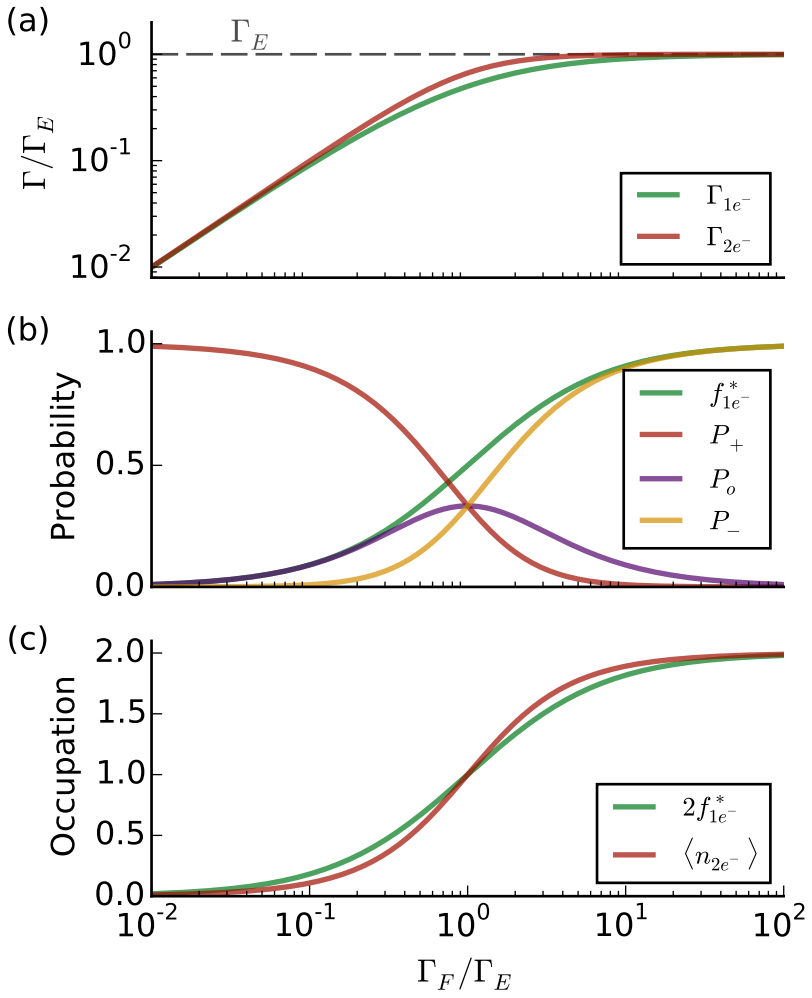


Figure 3.4: Comparison of a level which can only accommodate a single electron with a level that can accommodate two electrons. (a) Overall rates for current through a single level that can accommodate a single electron, or two electrons. (b) Non-equilibrium occupation, $f^*_{1e^-}$, of a one-electron level (which is the probability for the level to be occupied), and probabilities, P_+, P_o, and P_-, for the unoccupied, singly occupied, and doubly occupied states respectively of the two-electron level. (c) Twice the occupation of the one-electron level, and the expectation value of the occupation of the two-electron level, $\langle n_{2e^-} \rangle$.

negative states (singly and doubly occupied) as

$$P_o = \frac{\Gamma_{\mathrm{E}}\Gamma_{\mathrm{F}}}{\Gamma_{\mathrm{E}}^2 + \Gamma_{\mathrm{E}}\Gamma_{\mathrm{F}} + \Gamma_{\mathrm{F}}^2} \tag{3.4}$$

and

$$P_- = \frac{\Gamma_{\mathrm{F}}^2}{\Gamma_{\mathrm{E}}^2 + \Gamma_{\mathrm{E}}\Gamma_{\mathrm{F}} + \Gamma_{\mathrm{F}}^2}. \tag{3.5}$$

This allows us to define the expectation value of the DB's occupation as

$$\langle n_{2e^-} \rangle = P_o + 2P_-, \tag{3.6}$$

which transitions from 0 to 2 as the filling rate overtakes emptying.

Figure 3.4 compares the case where the DB can accommodate two electrons with the case where it can accommodate only one. Outside of the region where $\Gamma_{\mathrm{F}} \approx \Gamma_{\mathrm{E}}$, the two treatments are very similar. Figure3.4b shows that for $\Gamma_{\mathrm{F}} < \Gamma_{\mathrm{E}}$, the probability to be in the singly occupied state of the two-electron level, P_o, closely matches the non-equilibrium occupation of the one-electron level, $f^*_{1e^-}$. Likewise, for for $\Gamma_{\mathrm{F}} > \Gamma_{\mathrm{E}}$, the probability for the doubly occupied state closely matches $f^*_{1e^-}$. For this reason, we may use the simpler one-electron equations in thinking about non-equilibrium dynamics, as we will do in parts of this section. In particular, we will use the concept of a non-equilibrium occupation, f^*, with the understanding that twice this number (for the two electrons) is roughly the expectation value of the DB orbital's occupation, as shown in Figure 3.4c.

The illustration in Figure 3.5 shows six processes which transfer electrons between the tip, DB, and sample. The currents associated with each of these six processes are given in Table 3.1. We will use upper case letters to label currents, and lower case letters to refer to rates. For example, C_n^{STM} and $I_{\mathrm{DB-Si}}$ refer to currents, while c_n^{STM} and $i_{\mathrm{DB-Si}}$, refer to characteristic rates of electron transfer in units of Hz. One gets from the characteristic rate to the associated current by multiplying by the elementary charge, e, and in the case of currents which involve the DB one must take into account its occupation as well. For instance, the rate characterizing tunneling from the tip to the DB may be large as a result of large overlap between the DB orbital and the tip wavefunctions,

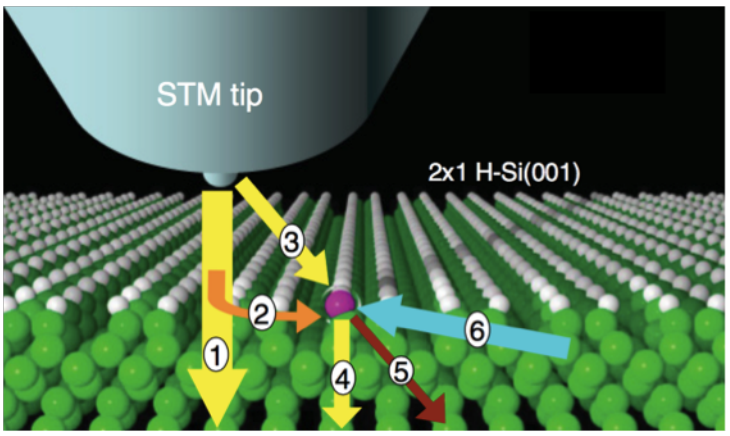


Figure 3.5: Schematic diagram illustrating the six dominant processes that contribute to the transfer of electrons between the tip, the DB, and the bulk silicon. (Figure from Livadaru *et al.* (2011).[1])

but if the DB orbital is nearly full (f^* near unity), the resulting current may still be small since the DB cannot accommodate any additional electrons.

1	$I_{\text{tip}-\text{Si}}$	$\rightarrow$	Direct tip-sample tunneling current.
2	C_n^{STM}	$\rightarrow$	Inelastic capture of electrons from the STM.
3	$I_{\text{tip}-\text{DB}}$	$\rightarrow$	Direct tip-DB tunneling current.
4	$I_{\text{DB}-\text{Si}}$	$\rightarrow$	Elastic tunneling from DB to CB.
5	E_n	$\rightarrow$	Thermal emission from DB to CB.
6	C_p	$\rightarrow$	Inelastic recombination with a hole in the VB.

Table 3.1: Currents associated with each of the six dominant processes of electron transfer illustrated in Figure 3.5.

3.3 Estimation of Rates

In this section we will briefly discuss theoretical treatments for quantitatively modelling each of the processes shown in Figure 3.5 and listed in Table 3.1. Three of these processes involve resonant tunneling: $I_{\text{tip}-\text{Si}}$, $I_{\text{tip}-\text{DB}}$, and $I_{\text{DB}-\text{Si}}$. These can be modelled following the formalism set out by Tersoff and Hamann,

as described in Section 2.1.2. What needs to be specified then, are the particular wavefunctions of the tip, sample, and DB, involved in resonant tunneling. The other three processes are inelastic. Two involve inelastic capture (C_n^{STM} and C_p), which depends on the capture cross-section of the DB wavefunction, and the other (E_n) involves thermal emission, which is of course dependent on temperature in addition to the cross-section. A more detailed treatment of each process can be found in Livadaru *et al.*.[1]

3.3.1 Tip-sample Tunneling

Following the Teroff-Hamann description of STM, the wavefunction for the tip can be expressed as

$$\psi^{\mathrm{tip}}_{E,\mathbf{K}}(\mathbf{r}) = \frac{1}{\sqrt{\Omega_{\mathrm{tip}}}} c_{0\mathrm{t}} \zeta_{\mathrm{tip}} R_0 \frac{\exp\left\{-\zeta_{\mathrm{tip}}(E)\left[|\mathbf{r}-\mathbf{r}_{\mathrm{tip}}| - R_0\right]\right\}}{\zeta_{\mathrm{tip}}(E)|\mathbf{r}-\mathbf{r}_{\mathrm{tip}}|} \tag{3.7}$$

for $|\mathbf{r}-\mathbf{r}_{\mathrm{tip}}| > R_0$, where $\mathbf{r}_{\mathrm{tip}}$ is the position of the tip center, R_0 is the tip radius, $c_{0\mathrm{t}}$ is a dimensionless constant of the order of unity, $\zeta_{\mathrm{tip}}(E) = \sqrt{2m(E_{\mathrm{vac}}-E)}/\hbar$ is a decay constant, and E_{vac} the vacuum level.

For the sample wavefunctions, we can assume a form based on a modified "jellium model",[39]

$$\psi^{\mathrm{Si}}_{E,\mathbf{K}}(\mathbf{r}) = -c_{\mathrm{Si}} \frac{k_z \exp\left\{-\sqrt{\zeta_{\mathrm{Si}}^2 + K^2}[z - H(\mathbf{R})]\right\}}{\sqrt{\zeta_{\mathrm{Si}}^2 + K^2} - ik_z} \exp(i\mathbf{K}\cdot\mathbf{R}), \tag{3.8}$$

where E is the eigenenergy of the state. $\mathbf{K} = (k_x, k_y)$ is the surface parallel wavevector and $\mathbf{R} = (x, y)$ is the surface parallel position, and k_z is the surface normal wavevector. $\zeta_{\mathrm{Si}} = \sqrt{2m(E_{\mathrm{vac}}-E)}/\hbar$ is a decay constant, $c_{\mathrm{Si}} = ic_{0\mathrm{s}}/\sqrt{\Omega_{\mathrm{Si}}}$ with $c_{0\mathrm{s}}$ a dimensionless constant of the order of unity, and Ω_{Si} is the sample volume. $H(\mathbf{R})$ is the corrugation of the sample surface. The density of states of the CB can be approximated within the effective mass approximation as

$$g^{\mathrm{CB}}_{\mathrm{Si}}(E) = \frac{8\pi\sqrt{2}}{\hbar^3} m_e^{3/2} \sqrt{E - E_{\mathrm{CB}}}, \tag{3.9}$$

where E_{CB} is the energy of the CB edge. For energies less than E_{CB} (but

greater than E_{VB}), the density of states is understood to be zero.

Since the tip-sample tunneling current is an integral over all energies from the sample Fermi level to the tip Fermi level, it is clear this current can be greatly affected by local band bending resulting from a charged DB. This band bending is reflected in a shift in the sample density of states, as the CB edge can be shifted upward or downward in energy, along with the entire sample density of states.

3.3.2 Vicinity Electron Capture

The rate of capture for a deep-level state is proportional to the velocity of the electron being captured,[40]

$$c_{\mathrm{n}} = \sigma_{\mathrm{n}} v_{\mathrm{n}} n, \tag{3.10}$$

where v_{n} is the group velocity, and σ_{n} is the capture cross-section. In equilibrium, at finite temperature, electrons sit at the bottom of the conduction band, and their energy distribution is well described by Maxwell-Boltzmann statistics, as long as $E_{\mathrm{CB}} - E_F^{\mathrm{Si}} \gg k_{\mathrm{B}}T$. This allows v_{n} to be replaced with v_{th}, the thermal velocity. The thermal velocity is the average group velocity for all electrons in the conduction band. Assuming equilibrium statistics and a parabolic conduction band minimum, the thermal velocity is given by $v_{\mathrm{th}} = \sqrt{8k_{\mathrm{B}}T/\pi m_e} \approx 10^7$ cm/s at room temperature, where m_e is the effective mass.

However, in calculating the excess capture rate, $c_{\mathrm{n}}^{\mathrm{STM}}$, due to electrons injected from the tip into the conduction band, we no longer have recourse to Maxwell-Boltzmann statistics. We therefore need to use the group velocities directly. For simplicity, we can make the assumption of an isotropic parabolic band, so that $E_{\mathbf{k}} = \hbar^2 k^2 / 2m_e$. In addition to simplifying calculations, this ensures that group velocities are parallel to their associated wavevectors.

The injected electrons can tunnel into eigenstates in the conduction band with energies below E_F^{tip}, however we expect that tunneling will occur most readily into states with high k_{z} (surface normal) and low $\mathbf{K}$ (surface-parallel) values. Thus their velocities are not evenly distributed across the range of polar angles, θ. We can solve for the angular distribution of tip-injected electrons

by considering the matrix element, $|M_{\mu\nu}(\mathbf{k})|^2$, for tunneling from tip states to sample states with wavevector, $\mathbf{k}$. The angular distribution of wavevectors, and hence also of velocities, of the injected electrons is then given by

$$D(\theta) = \mathcal{N} \int_0^{\sqrt{2m\left(\mu_{\rm tip} - E_{\rm CBM}\right)}/\hbar} k^2 \left|M_{\mu\nu}(\mathbf{k})\right|^2 dk \tag{3.11}$$

where

$$\mathcal{N} = \frac{2\pi}{\int_0^{2\pi}\int_0^{\pi/2}\int_0^{\sqrt{2m\left(\mu_{\rm tip} - E_{\rm CBM}\right)}/\hbar} k^2 \left|M_{\mu\nu}\right|^2 sin\theta dk d\theta d\phi}. \tag{3.12}$$

Ultrafast pump-probe reflectivity measurements of the Si(100) surface place the momentum relaxation time of free carriers at 32 fs.[41] Electrons traveling with speed $v_{\rm th}$ (much slower than the average velocity for injected electrons) will travel a distance of roughly 6 nm in this time. We therefore make the approximation that injected electrons retain their initial group velocities over the distance scales relevant to the present problem. The local excess electron density at each point in the silicon sample, due to the injected current from the STM tip when the DB is neutral, $I^{\rm DB^0}_{\rm tip-Si}$, is then given by

$$n^{\rm STM}(\mathbf{r}) = D(\theta)\frac{I^{\rm DB^0}_{\rm tip-Si}}{2\pi r^2 v} \tag{3.13}$$

where $\mathbf{r} = (r, \theta, \phi)$ being the distance vector from the STM tip apex, with the polar angle θ being measured from the tip axis pointing toward the silicon crystal.

Finally we account for the occupation of the DB with one electron, $1 - f^*_{\rm DB}$, and write the capture current by the DB level as

$$C^{\rm STM}_{\rm n} = (1 - f^*_{\rm DB}) \int_{\Omega_{\rm DB}} d\mathbf{r} \left|\psi_{\rm DB}(\mathbf{r})\right|^2 D(\theta) I^{\rm DB^0}_{\rm tip-Si} \frac{\sigma_{\rm n}}{2\pi r^2}, \tag{3.14}$$

where we assume that the capture cross section for an infinitesimal volume is $\left|\psi_{\rm DB}\right|^2 \sigma_{\rm n} d\mathbf{r}$. The capture rate $c^{\rm STM}_{\rm n}$ in $\rm s^{-1}$ is obtained by dividing $C^{\rm STM}_{\rm n}$ by the elementary charge and assuming a neutral DB, i.e. $f^*_{\rm DB} = 0$.

3.3.3 Tip-DB Tunneling Current

The tunneling current from the tip to the DB is calculated in the same way as the current from the tip to the sample, with the difference that there is a single DB level to tunnel into. In that case, the tip wavefunction can be modelled using the Slater-type orbital described by Equation 2.9. Direct tunneling from tip to DB is of course expected to have a very sensitive (exponential) dependence on the lateral distance from the DB to the tip apex.

Some interesting features of this process are left out of the present description. In particular, the DB orbital associated with the singly occupied DB is different from that of the doubly occupied DB, which is more spread out. Furthermore, as mentioned above, the realistic DB orbital is much more interesting and structured than our analytical treatment would suggest. This is evidenced by the striking shape of the DB orbital, as well as DFT calculations. This would add a non-trivial position dependence to the tip-DB tunneling rate.

3.3.4 DB-CB Tunneling

With sufficient upward band bending, it is possible for the energy level of the DB (in either charge state) to be raised above the normal CB edge. In that case, the DB may see resonant levels in the silicon CB some distance away, where band bending is less severe. It is then possible for electrons to tunnel through a roughly triangular barrier to these CB states. Like tip-sample tunneling and tip-DB tunneling, this resonant tunneling process can be treated using the Tersoff-Hamann approach. The conduction band wavefunctions are again described by Equation 3.8, and the DB wavefunction is again described as a Slater-type orbital, Equation 2.9.

Since TIBB is much sharper in the surface-normal direction, z, than it is in the surface-parallel direction, the shortest path for tunneling from the DB will usually be along z toward the bulk. The tunneling rate is qualitatively determined by the evanescent tail of the DB orbital through the triangular barrier formed by the tilted bandgap along this direction. The shape of this barrier can be expected to strongly depend on the applied bias, since higher

biases will result in greater band bending, and therefore a smaller triangular barrier. We therefore expect this tunneling rate to depend strongly on applied sample bias.

3.3.5 Thermal Emission of Electrons

An electron can be promoted from the DB energy level to the CB through thermal excitation. This escape rate depends on the barrier height, the cross-section for electron capture, and the temperature,

$$e_{\mathrm{n}} = \sigma_{\mathrm{n}} v_{\mathrm{n}} N_c \exp[-(E_{\mathrm{CB}} - E_{\mathrm{DB}})/k_{\mathrm{B}}T], \tag{3.15}$$

where the prefactor is an attempt frequency and N_c is the effective density of states at the bottom of the conduction band,

$$N_c = \frac{1}{\sqrt{2}} \left(\frac{m_{\mathrm{e}} k_{\mathrm{B}} T}{\pi \hbar^2} \right)^{3/2}. \tag{3.16}$$

While this process can play a significant (even dominant) role at room temperature, its temperature dependence means that it is rapidly extinguished as temperature decreases. At 4.2 K, this process is completely turned off.

3.3.6 Inelastic Recombination

A hole in the VB can also recombine with an electron localized at the DB, contributing to the DB-to-bulk current. The capture of holes is calculated analogously to the capture of electrons, according to

$$r_{\mathrm{rec}} = \sigma_{\mathrm{p}} p v_{\mathrm{p,th}}, \tag{3.17}$$

where σ_{p} is the cross-section for hole capture, p is the density of holes, and $v_{\mathrm{p,th}}$ is the thermal velocity of holes in the VB. Both p and $v_{\mathrm{p,th}}$ are functions of temperature. Additionally, p is a function of band bending, since holes become more likely to be found near the DB as bands are bent upward more. Like thermal emission, this process has a sensitive dependence on temperature, and

as a result is expected to be very small at low temperatures.

3.4 Summary

Given the rates described in the previous section, it is possible to theoretically investigate rates as a function of lateral tip position. This was done in detail for room-temperature,[1] and the results of that calculation are shown in Figure 3.6a. Many of the qualitative features discussed in the previous section (regarding dependence of rates on tip-position), are visible there. In particular, the strong dependence of i_{tip} on tip position is visible, as this rate starts out being negligible, but quickly rises as the tip approaches the DB, overtaking all other rates.

The occupation of the DB is expected to follow the form of Equation 3.2 (with the exception that the DB can accommodate two electrons, as described above). Realizing that the overall filling rate is just the sum of all filling rates, and likewise for the emptying rates, we can classify all the previously described rates into these two categories, and write the non-equilibrium occupation as

$$f^* = \frac{i_{\text{tip}} + i_{\text{Si}} + c_{\text{n}} + c_n^{\text{STM}}}{i_{\text{tip}} + i_{\text{Si}} + c_{\text{p}} + e_{\text{n}} + c_{\text{n}} + c_n^{\text{STM}}}. \tag{3.18}$$

This quantity is plotted in Figure 3.6b, which shows that as the tip approaches the DB, the filling fraction rises from 0 to 1, which in the case of an orbital which can accommodate two electrons, translates to a transition in occupation from 0 to 2 electrons. This signifies a change in the occupation of the DB from a situation of equilibrium with the sample, to a situation in which the electron dynamics at play in STM imaging have brought the DB occupation out of equilibrium.

This out-of equilibrium charge-state of the DB is what gives rise to the DB halo, as shown in Figure 3.6c. When the DB becomes negatively charged, its gating effect decreases the current from the tip directly to the extended states of the silicon CB. This creates the apparent depression around the DB, and explains the signature of a negative DB inside the halo. The calculation that generated this figure, however, only considered the transition from a neutral

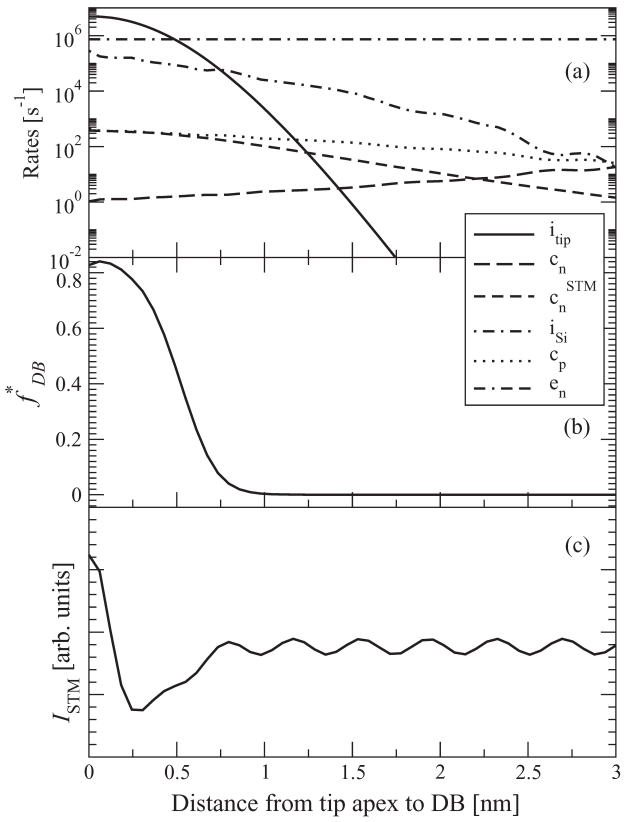


Figure 3.6: (a) Rates, (b) non-equilibrium occupation, and (c) tunneling current, all a function of lateral tip-DB separation. (Figure from Livadaru *et al.* (2011).[1])

DB to a negative DB, and therefore does not show the signature of a positive DB outside the halo region. Nonetheless, it is easy to see that by considering the positive charge state, and its associated band bending and gating, one would similarly see the signature of the positive charge state outside the halo region (that is, one would see an apparent topography that slopes upward toward the DB, before dropping rapidly into the halo).

Figure 3.7 shows an idealized topography, tracing the contour one might expect to see given perfect resolution and perfect signal-to-noise. When the tip is far from the DB, the DB is negatively charged, because of the doping of the sample. As the tip approaches, its band bending effect shifts the DB transition levels upward, eventually causing the DB to become neutral, and then negative.

Throughout this process, the DB remains in equilibrium with the rest of the sample. When the tip gets still closer, direct injection of electrons from the tip to the DB begins to dominate, and the DB is brought away from its equilibrium charge state and becomes negatively charged. This happens at the edge of the halo, which is sharp because of the sensitive dependence of i_{tip} on tip-DB separation.

There may be a very narrow region at the edge of the halo in which a neutral charge state may have a non-negligible occupation, as shown in Figure 3.4b. This is the point where filling rates roughly balance emptying rates, allowing the fraction of time spent in each of the DB charge states to be roughly comparable. However, within the assumptions used here — that there is a single filling rate and a single emptying rate to describe the system dynamics, independent of the DB charge state — the neutral state does not become dominant at the edge of the halo. Instead, it makes its appearance only transiently, as the charge state flickers unstably. When the flickering of the DB charge state is fast, it is averaged out by the measurements of current which

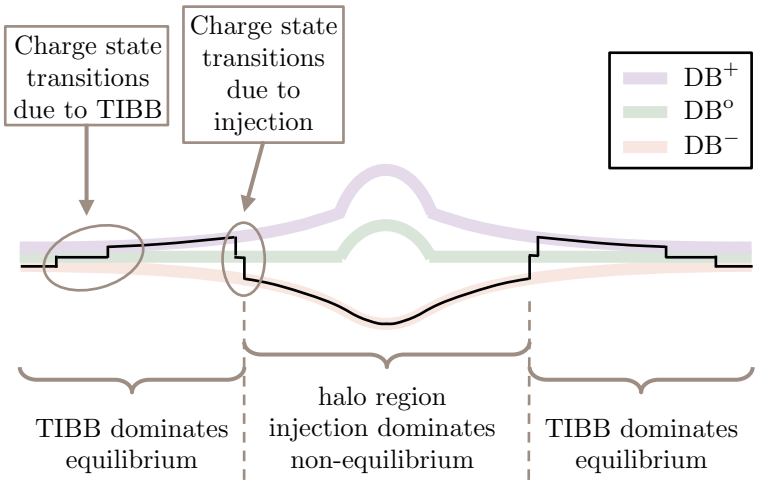


Figure 3.7: Schematic showing idealized topography across a dangling bond. The faint coloured lines represented the topography one would see if the DB could be held in each particular charge state. The observed topography reflects the dynamically changing DB charge state as the DB approaches.

occur on the relatively slow timescale of approximately one millisecond. The resulting topography drops into the halo rapidly, but smoothly.

In the next section, we will see that the temperature dependence of the rates described in this section allows us to slow the dynamics at the DB significantly by dropping the temperature to $\sim$ 4.2 K. This effectively turns off thermal emission (the dominant mechanism for emptying electrons from the DB to the bulk silicon at room temperature), and drastically reduces the density of free carriers in the CB and VB. As a result, resonant tunneling from the DB to CB levels becomes the dominant mechanism for emptying of electrons from the DB. Under these circumstances, the single electron dynamics can be slow enough to be directly observed in STM current measurements.

4 Single Electron Dynamics of DBs

This chapter describes recent experiments which analyzed the noise in tunneling current related to the single electron dynamics at the DB. These experiments provide a way of directly observing the dynamics described in the pervious chapter. This chapter is closely based on Taucer *et al.* (2014),[2] and includes several figures from that paper, as indicated in the captions.

Despite the central role of single-electron dynamics in STM imaging of DBs[1,36,37] as well as their importance in potential DB-based atom-scale devices,[42] until recently, they had not been directly observed in an STM experiment. Here, we discuss direct observation of single-electron charging dynamics of DBs. The dynamics are consistent with our model of non-equilibrium charging, in which the DB acts as the island of an SET, tunnel-coupled to the STM tip and to the silicon bulk. The variably charged DB has a gating effect on the tip-sample tunnel junction, so that the total tunneling current acts as a single-electron sensitive charge sensor. In this experiment, the DB does not act as a current-carrying state: a negligible fraction of the current passes through the DB. The tip Fermi level is at all times higher in energy than both transition levels. Thus, this experiment allows observation of the three DB charge states, whereas more familiar Scanning Tunneling Spectroscopy (STS) experiments would be fundamentally limited to observations of the two transition levels.

The experiments described in this chapter were performed using an Omicron LT-STM operated at 4.2 K. The tungsten tip was prepared by electrochemical etching followed by electron beam heating and field ion microscopy cleaning and sharpening.[43] The sample was cleaved from a 3-4 $\mathrm{m}\Omega \cdot \mathrm{cm}$ n-type As-doped Si(100) wafer, and was cleaned by heating several times to roughly

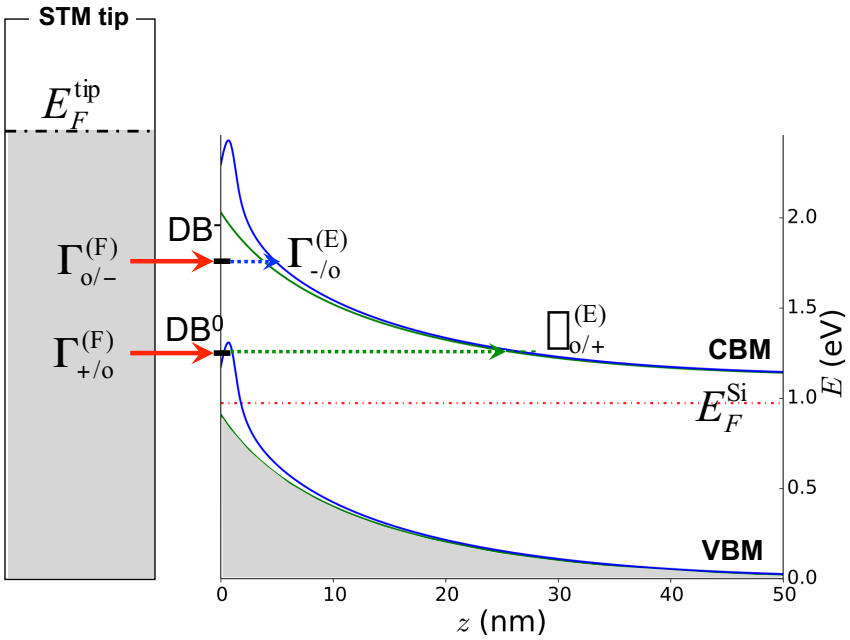


Figure 4.1: Band diagram representing the non-equilibrium dynamics of STM imaging of a DB. $E_{\mathrm{F}}^{\mathrm{tip}}$ and $E_{\mathrm{F}}^{\mathrm{Si}}$ label the tip and sample chemical potentials, respectively, and Γ labels filling and emptying processes. TIBB was calculated using the Semitip code[28] assuming typical STM experimental parameters and resulting bands are shown as green curves, corresponding to the case of DB$^{\mathrm{o}}$. For the case of DB^{-}, combined tip- and DB-induced band bending (shown as blue curves) are calculated using a Slater-type orbital for the DB[1] assuming also that dynamic screening effects in the sample are negligible. (Figure from Taucer *et al.* (2014).[2])

1250°C, and H-terminated at 330°C.[44] The high temperatures used to clean the sample are known to deplete the dopants near the surface.[45]

4.1 Observation of Single Electron Dynamics

As explained in Chapter 3, the DB halo can roughly be attributed to upward band bending near a negatively charged DB. The more detailed theory of STM of DBs, which captures the qualitative features of the topography in the vicinity of the DB, was put forth by Livadaru *et al.*.[1] In unoccupied state imaging, Tip-Induced Band Bending (TIBB) tends to empty nearby states (including

the DB). At the same time, electrons tunnel from the tip to the unoccupied sample energy levels. When an electron is injected into the DB, the dynamics which would re-establish thermal equilibrium in the sample can be relatively slow, and the equilibrium picture of STM no longer applies. According to this non-equilibrium picture of STM imaging of DBs, the charge state of the DB is determined by the competition of filling and emptying processes, as shown in Figure 4.1.

4.1.1 Current Instability at the Edge of the Halo

At low temperature, many thermal processes become negligible, and dynamics can be slow enough to be within the STM pre-amplifier bandwidth. Figure 4.2a shows a topographical unoccupied state image of a DB at 4.2 K. At these conditions, the edge of the halo is no longer sharp, but instead shows a distinctive streaky noise. When the tip is positioned in the halo region, and the tip height is held constant, the measurement of current as a function of time shows unusual jumps to discrete values, as seen in Figure 4.2c. Such current steps are absent when the tip is far from from any DBs. The histogram of current measurements shown in Figure 4.2b demonstrates that there are precisely three dominant current values. We identify these as corresponding to the negative (doubly occupied), neutral (singly occupied), and positive (unoccupied) charge states of the DB. Each charge state of the DB causes a different DB-induced band bending under the tip apex, and thereby creates a different current from tip to sample.

The electron dynamics represented in Figure 4.2b and c are for a particular tip position and voltage, but in general the dynamics and probabilities of the three charge states will depend on these parameters. Each panel in Figure 4.3 compactly shows a collection of histograms at different lateral positions crossing the edge of a DB halo, for a particular tip voltage. Colormap intensity is proportional to the number of counts at a particular current and tip-DB separation, and tip height is constant for all the data presented. At all voltages there is a periodic modulation in current as a function of position, due to the surface topography; as the tip moves laterally at constant height, the tip-

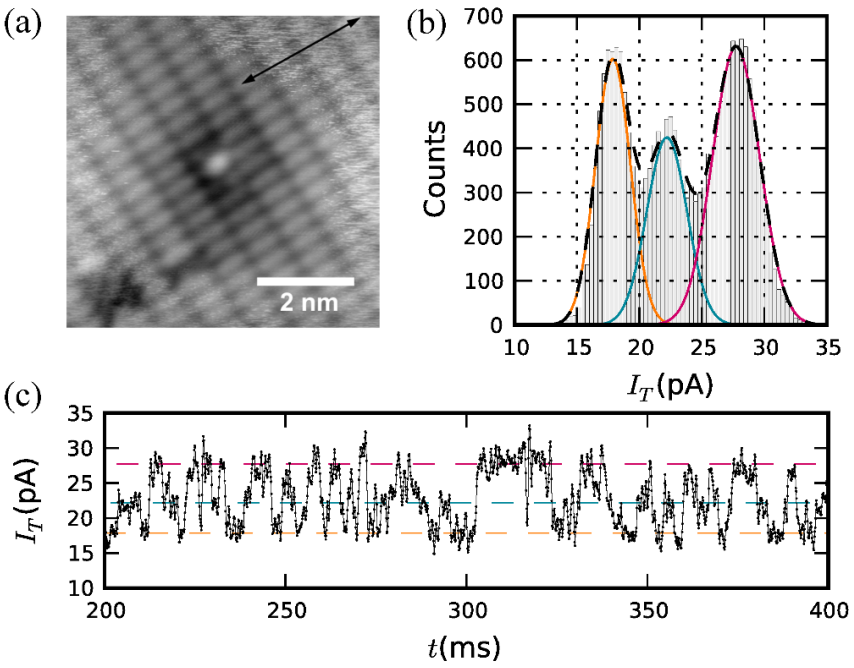


Figure 4.2: **(a)** Topographical STM image of a single DB taken with $V_S = 1.4$ V and $I_T = 20$ pA. The double-ended arrow shows the range of lateral positions used to acquire the data shown in Figure 4.3. **(b)** Histogram of current measurements with the tip at a constant height and a constant voltage of $V_S = 1.45$ V positioned 3.14nm from the DB. The peak at lowest current corresponds to the negative charge state, while the peaks at intermediate and highest current correspond to the neutral and positive charge states. **(c)** An example of a current-time trace. The sampling rate is 10 kHz and the entire trace (not shown) is 2s in length. (Figure from Taucer *et al.* (2014).[2])

sample distance is modulated because of the periodicity of the silicon surface. The striking feature is the appearance of current instability reflected by the broadening and/or existence of multiple current levels at particular voltages and positions.

At the lowest sample voltage (Figure 4.3a at 1.30 V), the DB is in a single charge state at all tip positions. The STM current decreases as the tip moves toward the DB (aside from the abovementioned periodic modulation due to topography), indicating upward band bending near the DB, consistent with

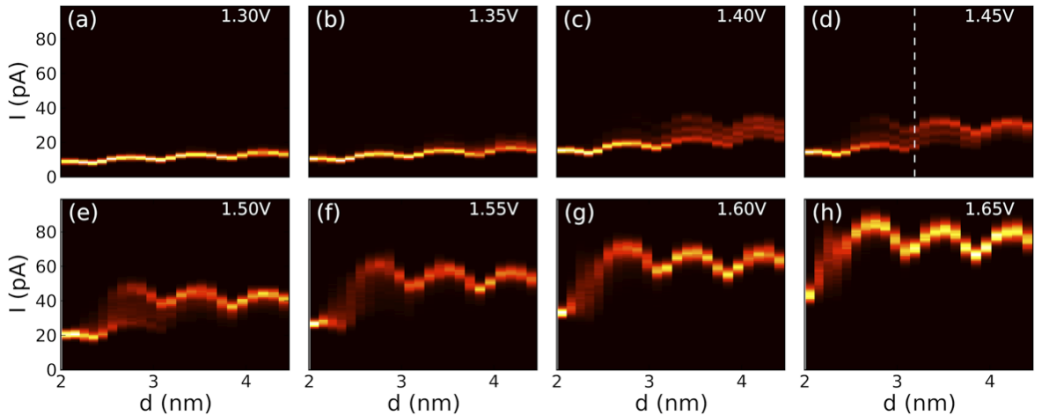


Figure 4.3: Colormaps showing frequency of current measurements as a function of tip-DB separation and current from (a) 1.30 V to (h) 1.65 V. Colormap intensity is proportional to the number of instances of a particular measurement of current at a particular position. Any vertical slice of any colormap gives a histogram whose total integral is 2 s $\times$ 10 kHz = 20000 samples. In particular, the dotted line in (d) corresponds to the data shown in Figure 4.2b. (Figure from Taucer *et al.* (2014).[2])

a negative charge state. As the tip bias is increased (Figure 4.3b-h), two additional charge states become visible, which we identify as the neutral and positive DB states. For voltages greater than 1.35 V, there is a transition region in which all three charge states are visible, with the positive DB charge state becoming dominant at larger tip-DB separations. Above 1.50 V, transitions occur on a timescale which competes with the data acquisition rate, so that the three states become blurred and eventually averaged. At 1.55 V and above, the high-current peak dominates for most tip positions, and here we see that current *increases* as the tip moves toward the DB, indicating *downward* band bending near the DB, consistent with a positive charge state. All traces show a low-current value at the smallest tip-DB separation because direct tunneling from the tip to the DB becomes dominant, in turn causing negative charging of the DB.

4.1.2 Transition Rates

Looking at Figure 4.2c we can see that the $I(t)$ traces contain dynamical information; the traces consist of plateaux of various lengths, and the dynamics can in principle be extracted by measuring their lengths as well as which states they transition to. This would give the transition rates, $\Gamma^{(\mathrm{E})}_{-/\mathrm{o}}$, $\Gamma^{(\mathrm{E})}_{\mathrm{o}/+}$, $\Gamma^{(\mathrm{F})}_{\mathrm{o}/-}$, and $\Gamma^{(\mathrm{F})}_{+/\mathrm{o}}$, for the kinetic scheme

$$\mathrm{DB}^{-} \underset{\Gamma^{(\mathrm{F})}_{\mathrm{o}/-}}{\overset{\Gamma^{(\mathrm{E})}_{-/\mathrm{o}}}{\rightleftharpoons}} \mathrm{DB}^{\mathrm{o}} \underset{\Gamma^{(\mathrm{F})}_{+/\mathrm{o}}}{\overset{\Gamma^{(\mathrm{E})}_{\mathrm{o}/+}}{\rightleftharpoons}} \mathrm{DB}^{+} \quad ,$$

which assumes that there is no direct transition between the negative state and positive state, thus neglecting any particular two-electron filling or emptying processes. The superscripts (F) and (E) indicate filling and emptying rates. Determining these rates by simply measuring the lengths of plateaux in Figure 4.2c turns out to be problematic, since the noise in the plateaux is comparable with their separation. Motivated by this, we take an approach developed by Hoffmann and Woodside[46] called signal-pair analysis, which considers the evolution of subsets of a dataset for a current-time trace chosen to initially belong to a particular charge state, fitting their evolving distributions using a dynamical model, and thereby extracting the transition rates between states even if their signals overlap significantly. This analysis combines earlier work on single-molecule fluorescence studies,[47] and a signal-pair correlation approach to analyzing structural dynamics of proteins.[46] The procedure is explained in more detail in Chapter 5.

The extracted filling and emptying rates for sample biases of 1.40 V, 1.45 V, and 1.50 V, are shown in Figure 4.4. At higher biases, dynamics could not be extracted because the transition rates were faster than the pre-amplifier bandwidth. At lower voltages, the DB tended to stay in the negative charge state at all tip positions, again making it impossible to extract the dynamics.

Figure 4.4a shows the dependence of filling rates on tip position for three different voltages. There is an exponential decay in the filling rate with increasing tip-DB separation, with values from roughly 3kHz to 50Hz, with no

clear systematic dependence of filling rates on voltage. This is consistent with the prediction that direct tunneling from tip to DB dominates the filling of the DB, assuming a constant density of tip states over the energy range of interest. The exponential fits to the filling rates of the neutral and negative charge states are shown as a black dashed-dotted line and a black solid line, respectively. Their decay rates are $k^{(\mathrm{F})}_{\mathrm{o}/-} = 1.91\ \mathrm{nm}^{-1}$ and $k^{(\mathrm{F})}_{+/\mathrm{o}} = 2.54\ \mathrm{nm}^{-1}$. We attribute the slower decay of $\Gamma^{(\mathrm{F})}_{\mathrm{o}/-}$ to the upward shift of the DB^- energy level with respect to that of DB^{o}, resulting in a smaller ionization potential, and a slower decay of the DB^- wavefunction into vacuum.

Figure 4.4b shows emptying rates. In contrast to Figure 4.4a, we see a strong voltage dependence and a very weak position dependence. The average emptying rate for each of the three voltages is shown as a horizontal dashed line. We find relatively flat emptying rates of 172 Hz, 434 Hz, and 1369 Hz for sample voltages of 1.40 V, 1.45 V, and 1.50 V respectively. While calculations for DBs at room temperature[1] found thermal emission of electrons to dominate emptying in unoccupied-state STM imaging, this process is virtually eliminated at 4.2 K. We instead consider the dominant mechanism at low temperature to be tunneling from the DB energy level to distant resonant Conduction Band (CB) levels. As the bias is increased, TIBB is also increased, while the associated barrier for an electron on the DB to tunnel to the CB becomes narrower. The weak dependence of emptying rates on lateral tip position is an indication that TIBB is relatively uniform on the scale considered here, as expected.

We can now see that the edge of the DB halo is the point at which filling rates overtake emptying rates. This corresponds to the intersection of the horizontal dashed lines (emptying) in Figure 4.4(b) with the exponential solid and dashed dotted lines (filling). As voltage is increased, the emptying rate, which is nearly flat with respect to position, increases while the filling rate remains an unchanged exponential. The point of intersection (edge of the halo) thus moves toward the DB. This is consistent with our routine observation of a DB halo size which decreases with increasing bias (not shown). Beyond the dependence of rates on bias and tip-DB lateral separation, we see that both

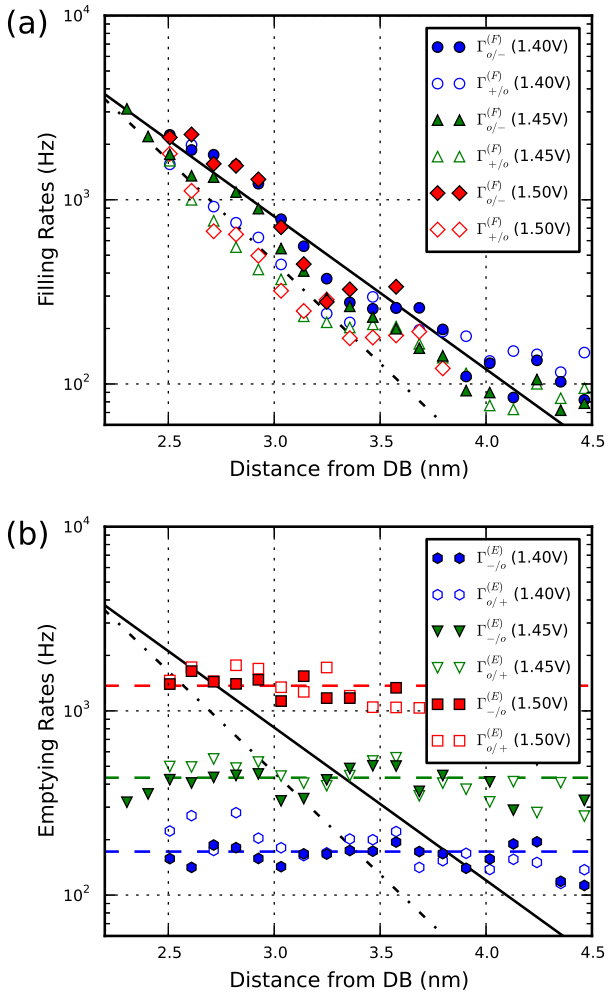


Figure 4.4: **(a)** Experimentally measured filling rates as a function of lateral tip distance from DB for three different voltages. The dashed dotted line indicates the exponential fit to the filling rate of the neutral DB energy level, while the solid line indicates the fit for the negative DB level. **(b)** Experimentally measured emptying rates as a function of lateral tip distance from DB for three different voltages. For each voltage, emptying rates have a weak dependence on tip position. The dashed coloured lines show the average emptying rate for each voltage, while the dashed-dotted and solid black lines show the same fits to the filling rates as shown in (a). (Figure from Taucer *et al.* (2014).[2])

filling rates are similar in magnitude, as are both emptying rates. At this time, a quantitative study is not permitted by the data we can acquire, but a more detailed discussion of the orbitals and band bending effects associated with DB^o and DB^- will be provided in the next section.

4.2 Further Considerations on Transition Rates

Detailed comparisons between the two filling rates and between the two emptying rates are beyond the scope of the data in this thesis. Further measurements and analysis of these rates have been undertaken by Roshan Achal. However, some noteworthy features of the data can be commented upon already. Specifically, we can address the fact that the two filling rates are of a similar magnitude, as are the two emptying rates.

4.2.1 Filling Rates

First, considering the filling rates, the DB^- orbital has a lower ionization energy than the DB^o orbital. One might at first expect that this would lead to a much larger filling rate for DB^-, since a lower ionization results in an expanded orbital, perhaps leading one to expect greater overlap between tip and DB wavefunctions. Upon careful consideration this turns out not to be the case. Since the DB is not acting as a current-carrying state, only a very small orbital overlap between the tip and DB wavefunctions is needed to get single electron charging events (around 100's of Hz). Although it is true that an orbital with a smaller binding energy will quite generally be less localized than one with a larger binding energy, it does not necessarily follow that the less localized orbital will have a greater overlap with the tip wavefunctions at a given lateral separation.

In Chapter 2, Figure 2.11 showed a heuristic depiction of the DB wavefunction, which despite its simplicity captures the essential features. Here, the DB is treated as a p-like orbital whose lobes decay according to the ionization potentials in each direction, defined by the previously described Slater-type orbital.

The lobe that extends into vacuum decays according to the ionization potential into vacuum (4.8 eV for the neutral DB and 4.5 eV for the negative DB), while the lobe that extends into the bulk decays according to the much smaller ionization potential into the bulk conduction band (0.77 eV for the neutral DB and 0.47 eV for the negative DB). Note that the difference in energy between the singly-occupied energy level and the doubly occupied energy level is underestimated for the sake of being able to plot the two orbitals with the same scale to see the qualitative features. The charging energy of $U = 0.3$ eV used in Figure 2.11 in reality is estimated closer to 0.5 eV. Still, this model is sufficient to capture the essential characteristics: the dependence of the exponential decay rates, and the relative weights of the two lobes on the DB energy in the bandgap.

Since the ionization potential into vacuum is much larger, than the ionization potential, the lobe in the silicon is much larger than the lobe in the vacuum. For mid-gap levels, the total electron density above the surface represents only about 1% of the whole orbital, or less. However, comparing the neutral orbital to the negative orbital shows a noteworthy difference. While the negative DB is indeed expanded overall as we should expect, we see that the lobe in the bulk expands at the expense of the lobe in the vacuum. The lobe in vacuum goes from making up 1.22% of the whole orbital for the neutral DB, to making up only 0.42% for the negative orbital. On careful consideration, this makes sense. The ionization energy into bulk is far more affected by the upward shift of the DB energy level than is the ionization energy into vacuum. For single electron transfer directly from tip to DB, the important factor is the degree to which the orbital extends upward into the vacuum. For the negative DB orbital, the decay of the wavefunction is slower, so that the DB orbital in some sense reaches out further, but the weight of the vacuum lobe is significantly smaller. So although the orbital is expanded overall, the overlap with the tip wavefunctions may not be increased. The main difference between the two cases should be the decay rate of overlap (and therefore of filling rates) with increasing lateral separation. This accounts for the different decay constants observed in Figure 4.4.

It is important also to note that the decay constants describing the decay of the DB wavefunction into vacuum are slower than what one would expect for the decay toward vacuum, where the barrier is roughly 4 or 5 eV. This is not altogether surprising, since the STM tip moves away from the DB *laterally*, and not in the direction of fastest decay of the wavefunction. The observed decay rate may be related to the angle of the tip at the DB position, a few nm from the tip apex. This angle would define the effective rate at which tip-DB separation increases as one increases the *lateral* separation.

There may be other possibilities to describe the observed decrease in the tip-DB tunneling rate. Our treatment of the filling processes, in the last chapter, considered direct overlap of the tip wavefunctions with the lobe of the DB orbital in vacuum, as well as the possibility of capture of hot electrons from the tip by the idealized Slater-type orbital. Capture of hot electrons was found to be negligible in comparison to direct tunneling. However, we need to keep in mind the limitations of the very idealized treatments used so far. First, the realistic orbital is very different and much more complicated than the simple p-like orbital we have been using. We know this both from DFT and from the remarkable shape of the halo which is seen with sharp tips. This may alter the capture of hot electrons, and may also affect direct tunneling in a non-trivial way. It is even possible that the portion of the DB wavefunction in the silicon (the more realistic analog of the silicon lobe) may extend to the surface, leading to direct tip-sample tunneling via the silicon "lobe" in the bulk. This can only be investigated by more accurate calculations (which need to capture the small exponential tails of wavefunctions), ideally combined with further measurements of single-electron dynamics.

4.2.2 Emptying Rates

We turn our attention now to emptying rates. Like with filling rates, there is a surprising similarity between the observed rates for emptying from the doubly and singly occupied orbitals. When the DB becomes negatively charged, suggesting at first a narrower confinement potential, the localized negative charge has a band bending effect on the nearby conduction band levels, presenting

that state with a higher and broader barrier. Figure 4.1 shows the CB edge for a neutral DB in green (including tip-induced band bending), and for a negative DB in blue (including tip-induced band bending and also DB-induced band bending). The blue curve shows that the barrier faced by the negatively charged DB is both higher in energy, and spatially wider than it would otherwise be. This increase in the height and width of the barrier is our current explanation for the comparable emptying rates.

Other mechanisms could play a role in the emptying of the DB. Recombination with holes in the valence band may play a role in emptying. Figure 4.1 also shows that TIBB and DB-induced band bending may bring the VB near or even above the sample Fermi level. This could allow holes to accumulate at the location of the DB, thus enabling recombination. While, in principle, this mechanism also contributes to the emptying rates, it is currently seen as negligible for two main reasons: (i) the space-charge layer associated with TIBB is severely depleted of mobile carriers, and (ii) the thermal velocity of holes at 4.2K is very small compared to room temperature. Lattice deformations resulting from changes in charge state, and associated phonon-coupling,[48] could also contribute to recombination dynamics, and hence to emptying rates. As described in Section 2.4.3, each change in the DB's charge state is expected to be quickly followed by a rearrangement in the lattice to accommodate the new charge. This change in the lattice configuration alters the electronic levels. It is not entirely clear how this would affect the rates of transitions between charge states, but it is very possible that this phenomenon plays some role in determining transition rates.

4.3 Relation to Room Temperature Results

Chapter 3 described the theory of STM imaging of DBs, particularly in the case of room temperature imaging. It was found that the topography in the vicinity of DBs was best explained using the concept of a non-equilibrium current through the DB. The results described in this chapter substantiate those concepts through the direct observation of the single electron dynamics

that give rise to non-equilibrium occupation of the DB. These results also provide a more complete fromework within which to understand further results at room temperature which were previously difficult to address.

Figure 4.5 shows the topography of a DB for three different sample biases for an n-type sample at 300 K. At low sample biases, as in Figure 4.5a, the dark halo which surrounds the dangling bond at +1.2 V can be understood as a natural consequence of the DB charge state changing as the tip approaches the DB, as described in Chapter 3. While the tip-DB separation remains greater than $\sim$ 1 nm, TIBB raises the doubly occupied DB level above the Fermi level, leaving the DB charge neutral. TIBB at the surface is less at 300 K than it is at 4.2 K, as a result of the enhanced screening. Since there is no DB-induced band bending for this neutral DB, the H-silicon images without topographical distortion over this region (see also the +1.2 V cross section in Fig. 4.5d). As the tip-DB separation reaches $\sim$ 1 nm, however, direct tunneling from the tip causes the DB's negative state to be filled faster than it can empty. The DB therefore takes on a negative charge state, leading to upward band bending, and the appearance of the dark halo in the vicinity of the DB.

As the bias is raised, the tip-sample separation increases to maintain the same tunneling current, decreasing tunneling from tip to DB, while at the same time increasing the field effect of the tip, and associated TIBB. As a result, the diameter of the dark halo gradually decreases (similar to Figure 3.3) until, as in Figure 4.5b, the dark halo disappears altogether. At this higher bias of + 1.8 V, the decrease in the fraction of current injected from the tip to the DB, as well as the increased emptying rate of the DB (driven by the increased TIBB at the DB), leads to a situation where the DB is on average neutral independent of the tip-DB separation. The dark halo is now completely absent (see also the + 1.8 V cross section in Figure 4.5d). While near the DB there may still be charging dynamics which are faster than the bandwidth of the preamplifier, the time-average of these shows a neutral DB.

When the bias is raised further to + 2.6 V (Figure 4.5c and the +2.6 V cross section in Figure 4.5d), a new and different halo emerges. In this case, the H-Si surface appears to slope upward as the tip approaches the DB, but abruptly

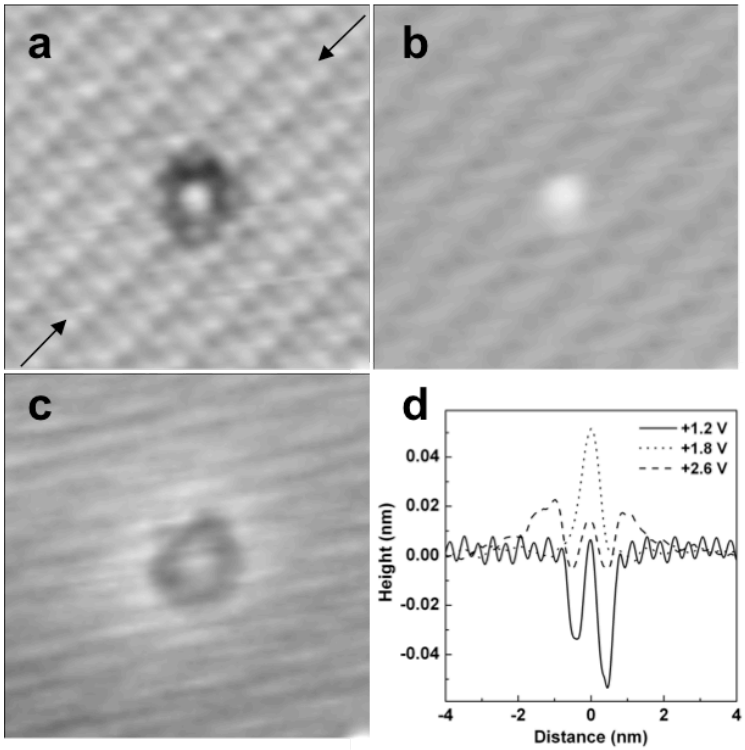


Figure 4.5: Constant-current empty-state STM imaging of DBs on n-type H:Si(100) at 300 K. (a) $V_S = +1.2$V. H-silicon within ~ 0.8nm of the DB images with depressed height. (b) $V_S = +1.8$V. DB images as a single bright protrusion. The height of the surrounding H-silicon is unperturbed. (c) $V_S = +2.6$V. DB images as a slight protrusion. H-silicon height within ~ 0.8nm of the DB is weakly perturbed. Beyond ~ 0.9nm, the H-silicon displays an abrupt increase in imaging height (~ 0.02nm) which decays with increasing distance from the DB centre. (d) Topographic cross sections (0.5 nm wide) extracted along the central H-silicon dimer row (indicated by black arrows in (a)) and across the DB centre. Tunnel current: 20 pA. Image areas: $\sim 6 \times 6\text{nm}^2$. This experimental data was acquired by Dr. Paul Piva. (Figure from Taucer *et al.* (2014).[2])

drops to a height which is comparable to the height of the unperturbed surface. We can understand this as resulting from a continuation of the trends discussed so far. The increased level of TIBB tends to empty the DB, leading to a positive

charge state when the tip is at an intermediate distance. This accounts for the brightening (i.e. increased imaging height) of the silicon in the vicinity of the DB. At even smaller tip-DB separations, direct tunneling from the tip to the DB becomes competitive and restores the DB to a neutral state on average, creating the new halo. Instead of imaging below the plane of the unperturbed H-silicon surface as in Figure 4.5a, the bottom of the halo ($\sim$ 1 nm from the DB) images with roughly the same height as the unperturbed surface far ($\sim$ 4 nm) from the DB.

4.4 Summary

In this chapter, we have shown that single-electron dynamics are directly observable in STM of single DBs when the tunnel junction between the tip and the sample acts as a single-electron sensitive charge detector. We can directly resolve the three possible charge states, negative, neutral, and positive, of the DB. The dynamics extracted from current traces are consistent with the non-equilibrium model of STM imaging of the silicon DB, in which the DB acts as an atomic quantum dot, tunnel-coupled both to the tip and to the bulk Si, with its occupation determined by the competition of filling from the tip and emptying to the bulk.

This measurement of single-electron dynamics should provide a means of quantitatively studying single-electron dynamics for DBs, multi-DB structures, and perhaps for other mid-gap defects at semiconductor surfaces. In particular, there is a great deal to be done in studying the position-dependence and bias-dependence of the filling and emptying rates of the DB. These can shed light on the shape and nature of the DB orbital, as well as its overlap with the bulk states of the CB in cases where TIBB is sufficient. Relatively straightforward extensions of this experiment may also demonstrate not only measurement, but also control of the DB charge state. Furthermore, there is no fundamental reason why the single atom charge state sensing demonstrated here cannot in future be implemented in an STM-free, lithographic structure.

5 Analysis of Random Telegraph Signals

5.1 Overview

Fluctuations in tunneling current have been observed in STM experiments in many different contexts, including charge state dynamics, as described in the previous chapter and also in studies of subsurface dopants in GaAs,[49–51] as well as in studies of the dynamics of molecules with different stable configurations,[52] and in atomic magnetic systems.[53] Undoubtedly, many other examples of such fluctuations in tunneling current exist. They are likely to occur any time random changes in the sample switch between states which have distinct conductivities, or which gate the tip-sample tunneling current, as in the case described here. Of course very similar situations exist in completely different contexts, such as in measurements of extension as a function of time with constant force applied to the ends of a strand of DNA, in optical tweezer measurements.[47] Despite the fact that what is being measured is nothing like tunneling current, the problem is identical from a data analysis perspective.

In many cases, dynamics can be extracted from time-dependent measurements by a simple thresholding analysis, in which one measures the lengths of plateaux corresponding to a given state, as well as which plateaux the system transitions to. Measuring all the plateaux lengths and the transitions in this way gives all the information about the dynamics of the system. Whenever the inherent noise in the signal associated with each state is much smaller than the separation between the signals corresponding to different states, this

method gives accurate results.

Another interesting method involves custom built analog circuitry, which takes the tunneling current as an input, and has outputs proportional to transition rates.[54] This method was applied to molecules on a surface, which switched randomly between two states.[52] This approach is attractive in that it gives a real-time measurement of dynamics, allowing one to map dynamics as one might map any other quantity in STM. Since it is essentially based on an implementation of a thresholding method, it has the same requirement of low noise compared to separation between states. In addition, it requires tuning of the threshold to the particular transition. As one imagines more states, and particularly if those states do not give a fixed position-independent signal (for example, the tunneling current associated with the negatively charged DB is not fixed, but rather becomes smaller as the tip approaches the DB), one sees that such circuits might require significant and careful tuning. More importantly, when noise competes with transitions between states, the output of such circuits would become unreliable.

In this chapter, I will describe the method that was used to analyze the telegraph data presented in the last chapter, and I will generalize it to the case of N states with arbitrary transitions. The disadvantage of this procedure is that it is somewhat complicated. It cannot be adequately described in a sentence or even a paragraph. The advantage is that it is robust and gives good results even for borderline cases, where noise nearly blurs the distinct states. An easy implementation of this technique requires a reliable algorithm. Significant progress toward this end has been made, but further automation and reliability would be needed for this technique to be more widely used.

5.1.1 Qualitative Description

Since we are thinking of STM the measured signal is current, our starting point is a measured trace of current as a function of time, $I(t)$, with a constant sampling rate, so that time can be discretized into time steps at times $t_i = i\Delta t$, where the sampling rate is $1/\Delta t$ and i is an integer. The measured currents can then be labeled I_i. Usually in STM, the noise is, roughly speaking,

randomly distributed around a mean value so that the distribution of measured currents is something like a gaussian. Each data point is randomly sampled from this distribution (assuming white noise). In this case there are essentially no dynamics of any interest.

The situation becomes interesting when the distribution of currents is no longer a single gaussian, but has multiple, say two, peaks. The overall distribution might then be the sum of two gaussian distributions. When one looks at the time trace, one sees that the currents around a particular moment are randomly distributed according to one of the gaussian distributions. After some time, the system might "click" into the other state, and suddenly currents are distributed according to the other distribution. When a long enough trace is measured, many such switching events are observed, so that the system spent part of its time in one state and part of its time in the other, and the overall distribution is the sum of the two gaussians.

If the distributions corresponding to each of the two states have no significant overlap, then given a single data point, I_i, one can say which state the system was in at time t_i. This is the case where thresholding works well. However, if there is significant overlap between the two distributions, it might not be possible, given a single data point, to say what state the system is in. For instance, if the data point happens to be in the region of overlap, then it can be equally likely to have been caused by either state of the system. This is the case where thresholding begins to be unreliable since it becomes difficult to distinguish random fluctuations within a state from random jumps between states. In principle, thresholding can be saved by, for example, using a median or mean filer to smooth the data, however this too runs into problems since smoothing runs the risk of smoothing out rapid transitions to the other state.

In the case of overlap between the two distributions, however, there is still hope of distinguishing the two states. We saw that a single point, I_i, was not enough to determine the system state with any confidence, but if we look at the data points *around* I_i, we see that they are very likely distributed according to one distribution or the other, as long as we look at a neighbourhood around t_i that is closer than the timescale associated with transitions between the two

states. So points near t_i are likely to reflect the state of the system at that time in their distribution.

Points further from t_i may have undergone a transition, and so the expected distribution of these further points will no longer reflect the distribution corresponding to the state of the system at time t_i, but will instead begin to reflect a chance for the system to be in the other state. Taking this further, if we consider a group of points *very* far from t_i, then *many* transitions will have taken place in the interim, so that it has no memory of its state at t_i. The expected distribution of these distant points will simply be the distribution for the whole time trace.

Very roughly speaking, this is the idea behind this analysis. We look at the evolution of distributions. Starting very near a data point, we might find that the distribution reflects a particular state. Some time later, the distribution no longer reflects that state, but begins to reflect different states. How far from the first data point do we need to get before we start to see that the distribution reflects states other than the initial state? Very roughly speaking, the timescale of this deviation is the time constant characterizing the transition from one state to the other. In actuality, we do not single out one data point, I_i, but instead follow this procedure for *any* data point within a chosen range $I_{\text{low}} < I_i \leq I_{\text{high}}$. We make these concepts precise in the next section.

5.2 Mathematical Description

5.2.1 Basic Concepts: Datasets, Subsets, Probabilities

We denote the whole dataset corresponding to a current trace as $\mathcal{D} \equiv \{I_i\}$, the set of all the currents, I_i. The total number of points in the data set is the cardinality of the set, written $|\mathcal{D}|$ (the cardinality is simply the number of elements in a set). Most of the analysis involved in this procedure involves considering certain subsets of the dataset, $\mathcal{S} \subseteq \mathcal{D}$. For instance, one can consider the subset $\mathcal{S} \equiv \{I_i : I_{\text{low}} < I_i \leq I_{\text{high}}\}$, that is, the set of all data points I_i such that I_i is between I_{low} and I_{high}, two arbitrarily chosen currents. One can use this to define the probability that a given condition is met for

a randomly chosen datapoint. For instance, the probability that a randomly chosen datapoint falls between the currents I_{low} and I_{high} is

$$P(I_{\text{low}} < I \leq I_{\text{high}}) = \frac{\left|\{I_i : I_{\text{low}} < I_i \leq I_{\text{high}}\}\right|}{|\mathcal{D}|}, \tag{5.1}$$

that is, the number of data points that satisfy the condition divided by the total number of data points*.

A histogram can be built by binning the data into ranges of width ΔI around values $I_n = n\Delta I$, where n labels the bin. The probability for bin n, per bin width, is given by

$$p_n = \frac{P\left(I_n - \frac{\Delta I}{2} < I \leq I_n + \frac{\Delta I}{2}\right)}{\Delta I} = \frac{\left|\left\{I_i : I_n - \frac{\Delta I}{2} < I_i \leq I_n + \frac{\Delta I}{2}\right\}\right|}{|\mathcal{D}|\ \Delta I}. \tag{5.2}$$

Imagining an idealized dataset with an infinite number of data points, we can take the limit as $\Delta I \to 0$, to arrive at the coninuous function

$$p(I) = \lim_{\Delta I \to 0} \frac{\left|\left\{I_i : I_n - \frac{\Delta I}{2} < I_i \leq I_n + \frac{\Delta I}{2}\right\}\right|}{|\mathcal{D}|\ \Delta I}, \tag{5.3}$$

which is called the Probability Density Function (PDF) and has units of inverse current. We use upper case P to denote probability, and lower case p to denote probability density. The discrete probability densities, p_n, and the continuous PDF, $p(I)$, by definition obey the following conditions

$$\sum_n \Delta I \cdot p_n = 1 \qquad \text{and} \qquad \int dI \cdot p(I) = 1\ . \tag{5.4}$$

Equation 5.3 is the PDF for the entire dataset, $\mathcal{D}$. We can define a PDF for any set at all, including a subset of the dataset, $\mathcal{S}$, with the slightly modified equation,

$$p(I; \mathcal{S}) = \lim_{\Delta I \to 0} \frac{\left|\left\{I_i \in \mathcal{S} : I_n - \frac{\Delta I}{2} < I_i \leq I_n + \frac{\Delta I}{2}\right\}\right|}{|\mathcal{S}|\ \Delta I}. \tag{5.5}$$

*Of course, the true probability is an idealization in which the total number of data points goes to infinity. In practice we have finite datasets, and any measurement is subject to error due to the finite sample size.

If the system under study only has a single static state, then, ideally, repeated measurements of the system would yield exactly the same value over and over. The PDF that describes this is a delta function. In reality, there are various sources of noise, specific to the system and the measuring apparatus, which cause a spread in the PDF around a mean value. The exact form of this spread depends on the nature of the noise, but in many cases it can be approximated by a gaussian function, and this serves as a good first guess (or simply as a way to think about the situation).

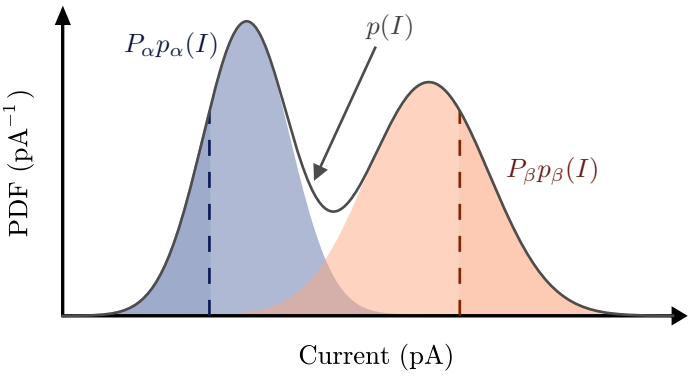


Figure 5.1: Probability density function for a two-state system, where the PDF for each state is a gaussian distribution. Data points to the left of the blue dashed line, coloured in a darker shade of blue, can be said to correspond to the system state α. Likewise for data points to the right of the brown dashed line for state β. In between, there is a region of overlap, where data points could correspond to either state.

Things become more interesting when the system can be in one of several states. If the system were locked in a particular state, α, we would expect a certain distribution of measured currents given by $p_\alpha(I)$. Locked in a different state, β, current measurements are distributed according to a different PDF, $p_\beta(I)$. These PDFs can be expressed in terms of conditional probability[†]. That

[†] A conditional probability, $P(A \mid B)$, is the probability of A given that B is the case. It can be expressed as

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}, \tag{5.6}$$

is, $p_\alpha(I) = p(I \mid \alpha)$, the probability density for measuring the current I, *given* that the system is in state α.

If the system has only these two states, then the PDF is $p(I) = P_\alpha p_\alpha(I) + P_\beta p_\beta(I)$, where P_α and P_β are the probabilities for the system to be in the states α and β, respectively. Such a PDF is depicted in Figure 5.1. Generalizing to an arbitrary number of states, the PDF is

$$p(I) = \sum_\alpha P_\alpha p_\alpha(I), \tag{5.7}$$

where the sum is over all states of the system. The integral of this PDF is equal to one, since the integral of each $p_\alpha(I)$ is one, and the sum of all the probabilities, P_α, is also one.

5.2.2 Fitting Histograms

Having measured a current trace, and looking to make sense of the data, the first thing we look at is the histogram of data, p_n. This is our experimental measure of the continuous PDF, $p(I)$. We then seek to extract from the histogram p_n some of the most important details to the analysis: the number of states, N_{states}; the PDF for each state, $p_\alpha(I)$; and the relative probabilities of each state, P_α. This is done by fitting the histogram using P_α as free parameters, as well as any free parameters introduced by the freedom in choosing the individual PDFs, $p_\alpha(I)$. Each $p_\alpha(I)$ is described by typically at least one free parameter (its position, *i.e.* mean, mode, median, or characteristic value), and perhaps others as well.

To use a concrete example, we could assume that the PDFs for the different states are gaussian functions, in which case each function $p_\alpha(I)$ has two free parameters, the mean, μ_α, and the width, σ_α. In that case, the fit of the histogram would have a total of $3N_{\text{states}}$ free parameters: P_α, μ_α, and σ_α for each state. This can quickly lead to an excess of free parameters, and when the fit becomes too "flexible," the resulting parameters can lose their meaning.

where $A \cap B$ stands for "A and B," or in terms of sets, the intersection of A and B. Equivalent expressions can be written for probability densities.

For that reason, it is important to constrain the fit to the histogram as much as possible. In the data analysis used in the last chapter, we used the fact that gaussian widths could be approximated as a linear function of their means. This eliminated several free parameters, giving more meaningful fits. This will be discussed in more detail in Section 5.3, along with a more robust and better-founded way to parametrize the PDFs.

Another difficulty in fitting histograms is determining the number of states. This can be trivial if it is a manual input, since the eye can often pick out the number of states directly. Most of the time, the number of states is simply the number of peaks. Sometimes, however, the peaks can be so close that they cannot be easily resolved, or if the probability of one state is much smaller than a neighbouring one, it may appear simply as a shoulder on a larger peak. Adding to the difficulty is the fact that it is almost always possible to improve the quality of a fit by adding another peak (thereby increasing the number of free parameters). With luck, it may be possible to fix the number of states simply on the basis of a physical argument. Otherwise, in order to automate the analysis, it needs to be done on the basis of statistical significance.

5.2.3 Evolution of Subsets

Once we have a good fit to the total PDF, $p(I)$, we know the overall probability for each state, P_α, as well as the corresponding PDF associated with each state, $p_\alpha(I)$. At this point we are ready to extract information about the dynamics between states. In this subsection, we will describe how we can track an evolving subset through the data set (this phrase will make more sense soon). The evolution of the subset should take place in a predictable way as long as transitions between the various states of the system are a Poisson process, consisting of randomly occurring transitions characterized by a single rate for each transition.

We start by choosing a subset of the data corresponding to a given range of data points. For simplicity, we consider a subset that is almost guaranteed to belong to a particular state. For example, the dark blue region of Figure 5.1 to the left of the dashed blue line is almost exclusively made up of data

points corresponding to state α, since it avoids the region of overlap. This is the distribution corresponding to the data points for which current is less than I_{blue}, indicated by the dashed blue line; that is, $\mathcal{S}(0) \equiv \{I_i : I_i \leq I_{\text{blue}}\}$. We can define a PDF for this subset, $p(I; \mathcal{S}(0))$, using Equation 5.5. This distribution is a truncated version of the total distribution shown in Figure 5.1, with the same shape as the part of the histogram shaded in slightly darker blue. It is not necessary to choose a subset that avoids the region of overlap, but it makes things a little easier to think about. We will come back to this point later.

The 0 in $\mathcal{S}(0)$ refers to a time delay of zero. We define a subset with a non-zero time delay, τ, as $\mathcal{S}(\tau) \equiv \{I(t_i + \tau) : I(t_i) \leq I_{\text{blue}}\}$, where τ must be a multiple of the sampling time Δt. Another way to say this is that $\mathcal{S}(\tau)$ is the subset of points that were measured at a time precisely τ later than the original data points that make up the set $\mathcal{S}(0)$. The PDF associated with $\mathcal{S}(\tau)$ is given by Equation 5.5, or, to make it explicit,

$$p(I; \mathcal{S}(\tau)) = \lim_{\Delta I \to 0} \frac{\left| \left\{ I_i \in \mathcal{S}(\tau) : I_n - \frac{\Delta I}{2} < I_i \leq I_n + \frac{\Delta I}{2} \right\} \right|}{|\mathcal{S}(\tau)| \ \Delta I}. \tag{5.8}$$

What relation do we expect there to be between $\mathcal{S}(\tau)$ and $\mathcal{S}(0)$? Clearly it depends on τ. If we take the smallest possible delay, $\tau = \Delta t$ — that is, the data points that occurred one time step after each of the data points for which current was less that I_{blue} — then we can be confident that the system will not have changed its state over this very small time interval. Nonetheless, assuming that the noise that gives rise to the distribution $p_\alpha(I)$ is white noise (so that each datapoint samples it independently), then $\mathcal{S}(\Delta t)$ will be randomly distributed according to the PDF for state α, which is $p_\alpha(I)$, the blue gaussian in Figure 5.1. We can state this mathematically by saying, $p(I; \mathcal{S}(\Delta t)) \approx p_\alpha(I)$.

As τ increases, there is an increasing chance, with each additional time step, that the system will have made a transition from state α to state β. Some intermediate time later, we can expect most data points to still be distributed according to $p_\alpha(I)$, but a growing number will have transitioned, and these

are distributed according to $p_\beta(I)$. The general form of the distribution is

$$p(I;\mathcal{S}(\tau)) \approx C_\alpha p_\alpha(I) + C_\beta p_\beta(I), \tag{5.9}$$

where C_α is a decreasing scalar and C_β is increasing, for small times, reflecting the transitions from α to β. Because of the similarity between Equation 5.9 and Equation 5.7, as well as the fact that the sum of C_α and C_β is always equal to one, we can think of the constants C_α and C_β as probabilities, and write them $P_\alpha(\tau)$ and $P_\beta(\tau)$. So we can think of the subset $\mathcal{S}(0)$ changing in time as probability "flows" from $P_\alpha(\tau)$ to $P_\beta(\tau)$. The rate at which $P_\beta(\tau)$ increases, for small times, is the transition rate from α to β. Things change for longer times, since it becomes possible to have transitions backward, from β to α. For longer times, or when the initial subset chosen does not correspond to a single state, or when there are more than two states involved, it is not as easy to determine the underlying transition rates from the changes in the probabilities, $P_\alpha(\tau)$. The next section will connect the changes in $P_\alpha(\tau)$ to the underlying rates.

To summarize, the distribution $p(I;\mathcal{S}(\tau))$ is specified by probabilities $P_\alpha(\tau)$, according to the equation,

$$p(I;\mathcal{S}(\tau)) = \sum_\alpha P_\alpha(\tau) p_\alpha(I), \tag{5.10}$$

since the distributions for each state, $p_\alpha(I)$, are known from the fit to the distribution of the full data set, $p(I)$. The distributions $p(I;\mathcal{S}(\tau))$ are extracted from the measured time trace, and the values of $P_\alpha(\tau)$ are found by fitting these distributions with Equation 5.10, using $P_\alpha(\tau)$ as free parameters. We then compare these measured values of $P_\alpha(\tau)$ with predictions based on calculations of dynamics (covered in the next section).

Arbitrary Initial Subsets Note that the delay-dependent probabilities extracted from the data are understood to depend on the initial subset that was singled out. In the example used previously, the initial subset, $\mathcal{S}(0)$, was chosen to be the set of data points with $I_i \leq I_{\text{blue}}$. With this subset, the initial

probabilities were approximately zero for the "pink" state and nearly one for the "blue" state. Had we chosen a different starting subset, the initial probabilities would have been different, and the time-dependent probabilities $P_\alpha(\tau)$ would also have been different. In general, the starting probabilities need not be concentrated in a single state.

For an initial subset specified by a range of currents, from I_{low} to I_{high}, we can extract the starting probability distribution straightforwardly using the equation,

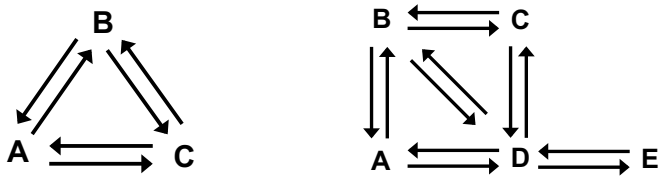
$$P_\alpha(0) = \frac{\int_{I_{\text{low}}}^{I_{\text{high}}} dI P_\alpha p_\alpha(I)}{\int_{I_{\text{low}}}^{I_{\text{high}}} dI p(I)}, \tag{5.11}$$

recalling that P_α and $p_\alpha(I)$ are known from the fit to the total distribution, $p(I)$.

5.2.4 N-state Dynamics

So far, we have seen how we can extract time-dependent (or rather delay-dependent) probabilities by considering subsets of the dataset that makes up a time trace. We now need to compare this to what we would expect for a particular "kinetic scheme," like the ones illustrated in Figure 5.2a and b for a three-state system and a five-state system respectively. That is, we will postulate a structure of connections between the various states of the system, specifying the transition rates connecting each state to each other state. Having done this, we can predict what the time dependent probabilities *should* be. If we can get this prediction to match the time-dependence extracted from the data, then we can be fairly certain that we have found the correct rates. In other words, the rates for transitions between states become free parameters used to fit the curves, $P_\alpha(\tau)$, defined in the previous section.

In general we can define a matrix of transition rates whose elements, $\Gamma_{\alpha\beta}$, represent the transition rate from state α to state β. This matrix has zeros along its diagonal, since we set $\Gamma_{\alpha\alpha} = 0$ by definition. This describes the most general kinetic scheme in which direct transitions are allowed between each state and each other state, as depicted in the three-state system shown in Figure 5.2a. More specific kinetic schemes, in which certain direct transitions



(a) (b)

Figure 5.2: Two examples of kinetic schemes for (a) a three-state system, and (b) a five-state system.

are assumed to not exist, can also be used, as depicted in Figure 5.2b. In this case, certain off-diagonal elements of the Γ matrix are assumed to be zero at the outset. Here, we will treat the most general case, in which all off diagonal elements can be non-zero.

Once we define the transition rates via the matrix Γ, we are in a position to calculate the flow of probability from state to state. In the last section, we discussed probabilities as a function of a specifically defined delay, τ. We can define, using the probabilities of each state, $P_\alpha(\tau)$, a "flux" matrix, Φ, whose elements are given by

$$\Phi_{\alpha\beta}(\tau) = P_\alpha(\tau)\Gamma_{\alpha\beta} - P_\beta(\tau)\Gamma_{\beta\alpha}. \tag{5.12}$$

Each element gives the "flow" of probability directly from state α to state β — like the highway traffic in one direction minus the highway traffic in the other. This matrix also has zeros along its diagonal, and is antisymmetric with respect to the transpose, $\Phi_{\alpha\beta} = -\Phi_{\beta\alpha}$. The flux matrix describes a directed graph, where vertices represent states and edges represent flux, as shown in figure 5.3.

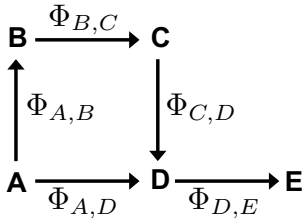


Figure 5.3: The flux matrix, Φ describes a directed graph.

In steady state, when $\tau \to \infty$, there is no change in the probability distribution in time. We describe steady state quantities with the superscript (ss). At first glance, this might seem to imply that all elements of the flux matrix must go to zero for large times — there should be no net highway traffic between different states. This condition is known as detailed balance, and is expected for a system in equilibrium. However, in this section, we hope to describe non-equilibrium as well as equilibrium dynamics, so we do not necessarily expect detailed balance to be obeyed in the steady state. Still, the requirement that probabilities are constant in steady state tells us something about the system. The flux between states needs to be balanced such that probability flux into a state from one place is balanced by flux out of the state towards others. For instance, Figure 5.2a shows a cyclic three-state kinetic scheme, where a steady state can be reached if probability flows continually in a clockwise or counter-clockwise direction, as long as the flux into each state is equal to the flux out. The true constraint turns out to be Kirchoff's law: the net flux of probability into any state is equal to the net flux out of that state. Mathematically,

$$\sum_{\beta} \Phi_{\alpha\beta}^{(\mathrm{ss})} = 0, \tag{5.13}$$

which means that the sum along any row or column of the $\Phi^{(\mathrm{ss})}$ matrix must be zero.

Both kinetic schemes in Figure 5.2 have a loop in them. It is therefore possible for both to have a steady state in which probability flows continually but in a balanced way. But whenever there is no loop, *i.e.* for *linear* kinetic schemes, then the condition of steady state implies the condition of detailed balance. All elements of Φ must tend to zero for large times. The kinetic scheme postulated in the last chapter was an example of such a linear kinetic scheme.

We should note that for N states, Equation 5.13 represents N different equations, but only $N-1$ independent equations. We can show that the last equation follows the others. That is, given $\sum_{\beta} \Phi_{\alpha\beta}^{(\mathrm{ss})} = 0$ for $\alpha \neq \gamma$, and $\Phi_{\alpha\alpha} = 0$

for all α, we can prove that $\sum_{\beta} \Phi^{(\mathrm{ss})}_{\gamma\beta} = 0$ as follows:

$$\sum_{\alpha\neq\gamma}\sum_{\beta} \Phi^{(\mathrm{ss})}_{\alpha\beta} = 0$$

$$\sum_{\beta}\sum_{\alpha\neq\gamma} \Phi^{(\mathrm{ss})}_{\alpha\beta} = 0$$

$$\sum_{\beta}\left(\sum_{\alpha} \Phi^{(\mathrm{ss})}_{\alpha\beta} - \Phi^{(\mathrm{ss})}_{\gamma\beta}\right) = 0$$

$$-\sum_{\beta}\left(\sum_{\alpha} \Phi^{(\mathrm{ss})}_{\beta\alpha} + \Phi^{(\mathrm{ss})}_{\gamma\beta}\right) = 0$$

$$\sum_{\alpha}\left(\sum_{\beta} \Phi^{(\mathrm{ss})}_{\alpha\beta} + \Phi^{(\mathrm{ss})}_{\gamma\alpha}\right) = 0$$

$$\sum_{\alpha\neq\gamma}\left(\sum_{\beta} \Phi^{(\mathrm{ss})}_{\alpha\beta} + \Phi^{(\mathrm{ss})}_{\gamma\alpha}\right) + \left(\sum_{\beta} \Phi^{(\mathrm{ss})}_{\gamma\beta} + \Phi^{(\mathrm{ss})}_{\gamma\gamma}\right) = 0.$$

The first term in the first set of brackets is zero since $\alpha \neq \gamma$, and the second term in the second set of brackets is zero by definition. So we have

$$\sum_{\alpha\neq\gamma} \Phi^{(\mathrm{ss})}_{\gamma\alpha} + \sum_{\beta} \Phi^{(\mathrm{ss})}_{\gamma\beta} = 0,$$

which implies

$$\sum_{\beta} \Phi^{(\mathrm{ss})}_{\gamma\beta} = 0.$$

This can be expressed conceptually in the following way: if Kirchoff's law is obeyed at all but one vertex of a graph, then it is necessarily obeyed at the final vertex as well.

So Equation 5.13, which describes the steady state, gives $N-1$ constraints on the fluxes $\Phi^{(\mathrm{ss})}_{\alpha\beta}$. This is important because the steady state probabilities are precisely the probabilities, P_α, which are found from the fit to the histogram for the entire data set. This means that we can use the fit to the entire dataset to constrain the possible values of the matrix elements $\Phi^{(\mathrm{ss})}_{\alpha\beta}$. But the fluxes are not the quantity of interest, and it would be more useful to directly constrain

the *rates*, $\Gamma_{\alpha\beta}$ which determine these fluxes.

To do this, we first need to choose which particular rates we wish to constrain in terms of the others. A convenient choice is to constrain the rates from one state to an adjacent one, where we label states sequentially, $\alpha = 1, 2, \cdots, N$. Here, I will choose to constrain the rates $\Gamma_{\alpha,\alpha-1}$, expressing these rates in terms of the others[‡]. In terms of the Γ matrix, this choice amounts to constraining the subdiagonal, shown in red here:

$$\Gamma = \begin{pmatrix} 0 & \Gamma_{1,2} & \Gamma_{1,3} & \Gamma_{1,4} & \cdots \\ \Gamma_{2,1} & 0 & \Gamma_{2,3} & \Gamma_{2,4} & \cdots \\ \Gamma_{3,1} & \Gamma_{3,2} & 0 & \Gamma_{3,4} & \cdots \\ \Gamma_{4,1} & \Gamma_{4,2} & \Gamma_{4,3} & 0 & \cdots \\ \vdots & \vdots & \vdots & \ddots & \ddots \end{pmatrix}. \tag{5.14}$$

In terms of transition rates, Equation 5.13 becomes

$$\sum_{\beta} \left(P_\alpha \Gamma_{\alpha\beta} - P_\beta \Gamma_{\beta\alpha} \right) = 0, \tag{5.15}$$

which can be rearranged to express the constraints as

$$\Gamma_{\alpha,\alpha-1} = \frac{P_{\alpha-1}\Gamma_{\alpha-1,\alpha} - \sum_{\beta \neq \alpha-1} \left(P_\alpha \Gamma_{\alpha\beta} - P_\beta \Gamma_{\beta\alpha} \right)}{P_\alpha}. \tag{5.16}$$

Finally, we derive the time-dependence of the probabilities by noting that the rate of change of the probability of state α is given at any time by the flux into the state minus the flux out of it. We express this more easily by defining

[‡]Note that here I am assuming non-cyclic labels. That is, I am not equating state $\alpha = N$ with a state $\alpha = 0$. This means that there is no transition rate, $\Gamma_{1,0}$ since there is no state $\alpha = 0$, so there are precisely $N - 1$ rates that can be written as $\Gamma_{\alpha,\alpha-1}$, corresponding to the $N - 1$ constraints alluded to earlier.

a matrix, $\mathbb{M}$, whose elements are

$$\begin{aligned} \mathbb{M}_{\alpha\beta} &= \Gamma_{\beta\alpha} \qquad \text{for} \quad \alpha \neq \beta \qquad (\text{off} - \text{diagonal}) \\ \mathbb{M}_{\alpha\alpha} &= -\sum_{\beta} \Gamma_{\alpha\beta} \qquad\qquad\quad (\text{diagonal}). \end{aligned} \tag{5.17}$$

Then,

$$\frac{d}{d\tau}\mathbf{P}(\tau) = \mathbb{M}\mathbf{P}(\tau), \tag{5.18}$$

where $\mathbf{P}(\tau)$ is the probability "vector," $\mathbf{P}(\tau) \equiv (P_1(\tau), P_2(\tau), \cdots, P_N(\tau))^{\mathsf{T}}$. Equation 5.18 describes a system of coupled differential equations for the evolution of the probabilities represented by $\mathbf{P}(\tau)$, which are decoupled by diagonalizing $\mathbb{M}$, yielding,

$$\mathbf{P}(\tau) = \mathbb{S}e^{\mathbb{J}\tau}\mathbb{S}^{-1}\mathbf{P}(0), \tag{5.19}$$

where $\mathbb{M} = \mathbb{S}\mathbb{J}\mathbb{S}^{-1}$. This equation is used to fit the time-dependent probabilities extracted from the data, giving the underlying rates, $\Gamma_{\alpha\beta}$.

5.2.5 Summary

The procedure for extracting characteristic rates from a time trace is as follows:

1. Generate a PDF (normalized histogram), $p(I)$, from the data in the time trace.

2. Fit this PDF using a sum of distributions, $p_\alpha(I)$, weighted by the probability P_α for each state, α.

3. Select one or more subsets of the data, corresponding to a range of currents, such as $I_{\text{low}} < I \leq I_{\text{high}}$.

4. For a given subset, the initial probabilities, $P_\alpha(0)$, are given by equation 5.11.

5. Generate a delay-dependent PDF using Equation 5.8.

6. Fit the delay-dependent PDF to extract the functions $P_\alpha(\tau)$ from the data.

7. Fit all the functions $P_\alpha(\tau)$ for a given subset using Equation 5.19, to find the rates $\Gamma_{\alpha\beta}$.

5.3 Current Noise in STM

The widths of the distributions that make up the histograms we have been discussing are determined by the character of the noise in tunneling current, which is dominated by two contributions: the intrinsic electronic pre-amplifier noise, and tip-height noise. The intrinsic noise of the STM pre-amplifier, δI_{pre}, is a constant. The noise in tip height, δz, is also a constant on a given day on a given machine, but in general depends on the mechanical coupling of the STM to its environment, and the mechanical noise in the environment (for instance, whether noisy pumps or fans are on).

The contribution of tip-height noise to the current noise, δI_z, is approximately $\delta I_z \approx \left|\frac{dI}{dz}\right| \delta z$. But since the current is an exponential function of tip height, we have $\left|\frac{dI}{dz}\right| \propto I$, which leads to the conclusion that the noise in tip height contributes a noise in current that is proportional to the mean current: $\delta I_z \propto \mu_I$. With these approximations, we are led to the conclusion that, for gaussian distributions, the width should be well approximated by a linear function of the mean.

This approach was used in the previous chapter to analyze the three-state noise due to the changing charge state of a DB. By constraining the gaussian widths, the number of free parameters in the triple gaussian fits was decreased, which allowed us to analyze a wider range of data (*e.g.* cases in which some peaks would otherwise be hard to resolve). In the data presented in the last chapter, good fits were found by constraining the widths to be the following linear function of the mean:

$$\sigma = 0.5 \text{ pA} + 0.05\mu. \tag{5.20}$$

We can go beyond the approximation of strictly gaussian distributions with a linear dependence of width on mean by considering more carefully the noise in current due to mechanical noise in tip height. We can start by assuming

that the distribution of tip heights is gaussian, so that the distribution of tip heights is given by

$$p(z) = \frac{1}{\sigma_z\sqrt{\pi}} e^{-(z-\mu_z)^2/\sigma_z^2}, \tag{5.21}$$

such that the probability of finding the tip in a range dz around z is $p(z)dz$. Thus, $p(z)$ is a PDF with units of inverse distance. Given that current is an exponential function of tip height,

$$I(z) = I_0 e^{-\kappa z}, \tag{5.22}$$

we can invert this relation to say that the tip height which gives a particular current is

$$z(I) = -\frac{1}{\kappa} \log\left(\frac{I}{I_0}\right). \tag{5.23}$$

The PDF of tip heights can be expressed in terms the current,

$$\begin{aligned} p(z(I)) &= \frac{1}{\sigma_z\sqrt{\pi}} e^{-(z(I)-\mu_z)^2/\sigma_z^2} \\ &= \frac{1}{\sigma_z\sqrt{\pi}} e^{-\left[-\frac{1}{\kappa}\log\left(\frac{I}{I_0}\right)-\mu_z\right]^2/\sigma_z^2}. \end{aligned} \tag{5.24}$$

We would like to find an expression for the distribution of currents, $p_z(I)$, with units of inverse current. We can find this expression using the relation between the distributions in z and in I,

$$p_z(I) = \frac{p(z)}{\left|\frac{dI}{dz}\right|}. \tag{5.25}$$

Noticing that the mean tip height, μ_z, can be expressed as a function of the mean current, μ_I, through Equation 5.23, we can write the distribution of currents as [§]

$$p_z(I) = \frac{\lambda}{\sigma_z I\sqrt{\pi}} \exp\left\{-\left[\frac{\log\left(\frac{I}{\mu_I}\right)}{\kappa\sigma_z}\right]^2\right\}. \tag{5.26}$$

This equation describes the expected probability density function of currents due to the uncertainty in tip height. The subscript z is there to remind us

[§] In fact, using Equation 5.23 to express μ_I as a function of μ_z is an approximation, since $I(z)$ is not linear. This is most likely a fairly benign assumption in most cases.

that this is the current noise that comes from mechanical noise (z noise).

In addition to this, there is a contribution to the noise from the electronic noise of the pre-amplifier. This noise exists even when there is no tunneling current at all, and we can express it as,

$$p_{\text{pre}}(I) = \frac{1}{\sigma_{\text{pre}}\sqrt{\pi}} e^{-I^2/\sigma_{\text{pre}}^2}, \tag{5.27}$$

where we have assumed that the noise is centred at $I = 0$. The total noise is a convolution of the noise from the preamplifier and the noise from the tip height,

$$p(I) = p_{\text{pre}}(I) * p_z(I). \tag{5.28}$$

In practice, the preamplifier noise sets a lower limit on the spread of the distribution, but when the tip is closer to the sample and the mean current is high, mechanical noise dominates, and the preamplifier noise barely has any effect. The distribution of the preamplifier noise is specified by a single parameter, σ_{pre}. The distribution of the mechanical noise, on the other hand, depends on μ_I, σ_z, and κ.

The shaded histograms in Figure 5.4 show measured distributions of current for different tip heights on the H-Si(100) surface, far from any DB's or other defects. These distributions are fit using Equation 5.28, and the fit is shown with the solid lines for each distribution. Each distribution in this figure is fit with mean as a free parameter. A single value for σ_{pre} and a single value for σ_z are used for all the peaks. The decay constant of current with increasing tip height, κ, is extracted from the corresponding $I(z)$ data, shown in the inset.

The power of Equation 5.28 is that most of the parameters can be determined beforehand and thereafter taken to be constant. These can be measured in regions where there is a single state, like the data in Figure 5.4. The $I(z)$ curve fixes the value of κ, while measured distributions with a single state can determine the values of σ_{pre} and σ_z. Once these parameters are known, the only free parameter that defines the PDF for a specific state of a multi-state system is the mean current, μ_I. This connects the current distribution to measurable quantities, whereas Equation 5.20 simply introduced additional

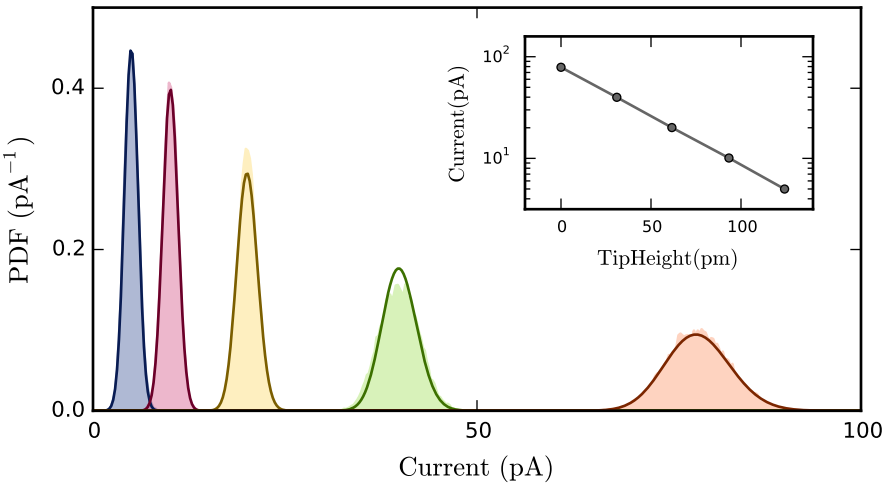


Figure 5.4: Measured distributions (PDFs) of current for five different tip heights, shown as shaded histograms. The solid lines show fits to these histograms, using Equation 5.28. The inset shows mean current as a function of relative tip height (with arbitrary zero). The fit to the inset gives $\kappa = 22.2$ nm^{-1}. The five histograms are fit using σ_z, σ_I, plus the five means of the five distributions, μ_I, as free parameters. The means are roughly 5, 10, 20, 40, and 80 pA, and the other parameters are $\sigma_z = 3.35$ pm, $\sigma_I = 3.39$ pA.

arbitrary free parameters.

5.4 Analysis of DB Charge State Dynamics

Figure 4.2c in the previous chapter showed a current trace, $I(t)$, of the tip-sample tunneling current at the edge of a DB halo. We will take this as an example of the generalized analysis method that we have described so far in this chapter.

At an appropriate tip position, it is possible to directly observe current levels corresponding to the negative, neutral, and positive charge states of the DB, and it is possible to determine the total fraction of time spent in each of these states. In other words, we can determine the probabilities for the DB to

be found in each of the three charge states, P_-, P_o, and P_+, whose sum must of course be one. We can represent these three probabilities more concisely with the probability vector, $\mathbf{P} \equiv (P_-, P_o, P_+)^\mathsf{T}$. The integral area of each of the three gaussian fits shown in Figure 4.2b of the manuscript is proportional to the probability for that charge state, which we refer to as the steady state probability, whose vector is $\mathbf{P}^{(ss)}$.

As described before, this first fitting procedure — fitting the multi-state histogram in Figure 4.2b using a sum of individual distributions — is important and involves some informed guesses at the outset. As described above, two main assumptions were made here. First, it was assumed that the distribution for each state was a gaussian. Second, in order to further confine the fit, it was assumed that the gaussian widths were linearly related to their means through Equation 5.20. These assumptions allowed good fits to nearly all the $I(t)$ traces where multiple states could be resolved. (Alternatively, we could have used a "control" dataset like the one shown in Figure 5.4 to find the characteristic parameters of the noise, and subsequently used Equation 5.28 for the distribution corresponding to each state.) Once we have fit the total histogram, we know a great deal about the system: not only do we know the steady state probabilities contained in the vector $\mathbf{P}^{(ss)}$, we also know the PDFs corresponding to each state, $p_\alpha(I)$ for $\alpha = \{-, 0, +\}$, gaussians in this case.

The histogram corresponding to the entire $I(t)$ trace, or the entire data set $\mathcal{D}$, is shown as a grey histogram in Figures 5.5a-d. This is exactly the histogram shown in Figure 4.2b. In Figure 5.5a, the orange part of the histogram shows the distribution corresponding to the subset of data points for which current is less than 19 pA, that is $\mathcal{S}(0) \equiv \{I(t_i) : I(t_i) \leq 19 \text{ pA}\}$. This is one particular "initial subset" which we single out. We can then consider the distribution of points which occur exactly a time τ later than the original subset, that is $\mathcal{S}(\tau) \equiv \{I(t_i + \tau) : I(t_i) \leq 19 \text{ pA}\}$. The orange histograms in Figures 5.5b-d show the distributions $p(I; \mathcal{S}(\tau))$ for τ equal to 0.5 ms, 2.0 ms, and 8.0 ms respectively[¶]. As τ increases, we see that the distribution first spreads to take

[¶] While the PDFs are normalized so that their integral is one, the plotted histograms in Figure 5.5 are not. They can be taken to be $|\mathcal{D}| \cdot p(I)$ for the grey histograms and $|\mathcal{S}| \cdot p(I; \mathcal{S})$ for the orange histograms. This reflects the "fraction" of $\mathcal{D}$ that is made up of $\mathcal{S}$.

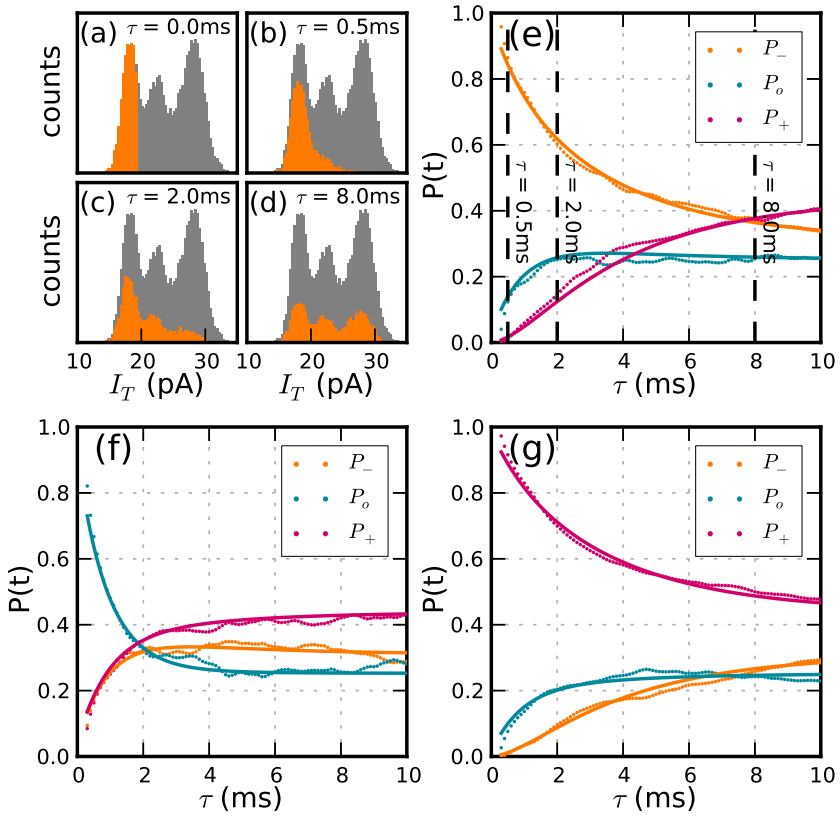


Figure 5.5: **(a-d)** The grey histogram is the histogram corresponding to the data set, $\mathcal{D}$, shown in Figures 4.2b and c. The superimposed orange histogram in (a) shows the subset, $\mathcal{S}(\tau = 0)$, chosen such that $I(t) \leq 19.0$ pA. The superimposed histograms in (b-d) show the subsets $\mathcal{S}(\tau = 0.5$ ms$)$, $\mathcal{S}(\tau = 2.0$ ms$)$, and $\mathcal{S}(\tau = 8.0$ ms$)$ respectively. **(e)** The three components of $\mathbf{P}(\tau)$ are plotted as a function of τ for the initial subset shown in (a), that is $\mathcal{S}(0) \equiv \{I(t) : I(t) \leq 19.0$ pA$\}$. The vertical dashed lines show the components of $\mathbf{P}$ corresponding to the subsets shown in (b-d). **(f)** Likewise, the three components of $\mathbf{P}(\tau)$ for a different initial subset $\mathcal{S}(0) \equiv \{I(t) : 20.6$ pA $\leq I(t) \leq 23.8$ pA$\}$. **(g)** The three components of $\mathbf{P}(\tau)$ for $\mathcal{S}(0) \equiv \{I(t) : 25.4$ pA $\leq I(t)\}$. (Figure from Taucer *et al.* (2014).[2])

roughly the shape of the negative charge state gaussian, and subsequently the amplitude of that peak decreases as the other two increase. At 8.0ms the

orange distribution, $\mathcal{S}(\tau = 8.0 \text{ ms})$ is approaching the steady state, which is to say that it becomes a scaled down version of the grey one, $\mathcal{D}$.

The probability vector, $\mathbf{P}(\tau)$, corresponding to each subset, $\mathcal{S}(\tau)$, is determined by the constrained triple gaussian fit described above, with $P_\alpha(\tau)$ as free parameters. Such a triple gaussian fit needs to be performed for each value of τ, giving the three components of $\mathbf{P}$ at the delay τ. This procedure generates the data points plotted in Figure 5.5e. Note that the subset, $\mathcal{S}(0)$, shown in Figure 5.5a was chosen so that $P_-(0) \approx 1$, so the evolution of the probability vector for this subset starts with $P_-(\tau)$ near one, and the other two near zero. The plot again shows the negative charge state probability dropping as the other two probabilities increase, and all three tending towards their steady state values. Figures 5.5f and g show the evolution of $\mathbf{P}(\tau)$ for subsets corresponding to the neutral and positive charge states, respectively. The solid lines in Figures 5.5e-g come from another fitting procedure, which I explain next.

The theoretical prediction for $\mathbf{P}(\tau)$ is based on a set of coupled differential equations, corresponding to the kinetic scheme

$$\mathrm{DB}^- \quad \underset{\Gamma^{(\mathrm{F})}_{\mathrm{o}/-}}{\overset{\Gamma^{(\mathrm{E})}_{-/\mathrm{o}}}{\rightleftharpoons}} \quad \mathrm{DB}^{\mathrm{o}} \quad \underset{\Gamma^{(\mathrm{F})}_{+/\mathrm{o}}}{\overset{\Gamma^{(\mathrm{E})}_{\mathrm{o}/+}}{\rightleftharpoons}} \quad \mathrm{DB}^+ \quad , \tag{5.29}$$

which can be concisely expressed as

$$\tfrac{d}{d\tau}\mathbf{P}(\tau) = \mathbb{M}\mathbf{P}(\tau) \quad ; \tag{5.30}$$

$$\mathbb{M} \equiv \begin{pmatrix} -\Gamma^{(\mathrm{E})}_{-/\mathrm{o}} & \Gamma^{(\mathrm{F})}_{\mathrm{o}/-} & 0 \\ \Gamma^{(\mathrm{E})}_{-/\mathrm{o}} & -\Gamma^{(\mathrm{E})}_{\mathrm{o}/+} - \Gamma^{(\mathrm{F})}_{\mathrm{o}/-} & \Gamma^{(\mathrm{F})}_{+/\mathrm{o}} \\ 0 & \Gamma^{(\mathrm{E})}_{\mathrm{o}/+} & -\Gamma^{(\mathrm{F})}_{+/\mathrm{o}} \end{pmatrix}.$$

Note that we have made the assumption here that direct transitions, $- \to +$ or $+ \to -$, can be neglected. This sets the corresponding matrix elements, $\Gamma_{-/+}$ and $\Gamma_{+/-}$, to zero.

Before fitting, we can apply one further constraint on the rates, based on the steady state probability vector, $\mathbf{P}^{(\mathrm{ss})}$. It follows from Equation 5.30 that the filling and emptying rates are related to the steady state probabilities through the two relations,

$$\frac{\Gamma^{(\mathrm{E})}_{-/\mathrm{o}}}{\Gamma^{(\mathrm{F})}_{\mathrm{o}/-}} = \frac{P^{(\mathrm{ss})}_{\mathrm{o}}}{P^{(\mathrm{ss})}_{-}} \qquad \text{and} \qquad \frac{\Gamma^{(\mathrm{E})}_{\mathrm{o}/+}}{\Gamma^{(\mathrm{F})}_{+/\mathrm{o}}} = \frac{P^{(\mathrm{ss})}_{+}}{P^{(\mathrm{ss})}_{\mathrm{o}}}, \tag{5.31}$$

which is the analog of Equation 5.16 for the kinetic scheme discussed here.

The set of coupled differential equations represented by Equation 5.30 is uncoupled by diagonalizing the matrix $\mathbb{M}$. Defining a unitary matrix, $\mathbb{S}$, we find a diagonal $\mathbb{J}$, such that $\mathbb{M} = \mathbb{S}\mathbb{J}\mathbb{S}^{-1}$. The time evolution of the probability vector is then given precisely by Equation 5.19. Any matrix of the form of $\mathbb{M}$ has at least one eigenvalue equal to zero with the other two less than or equal to zero. This means that each component of $\mathbf{P}(\tau)$ is comprised of a constant term (the steady state probability for that charge state) plus two decaying exponentials. The solid curves in Figures 5.5e-g are fits to the data using Equation 5.19, constrained by the relations 5.31. All nine curves are fit using only two free parameters, $\Gamma^{(\mathrm{F})}_{\mathrm{o}/-}$ and $\Gamma^{(\mathrm{F})}_{+/0}$. Thus these fits, along with the constraints of the steady state, give the four rates for the transitions between states.

This analysis was repeated for all current traces where multiple charge states could be resolved, leading to the rates plotted in Figure 4.4.

6 Dangling Bond Fabrication

Over distances of several nanometers, DBs can exhibit strong electron-electron interactions via Coulombic repulsion or attraction. In addition to this, more closely spaced DBs can exhibit tunnel-coupling, which permits the transfer of electrons directly from one DB to another. These two effects have been used to show that small DB structures can be polarized by external biases leading to localization of an electron on asymmetric sites of otherwise symmetric structures.[33] If DBs are brought closer together still, something similar to a chemical bond forms between them, as bonding and anti-bonding orbitals are formed. Clearly, any number of complex structures, comprising many DBs, are possible. Tight-binding Hamiltonians like the ones described in Chapter 1 can be embodied by tailored DB devices. Work in this area is still in its early stages, but engineering of energy levels and wavefunctions has been reported by several groups.[34,55–57]

This ability to control wavefunctions and elctron-electron interactions has already been widely explored and developed in the context of quantum dots, with applications from novel transport devices, to quantum computing. Single electrons can be trapped and manipulated in single- or multiple-quantum dot structures which allows control over occupation down to single electrons,[58,59] single electron charge detection,[60,61] and coherent control of both spatial wavefunctions[62,63] and spin states.[64–66] Schemes for employing quantum dot systems have been developed to the level of architectures for both classical[67,68] and quantum[69] information applications. A drawback of most quantum dot systems is the need for cryogenic temperatures, a consequence of the relatively small charging energies of the quantum dots. As quantum dots are miniaturized, charging energies are increased. Ultimately miniaturized quantum dots

are embodied in atomic impurities and atom-scale "defects", which are giving birth to a new arena for technological progress.

Recent work on embedded impurities in Si has delivered impressive single electron devices, demonstrating a single atom Single Electron Transistor (SET),[70] coherent spin control,[71] and optical addressing of single atoms.[72,73] Embedded phosphorus atoms can be embedded in silicon with nanometer precision by delta-doping. However, the inability to control the placement of impurities on a truly atomic scale is a fundamental limitation for some applications.[74] By contrast, Dangling Bonds (DBs) on the silicon surface can be fabricated with atomic precision,[33,55] making them an attractive candidate for atomic quantum dots. The commonalities between these two areas lead to a description of DBs as Atomic Silicon Quantum Dots (ASiQDs). There is already a body of theoretical work exploring the possibility of using DBs as building blocks for transport and logic devices.[12,13,42,75] The potential of using DBs to create functional device elements is only just being explored and understood,[34,55,76] and likewise fabrication is now being optimized and commercialized.[57,77,78]

This chapter will provide a brief description of some of progress in DB fabrication. The development of atomic silicon quantum dot technology will require reliable fabrication of DB structures, large and small, with nearly perfect atomic precision. At present, we can make small structures of a few atoms perfectly, and larger structures can be made with nearly atomic precision. We start this section by discussing image analysis, the process by which the periodicities of the surface can be extracted automatically from STM images. We follow this with a description of the process of hydrogen desorption, and finally, we discuss some of the patterns that have been created.

6.1 Image Analysis

For DB fabrication, the most important piece of information that can be extracted from an STM image is the periodicity and alignment of the surface. Although the periodicity of the H-Si(100) 2×1 surface is known exactly, errors

in the acquired data inevitably occur and need to be compensated for. Calibration of the piezoelectric scanners, for instance, is not necessarily constant, and needs to be continually re-measured and re-calibrated in order to ensure a good match to the lattice. Furthermore the angle of the surface unit cell with respect to the scanner coordinates is slightly different for each sample that is diced and loaded into the microscope. This chapter starts by considering the properties of continuous and discrete Fourier Transforms (FTs) in one and two dimensions, with an aim to automatically fitting the peaks corresponding to surface periodicities.

6.1.1 Continuous Fourier Transforms

The FT of an STM image contains the information about the periodicity of the surface. This section describes methods of extracting that information by analysis of STM images. Before discussing two dimensional Fourier transforms, we will briefly review the continuous and discrete Fourier transform in one dimension.[79]

We will use the bra and ket notation of quantum mechanics to describe functions in position and in frequency space. For a basis, we can use the set of eigenstates of position, $\{|x\rangle\}$. From these, we can construct the basis for the Fourier transform, $\{|k\rangle\}$, defined as oscillatory functions (plane waves in higher dimensions),

$$\langle x|k\rangle = \frac{1}{\sqrt{2\pi}} e^{ikx}. \tag{6.1}$$

Both basis sets obey the usual equations of orthonormality, $\langle x_i|x_j\rangle = \delta(x_i - x_j)$ and $\langle k_i|k_j\rangle = \delta(k_i - k_j)$.

We can express a function in terms of its Fourier components,

$$\begin{aligned} f(x) &= \langle x|f\rangle = \langle x| \left[\int dk|k\rangle\langle k|\right] |f\rangle = \int dk\langle x|k\rangle\langle k|f\rangle \\ &= \frac{1}{\sqrt{2\pi}} \int dk \widetilde{f}(k) e^{ikx}, \end{aligned} \tag{6.2}$$

or we can express the Fourier components in terms of the function,

$$\begin{aligned}\widetilde{f}(k) &= \langle k|f\rangle = \langle k|\left[\int dx|x\rangle\langle x|\right]|f\rangle = \int dx\langle k|x\rangle\langle x|f\rangle \\ &= \frac{1}{\sqrt{2\pi}}\int dx f(x)e^{-ikx}. \end{aligned} \tag{6.3}$$

Equations 6.2 and 6.3 are the inverse Fourier transform, and the Fourier transform, respectively. We denote the Fourier transform of a function, $f(x)$, as $\mathcal{F}[f(x)]$ or equivalently, $\widetilde{f}(k)$.

Note that the Fourier transform of the sum of two functions is equal to the sum of the Fourier transforms:

$$\begin{aligned}\mathcal{F}\left[f_1(x)+f_2(x)\right] &= \frac{1}{\sqrt{2\pi}}\int dx\left[f_1(x)+f_2(x)\right]e^{-ikx} \\ &= \frac{1}{\sqrt{2\pi}}\int dx f_1(x)e^{-ikx} + \frac{1}{\sqrt{2\pi}}\int dx f_2(x)e^{-ikx} \\ &= \mathcal{F}\left[f_1(x)\right] + \mathcal{F}\left[f_1(x)\right]. \end{aligned} \tag{6.4}$$

The Fourier transform of a delta function is trivial:

$$\mathcal{F}\left[\delta(x-x_0)\right] = \frac{1}{\sqrt{2\pi}}\int dx\delta(x-x_0)e^{-ikx} = \frac{1}{\sqrt{2\pi}}e^{-ikx_0}, \tag{6.5}$$

whose real and imaginary parts are even and odd, respectively:

$$\begin{aligned}\mathrm{Re}\left\{\mathcal{F}\left[\delta(x-x_0)\right]\right\} &= \frac{1}{\sqrt{2\pi}}\cos(kx_0) \\ \mathrm{Im}\left\{\mathcal{F}\left[\delta(x-x_0)\right]\right\} &= \frac{1}{\sqrt{2\pi}}\sin(kx_0)\ . \end{aligned} \tag{6.6}$$

Since any real function can be constructed as an infinite sum of delta functions, it follows that the Fourier transform of a real function will always have an even real part and an odd imaginary part. The STM images we intend to analyze, of course, contain *real* data only, so this will also apply to the FTs of STM images.

We will now briefly consider the one-dimensional equivalent of the image analysis required to extract the periodicity from an STM image. We consider

the Fourier transform of a periodic function,

$$f(x) = A\cos\left[k_0(x - x_0)\right], \tag{6.7}$$

which has an amplitude of A, a periodicity described by the wavevector k_0, and a maximum at x_0. How does the information contained in the simple function $f(x)$ show up in the Fourier transform? The Fourier transform can be worked out, and it gives

$$\widetilde{f}(k) = \frac{A\sqrt{2\pi}}{2}\left[e^{-ik_0x_0}\delta(k - k_0) + e^{ik_0x_0}\delta(k + k_0)\right]. \tag{6.8}$$

The Fourier transform exhibits two peaks at $k = \pm k_0$, with phases of $\mp k_0x_0$, and with equal amplitudes of $A\sqrt{\pi/2}$*. The periodicity of the function is encoded in the position of the two corresponding peaks; the amplitude of the periodic function is encoded in the amplitude of the peaks; and the position of the maximum is given by the phases of the peaks. Incidentally, all the information about the function is contained in only *one* of the peaks. This is the result of the fact, shown above, that the Fourier transform of any real function has an even real part and an odd imaginary part — the positive side of a function (where $x \geq 0$) dictates the negative side (where $x \leq 0$).

6.1.2 Discrete Fourier Transforms

In experimental settings, we deal with finite data sets, and as such, we need to consider the discrete Fourier transform, rather than the continuous one outlined above. This means that we have a situation where space is divided into discrete units of size Δ located at $x_n = n\Delta$; $i = 0\ , 1\ , \ldots\ , N - 1$. As before, a discrete function can be expressed in terms of its Fourier components,

$$f(x_n) = \frac{1}{\sqrt{N}}\sum_{j=0}^{N-1}\widetilde{f}(k_j)e^{ik_jx_n} \quad ; \quad x_n = n\Delta, \tag{6.9}$$

*Strictly speaking we should say that the *integral* of each peak is $A/2$, since the amplitude — in the sense of height — is infinite.

and the Fourier components can be expressed as

$$\widetilde{f}(k_j) = \frac{1}{\sqrt{N}} \sum_{n=0}^{N-1} f(x_n) e^{-ik_j x_n} \quad ; \quad k_j = \frac{2\pi j}{N\Delta}. \tag{6.10}$$

The discrete Fourier transform of a discretized periodic function is more involved than the Fourier transform of a continuous function. If the periodic function is

$$f(x_n) = Ae^{ikx_n}, \tag{6.11}$$

where an arbitrary phase can be included through A, then the Fourier transform is

$$\widetilde{f}(k_j) = \frac{A}{\sqrt{N}} \sum_{n} e^{i2\pi \frac{(z-j)}{N} n}, \tag{6.12}$$

where z is defined such that

$$k = \frac{2\pi z}{N\Delta}. \tag{6.13}$$

We can think of k as the "generating" wavevector, since it generates the periodicity in $f(x)$. z can be thought of as the value of k in units of $2\pi/N\Delta$, the spacing between the discretized wavevectors. When z happens to be an integer $0 \leq z \leq N-1$, Equation 6.12 is trivially evaluated and gives

$$\widetilde{f}(k_j) = \left\{ \begin{array}{ccc} \sqrt{N}A & \text{if} & j = z \\ \\ 0 & \text{if} & j \neq z \end{array} \right\}. \tag{6.14}$$

When z is not an integer, however, the we need to make an approximation.

Each sequential term in the sum in Equation 6.12 takes a step in phase of size $2\pi(z-j)/N$. When these steps are small we can replace the sum with an integral. More precisely, we can change to an integral whenever $(z-j) \ll N$. This approximation is therefore valid for values of k_j near k — that is, it is valid for Fourier components near the generating wavevector. In practice this means that it is valid near the peak in the Fourier transform. Making this

approximation, we see that

$$\sum_n e^{i2\pi\frac{(z-j)}{N}n} \approx \int_0^N e^{i2\pi\frac{(z-j)}{N}n} dn = \frac{-iN}{2\pi(z-j)}\left(e^{i2\pi(z-j)}-1\right) = Ne^{i\pi(z-j)}\,\mathrm{sinc}\left[\pi(z-j)\right], \quad (6.15)$$

so the Fourier transform can be expressed as

$$\widetilde{f}(k_j) = \sqrt{N}Ae^{i\pi(z-j)}\,\mathrm{sinc}\left[\pi(z-j)\right]. \quad (6.16)$$

The Fourier transform of a discretized periodic function, therefore, may have multiple non-zero values for the Fourier components. We can think of this in two distinct ways: firstly, we can say that this is because we are using a discrete set of periodicities which does not necessarily include the one which generated the discrete real-space function; or secondly, we can note that the discrete real-space function is actually not periodic, even though the function which generated it is.

For negative frequencies, the derivation is similar:

$$\mathcal{F}[Ae^{-ikx_n}] = \frac{A}{\sqrt{N}}\sum_n e^{-i2\pi\frac{(z+j)}{N}n}. \quad (6.17)$$

For the case where z is an integer, we have a very similar result to the one obtained above,

$$\widetilde{f}(k_j) = \left\{ \begin{array}{ccc} \sqrt{N}A & \text{if} & j = N-z \\ 0 & \text{if} & j \neq N-z \end{array} \right\}. \quad (6.18)$$

In the case where z is not an integer, there are often no values of j for which the steps in phase, $2\pi(z+j)/N$, are small ($\ll 2\pi$), since z and j are positive numbers. However, if the quantity $2\pi(z+j)/N$ is close to 2π, then the sum can again be replaced with an integral. If we let $N+\epsilon \equiv z+j$, where

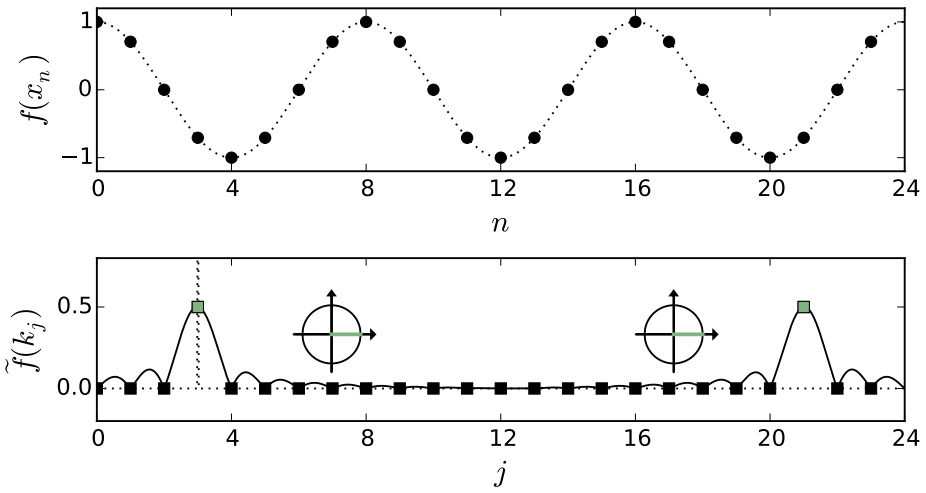


Figure 6.1: A periodic function, $f(x_n)$, and its 1D Fourier transform, $\tilde{f}(k_j)$, with a periodicity exactly matching a discrete Fourier component. The magnitude of the Fourier transform is plotted, and the phase of particular Fourier components are indicated by the insets in the lower panel. Dotted lines indicate the continuous function and its continuous Fourier transform.

ϵ is small, then we can write the sum as

$$\begin{aligned}\sum_n e^{-i2\pi\frac{(N+\epsilon)}{N}n} = \sum_n e^{-i2\pi\frac{\epsilon}{N}n} &\approx \int_0^N e^{-i2\pi\frac{\epsilon}{N}n}dn \\ &= Ne^{-i\pi\epsilon}\,\mathrm{sinc}\,(\pi\epsilon)\,, \end{aligned} \tag{6.19}$$

which allows us to write the Fourier transform as

$$\mathcal{F}[Ae^{-ikx_n}] = \sqrt{N}Ae^{i\pi(N-j-z)}\,\mathrm{sinc}[\pi(N-j-z)] \tag{6.20}$$

The Fourier transform of a real periodic function can be expressed by a sum of Equations 6.16 and 6.20:

$$\begin{aligned}\mathcal{F}\left\{A\cos[k(x_n - x_{\mathrm{sh}})]\right\} &= \frac{\sqrt{N}A}{2}e^{-ikx_{\mathrm{sh}}}e^{i\pi(z-j)}\,\mathrm{sinc}[\pi(z-j)] \\ &+ \frac{\sqrt{N}A}{2}e^{ikx_{\mathrm{sh}}}e^{i\pi(N-z-j)}\,\mathrm{sinc}[\pi(N-z-j)],\end{aligned} \tag{6.21}$$

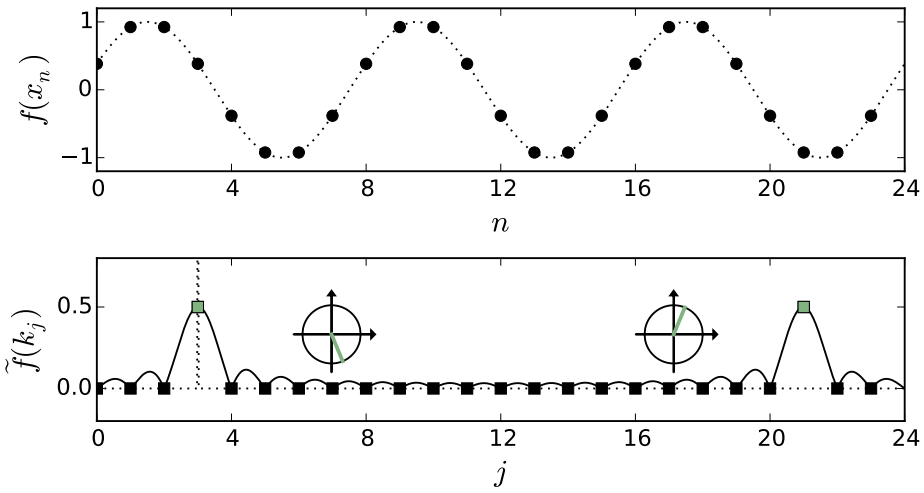


Figure 6.2: A periodic function and its 1D Fourier transform, with a periodicity exactly matching a discrete Fourier component. The offset in the function (upper panel) shows up as a phase in the corresponding peaks in the Fourier transform (insets of lower panel).

where x_{sh} represents the lateral shift of the periodic function. Equation 6.22 shows that a single periodicity in a real function gives rise to two peaks in the discrete FT. This was also the case for the continuous FT, where peaks in the FT showed up at $\pm k$, but in the case of the discrete FT, since there are no negative components in the dicrete FT, the "negative" peak instead shows up at $2\pi(N-z)/N\Delta$. The highest discrete components of the FT are in fact equivalent to negative wavevectors.

Figure 6.1 shows a discrete periodic function, whose generating wavevector corresponds to one of the Fourier component (meaning that z is an integer). The discrete FT has two non-zero components, at $j = z$ and at $j = N - z$. The inset of the FT shows that the phases of these two peaks are both zero. Figure 6.2 shows a different function, generated from the same wavevector, but shifted laterally. The magnitude of the discrete FT is identical to the unshifted function, but we see that the phases of the two peaks are no longer equal. In fact, they are opposite, as expected from the factors $e^{\pm ikx_{\text{sh}}}$ in Equation 6.22. We therefore see that the information about the periodicity is, roughly speaking, contained in the magnitude of the FT, and the information about

the phase of the periodic function is contained in the phase of the peaks.

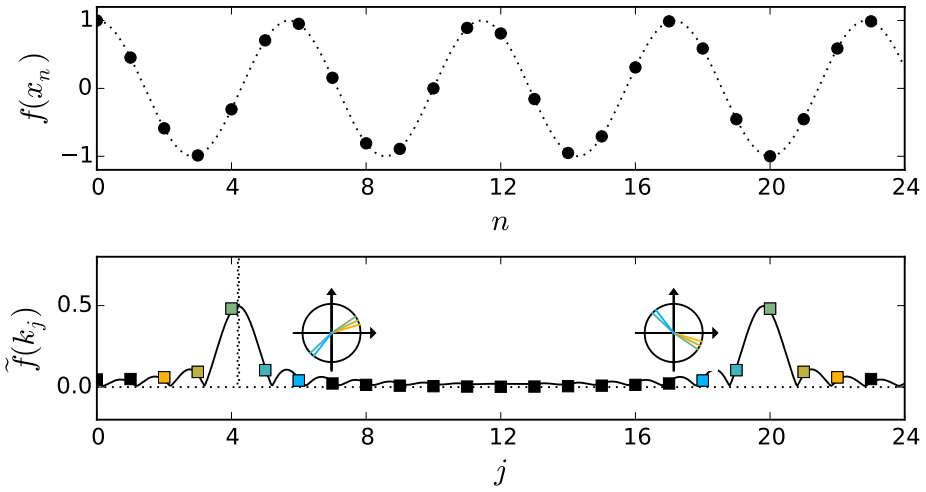


Figure 6.3: A periodic function and its 1D Fourier transform, with a periodicity which does not match a discrete Fourier component. The continuous function and its continuous Fourier transform are indicated by dotted lines. Because the peak in the continuous Fourier transform does not coincide with one of the discrete components, many discrete components have non-zero magnitudes. Phases of Fourier components near the two peaks are shown in the insets.

Figure 6.3 shows a pseudo-periodic function whose generating wavevector does not correspond to a Fourier component (meaning that z is not an integer), and which also experiences a shift, x_{sh}. First, we see that there are many non-zero components of the discrete FT, which are peaked near the generating wavevector, where j is close to z, but are also peaked near $N - z$. Secondly, we see that the phases of the components are no longer simply related to the shift. From one side of a peak to the other, the phase abruptly shifts by a phase of $\sim \pi$. Roughly speaking, this change in the phase is centered on the phase $\pm k x_{\text{sh}}$, bringing the phase from $\pm k x_{\text{sh}} - \pi/2$ to $\pm k x_{\text{sh}} + \pi/2$, for the two peaks.

Note that in Figures 6.1, 6.2, and 6.3, the discrete points represent the discrete function and the computed discrete FT. The solid black line in the FT

independently shows the predicted FT using Equation 6.22. We can see that in almost all cases, the discrete FT components fall precisely on the line defined by this equation, indicating the the discrete FT can be well approximated using sinc functions in this way.

6.1.3 Fitting a 1D Fourier Transform with Noise

We can fit noisy signals using the magnitude of the sinc function given by Equation 6.22 and the magnitude of the computed discrete FT. This works well in finding the periodicity if we start with a good guess (meaning, within about one Fourier component). However, it leaves out the phase of the periodic signal, and generally has difficulty finding the best fit for bad initial guesses. The difficulty in fitting comes from the bumpy nature of the magnitude of the sinc function. This means that when we algorithmically attempt to minimize the errors in a fit, we are looking for the minimum of a highly corrugated surface, where fits almost always "get worse before they get better," and the fit often settles in a local minimum.

When we fit using complex FT components, and keeping the phase information in Equation 6.22, fitting is less sensitive to the initial guess, but still requires the initial guess of k to be within a few Fourier components. Figure 6.4 shows a fit to a periodic signal with deliberately added noise. Despite the rather extreme noise, and a poor initial guess, a good fit is found. In this case, the fit also gives the correct value for the phase of the periodic signal.

6.1.4 Two-dimensional Discrete Fourier Transforms

As in the case of one-dimension, we can express two-dimensional discrete functions in terms of their Fourier components,

$$f(\mathbf{x}_{nm}) = \frac{1}{N} \sum_{i,j} \widetilde{f}(\mathbf{k}_{ij}) e^{i\mathbf{k}_{ij}\cdot\mathbf{x}_{nm}}, \tag{6.22}$$

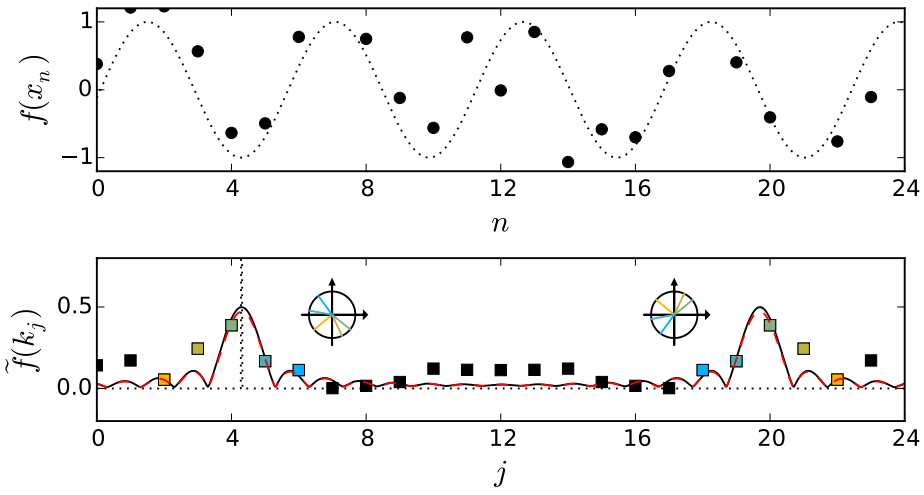


Figure 6.4: Fourier transform of a periodic function with simulated noise. The red dashed line in the lower panel shows the fit to the discrete values. Parameters for data generation were: $A = 1.0$, $x_{\text{shift}} = 1.5$, $z = 4.3$, and a noise level set at 0.3. Parameters for initial guess are $A = 0.5$, $x_{\text{shift}} = 0.0$, $z = 4.0$. Parameters found by fitting are: $A = 1.046$, $x_{\text{shift}} = 1.307$, $z = 4.250$.

and the components can be expressed in terms of the function as

$$\mathcal{F}\{f(\mathbf{x}_{nm})\} = \widetilde{f}(\mathbf{k}_{ij}) = \frac{1}{N}\sum_{n,m} f(\mathbf{x}_{nm}) e^{-i\mathbf{k}_{ij}\cdot\mathbf{x}_{nm}}, \tag{6.23}$$

which, again, are the inverse Fourier transform and the Fourier transform, respectively. Note that the two dimensional Fourier transform can be written in the form

$$\widetilde{f}(\mathbf{k}_{ij}) = \frac{1}{\sqrt{N}}\sum_{n}\left\{\frac{1}{\sqrt{N}}\sum_{m} f(\mathbf{x}_{nm}) e^{-ik_i x_n}\right\} e^{-ik_j x_m}, \tag{6.24}$$

which makes it clear that the two dimensional Fourier transform can be thought of as the one-dimensional Fourier transform along one dimension (say, rows) of the one-dimensional Fourier transform along the other (say, columns).

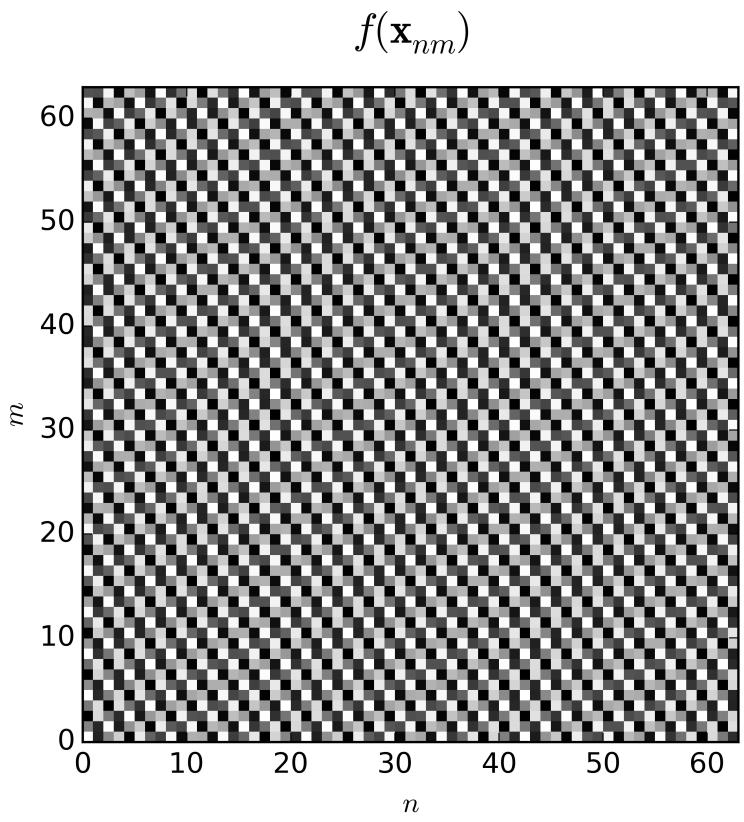


Figure 6.5: Simulated two-dimensional periodic image, with its two periodicities chosen to not correspond to an exact Fourier component.

In the specific case of a periodic function,

$$f(\mathbf{x}_{nm}) = Ae^{i\mathbf{k}_{ij}^{\circ}\cdot\mathbf{x}_{nm}}, \tag{6.25}$$

as shown for example in Figure 6.5, the two-dimensional Fourier transform

turns out to be the product of two one-dimensional Fourier transforms:

$$\begin{aligned}
\mathcal{F}\{Ae^{i\mathbf{k}^{\circ}_{ij}\cdot\mathbf{x}_{nm}}\} &= \widetilde{f}(\mathbf{k}_{ij}) = \frac{1}{N}\sum_{n,m}\left\{Ae^{i\mathbf{k}^{\circ}_{ij}\cdot\mathbf{x}_{nm}}\right\}e^{-i\mathbf{k}_{ij}\cdot\mathbf{x}_{nm}} \\
&= A\left\{\frac{1}{\sqrt{N}}\sum_{n}e^{ik^{\circ}_{i}x_n}e^{-ik_ix_n}\right\}\left\{\frac{1}{\sqrt{N}}\sum_{n}e^{ik^{\circ}_{j}x_m}e^{-ik_jx_m}\right\} \\
&= A\cdot\mathcal{F}\left\{e^{ik^{\circ}_{i}x_n}\right\}(k_i)\cdot\mathcal{F}\left\{e^{ik^{\circ}_{j}x_m}\right\}(k_j) \\
&\approx A\cdot\left\{\sqrt{N}e^{i\pi(z_i-i)}\,\mathrm{sinc}[\pi(z_i-i)]\right\} \\
&\qquad\times\left\{\sqrt{N}e^{i\pi(z_j-j)}\,\mathrm{sinc}[\pi(z_j-j)]\right\}
\end{aligned}$$

$$\mathcal{F}\{Ae^{i\mathbf{k}^{\circ}_{ij}\cdot\mathbf{x}_{nm}}\} \approx ANe^{i\pi(z_i+z_j-i-j)}\,\mathrm{sinc}[\pi(z_i-i)]\,\mathrm{sinc}[\pi(z_j-j)], \qquad (6.26)$$

where z_i and z_j are defined such that $k^{\circ}_i = 2\pi z_i/N\Delta$ and $k^{\circ}_j = 2\pi z_j/N\Delta$[†].

Figure 6.6 shows the magnitude of the computed FT, $\widetilde{f}(\mathbf{k}_{ij})$, of the 2D function shown in Figure 6.5, along with the predicted FT according to Equation 6.26, denoted $\widetilde{f}_{\mathrm{th}}(\mathbf{k}_{ij})$, in panels a and b. Panels c and d show the logs of these same functions. This shows that the 2D sinc function provides a good approximation to the magnitude of the exact discrete FT, except in regions far from the peaks, where the magnitude of the FT is very small. Figure 6.7 likewise compares the phase of the exact FT and the approximated one. Here, larger differences can be seen between. Still, the region very near each peak shows good agreement between these two representations of the two dimensional FT.

In practice, we have found that for real images including noise, fits to the two dimensional FT were most reliable when phase information was ignored, and the fit aimed to capture magnitude only. Just as in the case of the one dimensional FT, this results in a need for a very good initial guess. In practice, this is not a significant problem since a local maximum in the FT is often within one pixel of the underlying periodicity.

[†]There is an ambiguity here with regard to the symbol i, which can denote an integer index or the imaginary unit. I leave it to context to distinguish the two, but in general, the index can be recognized by its symmetry with respect to the index j.

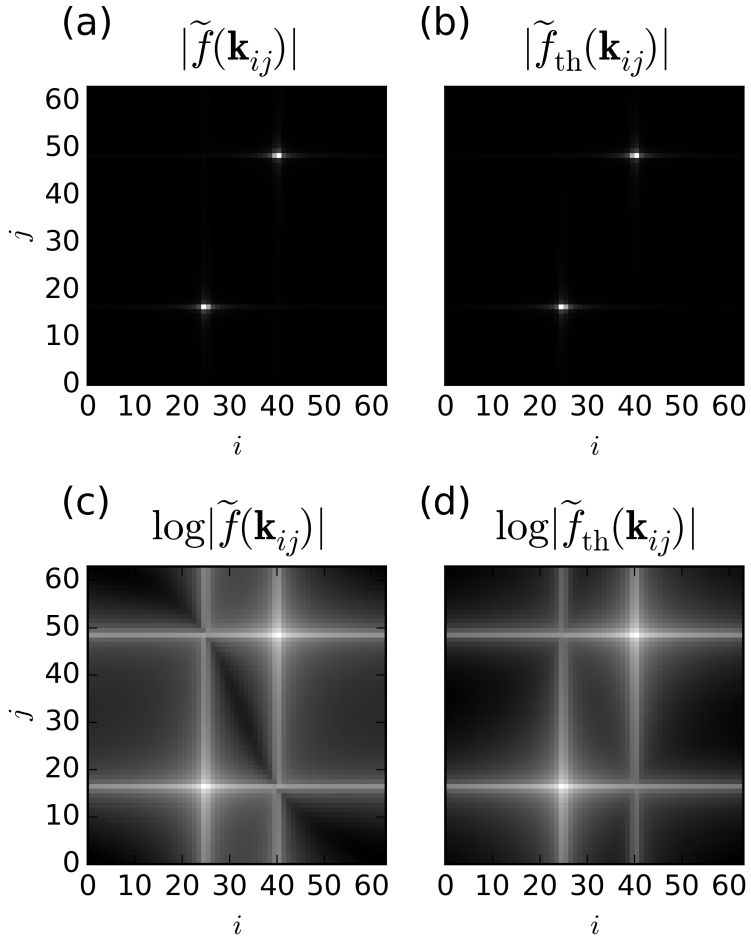


Figure 6.6: Two-dimensional Fourier transform of the periodic function shown in Figure 6.5 and its approximate representation, using Equation 6.26. (a) Computed Fourier transform of the image in Figure 6.5. (b) Approximate Fourier transform calculated using Equation 6.26. (c-d) Logarithms of the Fourier transform, computed and approximated.

6.2 Fabrication

DBs are fabricated using the STM tip by simple application of a pulse in voltage, which gives rise to a pulse in current. Typically, biases above ~ 1.6 V

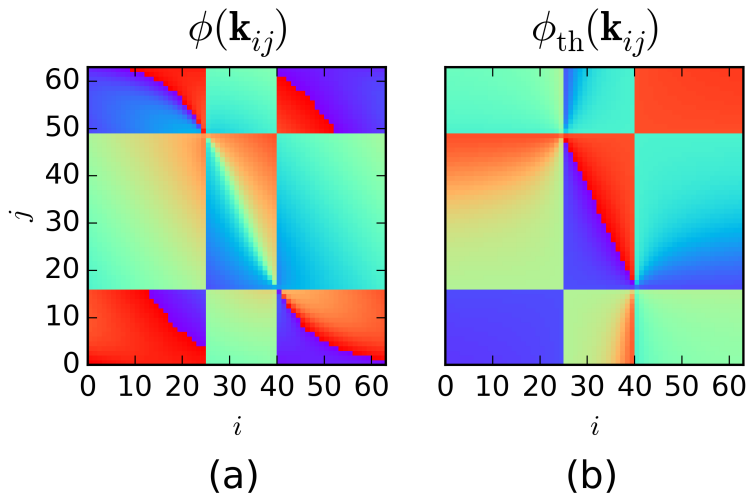


Figure 6.7: Phase of the two-dimensional Fourier transform, (a) computed, and (b) approximated using Equation 6.26.

are required to break the H-Si bond. In general, the STM tip is brought closer to the surface before the application of the pulse, to increase current during the pulse. Typically, the relevant parameters for DB fabrication are: pulse duration, τ_{pulse}, change in tip height (relative to a setpoint height), Δz_{pulse}, and pulse bias, V_{pulse}.

Figure 6.8 shows the result of this process of DB fabrication repeated at points arranged in a square grid of sites with nearest neighbours separated by 2 nm. For each line of this grid, the change in tip height for DB fabrication was gradually increased to bring the tip gradually closer to the sample. At the leftmost site, DB fabrication was attempted with no change in tip height, $\Delta z = 0$, so that the tip height during the pulse was determined solely by the feedback loop. At the rightmost site, DB fabrication was attempted with the tip brought closer to the sample by 200 pm. Aside from a brief period of time during the pulse, the feedback loop is left on. This experiment was performed for pulse biases of +2.4 V, +2.6 V, and +2.8 V.

The grids in Figure 6.8 illustrate the effects of tip height and tip bias on hydrogen desorption. The lack of DBs on the left side for all three pulse biases shows that the tip must be brought closer to the sample in order to remove

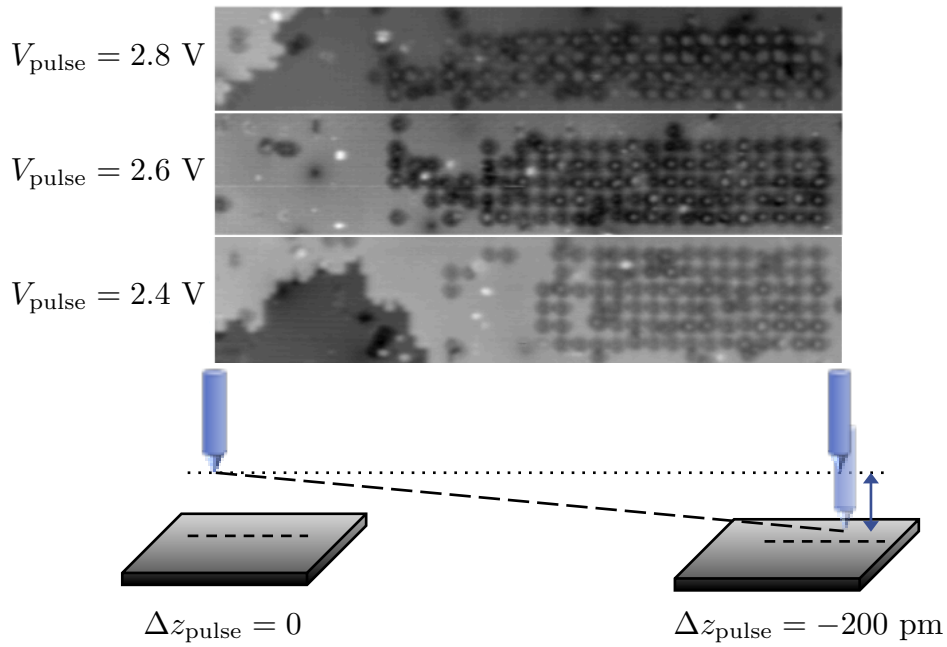


Figure 6.8: Grids of DBs fabricated with varying tip heights as the tip proceeds from left to right along each line, starting from $\Delta z = 0$ on the left and proceeding to $\Delta z = -200$ pm on the right. Tip displacements are defined relative to a tip height corresponding to the setpoint of $I_T = 40$ pA at $V_S = +2.0$ V.

hydrogen. As expected, desorption occurs more readily for higher pulse biases, indicated by the fact that the +2.8 V grid extends furthest to the left: high pulse biases allow removal of hydrogen at greater tip-sample separations.

A more interesting and less obvious observation is in the *quality* of desorption. Comparing the appearance of the DBs in the three panels of this figure, we see that the uniformity of the DBs is greatest for the lowest pulse biases, at +2.4 V. Furthermore, these are the darkest DBs. The DBs fabricated with higher voltage pulses appear more often as brighter and non-symmetrical protrusions. The reason for this is not that there is anything "wrong" with the DBs fabricated at higher biases. It is that multiple hydrogen atoms are more likely to be desorbed at higher biases. What appear to be single DBs in the upper two panels of Figure 6.8 are often actually tight clusters of multiple DBs. This illustrates an important fact for DB creation: it is in general preferable to desorb at low bias, with the tip close to the surface, rather than at higher

bias, with the tip further. The approach of the tip to the surface, however, if taken too far, can lead to changes in the atomic structure of the tip.

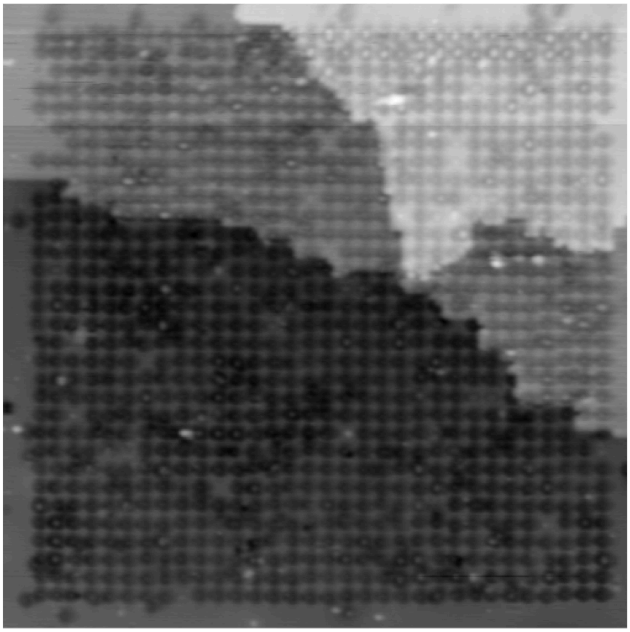


Figure 6.9: Grid of DBs at a 2nm pitch. Pulse bias was +2.4V and Δz_{pulse} was -150pm. Pulse width was 2 ms.

This illustrates one way to find good parameters for hydrogen desorption. For the tip and sample used in this experiment, optimal parameters for desorption are around $V_{\text{pulse}} = +2.4$ V and $\Delta z_{\text{pulse}} = -150$ pm. Figure 6.9 shows a grid of 1024 DBs created at these roughly optimized desorption conditions. The spacing between DBs is 2 nm. Single DBs are created in most cases, and in large portions of the grid, excellent uniformity is achieved. As described previously, most DBs which appear to be unusually bright and/or oddly shaped are in fact sites of multiple desorptions leading to clusters of DBs. The error rate in the number of DBs created (considering a single DB to be a success and zero or multiple DBs to be a failure) is in the range of 10-25%.

In Figure 6.9, there are placement errors in addition to the errors in the number of hydrogen atoms desorbed. That is, the created DBs were not created at exactly the sites of the 32×32 grid of pre-decided locations where desorption was attempted. In this case, this is entirely expected, since the grid of desorption sites was not defined with reference to the surface atoms, but

instead was arbitrarily chosen to have nearest neighbour spacing of 2 nm. By luck, 2 nm is close to the distance of five atomic spacings, equal to ~ 1.92 nm, so that portions of the grid may happen to align with the silicon surface. Since the desorption attempts most often do not correspond exactly to the locations of surface atoms, it would be impossible to *not* have placement error in this case. In order to have good alignment between the tip and the surface atoms, the image analysis methods described in the previous section need to be applied.

With analysis of the Fourier transforms described in the previous section, and care to minimize and account for other spurious effects such as drift, it is possible to achieve excellent alignment of the tip with the sample, as shown in Figure 6.10. This figure shows an attempt to desorb 12 hydrogen atoms in a line, with a spacing of exactly 1.536 nm, or four atoms over. Figure 6.10a shows 2 nm $\times$ 2 nm images recorded before and after each desorption attempt in order to establish whether or not good alignment was achieved for each desorption attempt. Desorption was attempted at the center of each image, at the intersection of the dashed lines. The larger STM image shown in Figure 6.10b shows the resulting line, circled in red. Part of a different attempt is visible in the bottom of this image.

This figure shows that much higher desorption success rates are possible with proper alignment between tip and sample. In this case 12 out of 12 desorption attempts resulted in single DBs being created at the intended site. (Note that the additional DB above and to the right of the third DB, as well as the additional DB below the eighth DB, were both pre-existing, as seen in the top right edge of the third "before" image as well as the eighth "before" image.) While this is a best case scenario, and does not indicate a genuine 100% success rate in all cases, it does show that much higher success rates are possible. A different attempt, visible in the lower part of Figure 6.10b shows a case with an error, both in number of DBs and location, at one of the sites. This is perhaps more representative of general statistics. With proper tip-surface alignment and good desorption parameters, success rates around 90% are typical.

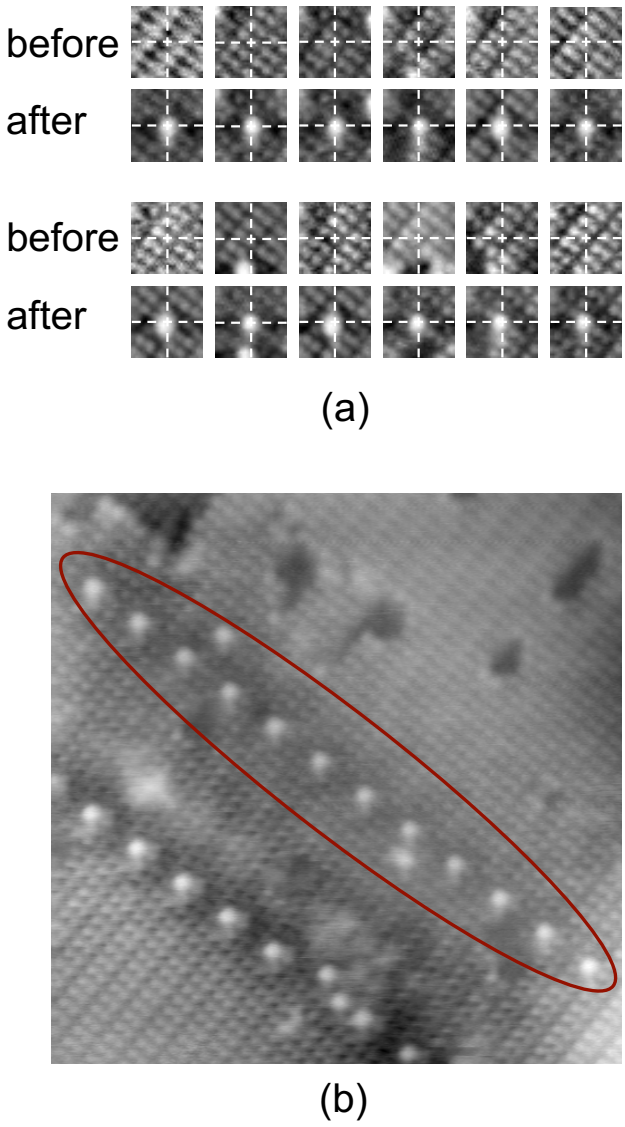


Figure 6.10: 1.0V 20pA images. (a) 2nm x 2nm images and (b) 15nm x 15nm image.

Also noteworthy in this figure is the challenge posed by tip changes. The lattice looks different depending on each tip. The first "before" picture gives the impression that we are aiming for the top left side of the dimer. In fact we are aiming for the bottom right side, but this tip happens to image the dimer rows in a way that exagerates the space between the atoms in a dimer, and under-emphasizes the space between dimer rows.

In general, near-perfect alignment between the tip and sample results in higher success rates for DB fabrication. In practice, great care needs to be

taken to achieve near-perfect alignment. An error in the initial placement of the tip for the first desorption misaligns the entire pattern relative to the lattice. Furthermore, even with perfect image analysis, drift can warp images, leading to incorrect unit vectors extracted from surface analysis, and of course drift can also cause misalignment later, when desorption is attempted. In addition to all this, hysteresis — or "creep" — in the scanner position adds another challenge to tip positioning. The result of all this is that extreme care needs to be taken to ensure near-perfect registry of the tip with the surface. At present, this requires a great deal of input from the STM operator.

6.3 Multi-DB Structures

The ability to fabricate DBs on the silicon surface opens very exciting avenues for creating technologies at the smallest scales. Conducting structures could be created, enabling transport-based devices, which could draw a fraction of the current of today's CMOS devices. An even more exciting prospect is the possibility of harnessing the very different properties of matter on this scale. It appears to be possible to make tailored wavefunctions by precisely controlling wavefunction overlap between neighbouring sites. Furthermore, the small scale of DB structures allows much stronger Coulombic coupling, opening the door for technologies based on strongly interacting electrons.

Figure 6.11, 6.12, and 6.13, show three of the simplest multi-DB structures that can be fabricated, a two-, three-, and four-DB line respectively. These three structures were created on a a single sample, separated by 20 to 60 nm from one another (meaning that they were isolated). This allowed these structures to be studied with a single STM tip, so that differences between these images and the corresponding topographical cross-sections, from one structure to another, are entirely due to the structures themselves, and not to any changes in the tip.

A noteable feature of these images is the presence of protrusions *between* the atomic sites of the DBs. We typically refer to these (rightly or wrongly) as "antinodes," since they look as if there were a maximum in the DB wavefunc-

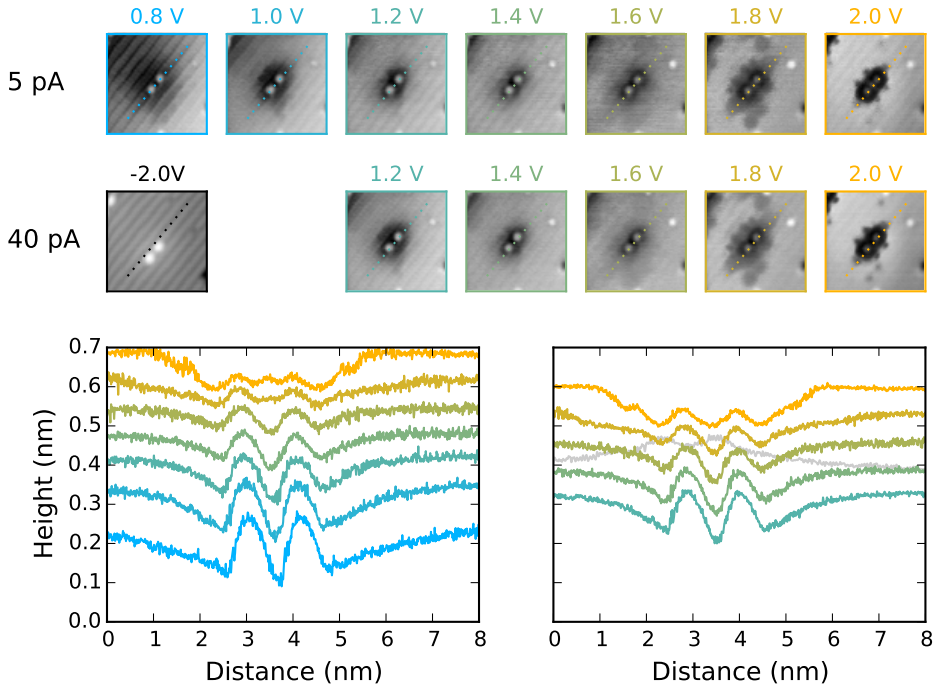


Figure 6.11: Two-DB structure. STM images and cross-sections.

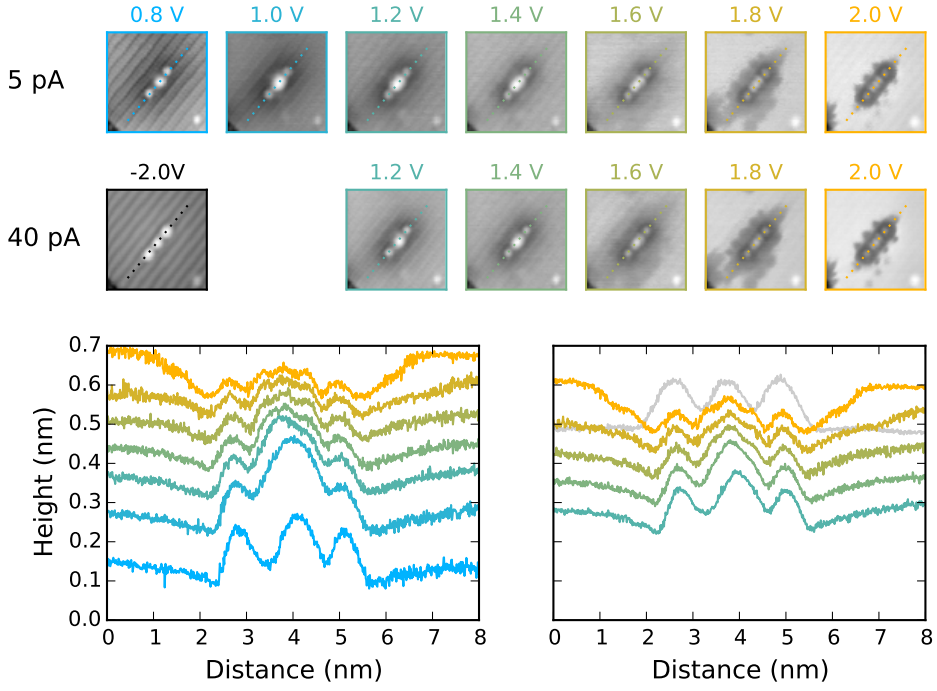


Figure 6.12: Three-DB structure. STM images and cross-sections.

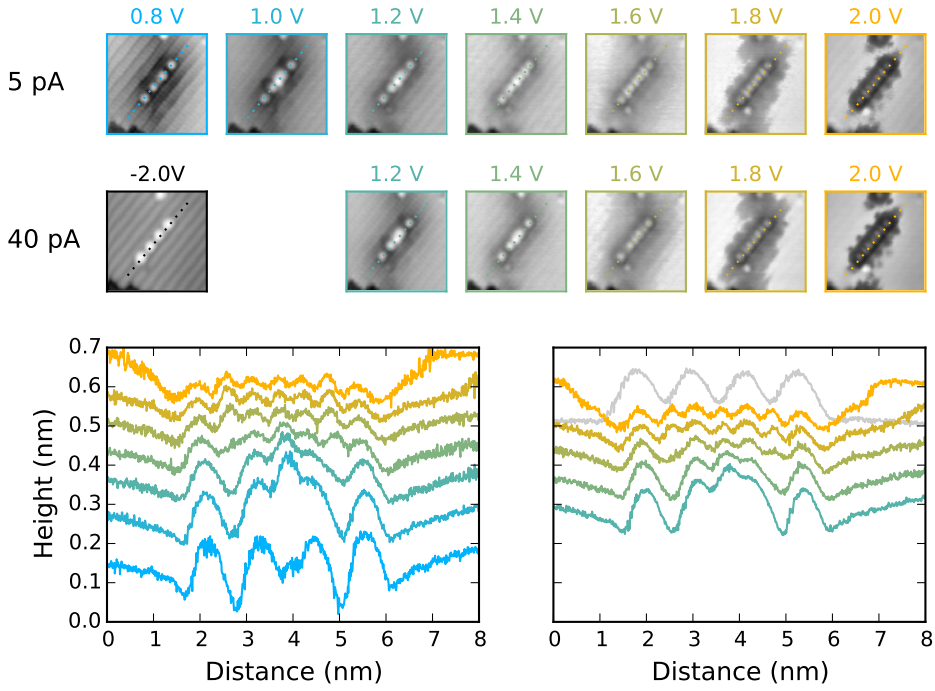


Figure 6.13: Four-DB structure. STM images and cross-sections.

tion, where one might expect a minimum. An explanation for these antinodes was provided by Schofield *et al.*,[34] who proposed that these features are due to the hybridization of excited states of the DB. This explanation is perfectly plausible, but, given that these features appear *inside* the DB halo, it is important to connect this explanation (or any other) to the non-equilibrium ideas of competing filling and emptying processes described in this thesis.

As these structures are imaged at varying biases, there are significant changes in the relative brightness of each atom, as well as in the appearance of the antinodes. In general antinodes become more visible as bias is increased. In the case of the four-DB structure, it is interesting that the central antinode appears at a much lower bias than the two outer anti-nodes.

Figure 6.14 shows a six-DB structure, comprised of a straight three-DB line and a kinked three-DB line. This structure illustrates an interesting feature of the antinodes, which is that they tend to be perpendicular to the line connecting two DBs. Thus we can see that the antinodes in the kinked three-DB structure are tilted relative to the lattice, while the antinodes in the straight

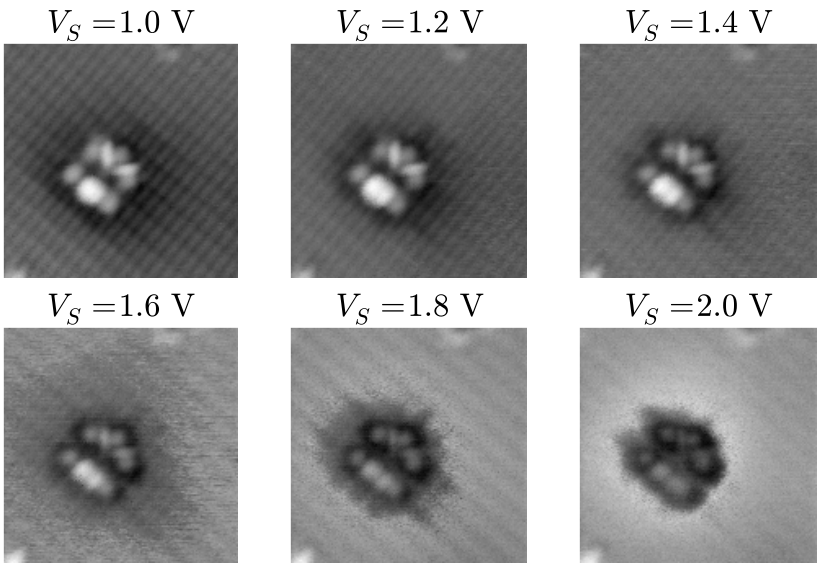


Figure 6.14: STM images of a six-DB structure at the six different biases indicated. $I_T = 40$ pA.

three-DB structure are aligned with it. Also notable in this figure is the fact that this structure collectively exhibits a halo, and at some biases this halo has the distinctive speckly noise that is characteristic of charge fluctuations. It would be interesting to study such fluctuations. Do they correspond to a changing charge state of the entire structure, or to the transfer of charge from one part of the structure to another?

In this thesis, we will not deal with the detailed topography of these structures, although this too is a fascinating topic in which there is clearly a great deal of work to be done. Instead, we simply present them as an illustration of the rich structures that can be made, even with only a few DBs.

6.4 Large Scale DB Patterning

Image analysis and automated desorption routines allow fabrication of DB structures from few DBs to many. The previous section showed some examples of small DB structures whose intricate topography is suggestive of rich electronic structure, and possibly complex dynamics during STM imaging.

Here, we discuss the possibility of using automated fabrication algorithms to pattern the H-terminated silicon surface on a much larger scale.

Figure 6.15 shows fabrication of a repeated pattern on different scales. Figures 6.15a and b show the repetition of the pattern in a 2×3 grid, and a 7×7 grid, respectively. While the precision of fabrication is still relatively rough, resulting in many atoms out of place, we see that the STM tip succeeds in making single DBs in most instances. This process can be scaled up to large scale patterning, as shown in Figure 6.15c, which shows the result of patterning over 24 hours, at a rate of slightly more than one desorption per second, resulting in the repetition of the atomic pattern over approximately $1.0\ \mu\text{m} \times 0.4\ \mu\text{m}$.

Figure 6.15d is a close up view of one unit of the repeated pattern after large scale patterning (the repeated pattern consists of four rotated quadrants, one of which is shown in Figure 6.15e). This STM image shows, first, that there remains a significant degree of inaccuracy in DB placement. Perhaps more important, however, is the resolution of the image, which is an indication of the condition of the STM tip. Even after the desorption of several tens of thousands of hydrogen atoms, the STM tip remains in excellent condition, still able to create single DBs, and still able to image with excellent resolution. This suggests that STM tips could be used to pattern on a large scale, without significant degradation of the STM tip.

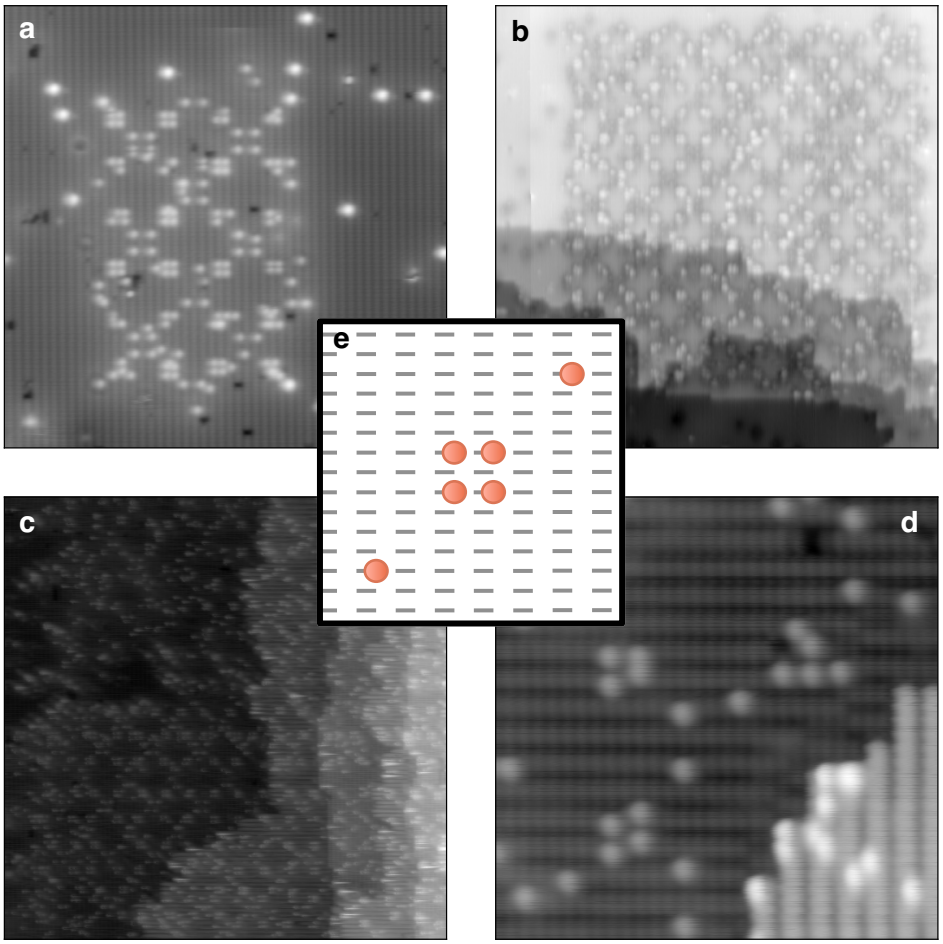


Figure 6.15: (a) 40×40 nm^2 STM image of a pattern repeated in a 2×3 grid. $V_S = -2$ V and $I_T = 100$ pA. (b) 80×80 nm^2 STM image of the same pattern repeated in a 7×7 grid. $V_S = 1.8$ V and $I_T = 100$ pA. (c) 80×80 nm^2 STM image after large scale patterning. The atomic pattern was repeated in a grid of size 1.0×0.4 μm^2, roughly, corresponding to fabrication of approximately 30 000 DBs. $V_S = -2$ V and $I_T = 100$ pA. (d) 12×12 nm^2 STM image after large scale patterning. $V_S = -2$ V and $I_T = 100$ pA. (e) Schematic diagram of one quadrant of the repeated atomic pattern. The full repeated pattern consists of four such configurations, each rotated by 90°.

7 Quantum-dot Cellular Automata

In this chapter, we discuss a potential application of silicon DBs: Quantum-dot Cellular Automata (QCA). We begin by reviewing some of the basic concepts of QCA. We then discuss the typical assumptions of QCA simulation, and proceed to examine the role of quantum correlations in QCA. Much of the discussion in this chapter makes use of the descriptions in Taucer *et al.* (2015).[3]

7.1 Introduction to QCA

QCA was first proposed by C. S. Lent in 1993.[80] It provides a basis for computation fundamentally different from the transistor-based technology that has dominated high-tech innovation for well over half a century.[81] While QCA makes use of quantum dots, and quantum tunneling between these dots, it is an architecture for *classical* computation, not to be confused with the entirely different (though related) field of quantum computing.

In QCA, information is encoded in cells made of quantum dots. A cell consists of four quantum dots arranged at the corners of a square, as shown in Figure 7.1. Each cell contains two mobile electrons, which can tunnel between adjacent dots, described by hopping constants, t_{ij}, but they cannot tunnel from cell to cell. Electrons of course experience mutual repulsion, described by an interaction energy, V_{ij}. In general, this repulsion causes the electrons to occupy opposite corners of the cell, forcing the configuration of the cell to fall into one of the two antipodal (or "diagonal") configurations.

In addition to the mutual repulsion between the two electrons inside a cell, there is a slightly weaker repulsion between the electrons of neighbouring cells, which tends to align their electronic configurations, when the cells are

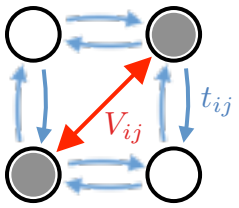


Figure 7.1: Single QCA cell.

placed side by side. A simple line of cells thus acts as a binary wire, transmitting information, as shown in Figure 7.2a. Other arrangements of cells lead to different interactions between the electronic configuration of each cell. For instance, diagonally placed QCA cells (corner-to-corner), cause an anti-alignment of their respective electronic configurations. This is the basis of the logical inverter shown in Figure 7.2b. Just like the binary wire, this gate operates only on the basis of electron-electron repulsion, but with different cell placement. Likewise, the arrangement of cells shown in Figure 7.2c is known as a majority gate, since the single output is determined by a "vote" of the three inputs. The binary wire, inverter, and majority gate are sufficient to enable the design of a universal computer capable of very low power operation.[82,83]

Computing speed can in principle be enormously improved by the incorporation of clocking,[84,85] where clock zones are used to sequentially send "bit packets" through a QCA array to allow for pipelining. A clock zone is a group of cells whose parameters can be tuned to either allow or disallow the group of cells to become polarized (*i.e.*, to take on one of the antipodal configurations). Within a clock zone, the cells can be rendered "active," or "inactive." Active cells are ones that can adopt one or the other antipodal configuration. Thus they are sensitive to the polarization of cells around them, and they are likewise able to affect cells nearby. Inactive cells are ones that, in one way or another, have been rendered insensitive to their environment. In simulations, this can be done by forcing electrons in a clock zone to delocalize, so that the cells are depolarized. It can also be done by forcing the two electrons into two out-of-plane dots, where they no longer have any effect on neighbouring cells.[84] In either case, the inactive cell is insensitive to its surroundings and cannot transmit information. A bit packet consists of a group of interacting cells in

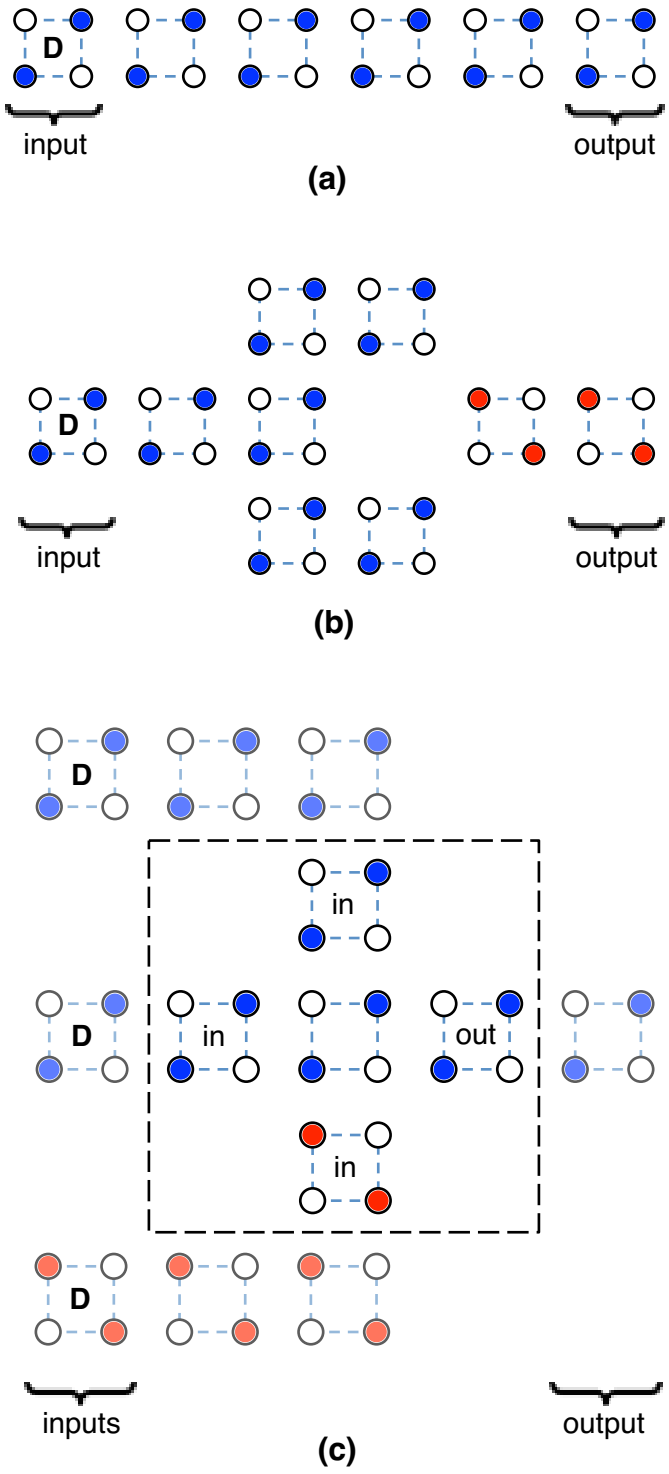


Figure 7.2: Basic QCA logic gates: (a) QCA wire. (b) Logical inverter. (c) Majority gate.

"active" clock zones, and can be moved through the circuit. As envisioned theoretically, many bit packets could be processed simultaneously in a single QCA circuit by modulating clock zone parameters appropriately.

QCA cells were demonstrated experimentally in metal-island quantum dots as early as 1997.[86] These metal-island QCA cells have been used to demonstrate transmission of information, logic gates,[87,88] and clocking.[89,90] In this implementation, clocking is realized by modulating the null-dot potential between dots. However, the relatively large size of the quantum dots means that energy levels are closely spaced, so metal-island devices must be kept at temperatures below $\sim 5K$[91,92] in order for quantum effects to be observable. As the dimensions of a QCA cell are reduced, the operating temperature increases, and at the molecular scale room temperature operation becomes possible. Further advantages of miniaturization include fast switching times and increased device density. For these reasons, molecular scale QCA has held great promise. Suitable candidates have been synthesized chemically,[93,94] and a QCA cell made of silicon DBs has also been realized.[33] DBs are an extremely promising route to QCA since they naturally exist at the atomic scale, and their presence on the Si(100) surface suggests avenues for integration into existing CMOS architectures. For QCA cells at the atomic scale, it is probably neither feasible nor desirable to have addressable control over the parameters of individual cells. Instead, large groups of cells could be addressed by external fields,[84,95,96] with the variation in these external fields producing clock zones.

A great deal of simulation and theory has been done on the topic of QCA.[81,97–101] This research has aimed to capture the qualitative and quantitative characteristics of QCA cells and of arrays of cells. Because of the difficulties inherent in solving the complete quantum mechanical problem, a number of simplifying approximations are typically made. These include a reduction of the Hilbert space to two states per cell,[98] a mean-field approach to the intercellular interactions,[81,97] and finally an assumption of exponential energy relaxation.[82,99] These approximations have allowed the field to progress tremendously, simulating devices from small groups of cells, to large-scale QCA processors. A great deal has already been learned about the nature and impor-

tance of quantum mechanical calculations that go beyond these approximations, including full and partial quantum correlations. Quantum correlations have been shown to affect dynamics in QCA wires[98] and circuits.[99] The degree of correlations and and magnitude of their effect has also been studied.

The work presented in this chapter applies standard equations from the literature to the simulation of clocked QCA, which have so far not been studied with the inclusion of intercellular correlations. We see that the inclusion of intercell correlations can significantly change the steady state of the system not only quantitatively, but qualitatively, even in the case of very simple systems, such as an unbiased line of QCA cells. Full quantum mechanical simulations, with the approximation of exponential relaxation to a thermal steady-state, predict an exponential loss of information even as bit packets propagate, as well as coherent oscillations whose period strongly depends on the size of the bit packet. This contrasts with mean-field simulations, which show perfect and indefinite propagation of information. The results presented here will have implications for molecular-scale QCA device design, and will highlight the need for implementation-specific theoretical treatments of the interaction of a QCA system with its environment.

In Section 7.2 we review the basic theory of QCA as well as some of the most common approximations used in QCA simulations. In Section 7.3 we consider the full Hamiltonian for a QCA line, and the characteristics of its solutions. In Section 7.4 we present the results of fully quantum mechanical simulations. Section 7.5 discusses the main results presented in this chapter, their scope, and their implications for QCA design, and finally Section 7.6 concludes.

7.2 Simulation of QCA Systems

The dynamic behaviour of QCA was first explored by Tougaw and Lent (1996),[98] who examined the time evolution of QCA cells by considering a basis set consisting of all sixteen possible states of a four-dot QCA cell populated by two electrons of opposite spin. In this sixteen-state approach, the authors con-

struct a Hubbard-type Hamiltonian, given as[98]

$$\begin{aligned}\hat{H} &= \sum_{i,\sigma,m} E_0 \hat{n}_{i,\sigma}(m) + \sum_{i>j,m,\sigma} t_{i,j}\left[\hat{a}^{\dagger}_{i,\sigma}(m)\hat{a}_{j,\sigma}(m) + \hat{a}^{\dagger}_{j,\sigma}(m)\hat{a}_{i,\sigma}(m)\right] \\ &+ \sum_{i,m} E_Q \hat{n}_{i,\uparrow}(m)\hat{n}_{i,\downarrow}(m) + \sum_{i>j,\sigma,\sigma',m} V_Q \frac{\hat{n}_{i,\sigma}(m)\hat{n}_{j,\sigma'}(m)}{|r_i(m) - r_j(m)|} \\ &+ \sum_{i,j,\sigma,\sigma',k>m} V_Q \frac{\hat{n}_{i,\sigma}(m)\hat{n}_{j,\sigma'}(k)}{|r_i(m) - r_j(k)|}, \end{aligned} \tag{7.1}$$

where the operator $\hat{a}_{i,\sigma}(m)$ ($\hat{a}^{\dagger}_{i,\sigma}(m)$) annihilates (creates) an electron on the i^{th} site of cell m with spin σ, the operator $\hat{n}_{i,\sigma}(m) \equiv \hat{a}^{\dagger}_{i,\sigma}(m)\hat{a}_{i,\sigma}(m)$ is the number operator for an electron on the i^{th} site of cell m with spin σ, and $V_Q = q_e^2/(4\pi\epsilon)$ is a constant where q_e is the charge of the electron and ϵ the electrical permittivity of the medium. The first term in Equation 7.1 represents the on-site energy of a dot. The second term describes the electron tunnelling, where $t_{i,j}$ is a hopping constant (with units of energy) between neighbouring sites i and j, determined from the structure of the potential barriers between the dots in the cell. The third term in Equation 7.1 accounts for the energetic cost, E_Q, of putting two electrons of opposite spin at the same site, and the final two terms are related to the Coulombic interactions between electrons in the same cell and in neighbouring cells, respectively. The polarization of each cell can then be found by evaluating

$$P_m = \frac{(\rho_1^m + \rho_3^m) - (\rho_2^m + \rho_4^m)}{\rho_1^m + \rho_2^m + \rho_3^m + \rho_4^m}, \tag{7.2}$$

where ρ_i^m is the expectation value of the number operator on the i^{th} site of cell m; *i.e.*, $\rho_i^m = \langle \hat{n}_i(m) \rangle$. Sites within a cell are labeled counter-clockwise starting from the top left. While this Hamiltonian considers the complete many-body configuration space, including correlation effects within and between cells, it becomes computationally intractable when used to model large systems. The exponential growth of the basis set makes it computationally prohibitive to model any circuit larger than just a few cells. We will now discuss three ubiquitous approximations used for studying QCA circuits and systems.

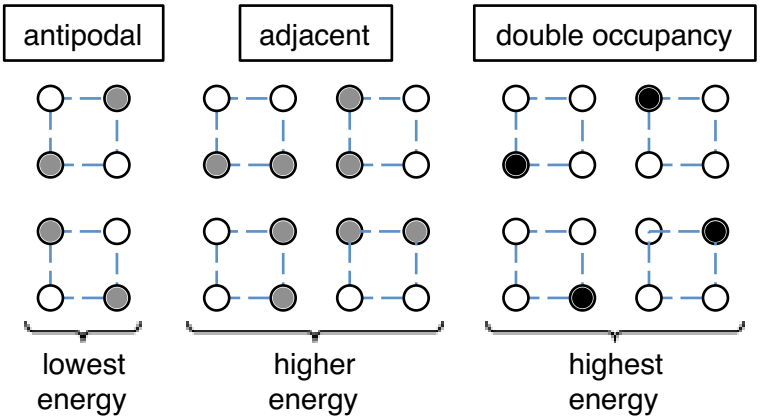
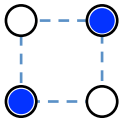


Figure 7.3: The various single QCA cell configurations.

7.2.1 Two-State Approximation

Figure 7.3 shows several different possibilities for the configuration of a QCA cell with two electrons. These can be broadly classified into antipodal configurations, where the electrons occupy opposite corners of the cell, adjacent configurations, where they occupy adjacent sites, and double occupancy configurations, where both electrons occupy the same site. Because of the small scale of a QCA cell, the electron-electron repulsion can be quite strong, leading to a large energy difference between these different types of configurations. As long as temperature is not too high, the QCA adopts the antipodal configurations, or a superposition of them, with high probability.

More rigorously, Tougaw and Lent (1996)[98] showed that the ground state of a single cell remains almost completely contained within a two-dimensional subspace of the full sixteen-dimensional Hilbert space. For sufficiently low temperatures, we can therefore expect the state of a cell or a line to be well described by a two-state approximation. For simplicity, we consider idealized cells with saturation polarizations of ± 1. A more realistic treatment would have the polarizations spanning a slightly smaller range, however the basic argument that follows would be unchanged. We therefore refer to our reduced basis as the polarization basis, and denote the two states as $|0\rangle$ and $|1\rangle$, with respective polarizations $P = -1$ and $P = +1$, as shown in Figure 7.4.

$|0\rangle \equiv$  ; $P = -1$

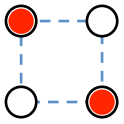
$|1\rangle \equiv$  ; $P = +1$

Figure 7.4: Schematic representation of the two diagonal states which form the basis in the two-state approximation.

The Hamiltonian in the polarization basis for a system of N interacting QCA cells, under the influence of driver cells, is then described by a $2^N \times 2^N$ Ising-like Hamiltonian:

$$\hat{H} = -\sum_{i=1}^{N} \gamma_i \hat{\sigma}_x(i) - \frac{1}{2}\sum_{i<j}^{N} E_k^{i,j} \hat{\sigma}_z(i)\hat{\sigma}_z(j) + \frac{1}{2}\sum_{D}\sum_{i=1}^{N} E_k^{i,D} P_D \hat{\sigma}_z(i), \qquad (7.3)$$

where γ_i is an effective tunneling energy, related to the hopping energy, $t_{i,j}$, in equation 7.1. $E_k^{i,j}$ is the so-called kink energy between cells i and j, and accounts for the energetic cost of two cells having opposite polarization. P_D labels the polarization of the driver cell labelled D, and $E_k^{i,D}$ is the kink energy between cell i and driver D. The driver cells provide a mechanism for input into the QCA circuit and have a polarization that can range from -1 to $+1$. The Pauli operators for the i^{th} cell, $\hat{\sigma}_a(i)$; $a = x, y, z$, represent the tensor product of N 2×2 identity operators, with the i^{th} identity operator replaced by the one of the Pauli matrices,

$$\hat{\sigma}_x = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \hat{\sigma}_y = \begin{pmatrix} 0 & i \\ -i & 0 \end{pmatrix}, \quad \hat{\sigma}_z = \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}.$$

For example, $\hat{\sigma}_y(2) \equiv \mathbb{1} \otimes \hat{\sigma}_y \otimes \mathbb{1} \otimes \ldots \otimes \mathbb{1}$. The polarization of a cell, i, can now be defined as $P_i = -\langle \hat{\sigma}_z(i) \rangle$. The first term in Equation 7.3 accounts for the kinetic energy of electrons, and tends to bring the cells to a superposition of

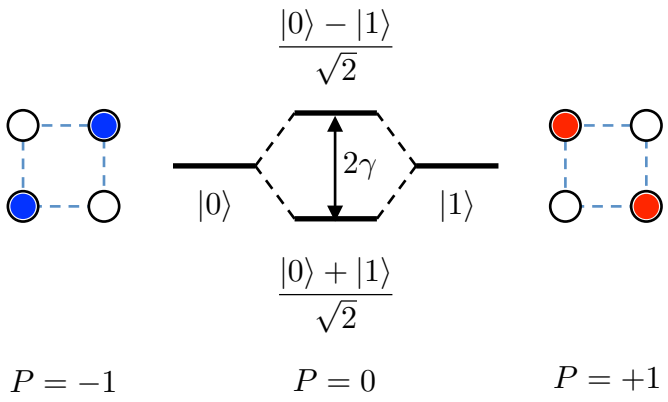


Figure 7.5: Analogy between QCA cell and covalent bond.

polarization states. The second and third terms account for the energy cost of having a cell misaligned with its neighbour, or with a driver cell respectively. The Hamiltonian in Equation 7.3 is very similar to the quantum Ising model with a transverse field, which has been extensively studied.[102,103]

In Section 7.3 we will use the two-state approximation to study systems of interacting QCA cells. This procedure, like the full sixteen-state one, naturally includes the effects of inter-cell entanglement, or quantum correlations. The price we pay for proper inclusion of quantum correlations is an exponential growth in the basis set with the number of interacting cells, N. Even with the two-state approximation, this limits its application to systems containing only a small number of interacting cells. Further approximations are used to solve the problem of scaling.

7.2.2 Intercellular Hartree Approximation

One way to eliminate the problem of exponential scaling is to ignore intercellular entanglement effects altogether and solve the Schrödinger equation for each individual cell separately. This method is known as the Intercellular Hartree Approximation (ICHA).[81,97] In this Hartree-type treatment, cells are coupled to one another through expectation values (polarizations) rather than operators. The Hamiltonian (in the polarization basis) for a single cell i, is then simply

$$\hat{H}_i = -\gamma_i \hat{\sigma}_x + \frac{1}{2} \sum_j E_k^{i,j} P_j \hat{\sigma}_z, \tag{7.4}$$

where P_j is the polarization of cell j. The polarization of cell i is found by evaluating, $P_i = -\langle \sigma_z \rangle$. Note that this equation is precisely Equation 1.8 used to describe the ionic bond in Chapter 1. In particular, the hopping constant between states, t, has been replaced by γ_i, and and the difference between the two on-site energies, Δ, has been replaced with $\sum_j E_k^{i,j} P_j$. This allows us to think of each cell as a relatively simple two-state quantum system, without dealing with the exponentially growing Hilbert space that would come from quantum interactions.

Because the solutions of one cell's Hamiltonian define parameters that enter its neighbours' Hamiltonians, the system of Schrödinger equations must be solved iteratively to obtain self-consistency. In calculating the state at a particular time, the initial guess is typically taken to be the state of the system at the previous time step.

The primary benefit of this approximation is that it only requires the diagonalization of N 2×2 Hamiltonians, which scales linearly with the number of cells in the system. Furthermore, if each cell remains in its ground state, the problem simplifies further since the polarization of any cell, i, can be evaluated analytically using,[104]

$$P_i = \frac{\frac{1}{2\gamma} \sum\limits_j E_k^{i,j} P_j}{\sqrt{1 + \left(\frac{1}{2\gamma} \sum\limits_j E_k^{i,j} P_j \right)^2}}. \tag{7.5}$$

Equation 7.5 produces the well-known nonlinear cell-to-cell response function shown in Figure 7.6.

While the ICHA is generally capable of arriving at the correct ground state of an array of QCA cells in a single clocking zone with a fixed driver, it can predict a latching mechanism within a group of cells that allows them to retain (and even obtain) a polarization in the absence of a perturbing cell. This latching turns out to be something that is added when we make a mean-field approximation. Compared to the "more exact" fully quantum mechanical calculation, this latching appears to be an inaccuracy of the ICHA. This goes beyond previously noted inaccuracies of the ICHA, such as in dynamics and

finite temperature behaviour, where fully coherent calculations of QCA wires were used to show that the many-cell excited states are needed to get quantitatively correct results.[98]

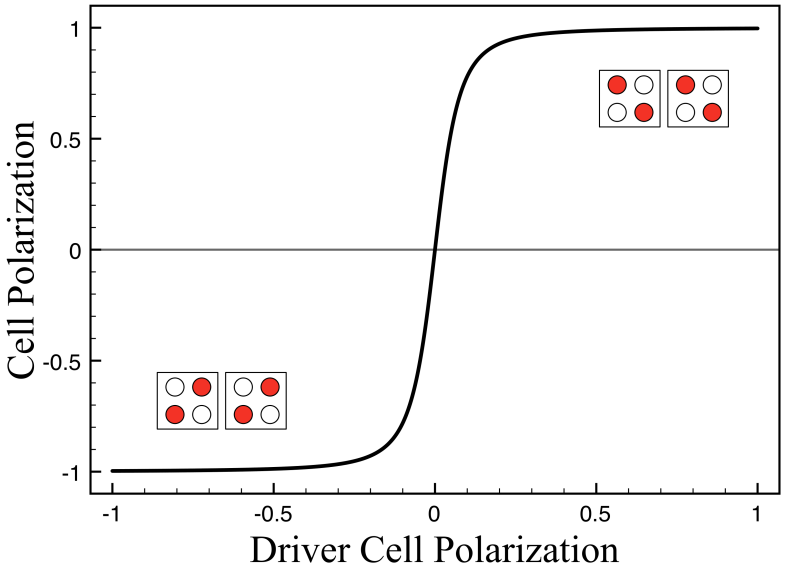


Figure 7.6: Nonlinear cell-to-cell response function for $\gamma \ll E_k$ and $k_B T \ll E_k$ (optimal conditions for QCA). The output cell is almost completely polarized for even a small driver polarization. (Figure from Taucer *et al.* (2015).[3])

To illustrate the effects of the ICHA on the calculated ground state of QCA arrays, consider the driven two-cell wire shown in Figure 7.7. Two simulations were conducted on the wire; the first using the ICHA, and a second using the more complete quantum mechanical treatment discussed in Section 7.2.1. For each simulation, the driver cell polarization is varied between $-1 \leq P \leq +1$, for three different values of E_k/γ. For each value of the driver cell polarization, the steady state response is calculated. The polarization of the first cell as a function of driver cell polarization is plotted in Figures 7.7a and b.

Let us first consider the results from the ICHA simulation. When using the ICHA, one has freedom in choosing an initial guess for the polarizations. The most common method in dynamic simulations is to use as an initial guess at each time step the solution from the previous time step. Though there are no time dynamics in our simulations, we can illustrate the effect of this

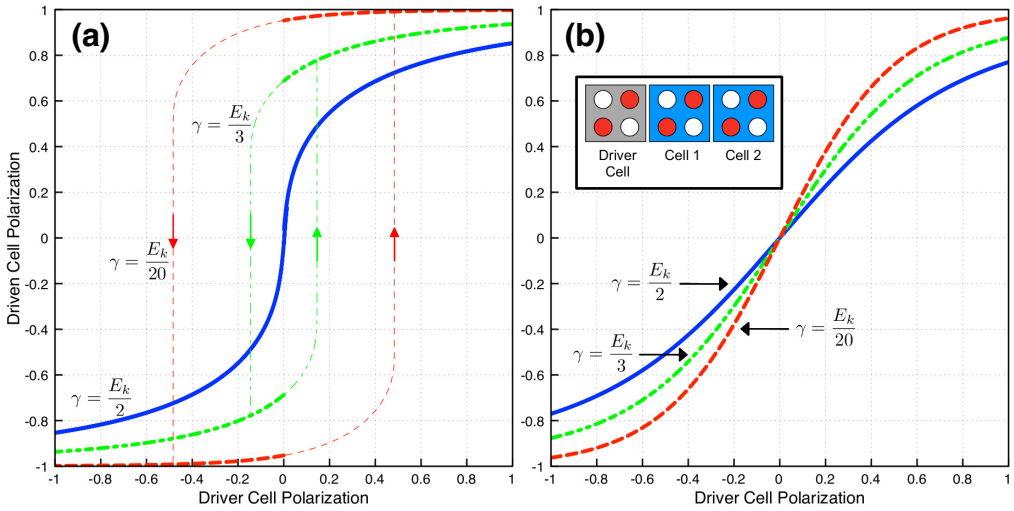


Figure 7.7: Simulations of a driven two-cell wire. (a) Polarization of Cell 1 as a function of the driver cell polarization, calculated using the ICHA. The ICHA predicts a hysteretic response indicating a memory effect. Where two solutions exist, the one with lowest energy is shown in bold. (b) Polarization of Cell 1 as a function of the driver cell polarization, calculated using the more complete quantum mechanical treatment described by equation 7.3. This more complete simulation shows no hysteresis. We have taken $k_B T \approx E_k/4$, which corresponds to $E_k \approx 100$ meV for room temperature. Inset shows the two-cell wire being simulated in both cases. (Figure adapted from Taucer *et al.* (2015).[3])

method by using the solution from the previous *iteration* as an initial guess for the current one. This amounts to the common sense approach of using the solution from a very similar Hamiltonian (i.e. very similar driver polarization) as an initial guess for some new Hamiltonian. With this method, Cell 1's response to the driver cell follows the hysteresis curve shown in Figure 7.7a. There is a retained polarization even as the driver polarization goes to zero, indicative of a memory effect.

When the driver cell is turned on, both the driven cells align themselves with the driver polarization. As the driver cell polarization is removed, the coupling between the two driven cells (through expectation values) allows them

to retain part of their polarization even as the driver cell polarization reaches zero. Only after a sufficiently strong driver polarization in the *opposite* direction do both cells switch polarizations. The strength of the driver polarization required to switch the driven cells depends solely on the ratio γ/E_k; the lower the tunnelling rate, the larger the residual polarization of the driven cells as the driver cell's polarization is removed.

The other method that can be used with the ICHA is to sample the space of polarizations in search of the self-consistent solution with the lowest energy. For certain driver cell polarizations, there are two self-consistent solutions. In Figure 7.7a, the lowest energy solution is shown in bold. If the lowest energy solution is always used, the ICHA predicts a discontinuity in the response of the driven cells at zero driver cell polarization. As the driver's polarization crosses zero, the cells respond by abruptly "snapping" to the other polarization state. The ICHA is not typically used in this way, however.

Simulations conducted using a more complete quantum mechanical treatment show no hysteresis, as shown in Figure 7.7b. Here, as the driver polarization is removed, Cell 1 also relaxes to zero polarization. Thus, there is a "depolarizing" effect that is not predicted when treating QCA systems using the ICHA. Also, the response is continuous, though still non-linear and with a slope greater than unity at the origin.

The ICHA has been widely used to show successful propagation of bit packets, and information processing, in clocked QCA wires[82] and circuits.[105–107] On the basis of the results shown here, it appears that this is the result of the use of the ICHA with the first method outlined above — that is, the one that yields a hysteresis curve. The ICHA represents, in a sense, the minimum inclusion of quantum mechanical effects and we will see in the following sections that this approximation can lead to errors when applied to clocked QCA, as compared with the many-cell Hamiltonian. It is possible to make a less drastic simplification of the system dynamics by including some, but not all, intercell correlations, for example by including only nearest-neighbour pair correlations.[99,100] Including some of the correlations has been shown to give the correct answer in some cases where the ICHA fails. In Section 7.4 we will

show a case in which the ICHA fails, and the inclusion of nearest neighbour correlations does *not* help.

7.2.3 Relaxation Time Approximation

A final approximation to be discussed here is the relaxation time approximation, which is commonly used in simulations of QCA dynamics. In the absence of energy dissipation and other decohering effects, a QCA array will evolve coherently according to the Liouville equation for the density matrix, $\hat{\rho}$:

$$\frac{d}{dt}\hat{\rho}(t) = \frac{1}{i\hbar}\left[\hat{H}(t), \hat{\rho}(t)\right], \tag{7.6}$$

which, for a pure state, is exactly equivalent to the Schrödinger equation. The Hamiltonian in Equation 7.6 may be time-dependent, e.g., when driver cells are switched or when cell parameters are changed to implement clocking.

Over fairly short time scales, quantum mechanical systems often fall to a thermal steady state.[82] If the QCA system is weakly-coupled to the environment, and the energy transfer between the system and environment is well-described by a Markov process, then at low temperatures, the simplest way to incorporate energy dissipation into a model of QCA dynamics is via the relaxation time approximation.[82,99,108,109] This is done by adding a dissipation term to Equation 7.6:

$$\frac{d}{dt}\hat{\rho}(t) = \frac{1}{i\hbar}\left[\hat{H}(t), \hat{\rho}(t)\right] - \frac{1}{\tau}\left(\hat{\rho}(t) - \hat{\rho}_{ss}\right), \tag{7.7}$$

where τ is a phenomenological time constant, and ρ_{ss} is the steady-state matrix defined as

$$\hat{\rho}_{ss} \equiv \frac{e^{-\hat{H}(t)/k_B T}}{\mathrm{Tr}\left\{e^{-\hat{H}(t)/k_B T}\right\}}. \tag{7.8}$$

Determining the steady state density matrix exactly is tantamount to solving the complete Schrödinger equation for the system, and thus comes up against all the difficulties mentioned above. It is therefore usually calculated using the

ICHA and the two-state approximation.

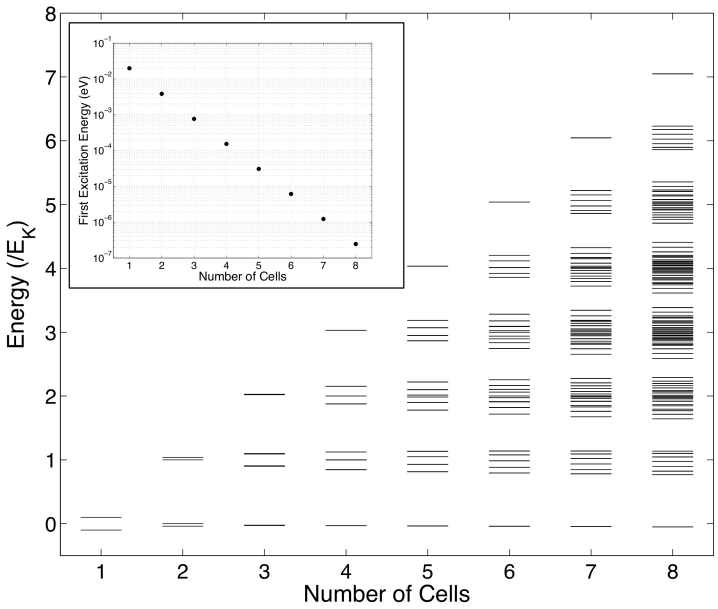
Equation 7.7 imposes an exponential approach of the density matrix towards its steady-state value, with a time constant τ. This is one example of a quantum master equation, or an equation of motion for the density matrix, which in this case is phenomenological. It would be useful to derive a quantum master equation based on the microscopic details of the system, its environment, and their interaction. This would show whether or not the form of Equation 7.7 is correct, and how the value of τ relates to microscopic parameters. However, a precise quantum master equation will in general be implementation dependent. In this paper, we treat an idealized QCA that is not tied to a specific implementation. Furthermore, regardless of the specifics of the quantum master equation, the density matrix in Equation 7.8 will very often represent the real steady state of atomic and molecular systems. We therefore attempt to calculate the correct steady state behaviour, acknowledging that the dynamics and the specific value of τ will be implementation-dependent.

7.3 Full Quantum Mechanical Calculations

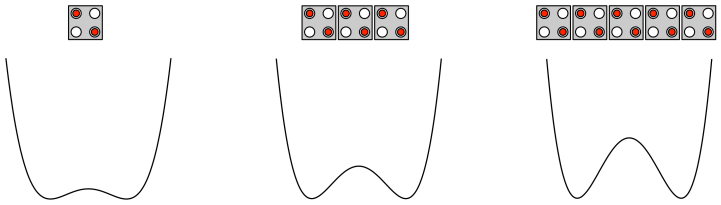
Having developed an appreciation for the effects of the ICHA and of the relaxation time approximation on the calculated dynamics of QCA arrays, we will briefly consider the qualitative features of some tractable systems of interacting QCA cells by once again considering the two-state Hamiltonian described in Equation 7.3.

The difficulty in solving Equation 7.3 can be reduced by assuming only nearest-neighbour coupling so that $E_k^{i,j} = E_k \delta_{i\pm1,j}$ and $E_k^{i,D} = 0$ for cells that are not adjacent to the driver labelled D. This does not imply that *correlations* beyond nearest neighbours are ignored, however. We also restrict our attention to linear chains of cells.

The spectrum for an unbiased line of N (ranging from 1 to 8) QCA cells, within the two-state approximation is shown in Figure 7.8a. For a line of N cells, there are 2^N eigenstates. Under the conditions required for QCA-based



(a) Line Spectra



(b) Analogous Double Well Systems

Figure 7.8: (a) Spectra of unbiased lines of interacting QCA cells, ranging in length from one to eight cells, with $\gamma/E_k = 0.1$. The spectra are solutions of Equation 7.3, with a constant $(N-1)E_k/2$ added for ease of interpretation. The inset shows the difference in energy between the two lowest energy levels for each line. (b) Conceptually, the array can be viewed as a single two state system with a barrier that increases with the number of cells. (Figure from Taucer *et al.* (2015).[3])

computing, that is $\gamma << E_k$, the energy levels come in clusters separated roughly by E_k. There are always two non-degenerate lowest-energy states. In the case of a single cell, these are in fact the only two states, and their

separation is exactly 2γ, just as in the case of the covalent bond, described in Chapter 1. The ground and excited state correspond to the symmetric ($|\psi_s\rangle = \frac{1}{\sqrt{2}}(|0\rangle + |1\rangle)$) and anti-symmetric ($|\psi_a\rangle = \frac{1}{\sqrt{2}}(|0\rangle - |1\rangle)$) combinations of the polarization states (with the former being the ground state). Note that the polarization basis vectors, $|0\rangle$ and $|1\rangle$, are *not* energy eigenstates of the undriven cell. If a single quantum measurement of $\hat{\sigma}_z$ is carried out on a QCA cell in its ground state (or in its first excited state for that matter), the outcome will yield either -1 or $+1$, with equal probability, *i.e.*, $P = 0$. In this sense, one can say that the ground state of a single unbiased cell carries no information. Only in the limit where $\gamma \to 0$ do the polarization basis vectors become valid energy eigenstates.

For longer lines, the separation between the two lowest-lying energy eigenstates becomes smaller and they represent entangled states. Specifically, they are the symmetric and anti-symmetric combinations of the state with all cell polarizations aligned along one diagonal, and the state with them aligned along the other. That is, $|\psi_{s,a}\rangle \approx \frac{1}{\sqrt{2}}(|000\ldots0\rangle \pm |111\ldots1\rangle)$. The equality becomes exact in the limit where $E_k/\gamma \to \infty$.

Again, note that the "aligned" states, $|000\ldots0\rangle$ and $|111\ldots1\rangle$, are not the energy eigenstates. The energy eigenstates are in fact superpositions of the aligned states, and the polarization of any cell in a line in its ground state (or first excited state) is $P = 0$. This indicates that the eigenstates do not carry information. The inset of Figure 7.8a shows that the splitting between these two lowest-lying states decreases exponentially. Specifically, each added cell causes the splitting to decrease by a factor of $\sim 2\gamma/E_k$, so that the first excitation energy, Δ, for a line of N cells is

$$\Delta(N) \approx E_k \left(\frac{2\gamma}{E_k}\right)^N . \tag{7.9}$$

Only in the limit where this splitting goes to zero (*i.e.* an infinite number of interacting cells), do the aligned states become energy eigenstates.[103] With the decrease in the splitting of the two lowest energy levels, comes an increase in the time required for coherent tunnelling from one polarization state to the other. For low temperatures, we can think of the group of cells as a

two-state system, analogous to a double well, as depicted in Figure 7.8b. As the length of the line increases, the barrier separating the two aligned states increases. It follows that longer lines acquire increased bistability. However, at finite temperature, thermal fluctuations may cause excitations and an eventual approach to the unpolarized steady state.

Figure 7.9 shows the unitary time evolution of unperturbed lines of one, three, and five cells initially in an aligned state in each case. All the cells in the wire oscillate, in unison, between the two polarization states. This simulation represents the limit of infinite relaxation time (*i.e.*, no energy dissipation or loss of phase coherence), and is meant to indicate the internal dynamics of the system. A realistic system will have its dynamics altered by its interaction with the environment. Depending on the interaction and its strength, this can lead to several behaviours[110] including stochastic dynamics, relaxation to the ground state (as discussed here), or a stabilization of the polarization, as suggested in Blair and Lent.[111]

Regarding the internal dynamics of the system, it is clear that longer lines exhibit increased bistability, meaning that the polarization state can be maintained for an arbitrary length of time by increasing the number of cells in the line, even in a completely coherent system. For single cells and small groups at the atomic scale, coherent oscillations will likely be much faster than the measurement time of a classical apparatus, and therefore will lead to a loss of classical information. If such sinusoidal oscillations do exist, it is unlikely that they would proceed with phase coherence, and so even the phase information would quickly be lost. For larger groups, the loss of information will likely be limited by decoherence and energy relaxation.

Finally, we emphasize that short unbiased lines, like individual cells, have a *unique* ground state. This contrasts with the commonly made assertion that cells and lines have two degenerate states,[112] namely, the aligned states. The ground state is an entangled state, and is therefore not accessible to the ICHA. Based on this observation, as well as the above-mentioned shortcomings of the ICHA in predicting the behaviour of even very short lines, we are led to a more in-depth investigation of QCA line dynamics, particularly with regard

to clock zones which are spatially separated from any driver cells.

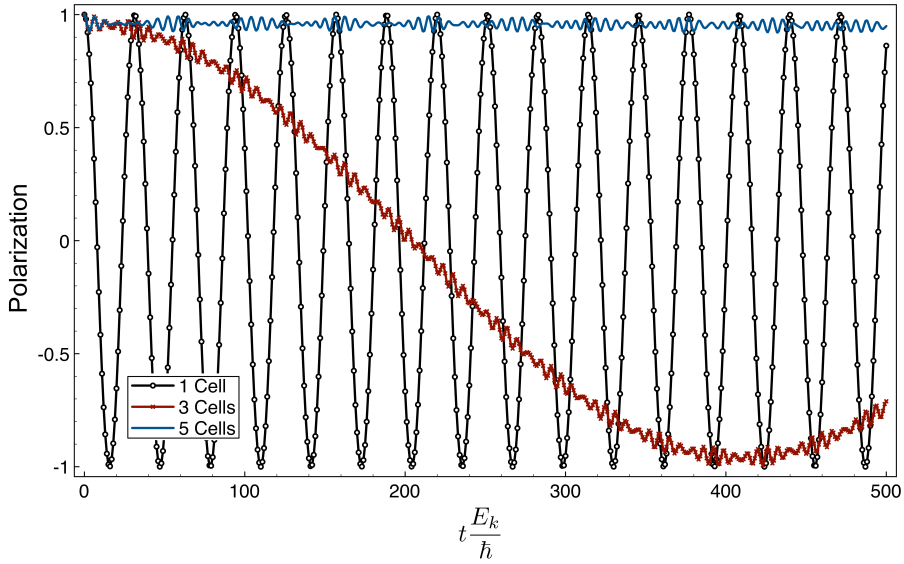


Figure 7.9: Coherent oscillations of the first cell in unperturbed one, three, and five cell lines. The cells are all initially in the $P = 1$ state. In each line, the cells oscillate together, so the polarizations of the first cell gives a good representation of the polarization of all its neighbours. The fast oscillations in the 3 and 5 cell lines are due to a small component of higher-energy states, which manifest as kinks propagating and reflecting through the line. $\gamma = 10$ meV and $E_k = 100$ meV. (Figure from Taucer *et al.* (2015).[3])

7.4 Loss of Polarization in Isolated Bit Packets

As discussed in Section 7.2.2, the ICHA predicts a latching mechanism within a line of interacting cells that allows them to polarize (and retain this polarization) in the absence of a fixed driver cell. This phenomenon is found to be an artifact of the ICHA, not necessarily representative of actual dynamics, at least as described by more complete Hamiltonians like Equation 7.3 and Equation 7.1. The result of this latching mechanism is seen in simulations that use the ICHA to model the propagation of information in QCA circuits. One such example is shown in Figure 7.10, from Timler and Lent (2003).[113]

This figure shows the result of a calculation done using the ICHA, showing lossless propagation of bit-packets down a seven-cell QCA wire. In this section, we contrast these simulations with new results that include intercellular correlations.

Consider a single, unbiased QCA cell. At low temperatures, the QCA cell will relax to its ground state, $|\psi_s\rangle = \frac{1}{\sqrt{2}}(|0\rangle + |1\rangle)$, and thus will have a polarization, $P = 0$. At higher temperatures, interactions with the environment can cause the QCA cell to be in a mixed state, and it is best described by a density matrix. If the steady state density matrix of the system is the one described in Equation 7.8, then it is easy to show that the expectation value of σ_z is zero, and therefore the polarization is also zero. For longer lines, the same reasoning applies, the only differences being that $|\psi_{s,a}\rangle$ represent the entangled states described in Section 7.3, and the splitting in their energies is decreased. It can be shown that the polarization of any cell in the unbiased line is zero if the line is in the steady state described by Equation 7.8.

Figures 7.11 and 7.12 show the results of two simulations of six cells in a one-dimensional chain, with periodic boundary conditions and nearest neighbour coupling only. γ is modulated between 1 meV and 200 meV for the first simulation, and between 1 meV and 1000 meV in the second. $E_k = 108.5$ meV in both simulations. The Hamiltonian in Equation 7.3 is used, so that intercellular correlations are completely included. In both simulations, the tunneling barriers in cells 1, 2, and 3 are initially high (meaning that γ is low) while the tunneling barriers in cells 4, 5, and 6 are low (so that γ is high). As time progresses, the tunneling barriers in cell 4 are raised, allowing it to polarize, and the barriers in cell 1 are lowered, which causes it to depolarize. Next, cell 5 has its barriers raised as the barriers in cell 2 are lowered, and so on. Because of the periodic boundary conditions, the bit packet moves cyclically through the six cells. Throughout the process, the Hamiltonian changes quasi-adiabatically, by which we mean that $d\gamma/dt \ll E_k^2/\hbar$. For our initial state, we create a nearly full negative polarization in the three active cells by taking the normalized sum of the two lowest energy eigenstates. The resulting cell polarizations are plotted as a function of time with and without dissipation in

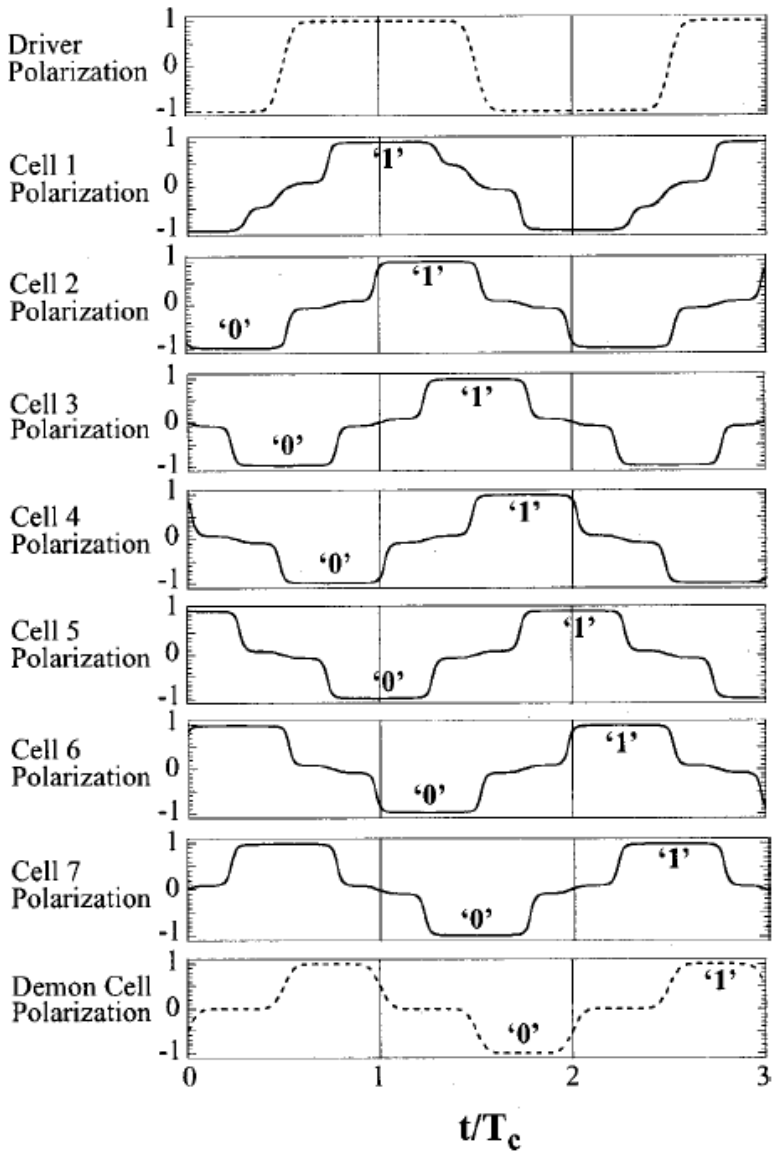


Figure 7.10: Propagation of information (alternating sequence of 0's and 1's) down a 7-cell QCA wire, simulated using the ICHA. For each cell, polarization is plotted as a function of time in units of the clock period. In addition to the seven clocked QCA cells, there is a driver cell acting as an input adjacent to Cell 1, and a "Demon Cell," which is a fictitious computational aid, added to prevent spurious reflections resulting from the finite wire length. Reprinted figure with permission from Timler and Lent, *Journal of Applied Physics* **94**, 1050, 2003.[113] Copyright (2003) AIP Publishing LLC.

each case. A third curve plotting the exponential decay, proportional to $e^{-t/\tau}$, is also shown in each plot as a reference.

In the simulation of Figure 7.11, as time progresses, the bit packet travels along the periodic line, and a polarization is maintained when there is no dissipation. Because the Hamiltonian is changed quasi-adiabatically, no kinks are created as the bit packet moves, and the bit packet evolves qualitatively the same as it would if it were stationary. Because the tunneling barriers are never lowered completely (γ has a maximum value of 200 meV), the cells never reach zero polarization. The interaction between the three active cells at a given time, plus the residual polarization in the "inactive" cells, slows the coherent oscillations to the point where the bit packet is fully polarized over the period of the simulation, as long as there is no dissipation. However, the steady state still has zero polarization in all cells, so the bit packet loses polarization exponentially when relaxation is included in the model. The slow oscillation frequency demonstrated in this simulation simply permits the cells to maintain their polarizations over several clock cycles.

In the simulation of Figure 7.12, since the tunnelling barriers are lowered enough to completely depolarize the cells (γ is now allowed as high as 1000 meV), fewer cells are "on" at any given time, and the coherent oscillations from negative to positive polarization state are sufficiently fast to be observed over the time scale of the simulation. As in Figure 7.11, the dissipation of energy does indeed bring the system exponentially to its steady state of zero polarization. In this simulation, because of the visible coherent oscillations, the polarization does not follow the decaying exponential as it does in Figure 7.11, but instead the amplitude of the oscillation is proportional to the exponential decay.

7.5 Discussion

The simulations in Section 7.4 show that the ICHA does not always provide a good approximation to the full quantum mechanical model of an isolated array of QCA cells. In the case of bit packets which are not coupled to a driving cell,

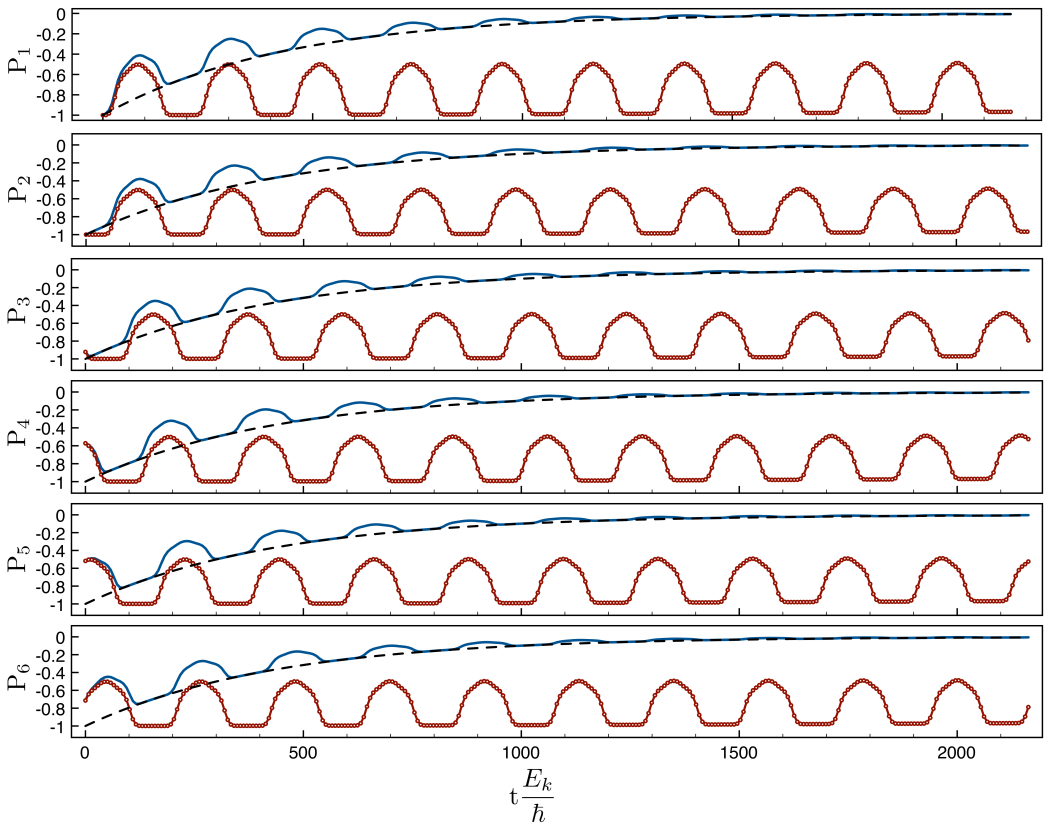


Figure 7.11: Simulations of a six cell line with periodic boundary conditions. The value of γ is adiabatically raised and then lowered between $1 < \gamma < 200$meV, with a period of $217\hbar/E_k$, with a different phase for each cell. The circled line shows the polarization of each cell in the absence of any energy dissipation. The solid line shows the polarization of each cell with an energy relaxation time of $434\hbar/E_k$. For reference, a decaying exponential as a function of τ is shown with the dashed line. (Figure from Taucer *et al.* (2015).[3])

the exponential relaxation causes any initial polarization to be lost. Within the framework of Equations 7.3, 7.7, and 7.8, the maximum time for a clocked computation will be limited by the loss of classical information either through energy relaxation or through oscillations of the type shown in Figures 5 and 7. In principle, coherent oscillations do not imply a loss of information (the information can be retrieved by a carefully timed measurement), however, the oscillations may be much more rapid than the duration of a measurement, *e.g.*

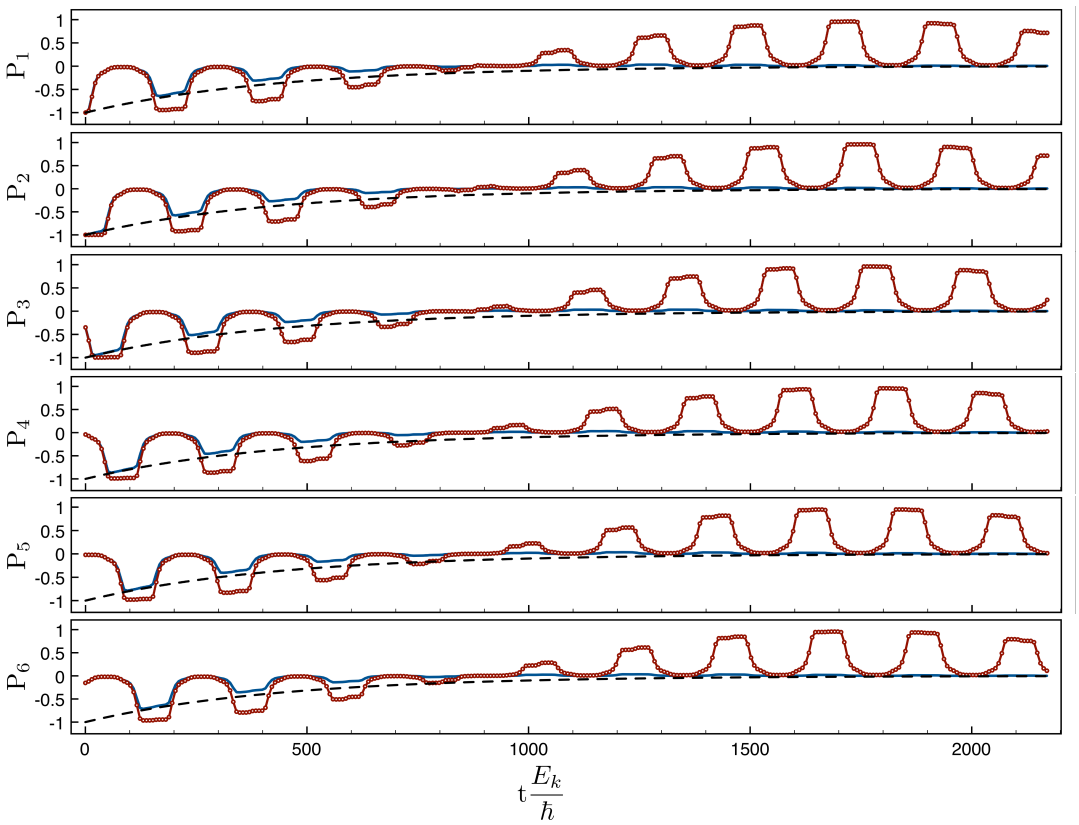


Figure 7.12: Simulations of a six cell line with periodic boundary conditions, using the same parameters as Figure 7.11, except that the value of γ varies in the range $1 < \gamma < 1000$meV, with a higher upper bound. Otherwise, line colours and styles are the same as in Figure 7.11. Half a period of a coherent oscillation can be seen over the course of the simulation, flipping the polarization state from negative to positive. (Figure from Taucer *et al.* (2015).[3])

in the case of a small number of cells, and will in general suffer phase decoherence, which will erase even this phase information. All the dynamic simulations presented here were done within the two-state approximation of equation 7.3, which is an approximation of equation 7.1. As mentioned previously, working directly with equation 7.1 is very computationally expensive. However, based on the conceptual understanding outlined in this chapter, in addition to other computational results on very small groups of cells (not shown here), it is clear that these new features of clocked QCA are not an artifact of the two-state

approximation, but are the result of the inclusion of intercellular correlations and the use of the relaxation time approximation.

Previous simulations have used the ICHA to predict the latching of lines and the successful propagation of clocked pulses across QCA arrays (even when tunnelling rates were varied by only one order of magnitude[82]). These have seen some degree of validation from experimental work on lithographically defined QCA-like systems.[89,90] The results presented here show that these experiments do not embody the dynamics of equations 7.3 and 7.7. The use of intermediate quantum dots and multiple tunnel junctions in these experiments makes it possible to tune the tunneling time from 10 ps to $\sim$ 3000 s.[89] This drastic change in tunneling rates amounts to a crossover from a regime where quantum tunneling is important to one where electrons are completely localized. The normal relaxation dynamics are suppressed because the system is strongly coupled to the environment through σ_z, which has the effect of localizing charge.[110] In molecular and atomic implementations, it is likely that the range of tunneling rates that allows this crossover from quantum dynamics to classical dynamics by directly modulating tunnel barriers, will not be achievable. A more promising approach to achieving such a crossover would be to change the internal dynamics of a bit-packet by changing its size. In section 7.3, we showed that longer lines exhibit slower dynamics and greater bistability. This suggests that for sufficiently large bit packets, there may be sufficient bistability that the coupling to the environment results in localization of charge, even with a limited range of tunneling rates available. The precise tunneling rate and bit packet size that will allow localization of charge with sufficiently slow relaxation dynamics will be determined by the nature and strength of the coupling of the environment to the QCA system, which will in turn compete with the QCA system's internal dynamics. A detailed analysis will need to take into account the microscopic details of a specific implementation.

The simulations performed here clearly identify depolarization due to quantum correlations as a critical issue for classical computation using clocked QCA at the molecular and atomic scale. However, it is important to note that this

loss of polarization occurs only when a set of cells becomes isolated from a perturbing influence such as a fixed driver cell, and thus, QCA systems operating within a single clocking zone will not experience such an effect.

It might be argued that for some QCA implementations the ICHA may actually provide a good assumption about the physics of the system, rather than an approximation to the many-cell Hamiltonian. This amounts to assuming that the effect of the interaction of the system with its environment is to completely eliminate quantum correlations between cells, while leaving superpositions and entanglement within a cell (that is, entanglement between the two electrons within the cell) undisturbed. Notwithstanding the arbitrary cutoff — allowing entanglement completely within each cell but not at all between cells — such an interaction is at least conceivable. But it would raise a new set of problems. In certain cases where the ICHA is known to "fail,"[100] such as the majority gate with uneven input legs which can become trapped in a metastable state,[99] we would then no longer have recourse to intercell correlations. The metastability of the state with the wrong logical output would be a genuine prediction of the model. A QCA system which obeys the ICHA will probably require fine-grained clocking to address this problem – an increasingly impractical solution as cells are miniaturized. Intercellular correlations, which we expect to play a role in the dynamics of molecular and atomic QCA systems, alleviate this problem, however the results that we have presented here show that they may present us with new problems which need to be considered in QCA theory and design.

7.6 Conclusion

In this chapter, we have assessed some limitations of conventional approaches to QCA simulation. Full quantum mechanical calculations show that the ground state of an unbiased cell, or of a line of cells, is a superposition of the two fully aligned states, and thus holds no polarization and carries no classical information. When the assumption of exponential relaxation to a thermal steady state is made, we find that cells, or groups of interacting cells, lose their

polarization over time unless they are influenced by a fixed driver. This is the case even if the cells start with an initial polarization before being decoupled from a fixed driver cell. This depolarization effect was not predicted in previous QCA simulations, which had often predicted a false latching mechanism among cells that would allow them to retain their polarization even in the absence of a driver cell. This discrepancy is related to the ICHA which neglects correlations and shows hysteresis in array polarization. Although lithographic QCA systems have managed to avoid this problem thanks to their inherently long tunnelling times, the molecular and atomic implementations of QCA required for room temperature operation will likely behave in a more purely quantum mechanical way, so that the solutions of the many-cell Hamiltonian need to be included. Only an appropriate and sufficiently strong interaction of a QCA array with its environment will make clocked QCA possible.

While these findings do not affect the original concept of ground state computing with QCA, they may require a reconsideration of QCA architecture for molecular and atomic QCA, specifically relating to clocking and memory devices. Because the simulations presented here are still for a very highly idealized model of QCA behaviour, and ignore, among other things, the specifics of the interaction with the environment, it would be wrong to conclude that clocking and memory are impossible in QCA. Effective clocking requires a tuneable change from quantum mechanical behaviour to classical. It remains to be shown how and if this can be achieved at the molecular scale. This underscores the need for a more sophisticated theory of QCA operation, which should include implementation-specific dynamics beyond the phenomenological relaxation time approximation.

8 Conclusion

This thesis has focused primarily on isolated dangling bonds on the hydrogen-terminated silicon surface. These DBs occur when a hydrogen atom is removed from the surface, or when a hydrogen atom happens to be missing from an otherwise terminated surface, leaving an unbonded, localized orbital with a mid-gap energy level. In principle, DBs can be variably occupied, depending on the Fermi energy relative to the DB transition levels. In empty state STM imaging, DBs are surrounded by a darkened region known as a halo. The DB halo was associated with a negative charge in early descriptions of STM of DBs. In Chapter 3, we saw that in fact there are a number of effects at play when the DB is imaged in empty state STM imaging, and the acquired STM image is far from being a simple picture of a negatively charged DB. The topography around a DB can only be understood in detail as the result of multiple factors: tip-induced band bending, direct injection of electrons, and processes which tend to bring the sample toward an equilibrium state. The result is a non-equilibrium picture of STM of DBs where filling and emptying processes compete to determine the charge state of the DB. The rates associated with these processes depend on applied bias as well as the tip position, so that the DB charge state is not a static feature of the sample, but rather responds dynamically as the tip scans the DB.

It was shown in Chapter 4 that these non-equilibrium dynamics, initially postulated to explain the topography of DBs in empty-state imaging, are in fact directly observable in STM imaging of DBs at low temperature (~ 4.2 K). The tip-sample tunnel junction, with the tip placed roughly at the edge of the DB halo, can act as a single-electron sensitive charge detector; the fluctuating charge of the DB has a gating effect on the tip-sample tunnel junction, causing

the tip-sample current to undergo random jumps when one measures current as a function of time. Analysis of the $I(t)$ traces reveal three distinct plateaux, corresponding to the three charge states of the DB, negative, neutral, and positive. The fraction of time spent in each of these charge states depends on the tip position and the tip bias. An analysis of the current-time traces reveals the underlying transition rates connecting the different charge states of the DB. These rates are therefore the rates of transfer of single electrons from tip to DB, and from DB to bulk. The single electron transition rates were found to depend on bias and tip position in a way that is consistent with injection of electrons by tip-DB tunneling, and emptying of electrons toward the bulk via tunneling from the DB level to the extended states of the conduction band. These observations broadly corroborate the model of STM imaging proposed to explain topography.

The analysis required to extract single electron rates from $I(t)$ traces required the application of novel analysis methods for STM data, which were covered in Chapter 5. Because of the inherent noise in sensitive STM measurements, the signals due to different states of the system may overlap, making it difficult or impossible to apply simple analysis methods, like thresholding. Instead an analysis method, based on techniques used in biophysics, was developed. This method allows extraction of the rates involved in multi-state dynamics even when states are only marginally resolvable in experiments.

While the first part of this thesis was concerned with the dynamics at play in STM imaging of DBs, the latter part of the thesis dealt with issues aimed at applications. Chapter 6 described fabrication of DBs, starting with the process of image analysis to extract the periodicities of the lattice. It went on to describe some of the small structures that can be created, as well as large-scale patterns, up to thousands of DBs. Chapter 7 described an exciting potential application of DBs, quantum-dot cellular automata. QCA is an alternative method of doing classical computation, which operates without the need for any macroscopic currents. The particular issue of quantum correlations was explored with regard to the dynamics of small systems, and their approach to thermal steady state. This is an issue which was largely ignored for the

large, lithographic QCA structures that have been fabricated in the past two decades, but as QCA is miniaturized to the atomic scale, either through atomic implementations based on silicon DBs, or through other proposed routes to molecular QCA, quantum effects will become more important, and the nature and behaviour of the system will need to be understood from a quantum mechanical perspective.

My hope is that the work described in this thesis might help in painting a clearer picture of these silicon DBs, as well as what precisely is going on when we image them in STM, and how they might be employed in emerging atom scale technologies. As with many academic subjects, the scope of this work is in some senses very specific and limited. Nonetheless, it is sometimes possible for a clear understanding of the specific to offer a clear view of the more general. Many of the ideas applied to DBs on the H-Si(100) surface likely apply also to other mid-gap states. As similar analyses are applied to different but related problems, a clearer picture of non-equilibrium dynamics will likely emerge.

Even within the system described here, there are a number of questions that remain unanswered, which we have the ability to address immediately. The quantitative data on single electron processes in empty state imaging, presented in this thesis, are hopefully compelling, and they are certainly consistent with many of our previous notions of non-equilibirum dynamics. But these measurements can certainly be improved upon, and far more detailed measurements could provide a wealth of information on the shape and extent of the DB orbitals (neutral and negative), as well as the processes that empty these into the bulk. Furthermore the measurements shown in this thesis are a very small step away from deliberate control of the charge state of a DB, which could open the door to a range of interesting experiments.

A study of the filled state spectroscopy of these same DBs on n-type silicon has already shown that similar non-equilibrium considerations are required to make sense of the current which flows from bulk to DB to tip.[38] In that case, the non-equilibrium dynamics tell us about the connection of DBs to the bulk silicon, as well as transport through the disordered electron donors near the

surface.

Non-equilibrium dynamics can also be studied at entirely different time scales by making use of all electronic pump-probe techniques,[53] which will offer a different and complementary view on these same issues. Such pump-probe experiments are still relatively novel, and have not, to date, been applied to charge dynamics. DBs present a system in which electron dynamics are known to extend from Hz, to kHz, and beyond (where we lose our ability to resolve them, because of the limit of the pre-amplifier bandwidth — a limit which is circumvented by pump-probe techniques). They therefore seem to be fertile ground for these first studies of fast charge dynamics. This too is being investigated at present.

The method used to analyze telegraph noise in tunneling currents was developed to address a very specific problem. Nonetheless, it is in principle very broadly applicable, not only within the field of STM, but also in any case where random fluctuations are observed through measurement of a noisy signal. The difficulty in adopting this method at present lies in the fact that it is complicated, or at least seems to be, which creates a barrier to adoption. More widespread use of this technique would require, first, a clear explanation of the technique, and second, ideally, a procedure for automating the analysis. Hopefully, this thesis has done something for the first. Efforts to address the second issue, of automation, are also underway.

It was shown in this thesis that large-scale DB patterns can be made with reasonable precision, and that structures of several atoms can be made with perfect atomic placement. There does not appear to be any fundamental impediment to nearly perfect patterning on the H-Si(100) surface at arbitrary scales (*e.g.* many thousands of atoms). Rather, there are a number of challenging issues, each of which may require a significant development effort to be overcome. Continued progress in the ability to fabricate DB structures will make the fabrication of increasingly complex structures increasingly routine. Demonstrations of technologically useful structures are possible already (an argument could be made that they have already been created), and undoubtedly many will be made in the coming years.

The ability to understand such structures is an entirely different subject, however. One outcome of the work presented here, which was perhaps not emphasized as much as it could have been, is that the STM tip, in some ways, gives an extremely distorted view of the sample — distorted, that is, by the effects of band bending and carrier injection/extraction. In particular, if we want to study and understand the potential for atomic technologies that rely on the interactions of single electrons in the sample, the STM is an odd choice of measuring apparatuses, since it is a tremendous disturbance to nearby electrons in the sample, creating a strong electric field, and perhaps worse, overwhelming the sample with a continuous stream of injected electrons or holes — roughly one billion per second. In addition to all of this, most STMs are limited in that they consist of a single probe. Issues of transport on the atomic scale can be studied indirectly,[114,115] but directly measuring the flow of current through atom-scale structures is usually not possible in STM.

There are exciting responses to these challenges, including the use of multiple-probe STM to study transport on the atomic scale in a direct way.[116] The issue of the perturbing effect of the STM, on the other hand, can be met by non-contact atomic force microscopy, which, like STM, gives atomically resolved information about the surface, but does not require the application of a bias, nor the flow of current between the tip and sample. These related techniques are sure to give new perspectives on the challenges facing the development of silicon DBs as a technology, and will likely provide a way of studying DB structures in a less perturbing way.

Of course, everything described so far is also part of a much broader effort. There is a feeling among many people that atom-scale technology is bound to play a role in the future of technology. This could take many forms, from transistor technology with dopants placed with atomic precision,[117] to single-atom single electron transistors,[70] to quantum computers based on the nuclear spin of embedded impurities in isotopically pure silicon.[118] Several materials are also candidates including not only silicon,[119,120] but also III-V semiconductors,[121] and physisorbed species on insulating layers grown on metal substrates.[53,122] Each system has different features of interest, and presents different challenges,

but all are interesting.

DBs are a versatile building block for atom-scale technologies at this early stage of the development of the field. They are also a necessary intermediate step in a process for embedding phosphorus donors in silicon, where STM tip-induced hydrogen desorption selectively enables the chemical adsorption of phosphine molecules at pre-determined sites, giving sub-nanometer control over donor placement. Furthermore, the study of DBs, it has been found recently, sheds light on the dynamics at play in the near-surface substrate, including dynamics associated with subsurface electron-donors.[38] Near-surface dopants are relevant to a number of technological applications, and also present a range of surprising behaviours that require further exploration. DBs provide a unique perspective on the issues associated with these near-surface dopants.

This thesis presents a very small part of the broad effort to realize atom-scale devices. Silicon DBs are promising candidates for such emerging technologies, and they are intimately connected with other silicon-based approaches. The surprisingly rich physics of these localized orbitals, studied in STM, has been one of the main topics of this thesis. As our understanding and capabilities grow, there is every reason to expect increasingly compelling demonstrations of single-electron control. The study of multi-DB structures also promises tremendous opportunities and undoubtedly many surprises along the path to technological applications.

Bibliography

[1] L. Livadaru, J. L. Pitters, M. Taucer, and R. A. Wolkow. "Theory of nonequilibrium single-electron dynamics in STM imaging of dangling bonds on a hydrogenated silicon surface". *Physical Review B* **84** (2011), p. 205416.

[2] M. Taucer, L. Livadaru, P. G. Piva, R. Achal, H. Labidi, J. L. Pitters, and R. A. Wolkow. "Single-Electron Dynamics of an Atomic Silicon Quantum Dot on the H-Si(100)-2x1 Surface". *Physical Review Letters* **112** (2014), p. 256801.

[3] M. Taucer, F. Karim, K. Walus, and R. A. Wolkow. "Consequences of Many-Cell Correlations in Clocked Quantum-Dot Cellular Automata". *IEEE Transactions on Nanotechnology* **14** (2015), pp. 638–647.

[4] W. H. Zurek. "Decoherence and the transition from quantum to classical". *Physics Today* **44** (1991), pp. 36–44.

[5] W. H. Zurek. "Decoherence, einselection, and the quantum origins of the classical". *Reviews of Modern Physics* **75** (2003), pp. 715–775.

[6] E. Joos and H. D. Zeh. "The emergence of classical properties through interaction with the environment". *Zeitschrift für Physik B Condensed Matter* **59** (1985), pp. 223–243.

[7] J. von Neumann. *Mathematical Foundations of Quantum Mechanics*. Princeton: Princeton University Press, 1955.

[8] S. Datta. "Electrical resistance: an atomistic view". *Nanotechnology* **15** (2004), S433–S451.

[9] N. W. Ashcroft and N. D. Mermin. *Solid State Physics*. First. Brooks Cole, 1976.

[10] C. Cohen-Tannoudji, B. Diu, and F. Laloe. *Quantum Mechanics*. Paris: John Wiley & Sons, 1977, pp. 72–74.

[11] E. W. Muller and K. Bahadur. "Field Ionization of Gases at a Metal Surface and the Resolution of the Field Ion Microscope". *Physical Review* **102** (1956), pp. 624–631.

[12] L. Livadaru, P. Xue, Z. Shaterzadeh-Yazdi, G. A. Dilabio, J. Mutus, J. L. Pitters, B. C. Sanders, and R. A. Wolkow. "Dangling-bond charge qubit on a silicon surface". *New Journal of Physics* **12** (2010), p. 083018.

[13] Z. Shaterzadeh-Yazdi, L. Livadaru, M. Taucer, J. Mutus, J. L. Pitters, R. A. Wolkow, and B. C. Sanders. "Characterizing the rate and coherence of single-electron tunneling between two dangling bonds on the surface of silicon". *Physical Review B* **89** (2014), p. 035315.

[14] J. J. Sakurai. *Modern Quantum Mechanics*. Ed. by S. F. Tuan. Addison-Wesley, 1994.

[15] J. Tersoff and D. R. Hamann. "Theory of the scanning tunneling microscope". *Physical Review B* **31** (1985), pp. 805–813.

[16] G. Binnig and H. Rohrer. "Scanning tunneling microscopy". *Helvetica Physica Acta* **55** (1982), pp. 726–735.

[17] G. Binnig, H. Rohrer, C. Gerber, and E. Weibel. "Tunneling through a controllable vacuum gap". *Applied Physics Letters* **40** (1982), pp. 178–180.

[18] G. Binnig, H. Rohrer, C. Gerber, and E. Weibel. "7x7 Reconstruction on Si(111) Resolved in Real Space". *Physical Review Letters* **50** (1983), pp. 120–123.

[19] G. Binnig, H. Rohrer, C. Gerber, and E. Weibel. "(111) facets as the origin of reconstructed Au(110) surfaces". *Surface Science Letters* **131** (1983), pp. L379–L384.

[20] G. Binnig, H. Rohrer, C. Gerber, and E. Stoll. "Real-space observation of the reconstruction of Au(100)". *Surface Science* **144** (1984), pp. 321–335.

[21] R. J. Hamers, R. M. Tromp, and J. E. Demuth. "Scanning tunneling microscopy of Si(001)". *Physical Review B* **34** (1986), pp. 5343–5357.

[22] R. M. Feenstra, J. A. Stroscio, J. Tersoff, and A. P. Fein. "Atom-Selective Imaging of the GaAs(110) Surface". *Physical Review Letters* **58** (1987), pp. 1192–1195.

[23] R. A. Wolkow. "Direct Observation of an Increase in Buckled Dimers on Si(001) at Low Temperature". *Physical Review Letters* **68** (1992), pp. 2636–2639.

[24] R. A. Wolkow and P. Avouris. "Atom-resolved surface chemistry using scanning tunneling microscopy". *Physical Review Letters* **60** (1988), pp. 1049–1052.

[25] P. Avouris and R. A. Wolkow. "Atom-resolved surface chemistry studied by scanning tunneling microscopy and spectroscopy". *Physical Review B* **39** (1989), pp. 5091–5100.

[26] D. M. Eigler and E. K. Schweizer. "Positioning single atoms with a scanning tunneling microscope". *Nature* **344** (1990), pp. 524–526.

[27] J. Bardeen. "Tunneling from a Many-Particle Point of View". *Physical Review Letters* **6** (1961), pp. 57–59.

[28] R. M. Feenstra. "Electrostatic potential for a hyperbolic probe tip near a semiconductor". *Journal of Vacuum Science & Technology B: Microelectronics and Nanometer Structures* **21** (2003), p. 2080.

[29] R. M. Feenstra, Y. Dong, M. P. Semtsiv, and W. T. Masselink. "Influence of tip-induced band bending on tunnelling spectra of semiconductor surfaces". *Nanotechnology* **18** (2006), p. 044015.

[30] L. C. Allen. "Interpolation Scheme for Energy Bands in Solids". *Physical Review* **98** (1955), pp. 993–996.

[31] M. L. Cohen and T. K. Bergstresser. "Band Structures and Pseudopotential Form Factors for Fourteen Semiconductors of the Diamond and Zinc-blende Structures". *Physical Review* **141** (1966), pp. 789–796.

[32] F. Herman. "The Electronic Energy Band Structure of Silicon and Germanium". *Proceedings of the IRE* **43** (1955), pp. 1703–1732.

[33] M. B. Haider, J. L. Pitters, G. A. DiLabio, L. Livadaru, J. Mutus, and R. A. Wolkow. "Controlled Coupling and Occupation of Silicon Atomic Quantum Dots at Room Temperature". *Physical Review Letters* **102** (2009), p. 046805.

[34] S. R. Schofield, P. Studer, C. F. Hirjibehedin, N. J. Curson, G. Aeppli, and D. R. Bowler. "Quantum engineering at the silicon surface using dangling bonds". *Nature Communications* **4** (2013), p. 1649.

[35] W. A. Harrison. "Surface Reconstruction on Semiconductors". *Surface Science* **55** (1976), pp. 1–19.

[36] M. Berthe, R. Stiufiuc, B. Grandidier, D. Deresmes, C. Delerue, and D. Stiévenard. "Probing the carrier capture rate of a single quantum level." *Science (New York, N.Y.)* **319** (2008), pp. 436–8.

[37] T. H. Nguyen, G. Mahieu, M. Berthe, B. Grandidier, C. Delerue, D. Stiévenard, and P. Ebert. "Coulomb Energy Determination of a Single Si Dangling Bond". *Physical Review Letters* **105** (2010), p. 226404.

[38] H. Labidi, M. Taucer, M. Rashidi, M. Koleini, L. Livadaru, J. Pitters, M. Cloutier, M. Salomons, and R. A. Wolkow. "Scanning tunneling spectroscopy reveals a silicon dangling bond charge state transition". *New Journal of Physics* **17** (2015), p. 073023.

[39] C. R. Leavens and G. C. Aers. "Tunneling current density within Tersoff and Hamann's theory of the scanning tunneling microscope". *Physical Review B* **38** (1988), pp. 7357–7364.

[40] W. Shockley and W. T. Read. "Statistics of the Recombination of Holes and Electrons". *Physical Review* **87** (1952), pp. 835–842.

[41] A. Sabbah and D. Riffe. "Femtosecond pump-probe reflectivity study of silicon carrier dynamics". *Physical Review B* **66** (2002), pp. 1–11.

[42] H. Kawai, F. Ample, Q. Wang, Y. K. Yeo, M. Saeys, and C. Joachim. "Dangling-bond logic gates on a Si(100)-(2x1)-H surface". *Journal of physics. Condensed matter : an Institute of Physics journal* **24** (2012), p. 095011.

[43] M. Rezeq, J. L. Pitters, and R. A. Wolkow. "Tungsten nanotip fabrication by spatially controlled field-assisted reaction with nitrogen". *The Journal of chemical physics* **124** (2006), p. 204716.

[44] J. J. Boland. "Scanning tunnelling microscopy of the interaction of hydrogen with silicon surfaces". *Advances in Physics* **42** (1993), pp. 129–171.

[45] J. L. Pitters, P. G. Piva, and R. A. Wolkow. "Dopant depletion in the near surface region of thermally prepared silicon (100) in UHV". *Journal of Vacuum Science & Technology B* **30** (2012), pp. 1–7.

[46] A. Hoffmann and M. T. Woodside. "Signal-pair correlation analysis of single-molecule trajectories." *Angewandte Chemie (International ed. in English)* **50** (2011), pp. 12643–6.

[47] A. Hoffmann, D. Nettels, J. Clark, A. Borgia, S. E. Radford, J. Clarke, and B. Schuler. "Quantifying heterogeneity and conformational dynamics from single molecule FRET of diffusing molecules: recurrence analysis of single particles (RASP)." *Physical chemistry chemical physics : PCCP* **13** (2011), pp. 1857–71.

[48] M. Berthe, A. Urbieta, L. Perdigão, B. Grandidier, D. Deresmes, C. Delerue, D. Stiévenard, R. Rurali, N. Lorente, L. Magaud, and P. Ordejón. "Electron Transport via Local Polarons at Interface Atoms". *Physical Review Letters* **97** (2006), p. 206801.

[49] K. Teichmann, M. Wenderoth, S. Loth, R. Ulbrich, J. Garleff, a. Wijnheijmer, and P. Koenraad. "Controlled Charge Switching on a Single Donor with a Scanning Tunneling Microscope". *Physical Review Letters* **101** (2008), p. 076103.

[50] K. Teichmann, M. Wenderoth, S. Loth, J. K. Garleff, A. P. Wijnheijmer, P. M. Koenraad, and R. G. Ulbrich. "Bistable charge configuration of donor systems near the GaAs(110) surfaces." *Nano letters* **11** (2011), pp. 3538–42.

[51] E. P. Smakman, P. L. J. Helgers, J. Verheyen, P. M. Koenraad, and R. Möller. "Tunable switching dynamics of a single Si dopant in GaAs(110)". *Physical Review B - Condensed Matter and Materials Physics* **90** (2014), pp. 1–5.

[52] J. Schaffert, M. C. Cottin, A. Sonntag, H. Karacuban, C. A. Bobisch, N. Lorente, J.-P. Gauyacq, and R. Möller. "Imaging the dynamics of individually adsorbed molecules". *Nat Mater* **12** (2013), pp. 223–227.

[53] S. Loth, S. Baumann, C. P. Lutz, D. M. Eigler, and A. J. Heinrich. "Bistability in atomic-scale antiferromagnets." *Science* **335** (2012), pp. 196–199.

[54] J. Schaffert, M. C. Cottin, A. Sonntag, H. Karacuban, D. Utzat, C. a. Bobisch, and R. Möller. "Scanning noise microscopy". *Review of Scientific Instruments* **84** (2013).

[55] J. L. Pitters, L. Livadaru, M. B. Haider, and R. A. Wolkow. "Tunnel coupled dangling bond structures on hydrogen terminated silicon surfaces." *The Journal of chemical physics* **134** (2011), p. 064712.

[56] B. Naydenov and J. J. Boland. "Engineering the electronic structure of surface dangling bond nanowires of different size and dimensionality." *Nanotechnology* **24** (2013), p. 275202.

[57] M. Kolmer, S. Godlewski, R. Zuzak, M. Wojtaszek, C. Rauer, A. Thuaire, J.-M. Hartmann, H. Moriceau, C. Joachim, and M. Szymonski. "Atomic scale fabrication of dangling bond structures on hydrogen passivated Si(001) wafers processed and nanopackaged in a clean room environment". *Applied Surface Science* (2013), pp. 1–7.

[58] M. Ciorga, a. Sachrajda, P. Hawrylak, C. Gould, P. Zawadzki, S. Jullian, Y. Feng, and Z. Wasilewski. "Addition spectrum of a lateral dot from Coulomb and spin-blockade spectroscopy". *Physical Review B* **61** (2000), R16315–R16318.

[59] C. B. Simmons, M. Thalakulam, N. Shaji, L. J. Klein, H. Qin, R. H. Blick, D. E. Savage, M. G. Lagally, S. N. Coppersmith, and M. A. Eriksson. "Single-electron quantum dot in Si/SiGe with integrated charge sensing". *Applied Physics Letters* **91** (2007), p. 213103.

[60] M. Field, C. Smith, M. Pepper, D. Ritchie, J. Frost, G. Jones, and D. Hasko. *Measurements of Coulomb blockade with a noninvasive voltage probe*. 1993.

[61] L. Gaudreau, S. A. Studenikin, A. S. Sachrajda, P. Zawadzki, A. Kam, J. Lapointe, M. Korkusinski, and P. Hawrylak. "Stability Diagram of a Few-Electron Triple Dot". *Physical Review Letters* **97** (2006), p. 036807.

[62] K. D. Petersson, J. R. Petta, H. Lu, and A. C. Gossard. "Quantum Coherence in a One-Electron Semiconductor Charge Qubit". *Physical Review Letters* **105** (2010), p. 246804.

[63] M. Korkusinski, I. P. Gimenez, P. Hawrylak, L. Gaudreau, S. Studenikin, and A. Sachrajda. "Topological Hunds rules and the electronic properties of a triple lateral quantum dot molecule". *Physical Review B* **75** (2007), p. 115301.

[64] L. Gaudreau, G. Granger, A. Kam, G. C. Aers, S. A. Studenikin, P. Zawadzki, M. Pioro-Ladrière, Z. R. Wasilewski, and A. S. Sachrajda. "Coherent control of three-spin states in a triple quantum dot". *Nature Physics* **8** (2011), pp. 54–58.

[65] M. Busl, G. Granger, L. Gaudreau, R. Sánchez, A. Kam, M. Pioro-Ladrière, S. A. Studenikin, P. Zawadzki, Z. R. Wasilewski, S. S. Sachrajda, and G. Platero. "Bipolar spin blockade and coherent state superpositions in a triple quantum dot." *Nature nanotechnology* **8** (2013), pp. 261–5.

[66] H. Ribeiro, G. Burkard, J. R. Petta, H. Lu, and A. C. Gossard. "Coherent Adiabatic Spin Control in the Presence of Charge Noise Using Tailored Pulses". *Physical Review Letters* **110** (2013), p. 086804.

[67] C. Lent, P. Tougaw, W. Porod, and G. Bernstein. "Quantum cellular automata". *Nanotechnology* **4** (1993), pp. 49–57.

[68] L. Lu, W. Liu, and M. O'Neill. "QCA systolic array design". *IEEE Transactions on Computers* **62** (2013), pp. 548–560.

[69] D. Loss and D. P. DiVincenzo. "Quantum computation with quantum dots". *Physical Review A* **57** (1998), pp. 120–126.

[70] M. Fuechsle, J. Miwa, S. Mahapatra, H. Ryu, S. Lee, O. Warschkow, L. C. L. Hollenberg, G. Klimeck, and M. Y. Simmons. "A single-atom transistor". *Nature nanotechnology* **7** (2012), pp. 1–5.

[71] A. Morello, J. J. Pla, F. A. Zwanenburg, K. W. Chan, K. Y. Tan, H. Huebl, M. Möttönen, C. D. Nugroho, C. Yang, J. A. van Donkelaar, A. D. C. Alves, D. N. Jamieson, C. C. Escott, L. C. L. Hollenberg, R. G. Clark, and A. S. Dzurak. "Single-shot readout of an electron spin in silicon." *Nature* **467** (2010), pp. 687–91.

[72] C. Yin, M. Rancic, G. G. de Boo, N. Stavrias, J. C. McCallum, M. J. Sellars, and S. Rogge. "Optical addressing of an individual erbium ion in silicon." *Nature* **497** (2013), pp. 91–4.

[73] E. P. Smakman, J. van Bree, and P. M. Koenraad. "Laser and voltage manipulation of bistable Si dopants in the GaAs (110) surface". *Physical Review B* **87** (2013), p. 085414.

[74] B. Kane. "Can We Build a Large-Scale Quantum Computer Using Semiconductor Materials ?" *MRS bulletin* **30** (2005), pp. 105–110.

[75] R. Robles, M. Kepenekian, S. Monturet, C. Joachim, and N. Lorente. "Energetics and stability of dangling-bond silicon wires on H passivated Si(100)." *Journal of physics. Condensed matter : an Institute of Physics journal* **24** (2012), p. 445004.

[76] A. Bellec, L. Chaput, G. Dujardin, D. Riedel, L. Stauffer, and P. Sonnet. "Reversible charge storage in a single silicon atom". *Physical Review B* **88** (2013), p. 241406.

[77] K. E. J. Goh, S. Chen, H. Xu, J. Ballard, J. N. Randall, and J. R. Von Ehr. "Using patterned H-resist for controlled three-dimensional growth of nanostructures". *Applied Physics Letters* **98** (2011), p. 163102.

[78] S. Chen, H. Xu, K. E. J. Goh, L. Liu, and J. N. Randall. "Patterning of sub-1 nm dangling-bond lines with atomic precision alignment on H:Si(100) surface at room temperature." *Nanotechnology* **23** (2012), p. 275301.

[79] R. N. Bracewell. *The Fourier Transform and its Applications*. New Delhi: Tata McGraw-Hill, 2003.

[80] C. S. Lent. "Quantum Cellular Automata". *Nanotechnology* **4** (1993), pp. 49–57.

[81] P. D. Tougaw and C. S. Lent. "Logical devices implemented using quantum cellular automata". *J. Appl. Phys.* **75** (1994), pp. 1818–1825.

[82] J. Timler and C. S. Lent. "Power gain and dissipation in quantum-dot cellular automata". *J. Appl. Phys.* **91** (2002), pp. 823–831.

[83] C. S. Lent, M. Liu, and Y. Lu. "Bennett clocking of quantum-dot cellular automata and the limits to binary logic scaling". *Nanotechnology* **17(16)** (2006), pp. 4240–4251.

[84] E. P. Blair and C. S. Lent. "An Architecture for Molecular Computing using Quantum-Dot Cellular Automata". *Proc. of the Third IEEE Conference on Nanotechnology*. 2003, pp. 402–405.

[85] I. Amlani, A. O. Orlov, R. K. Kummamuru, G. H. Bernstein, C. S. Lent, and G. L. Snider. "Experimental demonstration of leadless quantum-dot cellular automata cell". *Appl. Phys. Lett.* **77** (2000), pp. 738–740.

[86] A. O. Orlov, I. Amlani, G. H. Bernstein, C. S. Lent, and G. L. Snider. "Realization of a Functional Cell for Quantum-Dot Cellular Automata". *Science* **277** (1997), pp. 928–930.

[87] I. Amlani, A. O. Orlov, G. Toth, G. H. Bernstein, C. S. Lent, and G. L. Snider. "Digital Logic Gate Using Quantum-Dot Cellular Automata". *Science* **284** (1999), pp. 289–291.

[88] G. L. Snider, I. Amlan, A. Orlov, G. Toth, G. Bernstein, C. S. Lent, J. L. Merz, and W. Porod. "Quantum-dot cellular automata: Line and majority gate logic". *Jpn. J. of Applied Physics* **38** (1999), pp. 7227–7229.

[89] A. O. Orlov, R. K. Kummamuru, R. Ramasubramaniam, G. Toth, C. S. Lent, G. H. Bernstein, and G. L. Snider. "Experimental demonstration of a latch in clocked quantum-dot cellular automata". *Appl. Phys. Lett.* **78** (2001), pp. 1625–1627.

[90] A. O. Orlov, R. K. Kummamuru, R. Ramasubramaniam, C. S. Lent, G. H. Berstein, and G. L. Snider. "Clocked quantum-dot cellular automata shift register". *Surf. Sci.* **532–535** (2003), pp. 1193–1198.

[91] R. Kummamuru, A. O. Orlov, R. Ramasubramaniam, C. S. Lent, G. H. Bernstein, and G. L. Snider. "Operation of a quantum-dot cellular automata (QCA) shift register and analysis of errors". *IEEE Trans on Electron Dev.* (2003).

[92] A. O. Orlov, R. Kummamuru, J. Timler, C. S. Lent, G. L. Snider, and G. H. Bernstein. "Experimental studies of quantum-dot cellular automata devices". *Mesoscopic Tunneling Devices* (2004).

[93] J. Jiao, G. J. Long, F. Grandjean, A. M. Beatty, and T. P. Fehlner. "Building Blocks for the Molecular Expression of Quantum Cellular Automata. Isolation and Characterization of a Covalently Bonded Square Array of Two Ferrocenium and Two Ferrocene Complexes". *J. Am. Chem. Soc.* **125** (2003), pp. 7522–7523.

[94] Y. Lu, M. Liu, and C. S. Lent. "Molecular quantum-dot cellular automata: From structure to dynamics". *J. App. Phys.* **102** (2007).

[95] K. Hennessy and C. S. Lent. "Clocking of Molecular Quantum-Dot Cellular Automata". *J. Vac. Sci. Technol. B* **19** (2001), pp. 1752–1755.

[96] F. Karim, K. Walus, and A. Ivanov. "Analysis of Field-Driven Clocking for Molecular Quantum-Dot Cellular Automata". *J. of Comp. Elec.* **9** (2010), pp. 16–30.

[97] C. S. Lent and P. D. Tougaw. "Lines of interacting quantum-dot cells: A binary wire". *J. Appl. Phys.* **74** (1993), pp. 6227–6233.

[98] P. D. Tougaw and C. S. Lent. "Dynamic behavior of quantum cellular automata". *J. Appl. Phys.* **80** (1996), pp. 4722–4735.

[99] G. Tóth and C. S. Lent. "Role of correlation in the operation of quantum-dot cellular automata". *J. Appl. Phys.* **89** (2001), pp. 7943–7953.

[100] F. Karim, A. Navabi, K. Walus, and A. Ivanov. "Quantum Mechanical Simulation of QCA with a Reduced Hamiltonian". *Proc. of the 8th IEEE Conference on Nanotechnology*. 2008.

[101] K. Walus and G. A. Jullien. "Design tools for an emerging SoC technology: quantum-dot cellular automata". *Proc. IEEE* **94** (2006), pp. 1225–1244.

[102] E. Lieb, T. Schultz, and D. Mattis. "Two Soluble Models of an Antiferromagnetc Chain". *Ann. Phys.* **16** (1961), pp. 407–466.

[103] P. Pfeuty. "The One-Dimensional Ising Model with a Transverse Field". *Ann. Phys.* **57** (1970), pp. 79–90.

[104] C. S. Lent, P. D. Tougaw, and W. Porod. "Bistable saturation in coupled quantum dots for quantum cellular automata". *Appl. Phys. Lett.* **62** (1993), pp. 714–716.

[105] D. Tougaw and M. Khatun. "A Scalable Signal Distribution Network for Quantum-Dot Cellular Autoamata". *IEEE Transactions on Nanotechnology* **12** (2013), pp. 215–224.

[106] H. Cho and E. E. Swartzlander. "Adder and Multiplier Design in Quantum-Dot Cellular Automata". *IEEE Transactions on Computers* **58** (2009), pp. 721–727.

[107] J. R. Janulis, P. D. Tougaw, S. C. Henderson, and E. W. Johnson. "Serial Bit-Stream Analysis Using Quantum-Dot Cellular Automata". *IEEE Transactions on Nanotechnology* **3** (2004), pp. 158–164.

[108] U. Weiss. *Quantum Dissipative Systems*. Stuttgart, Germany: World Scientific, 2008.

[109] S. Srivastava, S. Sarkar, and S. Bhanja. "Estimation of Upper Bound of Power Dissipation in QCA Circuits". *IEEE Trans. on Nanotechnology* **8** (2009), pp. 116–127.

[110] A. J. Leggett, S. Chakravarty, A. T. Dorsey, M. P. A. Fisher, A. Garg, and W. Zwerger. "Dynamics of the dissipative two-state system". *Reviews of Modern Physics* **59** (1987), pp. 1–85.

[111] E. P. Blair and C. S. Lent. "Environmental decoherence stabilizes quantum-dot cellular automata". *Journal of Applied Physics* **113** (2013), p. 124302.

[112] E. P. Blair, M. Liu, and C. S. Lent. "Signal Energy in Quantum-Dot Cellular Automata Bit Packets". *J. Comput. Theor. Nanosci.* **8** (2011), pp. 972–982.

[113] J. Timler and C. S. Lent. "Maxwell's demon and quantum-dot cellular automata". *J. Appl. Phys.* **94** (2003), pp. 1050–1060.

[114] P. G. Piva, R. A. Wolkow, and G. Kirczenow. "Nonlocal conductance modulation by molecules: Scanning tunneling Microscopy of Substituted Styrene Heterostructures on H-Terminated Si(100)". *Physical Review Letters* **101** (2008), pp. 3–6.

[115] G. Kirczenow, P. G. Piva, and R. A. Wolkow. "Modulation of electrical conduction through individual molecules on silicon by the electrostatic fields of nearby polar molecules: Theory and experiment". *Physical Review B - Condensed Matter and Materials Physics* **80** (2009), pp. 1–21.

[116] B. V. C. Martins, M. Smeu, L. Livadaru, H. Guo, and R. A. Wolkow. "Conductivity of Si(111)-7x7: The role of a single atomic step". *Physical Review Letters* **112** (2014), pp. 1–5.

[117] P. M. Koenraad and M. E. Flatté. "Single dopants in semiconductors." *Nature materials* **10** (2011), pp. 91–100.

[118] B. E. Kane. "Silicon-Based Nuclear Spin Quantum Computer". *Nature* **393** (1998), pp. 133–137.

[119] R. A. Wolkow, L. Livadaru, J. L. Pitters, M. Taucer, P. G. Piva, M. Salomons, M. Cloutier, and B. V. C. Martins. "Silicon Atomic Quantum Dots Enable Beyond-CMOS Electronics". *Field-Coupled Nanocomputing*. Ed. by N. G. Anderson and S. Bhanja. Vol. 8280. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2014, pp. 33–58.

[120] S. R. Schofield, N. J. Curson, M. Y. Simmons, F. J. Ruess, T. Hallam, L. Oberbeck, and R. G. Clark. "Atomically Precise Placement of Single Dopants in Si". *Phys. Rev. Lett.* **91** (2003), p. 136104.

[121] D. Gohlke, R. Mishra, O. D. Restrepo, D. Lee, W. Windl, and J. Gupta. "Atomic-scale engineering of the electrostatic landscape of semiconductor surfaces". *Nano Letters* **13** (2013), pp. 2418–2422.

[122] L. Gross, F. Mohn, P. Liljeroth, J. Repp, F. J. Giessibl, and G. Meyer. "Measuring the charge state of an adatom with noncontact atomic force microscopy." *Science* **324** (2009), pp. 1428–1431.