ABSTRACT

Title of Dissertation: Electrostatic and Strain-Induced Quantum Dots in Silicon Nanostructures

Ted Thorbeck, Doctor of Philosophy, 2013

Dissertation Directed by: Dr. Neil Zimmerman, NIST

Professor Christopher Lobb, Department of Physics

Quantum dots (QDs) are nanometer scale regions that can trap charges. In this dissertation I describe my work on understanding the reproducibility of silicon QDs, and why unintentional QDs are so common.

I studied both the reproducibility and predictability of gate capacitances to intentional QDs. I found that, in our devices, electrostatic QDs have gate capacitances that are reproducible to within 10% and predictable using a capacitance simulator to within 20%.

I describe a technique that uses the gate capacitances to determine the locations of the unintentional QDs in a nanowire with a precision of a few nanometers. I do this by comparing the measured gate capacitances to simulated gate capacitances.

I suggest that strain from the gates or contacts may be the cause of many of the observed unintentional QDs. Strain can cause QDs because it changes the band structure, thus changing the energy of the conduction band and the valence band. I discuss the effects of strain in three common device architectures: a mesa-etched nanowire with poly-silicon gates, metal-gated bulk silicon, and chemically grown nanowires with metal contacts. Because strain can affect these very different architectures, I suggest that the strain in a QD can be as important as the electrostatics to understanding how a device works.

ELECTROSTATIC AND STRAIN-INDUCED
QUANTUM DOTS IN SILICON NANOSTRUCTURES

by

Ted Thorbeck

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2013

Advisory Committee:

Professor Christopher J. Lobb, Chair
Professor J. Robert Anderson
Professor Neil Goldsman
Professor Frederick C. Wellstood
Dr. Neil Zimmerman

© Copyright by

Ted Thorbeck

2013

Acknowledgements

I would to begin by thanking Neil Zimmerman for giving me the opportunity to work at NIST on silicon QDs. I have become a better physicist because of his guidance and feedback. His ability to find the weak point of any scientific argument has made this work much better. I would also thank Josh Pomeroy for all of his comments on my work both at team meeting, and early drafts of my papers.

I have enjoyed my time with the team of postdocs and grad students at NIST, in both Neil's team and Josh's team: Kevin Dwyer, Panu Koppinen, Russell Lake, and Justin Perron. I would like to thank Manolis Hourdakis for teaching me how to run the fridge and measure devices. I would especially like to thank Michael "Stew" Stewart for always being willing to talk, whether to answer a random question or listening to me complain. I have appreciated his advice and friendship.

I have gone to too many scientists at NIST to discuss various aspects of my research to mention everyone. I have been constantly surprised at how willing people were to talk to me. This is just one of the reasons I have been very glad to be able to do this work at NIST

I would like to thank Chris Lobb for being my second advisor. I have appreciated his advice and support over the years. I would like to thank Fred Wellstood, Bob Anderson and Neil Goldsman for serving on my committee and for giving me many helpful comments on it.

I would like to thank all of my housemates at 38th Ave, for giving me a great place to go back to at the end of every day. I would also like to thank all of my friends at MCF for their fellowship.

Finally, I would like to thank my parents, Orlan and Corella Thorbeck, for all of their love and support throughout the years.

Table of Contents

Chapter 1: Introduction 1

A. Motivation 1

1. Single-Electron Pump 1
2. Quantum Bits 3
3. Status 5

B. Device Layout and Gate Operation 5

1. Device Architecture 6
2. Gate Operation 8
   a) Upper Gate 8
   b) Lower Gates 10

C. Single and Double Quantum Dots 12

1. Motivation 12
2. Single QD 13
   a) Energetics 13
   b) Single Gate Scan 17
   c) Double Gate Scan 19
   d) Diamond Diagram 21
   e) Summary of Methods to Measure Capacitance 25
3. Double Quantum Dot 25
   a) Energetics 25
   b) Weak-Coupling Regime 27
   c) Intermediate-Coupling Regime 30
4. Applications of QDs 33
   a) Single-Electron Pump 34
   b) Qubits 35

D. Challenges for Silicon QDs 37

1. Reproducibility 38
2. Unintentional Quantum Dots 38

E. Advantages of Silicon 43
F. Outline of Dissertation 44
Chapter 2: Reproducibility and Predictability of Gate Capacitance to Intentional Quantum Dots 46
A. Preview 46
B. Motivation 47
C. Previous Work 48
D. Measurements 49
1. Table of All Measurements 49
2. Visual Presentation of Gate Capacitances 53
E. Simulations 56
1. Parallel Plate Method 56
2. Capacitance Simulator 58
F. Lessons Learned 59
1. Fabrication Implications 59
2. Four Sided Transport 60
3. Barrier Capacitances 62
G. Summary and Implications 63
1. Prediction of Highest Operating Temperature 64
2. Prediction of the Location of Unintentional QDs 65
Chapter 3: Determining the Locations of Unintentional Quantum Dots 68
A. Preview 68
B. Motivation and Previous Work 68
C. Unintentional Quantum Dots 69
1. The Data 69
2. Circuit Model 74
3. Capacitances and Resistances 78
4. Simulation of Hybrid Series-Parallel Model 82
D. Determining the Location of the Unintentional Quantum Dots 85
1. Qualitative Approach 85
2. Quantitative Approach 85
E. Implications and Conclusions 91

Chapter 4: Reviews of Stress, Strain and the Band Structure of Silicon **93**

A. Stress and Strain Review ... 93

1. Strain ... 93
2. Stress ... 94
3. Boundary Conditions ... 96
4. Hooke's Law ... 98

B. Silicon Band Structure ... 99

1. Conduction Band ... 100
   a) Band Structure ... 100
   b) Confinement ... 101
   c) Strain ... 103
2. Valence Band ... 104
   a) Band Structure ... 104
   b) Confinement ... 105
   c) Strain ... 106

C. Previous Work on Strain Effects in Nanostructures ... 106

1. Strain-Induced QDs in III-Vs ... 107
2. Intentional Strains in Silicon ... 107
3. Unintentional Strains in Silicon ... 107

D. Conclusions ... 108

Chapter 5: Strain-Induced Quantum Dots **110**

A. Preview ... 110

B. Mesa-Etched Nanowire ... 111

1. Dot B ... 112
2. Dot A ... 116

C. Metal-Gated Bulk Silicon ... 123

1. Metal Gates on Bulk Silicon ... 123
2. Poly-Silicon Gates on Bulk Silicon ... 126

D. Chemically-Grown Nanowire with Metal Contacts ... 128

E. Conclusions ... 131

Chapter 6: Conclusions and Future Work **133**

A. Summary ... 133

B. Future Work. 134
1. Barrier Capacitances 134
2. Device Fabrication 135
3. Valley States in Nanowires 135
4. Simulation Improvements 136
5. Measurement of the Strain 137
6. Mitigation of Strain-Induced QDs 137
7. Strain-Induced QD architecture 138
Appendix A: The Measurement Circuit 139
A. The Old Circuit 139
1. Circuit Diagram 139
2. The Room Temperature Electronics 139
3. Inside the Cryostat 142
B. The New Circuit 142
1. Motivation 142
2. The New Equipment 144
Appendix B: FASTCAP Tutorial 146
A. Motivation 146
B. How FASTCAP Works 146
C. Model and Meshing the Geometry 148
1. Two Level Hierarchy 148
2. Batch File 148
3. List File 149
4. Subfiles 152
5. Looking at the Model 153
D. Running the Simulation 154
E. Finished 158
Appendix C: COMSOL Multiphysics Tutorial 159
A. Preview 159
B. What COMSOL is Doing – One Dimension 159
C. What COMSOL is Doing – Two Dimensions 161
D. The Bimetallic Strip 165

E. COMSOL Simulation 167

F. Comparison 186

G. Material Properties 187

Appendix D: Additional Unintentional QD Data **188**

References **199**

Abbreviations

2DES:	Two Dimensional Electron System
AC:	Alternating Current
CB:	Conduction Band
CTE:	Coefficient of Thermal Expansion
DQD:	Double Quantum Dot
D:	Drain
LH:	Light Hole
LGC:	Lower Gate Center
LGD:	Lower Gate Drain
LGS:	Lower Gate Source
MOSFET:	Metal-Oxide-Semiconductor Field-Effect Transistor
NIST:	National Institute for Standards and Technology
HH:	Heavy Hole
QD:	Quantum Dot
S:	Source
SIMOX:	Separation by Implantation of Oxygen
UG:	Upper Gate
VB:	Valence Band

Chapter 1: Introduction

A. Motivation

Over the last two centuries, the ability to control the flow of electrons has led to countless inventions that have changed everyday life: from the telegraph, to the electric light and the transistor. But none of these inventions use the most basic fact about the electron: that is a discrete particle.

Over the last couple of decades, technology to manipulate single electrons has been developed. One way to do this is by trapping an electron in a silicon quantum dot (QD). A QD is a nanometer sized region that is confined in all three dimensions [1,2]. Single-electronics has led to new technologies, and has allowed us to rethink some basic ideas in physics. In this section I motivate the work in the rest of this dissertation by describing two of the ways in which experiments with single electrons in QDs are allowing us to rethink the definitions of the ampere and the bit.

1. Single-Electron Pump

The ampere, which is one of the seven SI (Le Système International d'Unités) base units, is defined as ‘‘that constant current which, if maintained in two straight parallel conductors of infinite length, of negligible circular cross section, and placed 1 meter apart in vacuum, would produce between these conductors a force equal to $2x10^{-7}$ newton per meter of length [1].” This has proven to be a difficult definition to realize. In practice, the ampere is represented by dividing the quantum standard for the volt by the quantum standard for the Ohm (based on the quantum Hall effect) [1,2]

$$A_{JV-QHR} = \frac{V_{JV}}{\Omega_{QHR}}. \quad 1.1.$$

The quantum standard for the volt comes from the Josephson effect

$$V_{JV} = nhf/2e, \quad 1.2.$$

where n is an integer, h is Planck's constant, f is frequency, and e is the electron charge. The quantum standard for the ohm comes from the quantum hall effect,

$$\Omega_{QHR} = h/e^2. \quad 1.3.$$

It would be much simpler to have a practical representation of the ampere that is based on the charge of an electron. Figure 1.1 shows a schematic for a single-electron pump, which can create a current one electron at a time.

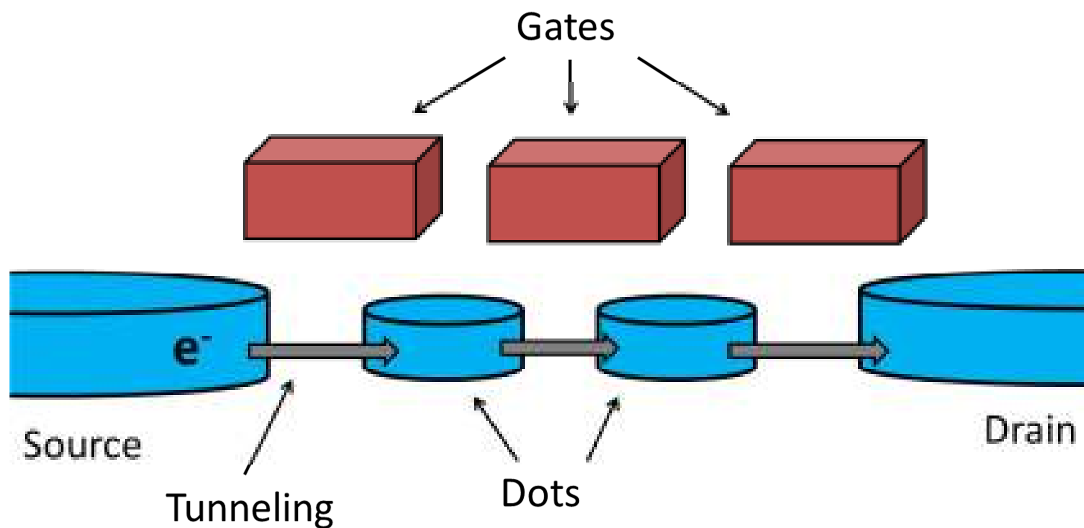


Figure 1.1. Schematic of a single-electron pump. In blue are two QDs in between source and drain electrodes. Grey arrows represent a path for single electrons to tunnel through the device: from the source to the left QD, to the right QD, and finally to the drain. The tunneling is controlled by nearby electrostatic gates (red).

A single-electron pump consists of two or more QDs in series between a source and a drain electrode. Electrons are induced to tunnel from the source to the left QD, then to right QD, and finally to the drain. Because electrons have to tunnel one-by-one, with small enough QDs and AC gate voltages, we can use this to clock electrons through at a frequency, *f*. Then the current through the single-electron pump is

$$I = ef. \tag{1.4.}$$

To compare a current using Eq. 1.4 with an ampere from Eq. 1.3, the pump must have a current of several hundred picoamps and an error rate of 0.1 parts per million or better [2,3]. This combination of high current and low error rate is beyond the capabilities of modern single-electron pumps [4]. Towards the end of this chapter, I will describe how the work in this dissertation could lead to a high-current, low-error-rate single-electron pump.

2. Quantum Bits

Quantum computing is exciting because there are some computation problems (such as factoring large numbers) for which a quantum algorithm has been found that is faster than any known classical algorithm [5]. Whereas a classical computer is based on classical bits, which must be zero or one, a quantum computer is based on quantum bits (qubits), which can be in a superposition of both 0 and 1 at the same time. To give a rough explanation for the power of a quantum computer, setting a qubit to be 0 and 1 at the same time means that a clever algorithm can test 0 and 1 simultaneously. The power of a quantum computer is due to entanglement of multiple qubits. With two entangled qubits, 0 through 3 can be tested simultaneously. With three entangled qubits, 0 through 7 can be tested simultaneously. Each additional qubit doubles the power of the quantum

computer. This exponential growth is what allows a small quantum computer to solve problems that the largest classical computers cannot solve.

There are many schemes for building a quantum computer using silicon QDs [6,7]. Here, I focus on a silicon double quantum dot (DQD) charge qubit [8–11]. This qubit requires two QDs (called left and right) that are close enough that an electron can tunnel between them. Now consider a single electron that is shared by the two QDs. The position of the electron is the qubit degree of freedom. If the electron is entirely on the left QD, then its wave function is $|\psi\rangle = |L\rangle$ [Fig. 1.2(a)]. If the wave function is entirely on the right QD, then $|\psi\rangle = |R\rangle$ [Fig. 1.2(b)]. We will use these two states as the basis states for our qubit. The Hamiltonian for this qubit is

$$H = \begin{pmatrix} \varepsilon_L & \mathrm{h}\Gamma \\ \mathrm{h}\Gamma & \varepsilon_R \end{pmatrix}, \qquad 1.5.$$

where $\varepsilon_{L(R)}$ is the energy of the left (right) QD, h is Planck's constant, and Γ is the tunnel frequency between the two QDs. Because the two QDs are tunnel coupled, the electron can also be in a superposition of being in the left and right QDs simultaneously. For example, if $\varepsilon_L = \varepsilon_R = 0$, then the eigenstates of the qubit are the symmetric [Fig. 1.2(c)] and anti-symmetric [Fig. 1.2(d)] combinations of $|L>$ and $|R>$.

Changing the gate voltages changes the energies of the left and right QDs, as well as the tunnel coupling between the dots. This can be used to coherently manipulate the qubit. I will go into more detail about coherent manipulations at the end of this chapter, after discussing the physics of QDs.

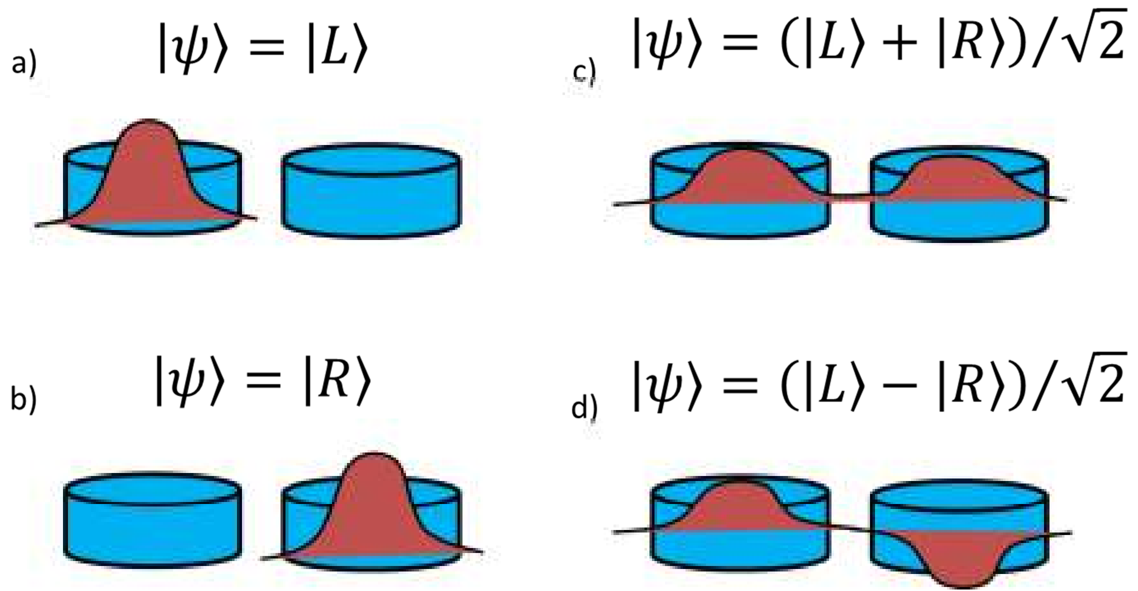


Figure 1.2. Wave functions for a DQD charge qubit. The qubit could be in the left (a) or right (b) QD. It might also be in the symmetric (c) or anti-symmetric (d) superpositions of left and right.

3. Status

Both single-electron pumps [2,4,12] and charge qubits [11,13] have been demonstrated in silicon QDs, but the devices have not performed as well as needed. Single-electron pumps cannot pump enough current to make a useful current standard. Charge qubits lose their coherence and become classical bits too quickly. In this dissertation I address some of the ways in which current silicon QDs are imperfect and how they can be made better.

B. Device Layout and Gate Operation

The devices I studied at NIST were made at NTT (Nippon Telegraph and Telephone) Basic Research Laboratories by Akira Fujiwara and coworkers [14]. At

NIST, our group has succeeded in fabricating devices with this architecture using a CMOS (complementary metal-oxide-semiconductor) compatible process flow [15].

1. Device Architecture

Figure 1.3 shows a schematic of the NTT device architecture, and Figure 1.4 shows two micrographs of the device.

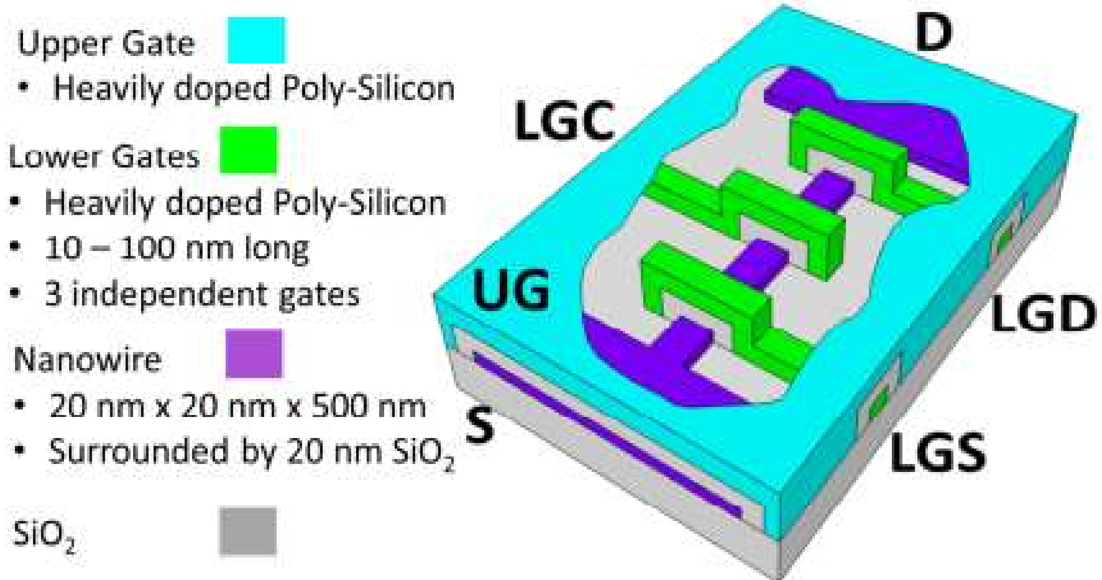


Figure 1.3. Schematic showing the nanowire (purple), the lower gates (green) and the upper gate (turquoise). The upper gate is partially cut-away to expose the nanowire and upper gate.

At the heart of the device is a silicon nanowire, which is mesa-etched from a silicon-on-insulator (SOI) wafer [14]. A typical nanowire is 20 nm tall, 20 nm wide and 500 nm long. Below the nanowire are a 200 nm thick buried silicon oxide layer (BOX) and a silicon handle wafer. 20 nm of thermally grown SiO_2 surround the nanowire.

Above the nanowire are two heavily-doped poly-silicon gate layers. The lower gate layer consists of three electrically independent gates that each covers a short portion of the length of the nanowire (typically 40 nm). The three lower gates are called lower gate source (LGS), lower gate center (LGC) and lower gate drain (LGD). 30 nm of thermally grown SiO_2 surround each lower gate. The upper gate (UG) layer is a single gate that covers the entire device.

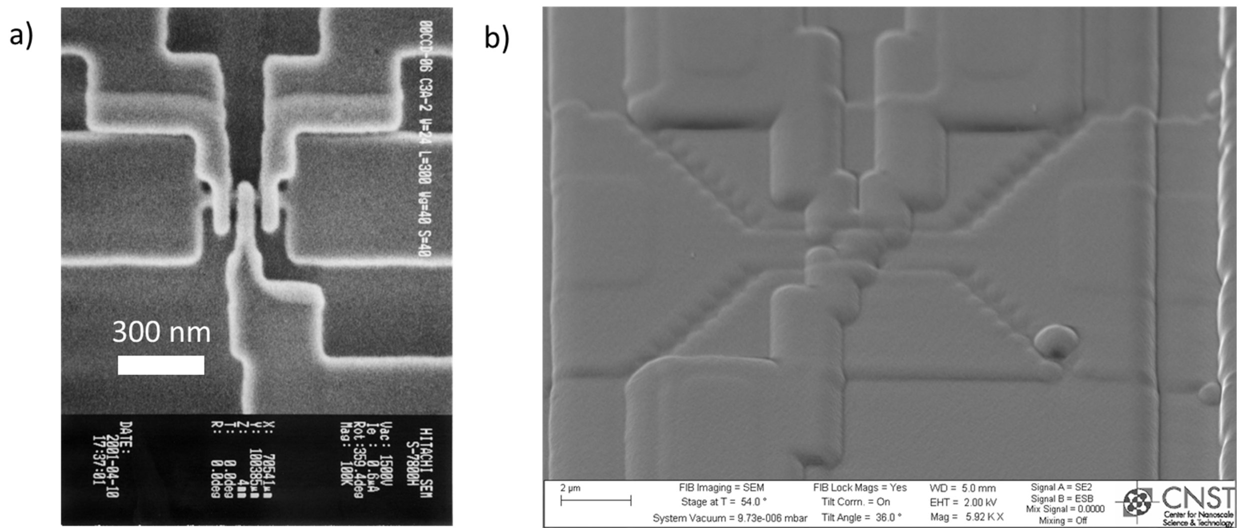


Figure 1.4. Micrographs of the device during (a) and after (b) fabrication. (a) Micrograph of the nanowire and lower gates before the upper gate was deposited; this micrograph was taken at NTT (courtesy of Akira Fujiwara). (b) Micrograph showing a wider view of the device after the upper gate was deposited and the fabrication was finished. Scale bar is 2 μm.

Because these devices are transistors, we also need to understand the doping. Several micrometers from the nanowire, along the direction of current, are the heavily n-

doped source and drain. The dopant density of the nanowire is unknown. The wafer was originally lightly boron doped ($<10^{15}$ /cm^3). However, the dopant density decreased during the SIMOX (Separation by IMplantation of Oxygen) process of turning a bulk silicon wafer into a SOI wafer. Studies of the threshold voltage of standard MOSFETS (metal-oxide-semiconductor field-effect transistor) at NTT suggest that the SOI silicon has a very low dopant density [16].

The NTT devices are similar to the current generation of silicon transistors used in commercial integrated circuits. Because the poly-silicon gate layers wrap around the nanowire, three of the four sides of the nanowire are close to the gate and can conduct. In the silicon industry, this is referred to as a tri-gate field effect transistor (FET) or finFET. Because three of the four sides of silicon can conduct, tri-gate FETs let a chip maker put more transistors on a chip without reducing the surface area of each transistor. This is why the silicon industry is transitioning from making planar MOSFET to tri-gate FETS and similar device architectures [17].

2. Gate Operation

a) Upper Gate

Now that I have described the layout of the device, I can describe its electrical operation. A positive voltage on the UG turns on conduction between source and drain. Figure 1.5(a) shows the current through the nanowire as a function of V_{UG}. The positive voltage on the UG draws electrons from the source and drain into the channel to form an inversion layer at the Si-SiO_2 interface [Fig. 1.5(b)]. Because at low temperatures kT is very small (kT = 86 μeV at 1 K), the thermal population of the conduction band will be

very small, unless the Fermi level is at or above the conduction band (E_C) [Fig. 1.5(c)]. To fully turn on the device, we typically apply between 1 V and 2 V on the UG to get several nanoamperes of current flowing through the device.

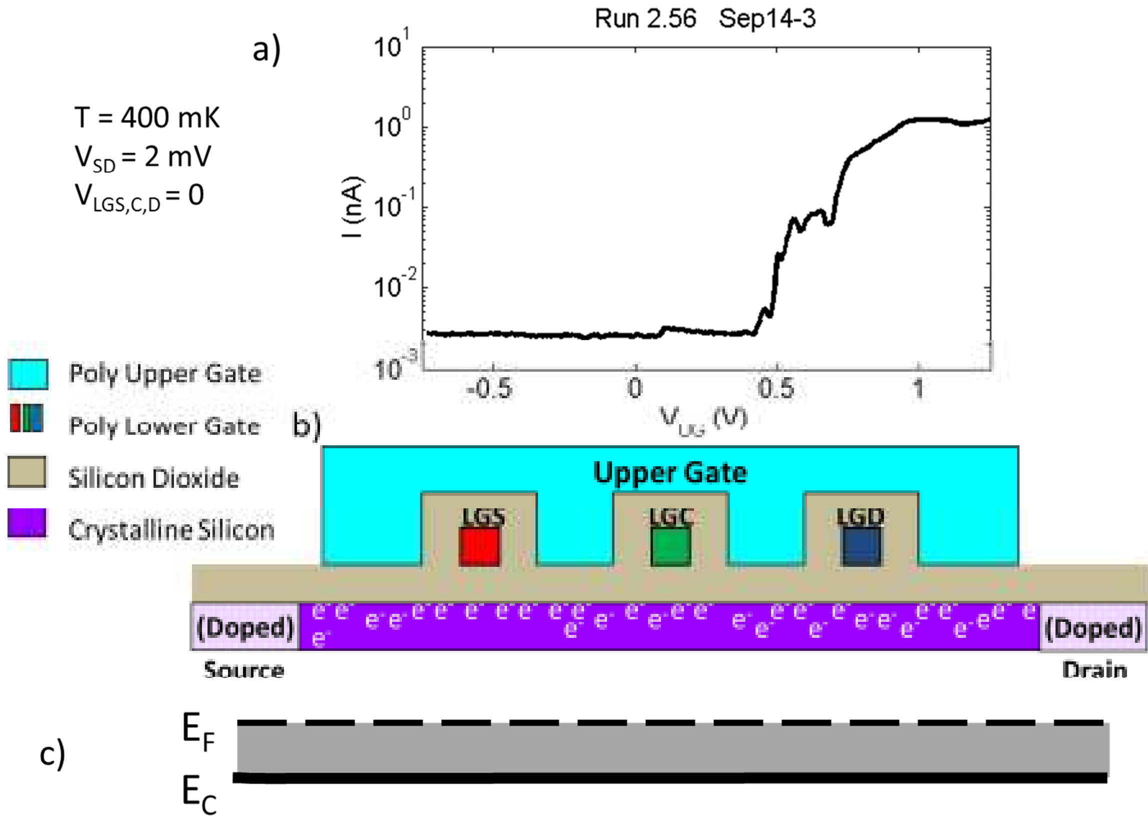


Figure 1.5. A positive voltage on the UG inverts the nanowire. (a) Current through the nanowire as a function of V_{UG} taken at T = 400 mK, V_{SD} = 2 mV and $V_{LGS,C,D}$ = 0. Data from device AF-CA2-U3D-3, run 2.56, Sep14_3. (b) Schematic of the device showing the electrons in an inversion layer in the channel. (c) Band diagram during inversion, showing the Fermi level above the conduction band.

b) Lower Gates

After turning on conduction with the UG, we use the lower gates to create tunnel barriers and to isolate QDs within the nanowire. In Figure 1.6(a) I show the current as a function of lower gate voltage for two different lower gates, LGS and LGD. Notice that the LGS and LGC curves are much smoother than the LGD curve. I show in Chapter 3 that the peaks in LGD curve are due to Coulomb blockade through unintentional QDs near LGD. I return to discuss unintentional QDs later in the chapter, but, for now, I focus on explaining the LGS and LGC curves.

Applying a negative voltage to LGS depletes the nanowire directly below the lower gate, as shown in Figure 1.6(b). Depleting the nanowire raises the conduction band above the Fermi level, as shown in the band diagram in Figure 1.6(c). This creates a tunnel barrier for electrons. Making V_{LGS} more negative raises the barrier, decreases the tunneling rate, and eventually shuts off conduction.

To create a QD we use two of the lower gates to create two tunnel barriers. The region in the nanowire between the two tunnel barriers is thus confined in all three dimensions, forming a QD. If LGS and LGD are used to create the tunnel barriers, then the QD is called a full QD. If LGS (LGD) and LGC create tunnel barriers, then the QD is called a source (drain) short QD. If all three lower gates form tunnel barriers simultaneously, then we can make two QDs.

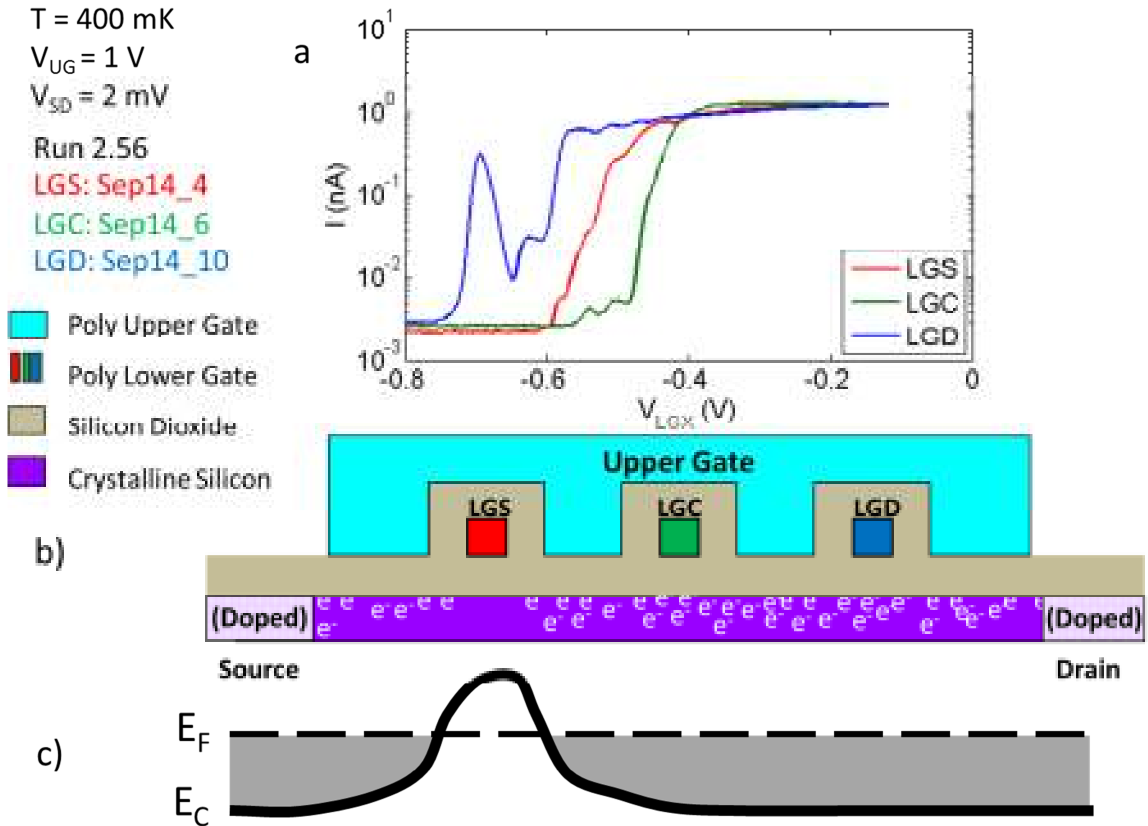


Figure 1.6. Lower gates are used to deplete the nanowire to form tunnel junctions. (a) Current through the nanowire versus lower gate voltage for two different lower gates in the same device, T = 400 mK, V_{UG} = 1 V and the other lower gates were set to V_{LGX} = 0. The peaks in in the LGD are due to unintentional QDs. Data from device AF-CA2-U3D-3. (b) Schematic of the device showing the nanowire depleted of electrons below LGS. (c) Band diagram showing depletion below the LGS, because the Fermi level is below the conduction band.

C. Single and Double Quantum Dots

1. Motivation

In subsequent chapters, I use the current through the device as a function of gate voltage to measure the capacitances from the gates to the QDs. In this section I develop the physics underlying this analysis.

Once a QD has been formed in a nanowire, the current through the nanowire as a function of gate voltage looks very different than it does in Figures 1.5 and 1.6. The reason is that once we form a QD in the nanowire, electrons must travel through that QD, for us to observe current. For an electron to tunnel onto the QD, it must overcome the repulsion from electrons already in the QD.

To begin, it helps to understand the energy scales involved. Consider how much energy it takes to put a second electron on a metal sphere, if the sphere already has one electron charge. For a sphere the size of a typical QD (radius R = 20 nm), surrounded by SiO_2, this energy is 18 meV. This is called the charging energy, and it is due to the repulsion between electrons.

$$E_{charging} = \frac{e^2}{4\pi\epsilon_{SiO_2}R} = 18\ meV \qquad 1.6.$$

where $\epsilon_{SiO_2} = 3.9\ \epsilon_0$. Typically we operate QDs at cryogenic temperatures, because the thermal energy is very small, kT = 86 μeV at 1 K. In this example, the charging energy is more than 200 times larger than the thermal energy. Any electron trying to travel through the QD will need to pay this energy cost. This is why forming a QD in the nanowire changes the transport so dramatically.

2. Single QD

a) Energetics

The physics of transport through a QD can be understood based on their energetics. A typical circuit for a single QD is shown in Figure 1.7(a) [18–20]. This circuit is called a single-electron transistor, and it consists of a QD connected to the source and drain by tunnel junctions and a capacitively coupled gate. A tunnel junction is represented in Fig. 1.7(b) and can be thought of as a resistor and capacitor combined in parallel. Because the only way for charge to get onto the QD is for an electron to tunnel, the charge on the QD can only change by an integer number. The charge on the QD is quantized (this terminology is somewhat unfortunate because, except for tunneling, no quantum mechanics is involved).

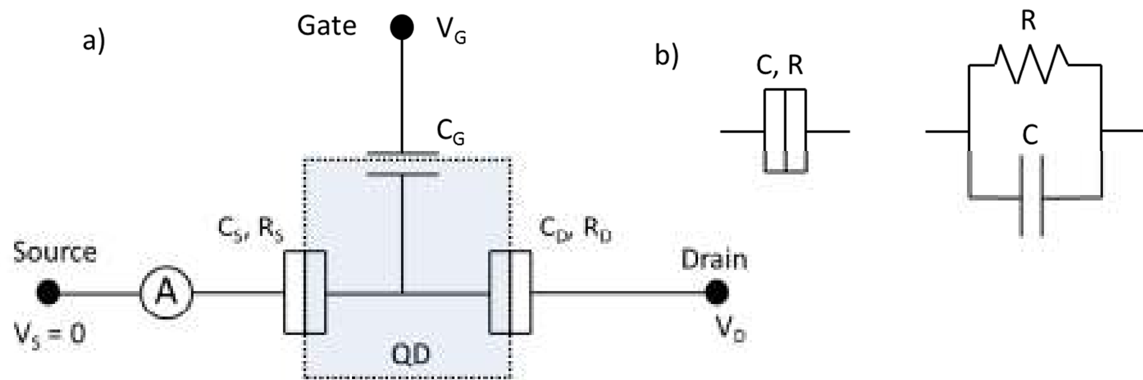


Figure 1.7. Single-electron transistor. (a) Circuit diagram for a single-electron transistor,. The QD consists of the region within the light blue box. (b) Two equivalent representations of an ultra-small tunnel junction, because a tunnel junction in a semiconductor device can be thought of as the parallel combination of a resistor and capacitor.

To understand the single-electron transistor, we need to consider the electrostatics of the QD. Just as in the example of a metal sphere, it takes energy to put an electron on the QD. Specifically, there are two energy costs that must be paid. First, the charge on the QD is spread out on all of the capacitors connected to the QD, and charging those capacitors requires energy. Second, applying voltages to the gate or drain moves charges through a potential difference, which means that work has been done by the voltage source. Including both terms, the energy in the circuit with *n* electrons on the QD is

$$E_n(V_G, V_{SD}) = \frac{(-(n-n_0)e + C_G V_G + C_D V_D)^2}{2C_\Sigma}, \quad 1.7.$$

where n_0 is a continuous (non-integer) offset charge that accounts for the background charge on the QD at $V_G = V_D = 0$. Also, $C_\Sigma = C_S + C_D + C_G$ is the total capacitance to the QD [18,19].The total energy is not, in my experience, the most useful quantity. I have found it much more convenient to think about how much energy it takes to add the nth electron to the QD, when n-1 electrons are already on the QD. This energy is called the chemical potential,

$$\mu(n, V_G, V_{SD}) = E_n(V_G, V_{SD}) - E_{n-1}(V_G, V_{SD})$$
$$= \frac{-e\left(-(n-n_0)e + C_G V_G + C_D V_D + {}^{e}/_{2}\right)}{C_\Sigma}. \quad 1.8.$$

Because the charge on the QD, $-ne$, is quantized, the chemical potential can only take discrete values, when V_G and V_D are fixed. Thus, only at specific values of V_G and V_D are electrons allowed to tunnel onto the QD. The charging energy E_Σ (most texts use E_C for the charging energy, but I use E_Σ to avoid confusion with the conduction band) is the spacing between the states on the QD on a chemical potential diagram.

$$\mu(n, V_G, V_{SD}) - \mu(n-1, V_G, V_{SD}) = \frac{e^2}{C_\Sigma} = E_\Sigma \qquad 1.9.$$

Because the charging energy does not depend on n, the chemical potentials on the QD consist of evenly spaced steps (Fig. 1.8). As can be seen in Figure 1.8, this means that on the QD has a discrete 'ladder' of chemical potentials, with each rung of the ladder, is one charging energy, E_Σ, higher than the previous rung.

In contrast, the source and drain leads have a continuous density of states that is filled below their Fermi levels. Because the source is grounded, it does not take energy to put an electron on the source, so $\mu_S = 0$. The drain is at a voltage V_D, so $\mu_D = -eV_D$. Figure 1.8 shows a diagram with the chemical potential of the QD and the Fermi levels of the source and drain.

Now that we understand the electrostatics of the QD, we can examine the current through the QD [18]. Let us assume, for the moment, that $kT < E_\Sigma$ and $e|V_D| < E_\Sigma$. Charge states on the QD below both the source and drain Fermi levels will be filled with an electron, because any empty states would be filled by an electron tunneling onto the QD from the source or drain. Similarly, all of the charge states above both the source and drain Fermi levels are empty, because an electron in any of these charge states would tunnel out of the QD. But if there is a charge state in between the source and drain Fermi levels, then an electron can tunnel from the source, into an empty state on the QD, then into an empty state on the drain. Then another electron can then repeat this process. Therefore, current from source to drain is allowed. However, if there are no states on the QD in between the source and drain Fermi level, then there is no path for current from source to drain. In this case the current is said to be Coulomb blockaded [18,19].

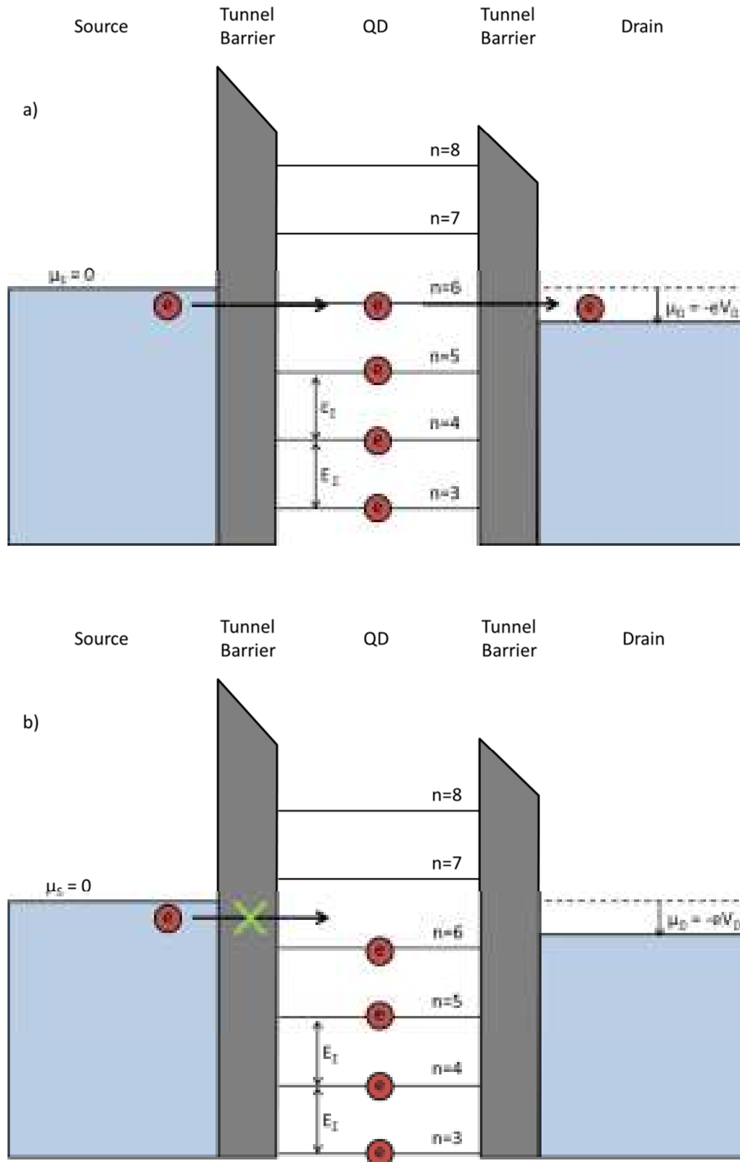


Figure 1.8. Chemical potential diagrams of a single-electron transistor. Source (left) and drain (right) are both filled below their Fermi levels (blue). The bias window is the difference between the source and drain Fermi levels. Tunnel barriers are grey. There are discrete energy levels on the QD. (a) Tunneling from the source to the QD to the drain is allowed, because there is a charge state on the QD in the bias window. (b) By changing the gate voltage we can bring the "ladder" of energy states up or down. Here, no state on the QD is in the bias window, so current is Coulomb blockaded.

If either $kT > E_\Sigma$ or $e|V_D| > E_\Sigma$, then the Coulomb blockade will be lifted. If $kT > E_\Sigma$, then a thermally excited electron can tunnel onto a state above both the source and drain Fermi levels. If $e|V_D| > E_\Sigma$, then there will be a state on the QD in between the source and drain Fermi levels for all values of V_G, so the current cannot be Coulomb blockaded.

b) Single Gate Scan

To understand how Coulomb blockade affects the current through the device as a function of gate voltage, let us start with the case in which $\mu(n, V_G, V_D)$ is in between the source and drain Fermi level. We will call this the bias window. As a voltage is varied on a gate, the chemical potential of each charge state on the QD will change. Increasing V_G will cause $\mu(n, V_G, V_D)$ to decrease. Soon $\mu(n, V_G, V_D)$ will exit the bias window. This turns the conduction from on to off. But as the gate voltage continues to increase, eventually $\mu(n+1, V_G, V_{SD})$ will enter the bias window, because

$$\mu(n, V_G, V_{SD}) = \mu\left(n+1, V_G + \frac{e}{C_G}, V_{SD}\right). \qquad 1.10.$$

This turns the conduction back on. Thus the current is thus a periodic function of gate voltage. For every e/C_G the gate voltage changes, the current will go from on to off to on again. An example of Coulomb blockade oscillations can be seen in Figure 1.9. Compare the periodic curve in Figure 1.9 with the monotonic (LGS) curve in Figure 1.6; this is how forming a QD in the nanowire can dramatically affect transport.

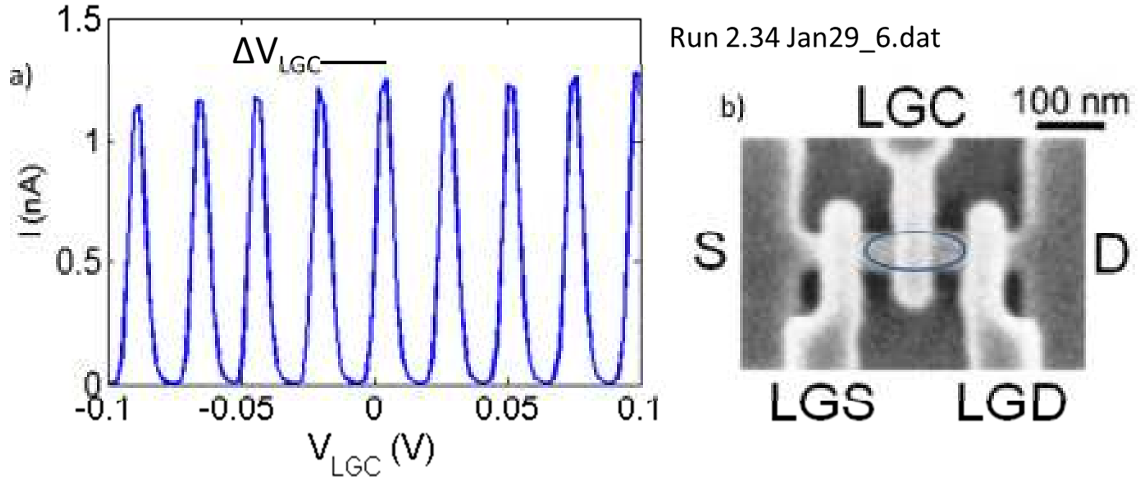


Figure 1.9. Coulomb blockade oscillations through a single QD. (a) Data from Run 2.34 Jan29_6.dat on device AF-CA3A3E-1 with T = 21 mK, V_{UG} = 2 V, V_{LGS} = -2.297, V_{LGD} = -1.874, and V_D = 1 mV. This data was taken by someone else in the group. (b) Micrograph of the device showing the location of the QD.

In this example, LGS and LGD are creating tunnel barriers (the full-QD), and the voltage on LGC is scanned. Because each peak is separated in gate voltage by $\Delta V_G = e/C_G$, we can measure the capacitance from the gate to the QD. For the Coulomb blockade oscillations shown in Fig. 1.9 the gate capacitance is

$$C_{LGC} = \frac{e}{\Delta V_{LGC}} = \frac{1.6 \times 10^{-19} C}{\big((0.098\, V) - (-0.094\, V)\big)/8} = 6.7 aF. \qquad 1.11.$$

This capacitance is on the order of attofarads (10^{-18} F) and can be measured with sub-attofarad precision. This is a very precise measurement of a very small capacitance. This precision will be very helpful in later chapters.

c) Double Gate Scan

So far we have considered what happens when a single gate is scanned, but our devices have four gates. Figure 1.10 shows what happens to the current through our devices as both V_{LGS} and V_{LGD} are scanned.

In Figure 1.10 we see two phenomena. (1) The plot is blue along the left side and the bottom of the plot, meaning there is no current, but in the upper right the plot is red, meaning there is a relatively large amount of current. (2) In the upper-right portion of the plot, we see periodic peaks in the current that take the shape of diagonal lines in this two gate scan. These two phenomena can be explained by considering that LGS and the LGD have two functions: (1) to form tunnel barriers in the silicon below and (2) to change the chemical potential of the QD.

First, negative voltages on LGS and LGD deplete the silicon below to form tunnel barriers. As the voltages on LGS and LGD get more negative, the tunnel barriers get bigger. This reduces the current through the device, and eventually shuts off the current. This is why for the most negative LGS or LGD voltages (along the left side and bottom of Fig. 1.10) there is no current. At the opposite end of the plot, the upper-right side of the plot, both LGS and LGD are at their least negative, the tunnel barriers are the smallest, and the current is the largest.

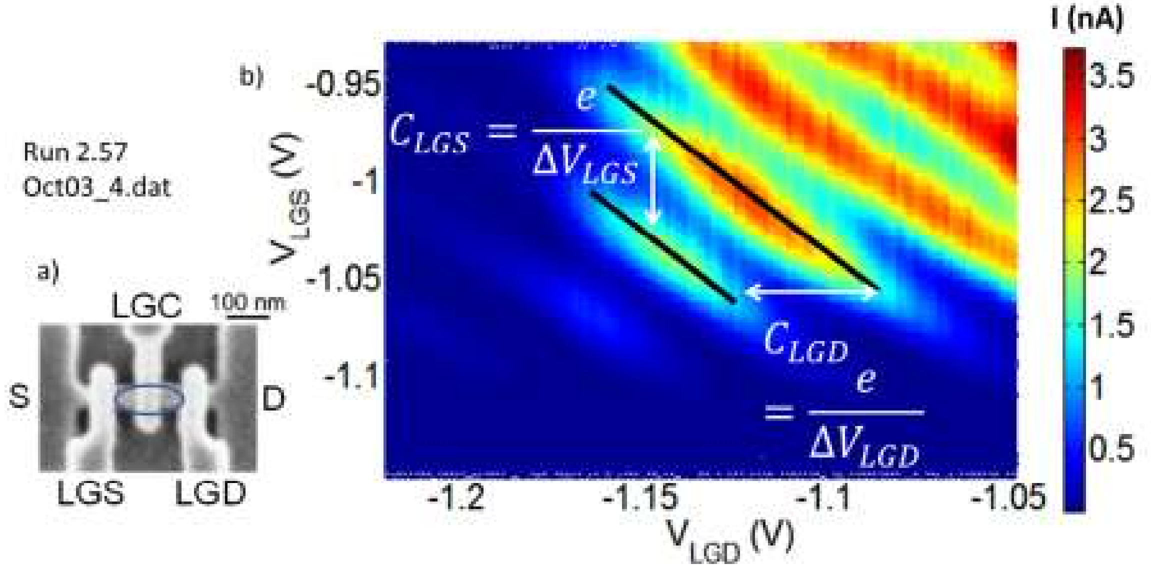


Figure 1.10 Current through a single-QD as LGS and LGD are scanned. (a) The QD is formed by tunnel barriers underneath LGS and LGD. (b) Diagonal current peaks as both V_{LGD} and V_{LGS} are scanned; the color scale represents current through the QD. Black lines are a guide for the eye. Data from run 2.57 Oct03_4.dat, T = 25 mK, V_{UG} = 2 V, V_{LGC} = 0, V_D = 1 mV with device AF-CA2-U3D-3.

Second, both LGS and LGD are capacitively coupled to the QD. To understand the effects of multiple gates, we can expand our equation (Eq. 1.10) for the chemical potential of a QD to include multiple gates,

$$\mu(n, V_{G1}, V_{G2}, V_{SD}) = \frac{-e\left(-(n-n_0)e + C_{G1}V_{G1} + C_{G2}V_{G2} + C_D V_D - {}^{e}/_{2}\right)}{C_\Sigma}. \qquad 1.12.$$

Because this is not specific to LGS and LGD, I used the generic V_{G1} and V_{G2} in this equation.

From Eq. 1.12, we now see that by sweeping either V_{G1} or V_{G2} individually, we obtain Coulomb blockade oscillations, just like the previous section. This can be seen in Figure 1.10 by taking a cut through the data along either the LGS or the LGD axis. The vertical spacing between peaks, ΔV_{LGS}, gives the capacitance from the gate LGS to the QD, $C_{LGS} = e/\Delta V_{LGS}$ Similarly, the horizontal spacing between peaks gives the capacitance from LGD to the QD, $C_{LGD} = e/\Delta V_{LGD}$.

In examining Eq. 1.12, notice that it is possible to change the voltages on both gates without changing the chemical potential of the QD. I like to think of this as balancing an increase of the voltage on one gate by decreasing the voltage on the other gate to maintain the same chemical potential,

$$\Delta V_{G1} = -\frac{C_{G2}}{C_{G1}}\Delta V_{G2}. \quad 1.13.$$

In Figure 1.10, the diagonal current peaks are an example of this kind of equipotential. Measuring the slope of these diagonal lines is another way to measure capacitance, or rather capacitance ratios, which will also be used in later chapters of this dissertation.

d) Diamond Diagram

Finally, I also used another two dimensional scan to study the QDs: a "diamond" plot, in which both V_G and V_D are scanned. Figure 1.11 is a schematic explanation of the diamond diagram. To understand a diamond diagram, recall that to see current through a QD there must be a state with a chemical potential below the source Fermi level and above the drain Fermi level (assuming that $0 < V_D < \frac{e}{c}$). These constraints can be expressed as,

$$\mu_S \geq \mu(n, V_G, V_{SD})$$
$$\mu(n, V_G, V_{SD}) \geq \mu_D. \quad 1.14.$$

Next, we plug in Eq. 1.10 for the chemical potentials and find the constraints,

$$0 \geq \frac{-e\left(-(n-n_0)e + C_G V_G + C_D V_D - {}^{e}/_{2}\right)}{C_\Sigma}$$
$$\frac{-e\left(-(n-n_0)e + C_G V_G + C_D V_D - {}^{e}/_{2}\right)}{C_\Sigma} \geq -eV_D. \quad 1.15.$$

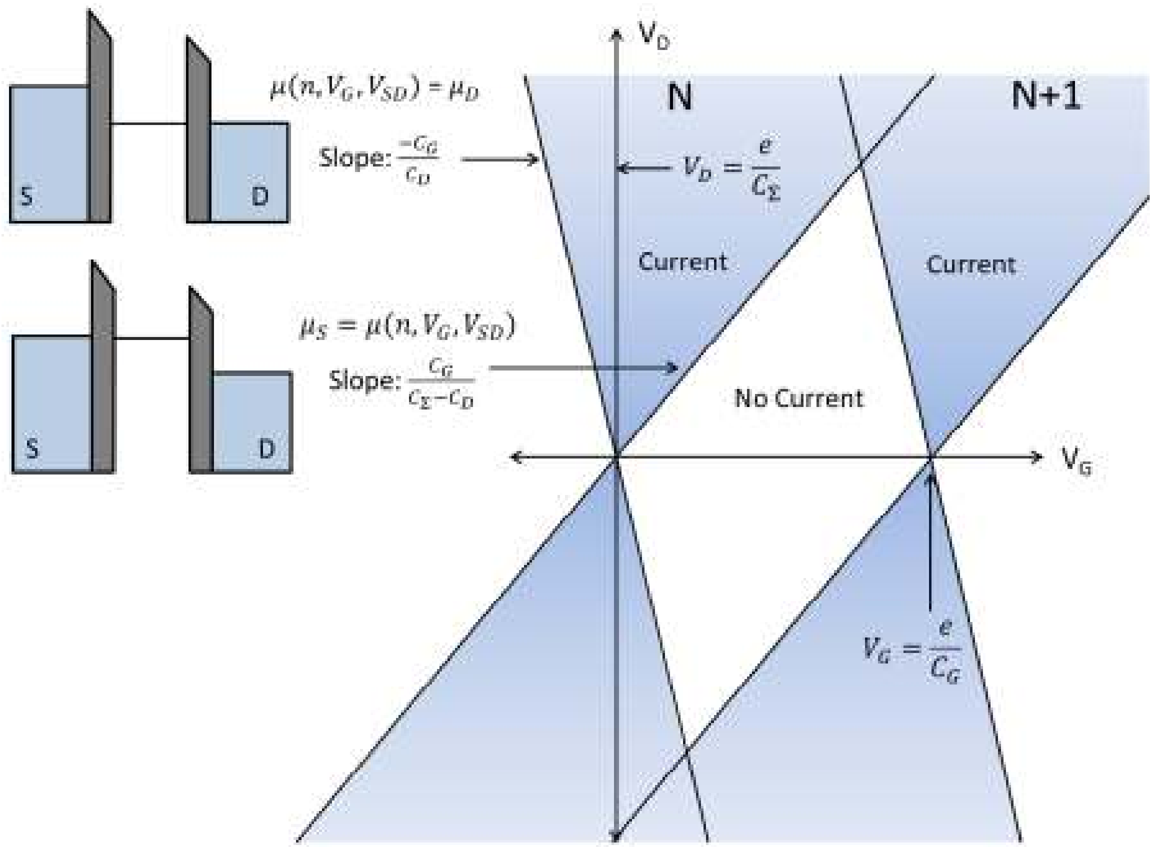


Figure 1.11 A schematic explanation of diamond diagrams. The right-hand side represents regions of allowed current through the QD as a function of V_G and V_D. The left-hand side shows chemical potential diagrams where the chemical potential of the QD is equal to either the source or drain Fermi levels.

To make the math easier, I will set $n_0 = 1/2$. Because n_0 is an offset charge, this only changes the x-intercept of the two inequalities. Simplifying these equations, we find

$$0 \geq -(-ne + C_G V_G + C_D V_D)$$
$$-(-ne + C_G V_G + C_D V_D) \geq -C_\Sigma V_D. \qquad 1.16.$$

Collecting terms yields,

$$V_D \geq -\frac{C_G}{C_D}\left(V_G - \frac{ne}{C_G}\right)$$
$$V_D \geq \frac{C_G}{C_\Sigma - C_D}\left(V_G - \frac{ne}{C_G}\right). \qquad 1.17.$$

Thus, for each n we have two inequalities, both of which must be satisfied for current to flow. Both inequalities are represented as black lines on Figure 1.11. The shaded region between both lines satisfies both inequalities, so in these regions current is allowed. Moving along the x-axis by e/C_G, we see a separate cone that represents the next charge state of the QD. As V_D increases, eventually the two cones intersect, once $eV_D > E_\Sigma$. For $V_D < 0$, there are two similar inequalities that must be satisfied. The diamond shaped region between the cones, where no current flows, gives this plot its name.

The slopes of the lines that form the edges of the diamonds can be used to measure the barrier capacitances. Examining the negative slope first, it depends on both C_G and C_D. Because we have already measured C_G, this allows us to measure the barrier capacitance, C_D. Similarly, the positive sloped line depends on C_G, C_D and C_Σ. This allows us to calculate C_Σ. Because C_Σ contains C_S, we can calculate C_S. We thus have the ability to measure all of the capacitances in the single-electron transistor circuit shown in Fig. 1.7.

Figure 1.12 shows a diamond diagram for the full-QD. We can use the slopes to calculate C_D:

$$Slope = -\frac{C_G}{C_D}. \quad 1.18.$$

In this device C_{LGC} = 6.7 aF, and the slope of the negative line is -0.75. Therefore, C_D = 8.9 aF.

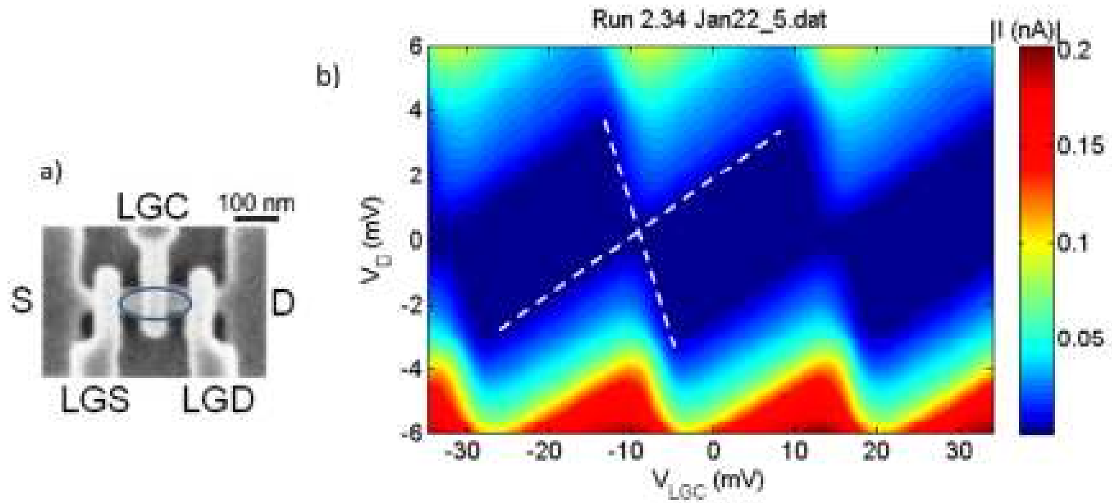


Figure 1.12 Diamond Data. (a) Full QD highlighted on a micrograph. (b) Current through QD as a function of V_D and V_{LGC}. Data from Run 2.34, Jan22_5.dat on device AF-CA3-A3E-1, T = 1.57 K, V_{UG} = 2 V, V_{LGS} = -2.483 V and V_{LGD} = -1.957 V. This data was taken by someone else in the group. Dashed lines show the positive and negative sloped edges of the diamond.

e) *Summary of Methods to Measure Capacitance*

In the previous sections I introduced several methods of measuring capacitances. These methods are summarized in Table 1.1.

Table 1.1 Summary of methods to measure capacitance

Capacitance	Data	Measured	Figures	Equation
Gate Capacitance (C_G)	Single or Double Gate Scan	Period in Gate Voltage	1.9 & 1.10	1.11
Gate Capacitance Ratio (C_{G1}/C_{G2})	Double Gate Scan	Slope	1.10	1.13
Barrier Capacitance – Drain (C_D)	Diamond	Negative Slope	1.11 & 1.12	1.17(a)
Barrier Capacitance – Source (C_S)	Diamond	Positive Slope	1.11 & 1.12	1.17(b)

3. Double Quantum Dot

a) *Energetics*

Having discussed a single QD, I now move on to the double quantum dot (DQD) [21]. Figure 1.13 shows a circuit diagram for the DQD. Note that I have not included any cross capacitances between V_{GR} and the left QD or vice versa.

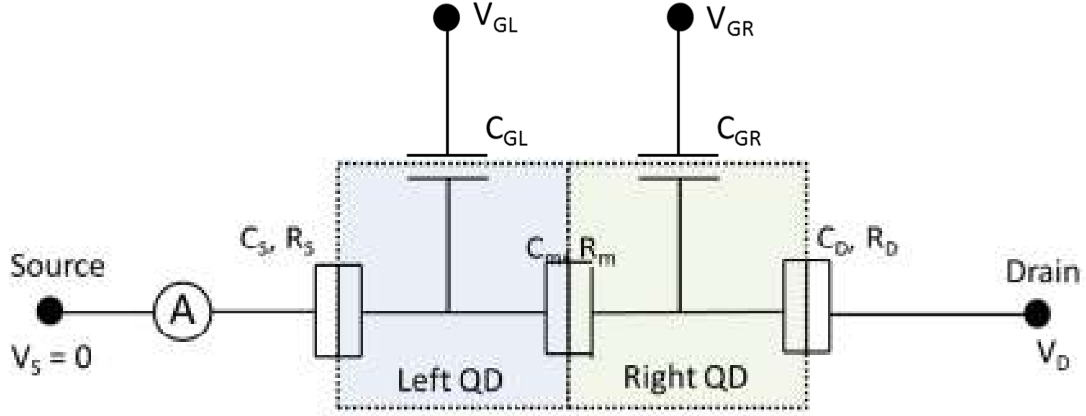


Figure 1.13 Circuit diagram for a DQD using the same notation as Fig. 1.7.

I solve the DQD problem the same way as the single QD problem, by writing the energy of DQD. I have set the offset charge on both dots, $-en_{0L}$ and $-en_{0R}$, to 0.

$$E(n_L, n_R V_{GL}, V_{GR}) = \frac{(-(n_L)e + C_{GL}V_{GL} + C_M V_R)^2}{2C_{\Sigma L}} + \frac{(-(n_R)e + C_{GR}V_{GR} - C_M V_L)^2}{2C_{\Sigma R}} \quad 1.19.$$

The charge and voltage on the left (right) dot are $-en_{0L(R)}$ and $V_{L(R)}$. Next, I need to define the individual chemical potentials for the left and right QDs:

$$\mu_L(n_L, n_R, V_{GL}, V_{GR}) = E(n_L, n_R V_{GL}, V_{GR}) - E(n_L - 1, n_R, V_{GL}, V_{GR})$$
$$\mu_R(n_L, n_R V_{GL}, V_{GR}) = E(n_L, n_R, V_{GL}, V_{GR}) - E(n_L, n_R - 1, V_{GL}, V_{GR}). \quad 1.20.$$

The charge on a QD and voltage on the quantum dots are related by:

$$-en_L + C_{GL}V_{GL} = C_{\Sigma L}V_{\Sigma L} - C_{\mathrm{m}}V_{\mathrm{R}}$$
$$-en_R + C_{GR}V_{GR} + C_D V_D = C_{\Sigma R}V_R - C_{\mathrm{m}}V_{\mathrm{R}}. \quad 1.21.$$

With some algebra I obtain,

$$\mu_L(n_L, n_R, V_{GL}, V_{GR}) = E_L\left(n_L - \frac{C_{GL}V_{GL}}{e} - {}^1\!/_2\right) + E_m\left(n_R - \frac{C_{GR}V_{GR}}{e}\right) \quad 1.22.$$

$$\mu_L(n_L, n_R, V_{GL}, V_{GR}) = E_R\left(n_R - \frac{C_{GR}V_{GR}}{e} - {}^1\!/_2\right) + E_m\left(n_L - \frac{C_{GL}V_{GL}}{e}\right).$$

$$\text{where } E_L = \frac{e^2}{C_{\Sigma L}}\frac{1}{1-\frac{C_m{}^2}{C_{\Sigma L}C_{\Sigma R}}},\ E_R = \frac{e^2}{C_{\Sigma R}}\frac{1}{1-\frac{C_m{}^2}{C_{\Sigma L}C_{\Sigma R}}},$$

$$\text{and } E_m = \frac{e^2}{C_m}\left(\frac{1}{C_{\Sigma L}C_{\Sigma R}/C_m^2 - 1}\right).$$

To make the following discussion easier, in this section I only discuss the case $V_D = 0$. But I highly recommend using the review by W.G. van der Wiel, *et al.*, [21] to go through the case for finite V_D.

b) Weak-Coupling Regime

First, I consider the weak-coupling regime, where the coupling capacitance, $C_m << C_{\Sigma L}, C_{\Sigma R}$. In this regime what happens on one dot does not affect the other. Changing the charge on the left QD does not change the charge state of the right QD. Rewriting the chemical potentials of the two QDs for this case yields

$$\begin{aligned} \mu_L(n_L, V_{GL}) &= E_L\left(n_L - \frac{C_{GL}V_{GL}}{e} - {}^1\!/_2\right) \\ \mu_L(n_R, V_{GL}) &= E_R\left(n_R - \frac{C_{GR}V_{GR}}{e} - {}^1\!/_2\right). \end{aligned} \quad 1.23.$$

The chemical potentials let us determine for what V_{GL} and V_{GR} current through the DQD is possible. Again, I restrict myself to discussing the case of $V_D = 0$, so $\mu_S = \mu_D = 0$. Because the current has to flow from the source to the left QD to the right QD to the drain, the chemical potentials must line up as

$$\begin{aligned} \mu_S &= \mu_L(n_L, V_{GL}) \\ \mu_L(n_L, V_{GL}) &= \mu_R(n_R, V_{GR}) \end{aligned} \quad 1.24.$$

$$\mu_R(n_R, V_{GR}) = \mu_D.$$

Because $\mu_S = \mu_D = 0$, the second equation is redundant. Plugging in the chemical potentials I find

$$\mu_L(n_L, V_{GL}) = E_L\left(n_L - \frac{C_{GL}V_{GL}}{e} - 1/2\right) = 0$$
$$\mu_R(n_R, V_{GR}) = E_R\left(n_R - \frac{C_{GR}V_{GR}}{e} - 1/2\right) = 0. \qquad 1.25.$$

Both of these requirements must be satisfied to allow current to flow. To understand these equalities, I plot them on a graph of V_{GL} versus V_{GR} in Figure 1.14. The solutions to the first equation will be a periodic set of horizontal lines, and the solutions to the second equation will be a set of vertical lines. Only where those two lines intersect will both equations be satisfied, allowing current to flow. Together the vertical and horizontal lines look like a 'square array', (Fig. 1.14(a)) giving this plot its name.

To form a square array in our devices all three of the lower gates must be used to form tunnel barriers [Fig. 1.15(a)]. Figure 1.15(b) shows an example of the current as both V_{LGS} and V_{LGD} (as V_{GL} and V_{GR}) are scanned. As expected from Figure 1.14(a) the current peaks look like the vertices of a square array.

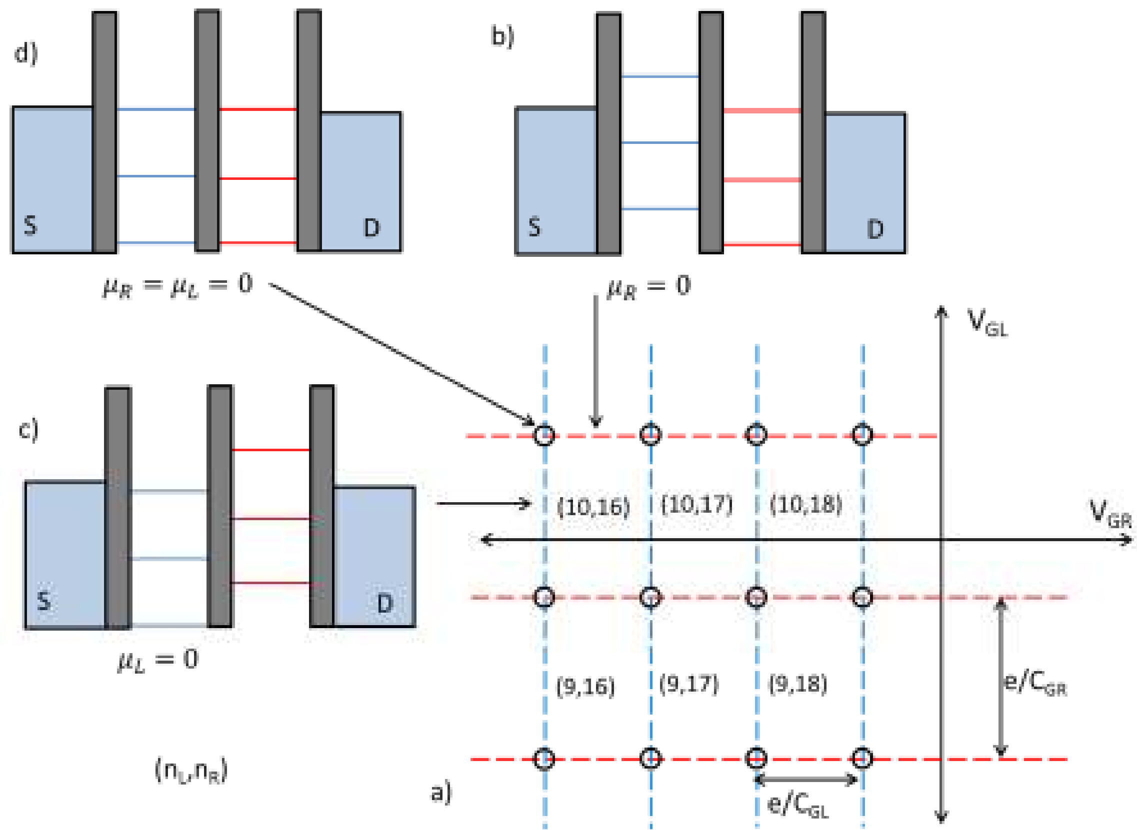


Figure 1.14. (a) Schematic showing where current is possible through a DQD in the weak-coupling regime. Along the red dashed lines $\mu_R = 0$, which is shown schematically in (b). Along the blue dashed line $\mu_L = 0$, (c). Only where $\mu_R = \mu_L = 0$, which occurs at the intersection of the blue and red lines that are represented by the black circles and shown in (d), will current be able to flow. (n_L, n_R) represents the charge on the on the left and right QDs.

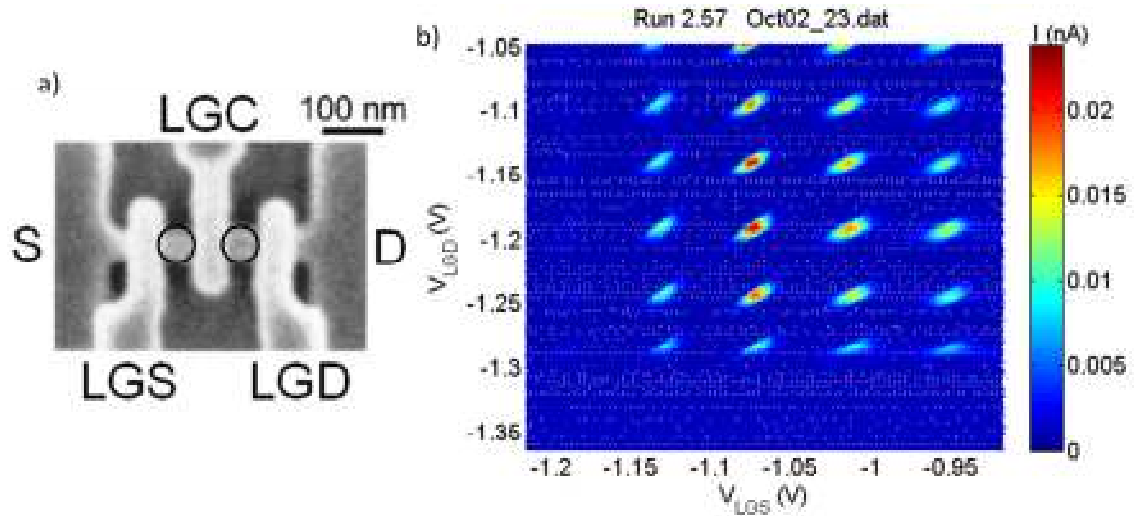


Figure 1.15. Current through a DQD in the weak-coupling regime. (a) Using all three lower gates to form tunnel barriers, I created both the source-short QD and the drain-short QD (blue circles). (b) Plot of current versus V_{LGD} and V_{LGS} showing the square array. Data from Run 2.57, Oct02_23.dat: T = 25 mK, V_{UG} = 2 V, V_{LGC} = -0.94, V_D = 1 mV, in device AF-CA2-U3D-3.

c) Intermediate-Coupling Regime

We can move from the weak-coupling regime to the intermediate-coupling regime by making the voltage on V_{LGC} less negative. In the intermediate-coupling regime, the coupling capacitance C_m is no longer much smaller than $C_{\Sigma L}$ and $C_{\Sigma R}$. This means that what happens on the left QD affects the right QD and vice versa.

The requirements for an electron to tunnel from source to drain are the same as the weak-coupling regime. To get from the source to the drain, an electron must tunnel three times. For $V_D = 0$, the chemical potential requirements for these three tunneling events are:

$$
\begin{aligned}
0 = \mu_S &= \mu_L(n_L, n_R, V_{GL}, V_{GR}) && \text{(blue)} \\
\mu_L(n_L, n_R, V_{GL}, V_{GR}) &= \mu_R(n_L, n_R, V_{GL}, V_{GR}) && \text{(green)} \\
0 = \mu_D &= \mu_R(n_L, n_R, V_{GL}, V_{GR}). && \text{(red)}
\end{aligned}
\qquad 1.26.
$$

These three equations are the same as Eq. 1.24, except that now μ_L is also function of the charge and voltage of the right QD ($-en_R$, V_R). I plot these three requirements in Figure 1.16. Figure 1.16 looks quite different than Figure 1.14. Because $\mu_L(n_L, n_R, V_{GL}, V_{GR})$ is a function of both V_{GL} and V_{GR}, the set of lines representing $\mu_L(n_L, n_R, V_{GL}, V_{GR}) = 0$ now have a slope. Because of the dependence on n_R, the lines are no longer continuous; instead, the line jumps where n_R changes.

Now I can use Figure 1.16 to understand current through the DQD in the intermediate-coupling regime. Let us start with the (n_L, n_R) charge configuration. Along the blue lines in Fig. 1.16, an electron can tunnel from the source to the left QD. Along the green lines, an electron can tunnel from the left QD to the right QD. Along the red lines, an electron can tunnel from the right QD to the drain. Only where these three lines intersect is current allowed to flow from source to drain. These intersections are called triple points. In the intermediate-coupling regime, the triple points form the vertices of a hexagon, so this is called a hexagon diagram or a honeycomb diagram.

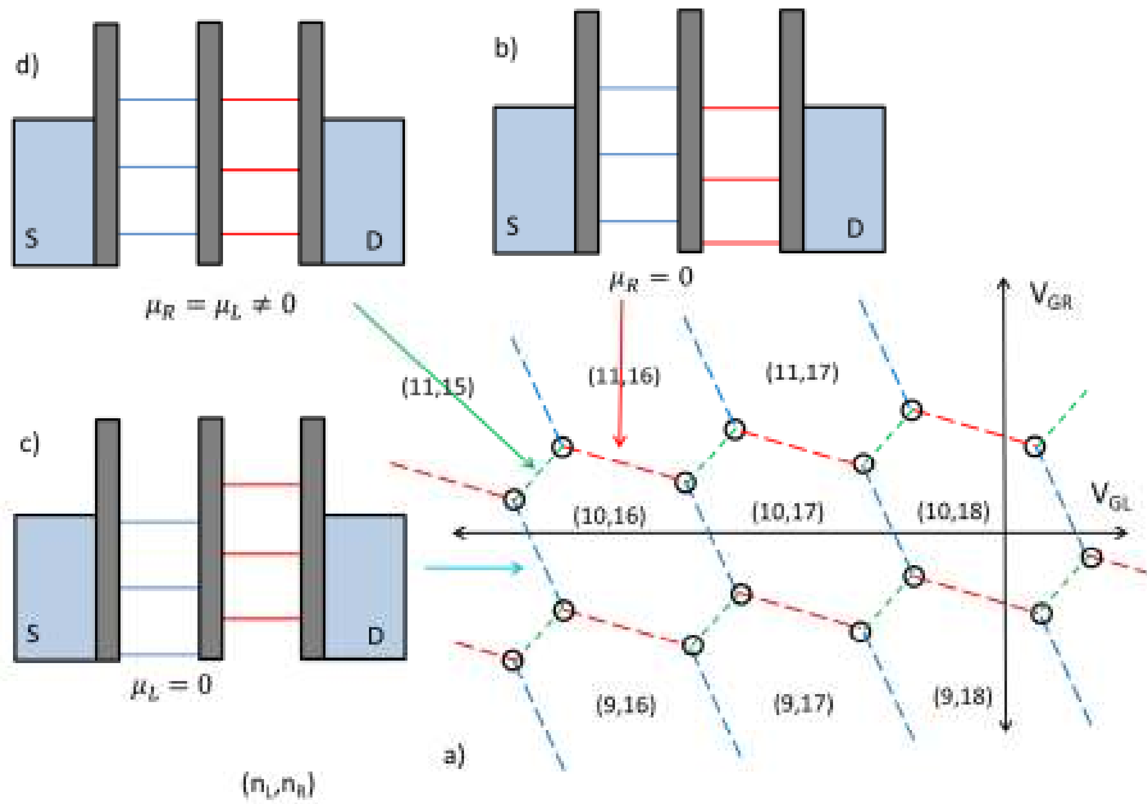


Figure 1.16. (a) Schematic showing where current through a DQD is possible in the intermediate-coupling regime. Along the red dashed lines $\mu_R = 0$, which is shown schematically in (b) . Along the blue dashed line $\mu_L = 0$, (c). Only where $\mu_R = \mu_L = 0$, which occurs at the intersection of the blue and red lines that are represented by the black circles, will current be able to flow. The green dashed lines show where $\mu_R = \mu_L \neq 0$, (d). (n_L, n_R) represent the charge on the on the left and right QDs.

In Figure 1.17 I show a honeycomb diagram, with a hexagon drawn on for clarity.

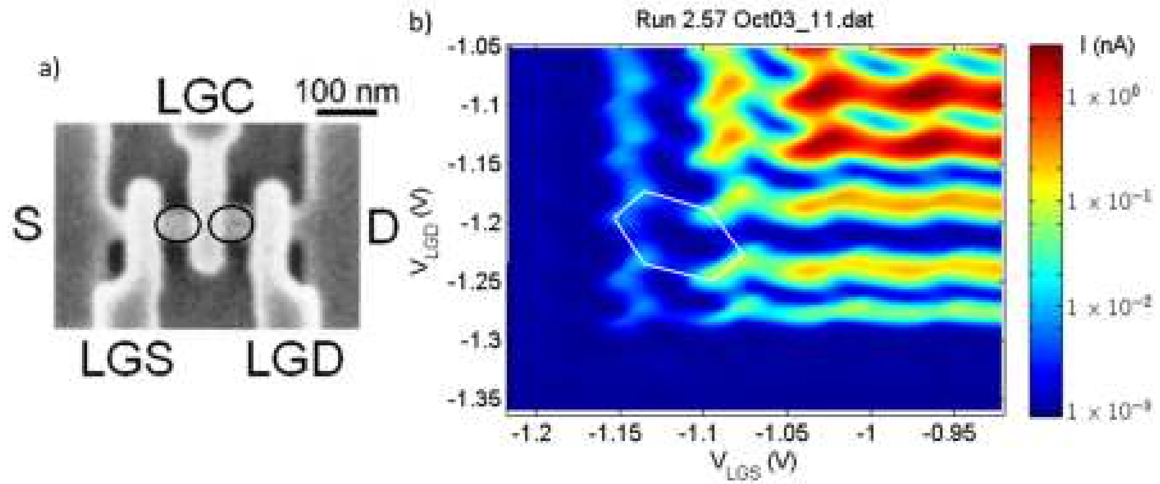


Figure 1.17. DQD in the intermediate-coupling regime. (a) The blue circles show the locations of the two QDs in the device. (b) Current through the DQD in the intermediate-coupling regime, showing hexagons. Data from Run 2.57 Oct03_11.data, T = 25 mK, V_{UG} = 2 V, V_{LGC} = -0.83 V, V_D = 1mV with device AF-CA2-U3D-3.

As V_{LGC} becomes even less negative, the coupling capacitance C_m continues to increase, and the coupling between the two QDs grows stronger. Eventually, the two QDs will merge into a single QD, and the current peaks form continuous diagonal lines, as seen in Figure 1.10. This is called the strong-coupling regime.

4. Applications of QDs

Now that I have explained how DQDs operate, I can return to the applications that I mentioned at the beginning of this chapter (single-electron pumps and charge qubits), and explain how they can be accomplished.

a) Single-Electron Pump

In the previous section I showed how current can flow through a DQD. Based on this understanding it is easy to see how to pump electrons through at a certain frequency (Figure 1.18).

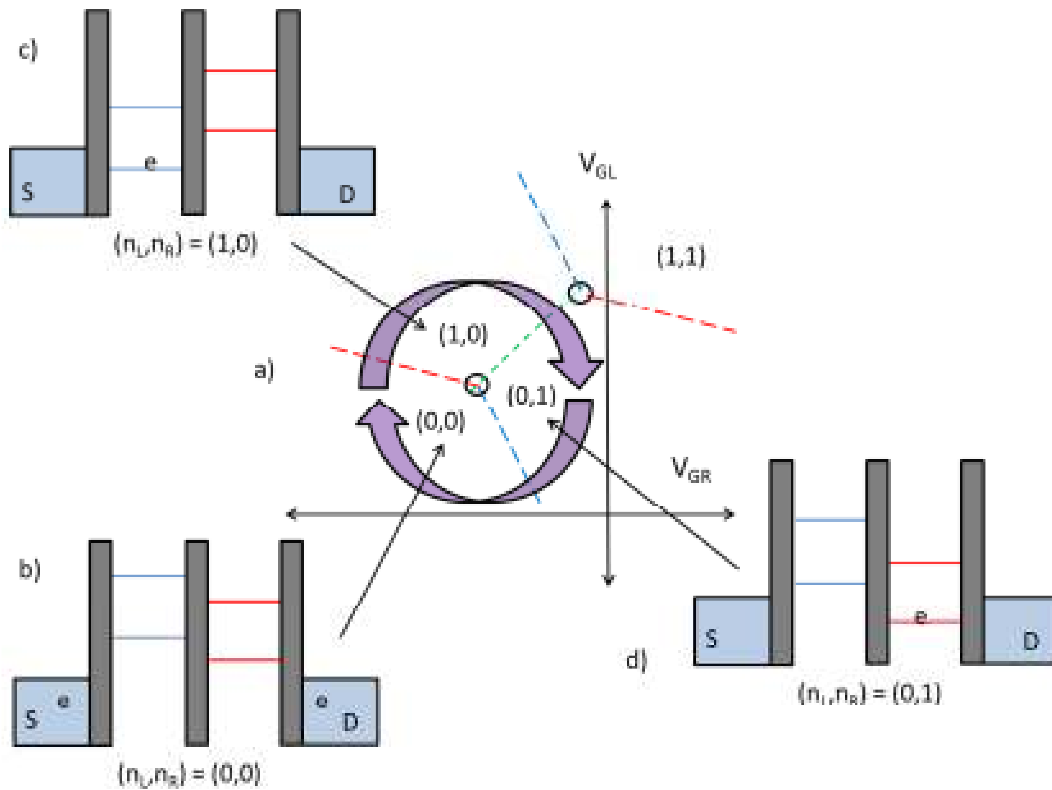


Figure 1.18. Schematic of charge pumping cycle in a DQD. (a) The purple arrows show the path in gate voltage through the honeycomb diagram during one pump cycle. The cycle starts in the (0,0) charge configuration (b). By going to the (1,0) charge configuration, one electron tunnels onto the left QD (c). The electron then tunnels to the right QD (c) and finally to the drain (b).

In the single-electron pump, sinusoidal voltages that are 90° degrees out of phase are applied to V_{GR} and V_{GL}. When plotted on top of the honeycomb diagram in Figure 1.18, we see that the pump cycle goes through a circle in gate space that is centered on a triple point. Starting from the (0,0) charge configuration [Fig. 1.18(b)], we first move to the (1,0) charge configuration [Fig. 1.18(c)]. This causes an electron to tunnel from the source to the left QD. We next go to the (0,1) charge configuration [Fig. 1.18(d)], so an electron tunnels onto the right QD. Finally, we return to the (0,0) charge configuration, and an electron tunnels onto the drain [Fig. 1.18(d)]. Each cycle transfers exactly one electron from source to drain. Note that because the amplitude of the gate voltage drives is small, of order 10 mV, the gate voltage drives do not significantly change the barrier resistances.

In practice we must pump at a much lower frequency than the tunneling rates to achieve a low error rate, which limits the frequency of a single-electron pump. This also limits the current of a single-electron pump, because the current is proportional to the frequency, At present, the current through a single-electron pump is too small to use as a practical current standard [2,4]. One method to increase the current without sacrificing the error rate is to operate many single-electron pumps in parallel. I will return to this point when I discuss the challenges facing silicon QDs.

b) Qubits

Charge qubits are the other application for silicon QDs that I mentioned at the beginning of this chapter [9,11,13]. Figure 1.19 shows a simple method of operating a DQD to create a coherent charge superposition.

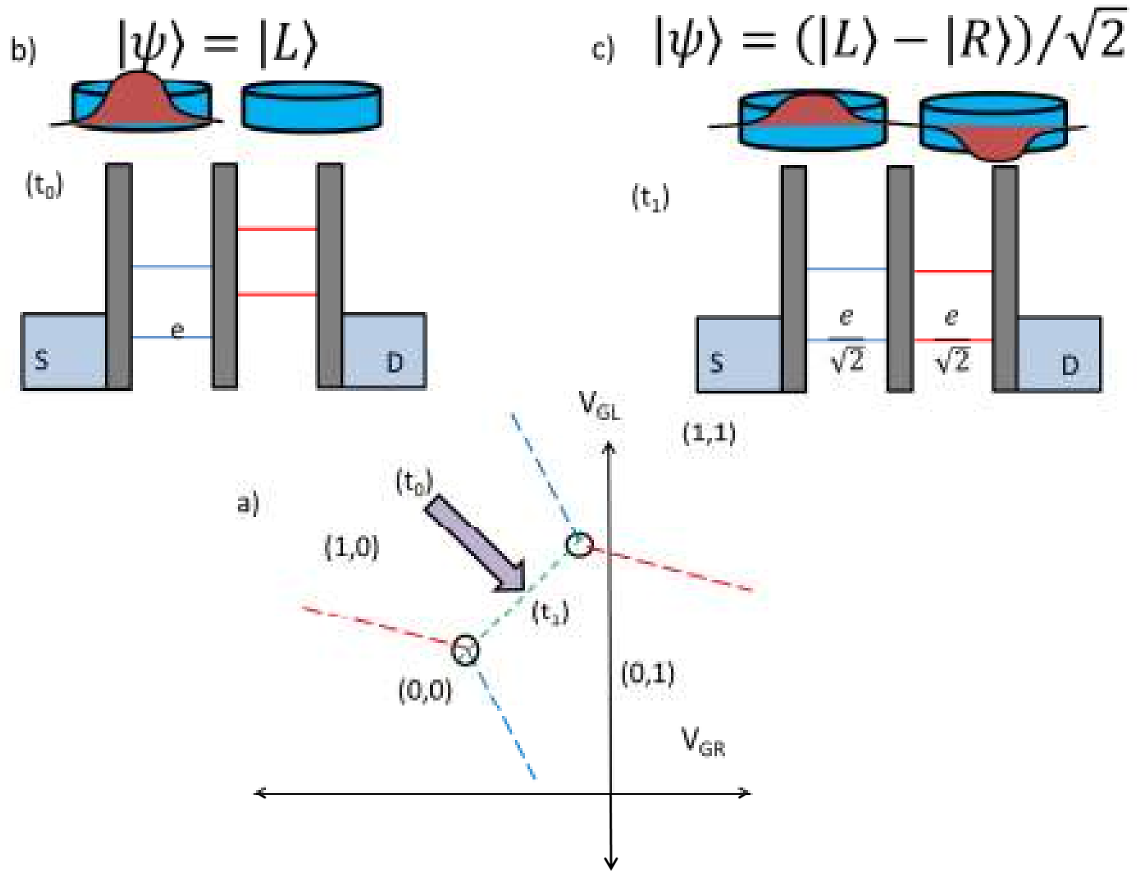


Figure 1.19. Charge qubit operation of a DQD. (a) Adiabatically move from deep in the (1,0) regime to where the (1,0) and (0,1) charge states are equal in energy (along the green line). (b) At time t_0, we start deep in the (1,0) charge state and the electron is in the left QD. (c) At time t_1, the electron is in the anti-symmetric combination of being in the left QD and right QD simultaneously.

We start deep in the (1,0) charge state at time t_0 [Fig. 1.15(a)]. The initial wave function is $|\psi(t_0)\rangle = |L\rangle$. By changing the gate voltages adiabatically, we move to the condition $\mu_L(1{,}0, V_{GL}, V_{GR}) = \mu_R(0{,}1, V_{GL}, V_{GR})$. Because this is an adiabatic process, the wave function remains in the ground state, and attime t_1, the wave function is $|\psi(t_1)\rangle =$

$(|L\rangle - |R\rangle)/\sqrt{2}$, which is the anti-symmetric superposition of being in the left and right QD at the same time.

To make a useful quantum computer will require many qubits. To operate many qubits will probably require parallel operation, just like the single-electron pumps.

D. Challenges for Silicon QDs

There have been just a few experimental demonstrations of both charge pumps [12,22] and charge qubits [11,13] in silicon QDs. One reason for the dearth of demonstrations is that most silicon QD devices do not behave as expected. It is common for silicon QD devices to have *unintentional* QDs in addition to the *intentional* QDs. In this section I address two questions related to the unintentional QDs: (1) How do we know that any of the QDs are intentional? (2) What are the signatures of unintentional QDs in current through a device?

Having an unintentional QD does not make a device useless. The first reports of Coulomb blockade oscillations in silicon were due to unintentional QDs [23–26], and the first three demonstrations of Pauli spin blockade in silicon were in devices in which at least one of the QDs was unintentional [27–29]. But having an unintentional QD makes a device harder to operate. Each additional QD is another chemical potential and another tunnel barrier that must be controlled in a device that has a limited number of gates. With only one or two unintentional QDs, we might no longer have enough control to perform the desired experiment.

1. Reproducibility

How can we distinguish between intentional and unintentional QDs? If a QD is intentional, it should have gate capacitances that are reproduced in nominally identical devices, and it should have gate capacitances that can be predicted by a capacitance simulator given the geometry and fabrication parameters. I discuss the reproducibility and predictability of the gate capacitances in my devices in Chapter 2.

Also, the applications for silicon QDs I discussed earlier will require uniform gate capacitances. For both single-electron pumps and charge qubits, we will probably need to operate the devices in parallel, with the same drive voltages applied to multiple devices. Uniform gate capacitances are needed so that the same voltages have the same effect on multiple devices. Uniform gate capacitances are necessary but not sufficient to let us operate the devices in parallel. An example of something else that would be required is uniform threshold voltages.

2. Unintentional Quantum Dots

What is the signature of an unintentional QD? In Figure 1.6, I showed an example of the current through a DQD as a function of V_{LGD} that is bumpier than the current as a function of V_{LGS}. The LGD curve is bumpier because there are two unintentional QDs located in the nanowire near the LGD. I will prove this in Chapter 3. About a third of the lower gates I have studied have an unintentional QD associated with them. These unintentional QDs are not just a problem in our devices; they are endemic in the field of silicon QDs. In the rest of this section, I will show unintentional QDs in three device architectures that are similar to mine.

Figure 1.20 shows a device architecture that is very similar to the one I studied. This device architecture, made at UNSW, has metal gates on top of bulk silicon [30]. Electrically, these devices operate just like the devices I studied. The G gate inverts the silicon (like UG in my devices), and B1 and B2 (like LGS and LGD) are used to deplete the silicon to create tunnel barriers. Whereas in our devices the lateral confinement is created by the nanowire, here the lateral confinement in the bulk silicon is created by patterning the G (upper) gate to be 50 nm wide. Another difference, that will become important in Chapter 5, is that the gates are made with aluminum rather than poly-silicon.

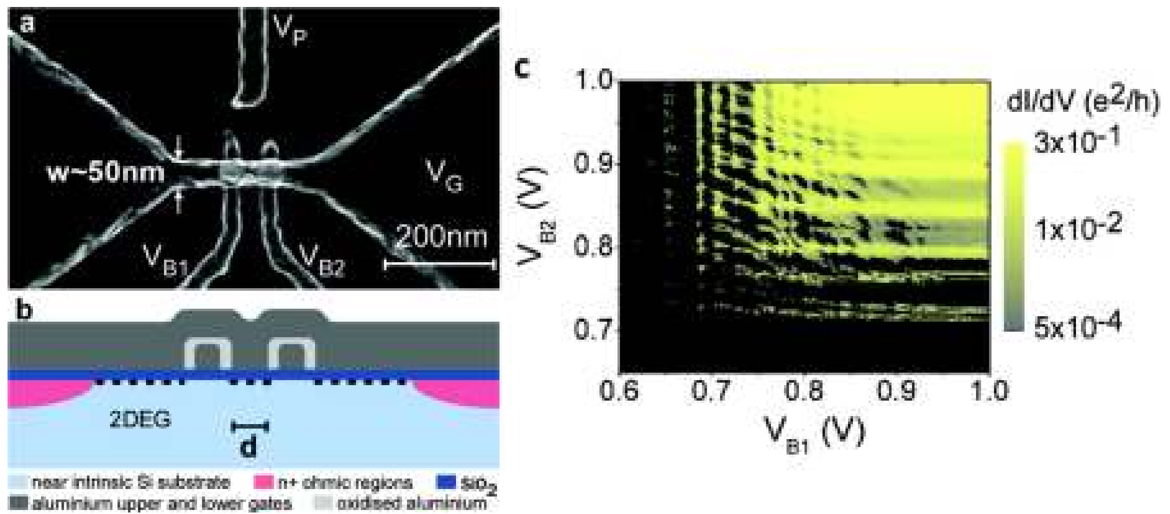


Figure 1.20. Unintentional QDs in a device with two layers of patterned metal gates on bulk silicon from UNSW. (a) Micrograph of the device. (b) Schematic cross section showing device operation. The G gate inverts the silicon, and B1 and B2 deplete the silicon to form tunnel barriers. (c) Conductance through the device as a function of B1 and B2, showing at least three QDs, two of which are unintentional. Adapted with permission from S.J. Angus, *et al*. Nano Lett. **7** pp 2051 [30]. Copyright (2007) American Chemical Society.

In Figure 1.20(c) we see the conductance as a function of V_{B1} and V_{B2}. Because these gates are equivalent to LGS and LGD in the devices I studied and this device should only have a single QD, the peaks in conductance should look like a single a single set of parallel, diagonal lines (as in Fig. 1.10). Instead, there are at least three sets of parallel peaks, with each set having a different slopes. This means that there are at least three QDs in this device, two of them unintentional. The set of horizontal peaks are caused by an unintentional QD near B2, and the set of vertical peaks are caused by an unintentional QD near B1. Angus, *et al*, attribute the unintentional QDs to disorder, but they do not determine the cause of the disorder, although they suggest interface traps might be the cause. Despite the unintentional QDs, this device architecture has been used to do several experiments, such as a single phosphorus donor qubit demonstration [31] and a single-electron pump [21]. Unintentional QDs were identified as a potential source of errors in their single-electron pump experiment [22].

Figure 1.21 shows another device architecture with metal gates on bulk silicon. This device was made in the Electrical Engineering Department at the University of Maryland, and it was used in one of the first demonstrations of Pauli spin blockade in silicon QDs [29]. As in our devices, there are two gate layers. The top-gate (like our UG) is a global gate which inverts the silicon below. The lower gates deplete the silicon below to create the lateral confinement. If only gates C and D are used, then only a single tunnel junction should form. However, Figure 1.21(d) shows the current through the device as the voltage applied to both gates C and D is swept. The oscillations in Figure 1.21(d) are due to an unintentional QD. The location of the unintentional QD is shown in Figure 1.21(c). The conductance as a function of voltage applied to both gates A and B

shows a different unintentional QD. The authors suggest that interface traps are the origin of the unintentional QDs.

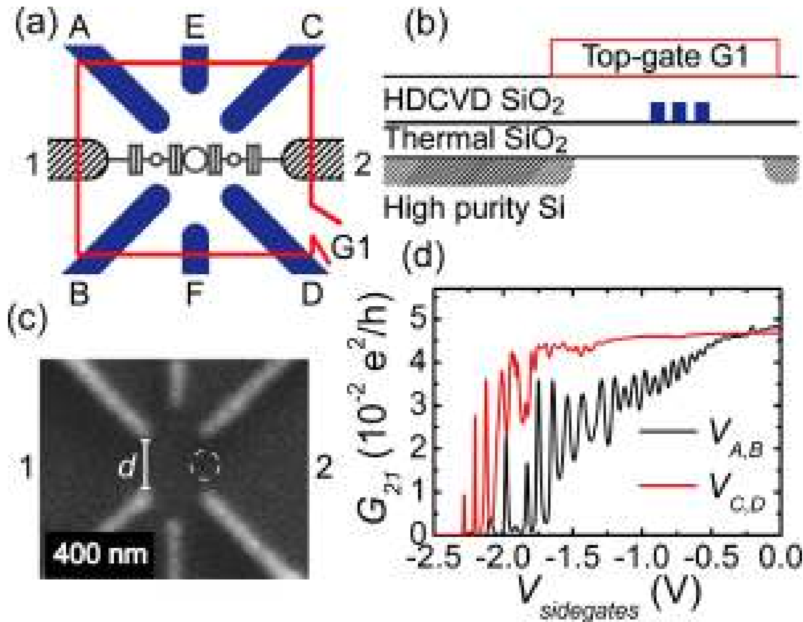


Figure 1.21. Unintentional QDs in a device with a global upper gate and patterned lower gates from B. Hu and C. H. Yang at the University of Maryland. (a) Schematic of the device. The large circle in the center is the intentional QD, the two small circles are the unintentional QDs. (b) Schematic of the device, as seen from the side. (c) Micrograph of the device during fabrication. The circle shows the location of an unintentional QD. (d) Red and black curves show conductance as two of the lower gates are used to deplete the silicon. The oscillations are due to an unintentional QD. Reprinted with permission from B. Hu and C. H. Yang PRB **80**, 075310 (2009) [29]. Copyright (2009) by the American Physical Society.

The third architecture I discuss is from Sandia National Laboratory [Fig. 1.22]. These bulk silicon devices have a metal upper gate and poly-silicon lower gates [32,33]. Again, the upper gate inverts the silicon, and the lower gates deplete the silicon. In Figure 1.22(c) the current through the constriction between gates D and E shows Coulomb blockade peaks due to unintentional QDs. They suggest that interface traps are the origin of the unintentional QDs.

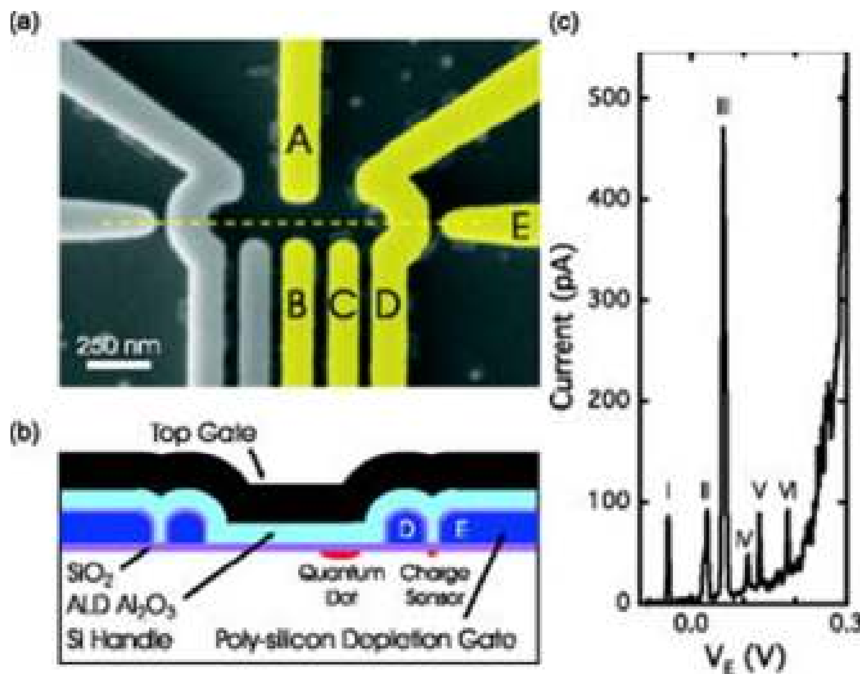


Figure 1.22. A device with metal top gate and poly-silicon lower gates on top of bulk silicon from Sandia Nat. Lab. (a) Micrograph of the device before the top gate was deposited. (b) Schematic cross section of the device through the dashed yellow line in (a). (c) Current through the gap between gates D and E, as a function of V_E, showing unintentional QDs. Adapted with permission from E.P. Nordberg, *et al*, APL **95**, 202102, (2009) Ref. [33] . Copyright (2009) American Chemical Society.

These are just three examples of devices from the Si-SiO_2 QD community, but this shows that it is difficult to make a tunnel barrier that does not have any unintentional QDs. While these groups attributed the unintentional QDs to interface traps, in Chapter 5 I will suggest that many of the unintentional QDs both in our devices and in others could be due to strain.

E. Advantages of Silicon

Having described the problems with silicon QDs, I want to explain why silicon is still a good material for QDs. The first advantage of silicon is that it is the dominant material for making transistors. Not only can we use tools developed by the microelectronics industry to make our devices, but we also use the decades of knowledge about practical techniques for making high-quality, low-defect devices.

A second reason for using silicon is that several studies at NIST have demonstrated the stability of silicon QD devices [34–37]. Instabilities would show up as a change, or offset, to the Coulomb blockade oscillations as a function of time.

The third reason why silicon is a good material for QDs is specific to charge qubits. Charge qubits in silicon have been demonstrated to have longer coherences times than charge qubits in other semiconducting systems [11,13]. One reason for the longer coherence times is that silicon has inversion symmetry, unlike GaAs and other III-Vs. Therefore, it does not have piezo-electric coupling. This reduces the electron-phonon coupling in silicon and extends the coherence time [38].

Another type of qubit that can be made in silicon is a spin qubit [7,39]. Because this is a more complicated qubit, I will not describe it here. But there are additional

benefits to working in silicon for spin qubits, as compared to other material systems, such as GaAs. First, the hyperfine coupling is smaller, because only 5 % of the atoms are an isotope with a nuclear spin (^{29}Si), as compared with 100 % of the atoms in GaAs. The concentration of ^{29}Si can be reduced through isotopic enrichment. Second, the spin-orbit coupling is smaller in silicon than in GaAs, because silicon has a smaller atomic number than gallium or arsenide. Also the inversion symmetry of the silicon lattice means there is no Dresselhaus spin-orbit coupling. For more details, good reviews are provided in references [6,7,38].

F. Outline of Dissertation

In the rest of this dissertation I discuss my work to understand the reproducibility of silicon QDs and why unintentional QDs are so common in silicon QD devices. In Chapter 2, I focus on the reproducibility of the gate capacitances to intentional QDs. I show that silicon QDs can be made with reproducible gate capacitances, and I also that I can use a capacitance simulator to predict the gate capacitances. After Chapter 2, I move on to discussing the problem of the unintentional QDs. In Chapter 3, I show the effect of unintentional QDs in our devices, and then I describe a technique to determine the location of those QDs with nanometer precision using the capacitance simulator from Chapter 2.

In Chapter 4, I review the basics of stress, strain and the silicon band structure. In Chapter 5, I simulate the strain in three different devices to show how strain could be the cause of many of those unintentional QDs. I will argue that for typical parameters strain-induced QDs should be expected in many different materials and geometries of QD devices.

Finally, in Chapter 6, I conclude by giving an outlook of possibilities for future work, including the future of strain-induced QDs.

Chapter 2: Reproducibility and Predictability of Gate Capacitance to Intentional Quantum Dots

Based on "Simulating Capacitances to Silicon Quantum Dots: Breakdown of the Parallel Plate Capacitor Model," by Ted Thorbeck, Akira Fujiwara, and Neil M. Zimmerman, published in IEEE Trans. Nano. **11** 1536, (2012).

A. Preview

In the previous Chapter I argued that a practical charge pump or charge-qubit based quantum computer will require reproducible and predictable gate capacitances. The reproducibility and predictability of the gate capacitances will be quantified in this Chapter. First, nominally identical devices should have reproducible gate capacitances. The gate capacitances in our devices are reproducible to within 10% for devices with many different sizes. Second, gate capacitances should scale with the size of the QD. In our devices the gate capacitances do scale with the size of the device as determined from fabrication parameters. Third, we should be able to predict the gate capacitances based on the size of the devices from the parameters used during fabrication. The gate capacitance in our devices can be predicted to within 20% using the fabrication parameters.

In performing this analysis, I gained several new insights into the operation of these devices. One insight is that, for small devices in this geometry, fringing electric fields can dominate the gate capacitance. The fringing fields cause the parallel plate

method, a commonly used method to estimate gate capacitances, to break down. But I have developed a method to improve the estimate by including fringing fields.

B. Motivation

The first motivation for studying the reproducibility and predictability of the gate capacitances is that we can determine if the measured QDs are intentional. In this Chapter we will see that the measured gate capacitance match the predicted capacitance, giving us confidence that these QDs are intentional. In the next Chapter, I will show QDs that do not match these predictions for the gate capacitances, so those QDs are unintentional. But I will use the same capacitance simulator that I used in this Chapter to determine the locations of those unintentional QDs.

The second motivation for studying the reproducibility and predictability of the gate capacitances is that it can help us make devices that are more useful by, for example, minimizing the cross capacitances or by correctly predicting the maximum operating temperature of a QD device. As discussed in the previous Chapter, to observe Coulomb blockade the charging energy of the QD must be larger than the thermal energy, $E_{\Sigma} = \frac{e^2}{C_{\Sigma}} > kT$. For these devices, a typical gate capacitance is between 2 aF and 30 aF, and a typical barrier capacitances is between 10 and 50 aF [40]. So $C_{\Sigma} = 40\ aF$ is a reasonable estimate for the smallest total capacitance in this architecture. This means that E_{Σ}= 4 meV (or 50 K in thermal units). This device will show robust Coulomb blockade at liquid helium temperature (4.2 K), but not at liquid nitrogen temperature (77 K).

We would like to design devices that operate at liquid nitrogen temperatures. A capacitance simulator would let us predict the operating temperature before we made the device.

C. Previous Work

Work was done at NIST, prior to my arrival, on the reproducibility of gate capacitances. This work showed reproducible gate capacitances for a single set of 3 lithographically similar devices [14]. In this Chapter, all of the devices we have studied over the years will be discussed. With all of this data, I can show that nominally identical devices have reproducible gate capacitances for a variety of device dimensions. Also, this is the first time the reproducibility of gate capacitances has been described numerically.

Gate capacitances that scale with a single fabrication parameters have been shown previously [41–43]. Here, I show how multiple fabrication parameters can be combined to scale with the gate capacitance.

There has been previous work on predicting gate capacitances using device dimensions from fabrication [44,45]. The simplest method is to treat the QD and the gate as the two plates of a parallel plate capacitor [30,41,42,46]. I will show how this approximation breaks down, and I will show how it can be improved.

Some groups have done numerical simulations of the device to predict gate capacitances. However, most of these simulations were of only a single device, whereas I do it for all of our devices. Often these other simulations require several fitting

parameters to match the measured capacitances, but my simulations do not require any fitting parameters.

D. Measurements

Over the years of studying these devices we measured many gate capacitances in these devices. There was no intention to do a comprehensive set of gate capacitance measurements, so for many of the devices we only measured a few of the gate capacitances. In measuring gate capacitances from the old data, I have benefitted from the measurements of several people including Neil Zimmerman, Akira Fujiwara, Manolis Hourdakis and Stuart Martin.

1. Table of All Measurements

Table 2.1, at the end of this Chapter, is a compendium of all of the gate capacitances we have measured in these devices at NIST. The table is arranged by the run number and the device label. To understand Table 2.1, our device labeling must first be understood. A typical device label is “AF-CA2U3D-4”. The “AF” corresponds to Akira Fujiwara, who led the team at NTT which fabricated the devices and brought them to NIST. “CA2” or “CA3” corresponds to which wafer from which the device comes. Each wafer is divided into 13 identical sections, labeled with a letter (A-J,L,R,U); this dye came from the U section. Each section of the wafer has 36 dies in it; this die came from row 3 column D of the die. Each die is about 2 mm by 2mm. A single die, or chip, is seen in Figure 2.1 (a) and (b). On each chip are 4 devices, this device is number 4. In each run we would measure one or two devices from a single die. A single device is seen in Figure 2.1 (c) and (d).

Each device has 5 columns to identify the fabrication parameters. These parameters are shown in Figure 2.2. The nanowire has thickness t_{Si}, width w_{Si}, and length L_{Si}. The nanowire is surrounded by 20 nm of thermally grown silicon dioxide, t_{ox1}. The length of the lower gates is L_{LG}. The lower gates are surrounded by another thermally grown silicon dioxide, $t_{ox,2}$ = 30 nm. Finally, the UG length, L_{UG}, is defined as the length between lower gates that is filled with UG poly-silicon.

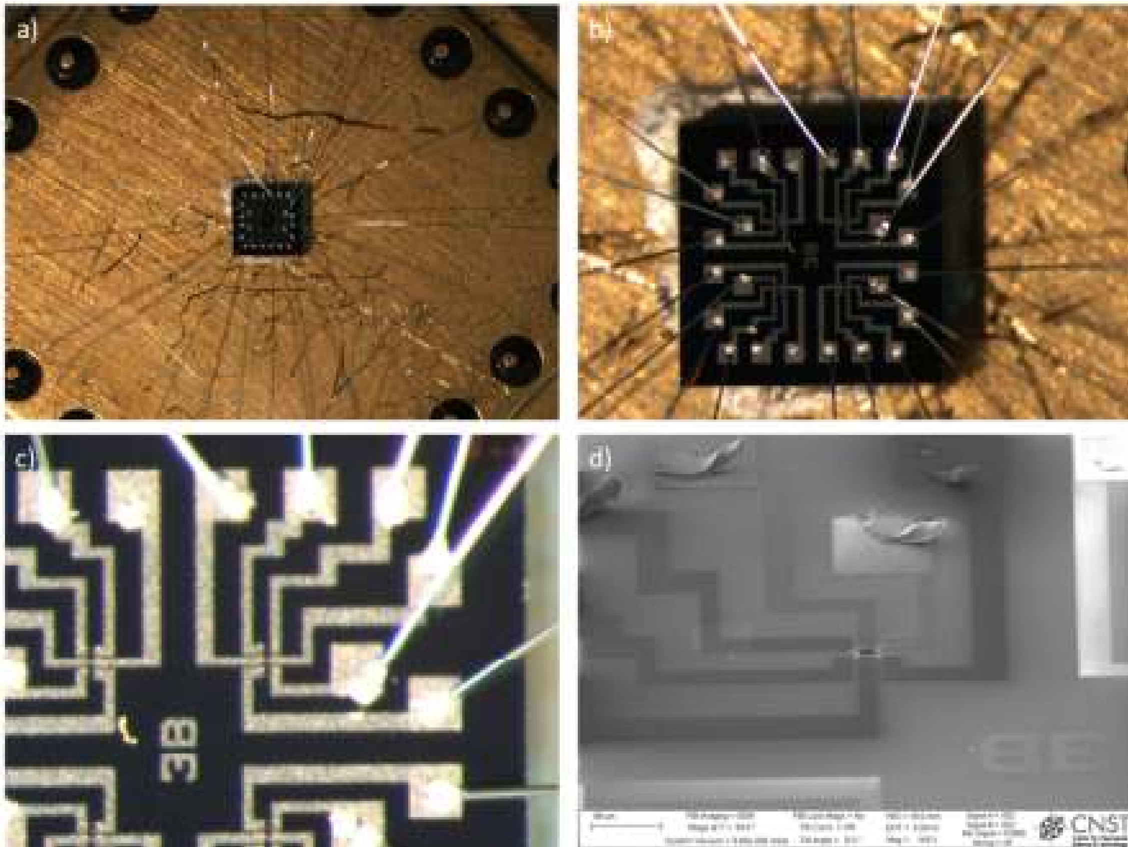


Figure 2.1 Pictures of a chip. (a) A chip on top of a chip header. The chip is 2 mm by 2 mm. (b) The chip at higher magnification. (c) One device on a chip. (d) Low magnification scanning electron micrograph of a device with broken wire bonds. The scale bar in this micrograph is 100 μm.

In each device we can measure up to 12 gate capacitances, from each of four gates to each of three QDs. The four gates are labeled UG, LGS, LGC and LGD. The large QD, which forms between LGS and LGD, is labeled the "full QD". The short QD between LGS and LGC is labeled "S-short QD", and the short QD between LGC and LGD is labeled "D-short QD." In Table 2.1 the gate capacitance from UG to the S-short QD is labeled UG-S-short.

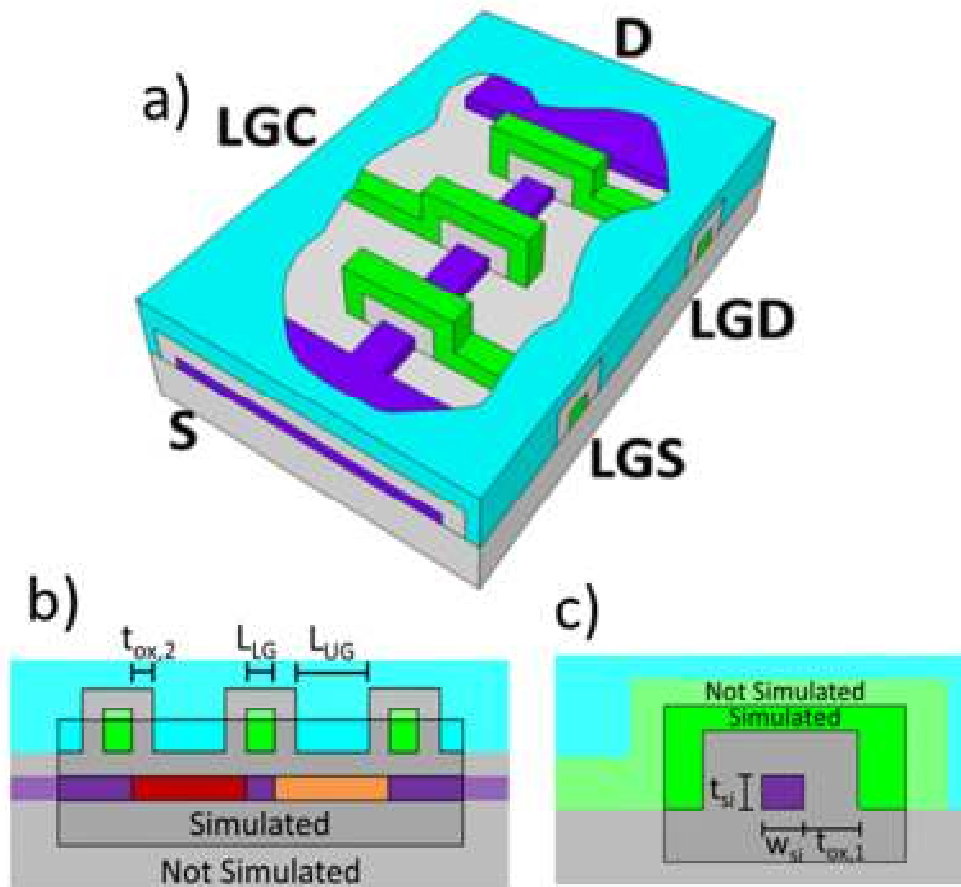


Figure 2.2. Schematics of the device showing fabrication parameters and simulated volume. (a) Cutaway schematic of the device. (b) Cross section of the device along the direction of the nanowire. (c) Cross section perpendicular to the nanowire.

Each capacitance measurement is also labeled with the data file that contains the measurement. As an example, Aug1_15 means the data was the 15th data set taken on August 1.

Table 2.1 contains many different devices with many different dimensions. To make a table that is easier to examine, I took two nominally identical devices from Table 2.1 and put them in Table 2.2. There are many notable features in Table 2.2. First, the measured capacitances make intuitive sense. For the full QD, the capacitances from LGS (C_{LGS} = 2.7 aF in device 1) and LGD (C_{LGD} = 2.5 aF in device 1) are similar, as they should be by symmetry. Likewise, for the S-Short QD, the capacitance from LGS (C_{LGS} = 2.3 aF in device 1) and LGC (C_{LGC} = 2.8 aF in device 1) are similar, as we would expect from symmetry. Also for the S-short QDs, the capacitances from LGS and LGC are both much larger than the capacitance from LGD (C_{LGD} = 0.1 aF in device 1), which is screened by the UG. Finally, because the full QD can be thought of as the sum of the S-short and D-short QDs, the capacitances from UG to the full QD (C_{UG} = 22 aF in device 1) is equal to the sum of capacitances from UG to S-short (C_{UG} = 11 aF in device 1) and D-short (C_{UG} = 11 aF in device 1).

In Table 2.2 we see that the gate capacitances are reproducible in two devices. But to understand the reproducibility of all the gate capacitances in Table 2.1, we need a numerical method of describing the reproducibility. To do this I will introduce a parameter called the deviation. The deviation equals the absolute value of the difference between an individual capacitance measurement and the average of all measured capacitances for identical devices, divided by the average. For the data in Table 2.2, all of the deviations are less than 15% and the average deviation is 6%.

Table 2.2. Measured capacitances for two nominally identical devices and as well as simulated capacitances for the same geometry. Device 1 is AF-CA2F3E-1, and device 2 is AF-CA2R3E-1. For comparison to Figures 2.3, the area of the nanowire directly below LGC is 940 nm^2; for comparison to Figure 2.4, the area of the short QD directly below the UG is 3760 nm^2.

Cap (aF)	UG			LGS			LGC			LGD		
	Dev 1	Dev 2	Sim	Dev 1	Dev 2	Sim	Dev 1	Dev 2	Sim	Dev1	Dev 2	Sim
Full QD	22	22	25	2.7	3.0	2.6	6.2	6.0	6.4	2.5	2.8	2.6
Source-Short QD	11	10	12	2.3	2.8	2.6	2.8	2.6	2.6	0.1	0.1	0.1
Drain-Short QD	11	11	12	0.1	0.1	0.1	3.1	2.8	2.6	2.4	2.5	2.6

2. Visual Presentation of Gate Capacitances

To examine the reproducibility and predictability of the gate capacitances, I will use the gate capacitances for which we have the most data. The gate capacitance we have measured the most often is LGC-full, which is shown Figure 2.3. Because S-short and D-short are nominally identical, the capacitances from UG to S-short and D-short are combined into a single set of data, which is shown in Figure 2.4. I will use these plots to make two points: (1) the gate capacitances are reproducible between nominally identical devices and (2) the gate capacitances scale with the fabrication parameters.

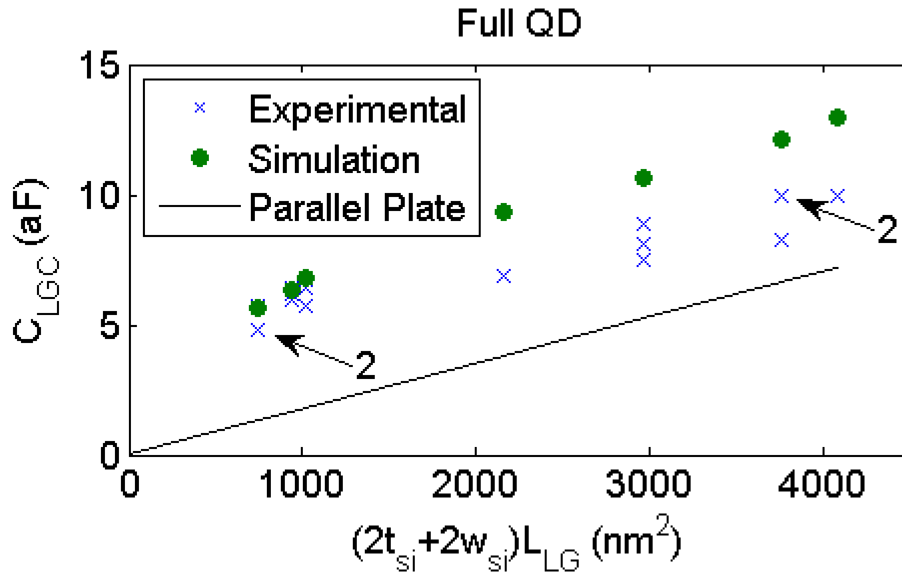


Figure 2.3. Measured and simulated gate capacitances from LGC to the full QD showing the reproducibility and predictability of the gate capacitances. Measured gate capacitances are blue x's, simulated gate capacitance are green circles, and the planar capacitance model is the black line. Arrows with a number correspond to multiple data points on top of each other. The horizontal axis represents the area of the nanowire directly below the lower gate, $(2t_{si}+2w_{si})L_{LG}$. Uncertainties are not shown, but are typically smaller than the plotting symbols.

Figure 2.3 shows the measured capacitance from LGC to the full QD for 13 different devices. The devices are parameterized by the surface area of the nanowire directly underneath the lower gate. This area is given by the perimeter of the nanowire, $2t_{si}+2w_{si}$, multiplied by the length of the lower gate, L_{LG}. In the Figure the blue x's are the measured capacitances, the green circles are the simulated capacitances, the black line

is a calculated capacitance using the planar capacitance model. I will return to the simulated and calculated capacitances later.

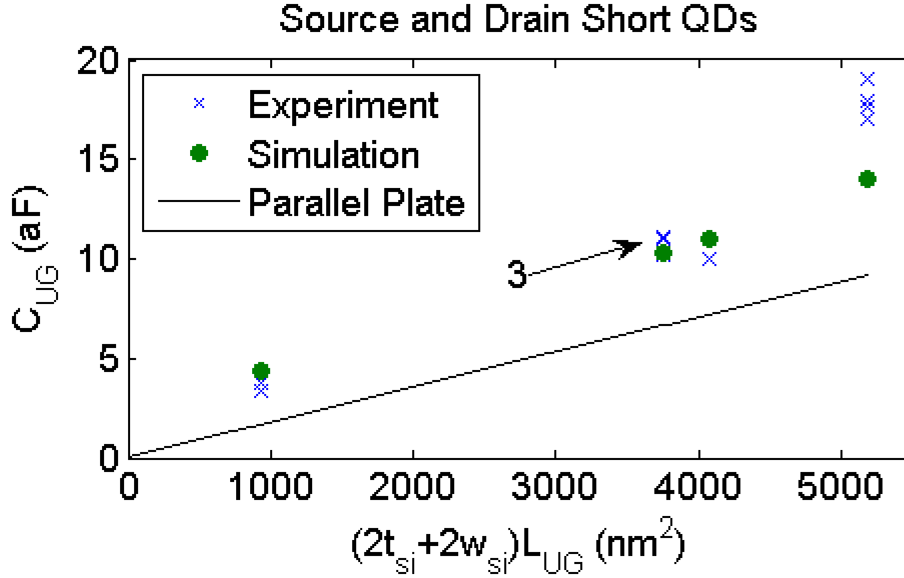


Figure 2.4. Measured and simulated gate capacitances from the UG to either the source or drain short QD showing the reproducibility and predictability of the gate capacitances. Area is parameterized by the nanowire below the UG. Symbols used are the same as in Figure 2.3.

The capacitances from the UG to the short QDs (both source and drain) are shown in Figure 2.4. The devices are parameterized by the perimeter of the nanowire, $2t_{si}+2w_{si}$, multiplied by the length of the upper gate, L_{UG}.

Figures 2.3 and 2.4 make two of the key points of this Chapter. First, notice that in Figures 2.3 and 2.4, for devices with the same surface area, the spread of measured capacitances is small. This shows that nominally identical devices have reproducible

gate capacitances. For Figure 2.3 the maximum deviation is 9% and the average deviation is 7%. For Figure 2.4 the maximum deviation is 6 % and the average deviation is 5%. Second, notice that in Figures 2.3 and 2.4 the spread of capacitances for a single surface area is much smaller than the total range of capacitances. This means that the gate capacitances are scaling with the fabrication parameters.

E. Simulations

Having shown that the gate capacitances are reproducible and scale with the fabrication parameters, I will now use the fabrication parameters to numerically predict the gate capacitances. I will start with the parallel plate method. This method is commonly used because it is quick, but it does not accurately predict gate capacitances, especially for small devices. Then I discuss how to easily improve the parallel plate method by including fringing fields. Finally, I use a capacitance simulator to predict the capacitances, which can predict the gate capacitances to within 20% without fitting parameters.

1. Parallel Plate Method

The parallel plate method is the simplest way to predict the gate capacitances. We use the surface area of the nanowire directly below the gate of interest as the area of a parallel plate capacitor. Therefore the capacitance is

$$C = \epsilon_{SiO_2}(2t_{si} + 2w_{si})L_{LG}/t_{ox,1}. \qquad 2.1.$$

where $\epsilon_{SiO_2} = 3.9\epsilon_0$ and the other variables are defined in Figure 2.2. Capacitances calculated using the parallel plate method are shown by a black line in Figures 2.3 and 2.4. Notice that the parallel plate method does an adequate job of predicting the slope of

the measured gate capacitances, but it fails to predict the y-intercept of the measured capacitances. The difference between the predicted and measured capacitances is most significant for the smallest devices. This is unfortunate because the smallest QDs have the highest operating temperature. Because the parallel plate method underestimates the capacitances, it would lead us to overestimate the maximum operating temperature. For the smallest devices we have studied, the average measured capacitance is C_{LGC} = 5.1 aF, but the parallel plate method only predicts a capacitance of C_{LGC} = 1.0 aF.

What is the parallel plate method missing? It is missing the fringing fields from the gate to areas of the QD not directly below the gate. In Figure 2.5 I have drawn the field lines from LGC to the nanowire. The solid lines represent the fraction of total field lines that are captured in the parallel plate model. The dashed lines represent the additional fringing electrical field lines.

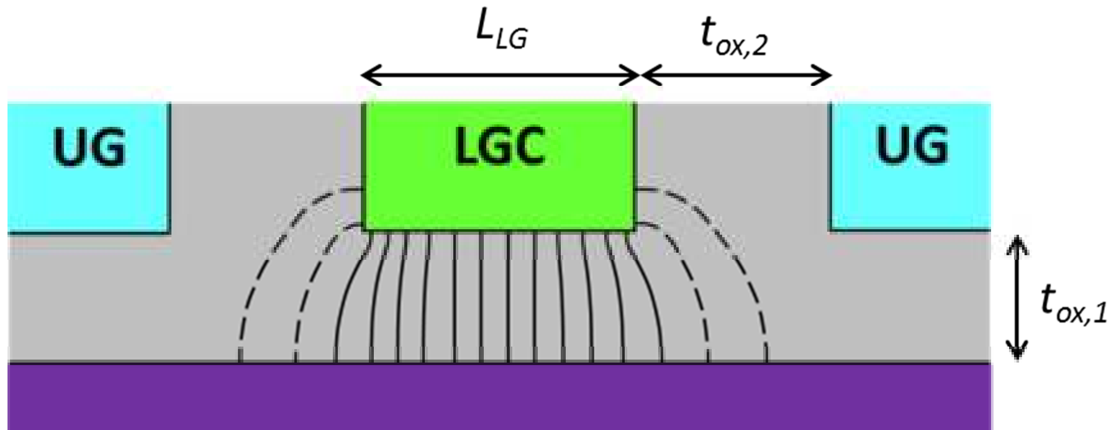


Fig 2.5 Schematic of the electric field lines from LGC to Full QD. The solid lines represent the portion of field lines that are calculated in the parallel plate model. Dashed lines represent the fringing field lines that are not capture by the parallel plate model.

Because the fringing field lines fall off with a characteristic length given by the separation of the nanowire and the gates, we can amend the parallel plate method to include the fringing fields using:

$$C = \epsilon_{SiO_2}(2t_{si} + 2w_{si})(L_{LG} + t_{ox,1})/t_{ox,1}. \quad 2.2.$$

This corrected estimate for the parallel plate method increases the estimate of the capacitance for previously mentioned smallest device from C_{LGC} = 1.0 aF to C_{LGC} = 3.9 aF, which is a much better estimate of the actual average measured capacitance of C_{LGC} = 5.1 aF. For all of the devices the average deviation is 14 %.

2. Capacitance Simulator

We need to use a capacitance simulator to do better than the corrected parallel plate model. I used FASTCAP, an electromagnetic simulator [47]. A detailed guide to doing the FASTCAP simulations is included in Appendix B.

The devices were simulated according the parameters used in fabrication: t_{si}, w_{si}, L_{LG}, L_{UG}, t_{ox1}, t_{ox2}. The nanowire was treated as a metal. The QDs were assumed to terminate at the near end of the lower gate creating the tunnel barrier, as shown in Figure 2.2(b). In the simulation all structures were terminated 50 nm away from the QD, because structures more than 50 nm away had little impact on the simulated capacitance. This physically means structures more than 50 nm away from the QD are screened by the gates. No fitting parameters were used in the simulation.

The FASTCAP simulations do a better job of predicting the capacitances than the planar capacitance model, especially for smaller devices. The results of the FASTCAP simulation are shown as green circles in Figures 2.3 and 2.4. For the smallest devices in

Figure 2.3, the average capacitance is 5.1 aF, while the simulation predicts 5.6 aF. This is a better prediction than corrected parallel plate method (Eq. 2.2), which predicted 3.9 aF, and a much better prediction than the uncorrected parallel plate method, which predicted 1.0 aF.

To quantify how well the simulation predicts the capacitances, I define the average deviation of the simulation, which is the absolute value of the difference between the average measured capacitance and the simulation divided by the average capacitance. The average deviation of the simulation for C_{LGC} is 17%, and the average deviation for C_{UG} is 14%.

All of the files for the FASTCAP simulation are in Guestroom PC C:\Ted\Program\FastCap\save\, and are saved in folders by device type.

F. Lessons Learned

Now I can take the all of the data I have on the capacitances, both experimental and predicted, and use it to extract lessons about the fabrication or operation of these devices.

1. Fabrication Implications

Now that we have all of this data on the measured and simulated capacitances, we can begin to compare the two to learn more about the fabrication of the devices. If the dimensions of the device are systematically different from the fabrication parameters, this should cause a systematic difference between the measured capacitances and the simulated capacitances. This feedback could help us to improve future devices.

As an example, as can be seen in Figure 2.3, the simulated capacitances for C_{LGC} are larger than the measured capacitance for the biggest devices. This observation could be explained if, for those devices, the thickness of the oxide isolating the UG from the LGs, $t_{ox,2}$, were larger than expected. Because this is a thermal oxide, a thicker $t_{ox,2}$ would make L_{LG} smaller than expected. This would lead us to predict capacitances that are too large. This trend might be explained by stress buildup during the thermal oxidation of the lower gates. For devices with longer lower gates, less stress might build up, which would result in more oxide growth. For the longer lower gates, this would make L_{LG} smaller than expected and $t_{ox,2}$ bigger than expected [16].

2. Four Sided Transport

When I first plotted the data in Figure 2.4, I plotted it as shown in Figure 2.6. In Figure 2.6 I used $(2t_{si} + w_{si})$ as the x-axis. I used this because I assumed that the QD would occupy only the three sides of the nanowire closest to the upper gate. In other words, I assumed that there was no inversion layer on the bottom of the nanowire. When the data are plotted in this way, the planar capacitance model does a worse job of predicting the measured gate capacitance; even the slope is wrong. So I began to wonder if the nanowire could be inverted along all four sides. I recalculated the planar capacitances using all four sides of the nanowire (as shown in Fig. 2.4), and now the planar capacitance model successfully predicts the slope of the measured data.

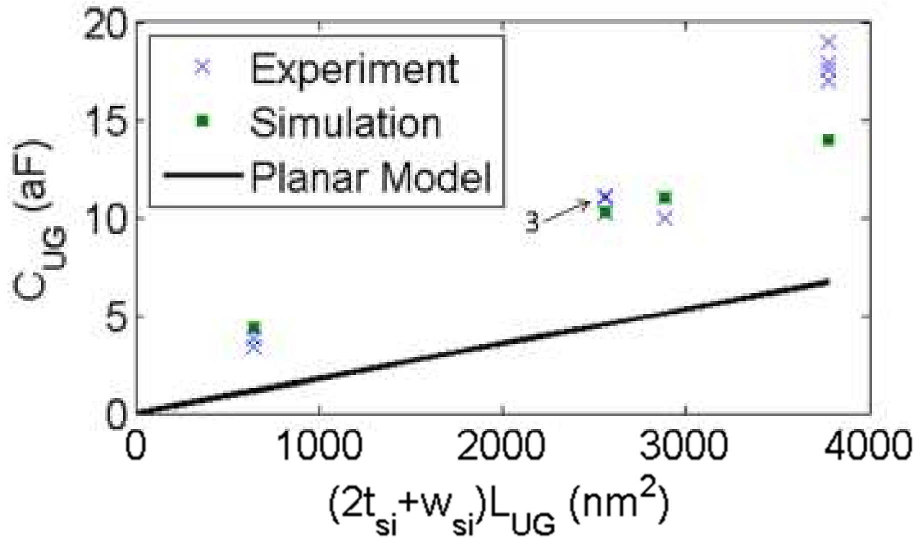


Figure 2.6. Capacitance to the UG from the full QD, where the area of the device does not include the bottom of the nanowire, using the same symbols as Figure 2.3.

How could the UG, which only surrounds three sides of the nanowire, invert all four sides? Many electric field lines would have to go from the UG to the bottom of the nanowire. To determine if this was reasonable, I simulated a 2D cross section of the nanowire and the UG in COMSOL (a finite element simulator used in Chapter 5 and Appendix C). In Figure 2.7 I show the calculated electric fields, for $V_{UG} = 1$ V and $V_{NW} = 0$ and $t_{si} = w_{si} = t_{ox,1} = 20$ nm. Confirming my expectations, the electric field strength is only 30% weaker on the bottom of the nanowire than on the top. Therefore, the charge density on the bottom of the nanowire was only 30% smaller on the bottom of the nanowire compared with the top of the nanowire. Because we typically operate these devices 0.5 V to 1.5 V above threshold, I think that it is reasonable to expect that an inversion layer can form on all four sides of the nanowire. I cannot rule out an alternative

suggestion that if there is not an inversion layer on the bottom of the nanowire, electric field lines might go through the bottom of the nanowire and terminate along the inversion layers on the top and sidewalls of the nanowire.

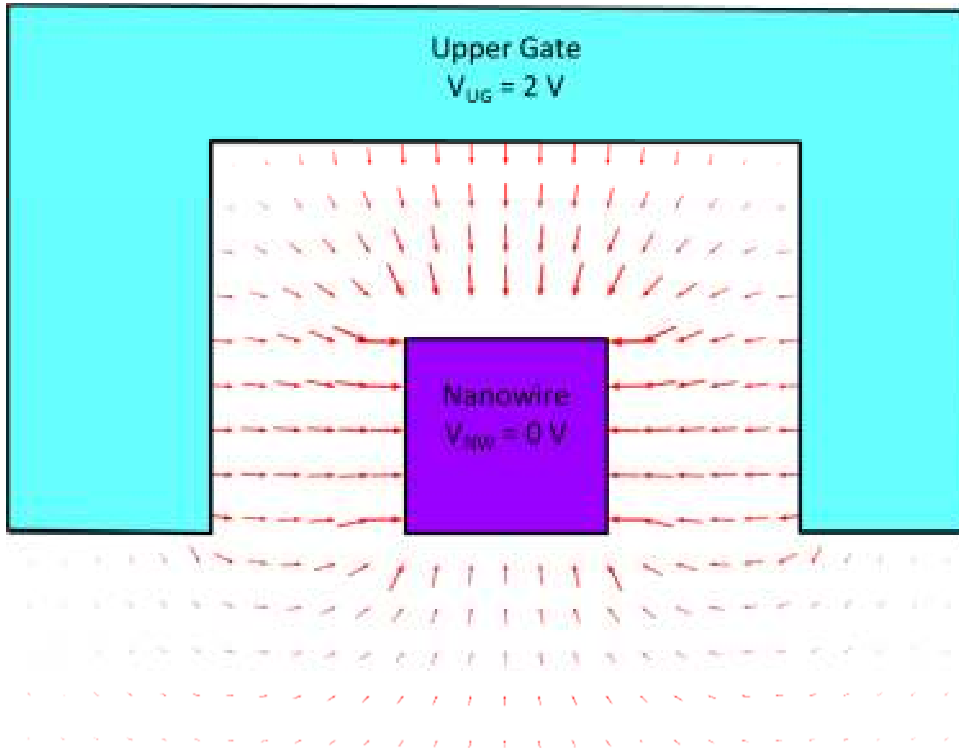


Figure 2.7 Electric field lines between the UG (V_{UG} = 2 V) and the nanowire (V_{NW} = 0), showing electric field lines going to the bottom of the nanowire.

3. Barrier Capacitances

Gate capacitances are not the only capacitances to the QD. The source and drain are also capacitively coupled to the QD through the barrier capacitances. I have been unable to simulate the barrier capacitances. This is unfortunate because the barrier capacitances are often larger than the gate capacitances. The two major challenges to

predict the barrier capacitances in these devices are understanding the gate voltage dependence of the barrier capacitance and predicting an accurate barrier capacitance.

First, the barrier capacitances are a function of gate voltage [40,48]. In a previous study at NIST [40], the barrier capacitance was shown to vary from 15 aF to 50 aF depending on the gate voltage. (Simultaneously, the gate is changing the barrier resistance by orders of magnitude.) The gate must either be changing the length of the tunnel barrier or the dielectric constant. Because the gate capacitances do not change as a function of voltage, I do not think the length of the barrier is changing significantly. So I suspect that a changing dielectric constant is primarily responsible for the change in barrier capacitance. Unfortunately, FASTCAP is not capable of incorporating a gate voltage dependent dielectric constant. I will discuss what could be done to better understand and predict the barrier capacitance in Chapter 6.

Second, the barrier capacitances I simulate are too small. When I tried to predict the barrier for the device mentioned above I predicted a barrier capacitance of only 4 aF. In contrast, the smallest measured barrier capacitance was 15 aF.

G. Summary and Implications

At the beginning of this Chapter I set out three goals, and all three goals have now been met. First, I showed that nominally identical devices have reproducible gate capacitances to within 10%. Second, I showed that the gate capacitances scale with the size of the device, as determined from the fabrication parameters. Third, I was able to numerically predict the gate capacitance to within 20%, without any fitting parameters.

These results answer the question from the previous Chapter of whether these QDs are intentional or unintentional. They are intentional; if they were unintentional QDs, then none of these three goals would have been met.

Furthermore, with the ability to predict gate capacitances using the fabrication parameters, I can use of FASTCAP to design new devices or to better understand the unintentional QDs.

1. Prediction of Highest Operating Temperature

The devices I studied show Coulomb blockade at 4K, but we would like to operate our devices at higher temperatures, such as liquid nitrogen temperature (77 K). I have tried to operate our existing devices at 77 K, but I never saw Coulomb blockade. Could we raise the operating temperature to 77K, by making a smaller device?

I can simulate the smallest device that I think could be fabricated in this device architecture. I will use L_{LG} = 10 nm, L_{UG} = 10 nm, w_{si} = 10 nm and $t_{ox,1}$ = 20 nm, because those are the smallest values used in the NTT devices. I will also reduce the thickness of the silicon to t_{si} = 10 nm, because it is possible to get SOI silicon that thin. Finally, I reduce $t_{ox,2}$ to 20 nm, to make it as thin as $t_{ox,1}$. From FASTCAP I get C_{UG} = 3.2 aF, C_{LGS} = C_{LGC} = 1.2 aF and C_{LGD} = 0.1 aF for the source short QD.

To calculate the total capacitance, I also need the barrier capacitances. I cannot simulate the barrier capacitances, so I need to estimate the barrier capacitance in another way. Because the nanowire is much smaller in this device, I will assume that the minimum barrier capacitance is half of what is was before (8 aF). This gives us a total capacitance of 22 aF. This gives a charging energy of 7.3 meV, which is equivalent to 84

K. From this, it should be possible to observe Coulomb blockade at liquid nitrogen temperature.

The FASTCAP file for the high temperature device is in Guestroom PC C:\Ted\Program\FastCap\save\really small device\half island20 nm.lst

2. Prediction of the Location of Unintentional QDs

As I explained in the previous Chapter, we frequently observe unintentional QDs in these devices. I can measure the gate capacitance to the unintentional QDs. This led me to develop a technique to run the capacitance simulator backwards. By starting with the capacitances of the unintentional QDs, I could determine the location of the unintentional QDs with a precision of a few nanometers. I will describe this technique in the next Chapter.

Table 2.1 Compendium of all measured gate capacitances. The first four columns identify the date, run number, device, and motivation for a set of data. All 3B devices are in green, 3C devices are in red, 3D devices in blue, and 3E devices in black. The next five columns identify dimensions of the device. The

				Dimensions in nm									
SET-PC:C\Data\09_11 log of tunable devices.xls									gate C_i (aF); fname				
Date	Run	Device	Motivation	t_{si}	w_{si}	L_{si}	L_{LG}	L_{UG}	UG-full	LGS-full	LGC-full	LGD-full	Run
10/03	2.34	AF-CA3A3E-3	broke										2.34
12/03	2.34	AF-CA3A3E-1	SETT, Cg VII-61	21	30	360	10	40	22; Jan22_6	3.2; Jan22_8	6.7; Jan14_19	2.5; Jan22_7	2.34
2/04	2.35	AF-CA2F3E-1	SETT, degradation	17	30	360	10	40	22; Feb17_2	2.7; Feb17_4	6.2; Feb17_1	2.5; Feb19_1	2.35
4/04	2.36	AF-CA2R3E-1	asymmetric	17	30	360	10	40	22; May3_12	3.0; May3_14	6.0; May3_8	2.8; May3_13	2.36
6/04	2.38	AF-CA2F3E-3	shuttle, leakage	17	30	450	40	40			10; Jan17_2		2.38
		AF-CA2R3E-3		17	30	450	40	40			8.3; Jun30_2		
11/04	2.39	AF-CA3D3E-1	shuttle, leakage	21	30	360	10	40			5.7; Jan8_2		2.39
		AF-CA3D3E-3		21	30	450	40	40			10; Dec3_2		
3/05	2.40	AF-CA3L3E-1	shuttle, leakage	21	30	360	10	40			6.4; Mar7_3		2.40
		AF-CA3L3E-3		21	30	450	40	40					
	2.40B	AF-CA2U3E-1	shuttle, leakage	17	30	360	10	40			6.4; Apr14_1		2.40B
		AF-CA2U3E-3		17	30	450	40	40					
4/05	2.41	AF-CA2U3C-2	shuttle, leakage	17	20	360	10	40			4.8; Apr29_2		2.41
		AF-CA2U3C-4		17	20	450	40	40					
3/06	2.46, .47	AF-CA2R3D-1	Q0(t)	17	20	420	10	70			4.8; Apr27_4		2.46, .47
		AF-CA2R3D-2		17	30	300	10	10					
6/06	2.48	AF-CA2R3C-2	Q0(t)	17	20	360	10	40			5.7; Jun8_6		2.48
7/06	2.51	AF-CA2C3C-3	Q0(t)	17	20	390	40	10			8.9; Aug1_15		2.51
9/07	2.56, .57	AF-CA2U3D-3	size quant, high T	17	20	510	40	70		2.8; Oct03_4	8.1; Sep14_11	3.2; Oct3_4	2.56, .57
1/10	2.58	AF-CA2C3B-2		17	10	420	10	70	36; Mar8_1	3.5; Mar26_5		2.8; Mar26_5	2.58
		AF-CA2C3B-3		17	10	450	40	40			6.9; Jan22_17		
1/11	2.62	AF-CA2C3D-3		17	20	510	40	70	36; Jan15_12	2.2; Jan 15_2	7.5; Jan15_1	3.8; Jan 15_3	2.62
		AF-CA2C3D-4		17	30	390	40	10	8.8; Jan16_9	4.3; Jan16_7	10; Jan16_5	6.5; Jan16_6	

		gate C_i (aF); fname				gate C_i (aF); fname			
Run	Device	UG-S-short	LGS-S short	LGC-S short	LGD- S short	UG- D short	LGS-D short	LGC-D short	LGD-D short
2.34	AF-CA3A3E-3								
2.34	AF-CA3A3E-1	10; Jan23_7	2.8; Jan23_1		0.08 Jan23_6	11; Jan23_8	0.09; Jan23_3		2.4; Jan23_4
2.35	AF-CA2F3E-1	11; Feb19_6	2.3; Feb19_4	2.8; Feb19_11	0.14; Feb19_5	11; Feb19_7	0.12; Feb19_3	3.1; Feb19_9	2.4; Feb19_7
2.36	AF-CA2R3E-1	10.2; May3_19	2.8; May3_16	2.6; May3_22	0.082; May3_17	11.1 May3_20	0.07; May3_15	2.8; May3_23	2.53; May3_18
2.38	AF-CA2F3E-3								
	AF-CA2R3E-3								
2.39	AF-CA3D3E-1								
	AF-CA3D3E-3								
2.40	AF-CA3L3E-1								
	AF-CA3L3E-3								
2.40B	AF-CA2U3E-1							2.7; Mar25_4	2.1; Mar28_1
	AF-CA2U3E-3								
2.41	AF-CA2U3C-2								
	AF-CA2U3C-4								
2.46, .47	AF-CA2R3D-1								
	AF-CA2R3D-2								
2.48	AF-CA2R3C-2								
2.51	AF-CA2C3C-3								
2.56, .57	AF-CA2U3D-3	19; Sep14_21	2.9; Dec20_32	4.2; Dec20_37		17; Sep18_4			2.7; Oct03_11
2.58	AF-CA2C3B-2		3.3; Mar29_5						
	AF-CA2C3B-3								
2.62	AF-CA2C3D-3	17.6; Jan15_8	2.0; Jan 15_5	3.8;Jan15_6		17.9; Jan15_9		3.2; Jan15_10	3.5; Jan15_11
	AF-CA2C3D-4	3.4;Jan17_8	2.4; J1n17_11	3.5; Jan 17_12		3.8; Jan17_13		3.8; Jan17_16	6.5;Jan17_17

Chapter 3: Determining the Locations of Unintentional Quantum Dots

Based on "Determining the location and cause of unintentional quantum dots in a nanowire," by Ted Thorbeck and Neil M. Zimmerman, published in Journal of Applied Physics **111**, 064309, (2012).

A. Preview

Unintentional QDs are a problem for us and many other groups working in silicon QD devices. In this Chapter, I will show effects of the unintentional QDs on the current through our devices. I show that the two QDs are arranged in a new way that I call the hybrid series-parallel model. I also show how we measure the gate capacitances to the unintentional QDs. Then I will determine the location of unintentional QDs in the nanowire with nanometer precision by comparing the measured gate capacitances to a FASTCAP simulation.

B. Motivation and Previous Work

We want to determine the locations of the unintentional QDs because knowing their location might give us clues about their cause. In Chapter 5 I will suggest that strain might be the cause of the unintentional QDs. So in Chapter 5 I will simulate the strain in the device, and I will compare the strain in the device to the location of the QD, as determined in this Chapter.

There is another motivation for the work in this Chapter. Our transistors look like modern transistors, and modern transistors have reached the point where a single dopant atom can affect the transport characteristics [49]. At low temperatures a single dopant

can become a QD [50]. By measuring the capacitances to the dopant and using the techniques I describe in this Chapter, we might be able to determine the location of a single dopant atom. However, there are significant additional challenges to working with dopants as compared to other QDs. As I was working on the research in this Chapter, similar work was being pursued at Sandia National Laboratory [51]. In their work the locations of donors in a point contact were "triangulated". Because their work was on donors, which only have a few charge states, gate capacitances could not be measured. Instead, they compared capacitance ratios, gate capacitance to barrier capacitance, to a simulation. Because in our devices the barrier capacitances are a function of gate voltages, I avoided using the barrier capacitances when determining the locations of the QDs.

The hybrid series-parallel model that I will develop in this Chapter had not been previously reported. However, it can explain features of previously published data [46]. While the paper this Chapter is based on was being published, another group reported something similar [52]. Their paper, which discussed transport through donors in a finFET, described this as "switching quantum transport," because the gate voltages change the conduction path through the device.

C. Unintentional Quantum Dots

1. The Data

In Chapter 1 I showed a plot of current as a function of each of V_{LGD}, V_{LGC} and V_{LGS} (reprinted in Fig. 3.1). I explained that the LGS and LGC curves are what we expect for a tunnel barrier and that the LGD curve was bumpier due to unintentional QDs.

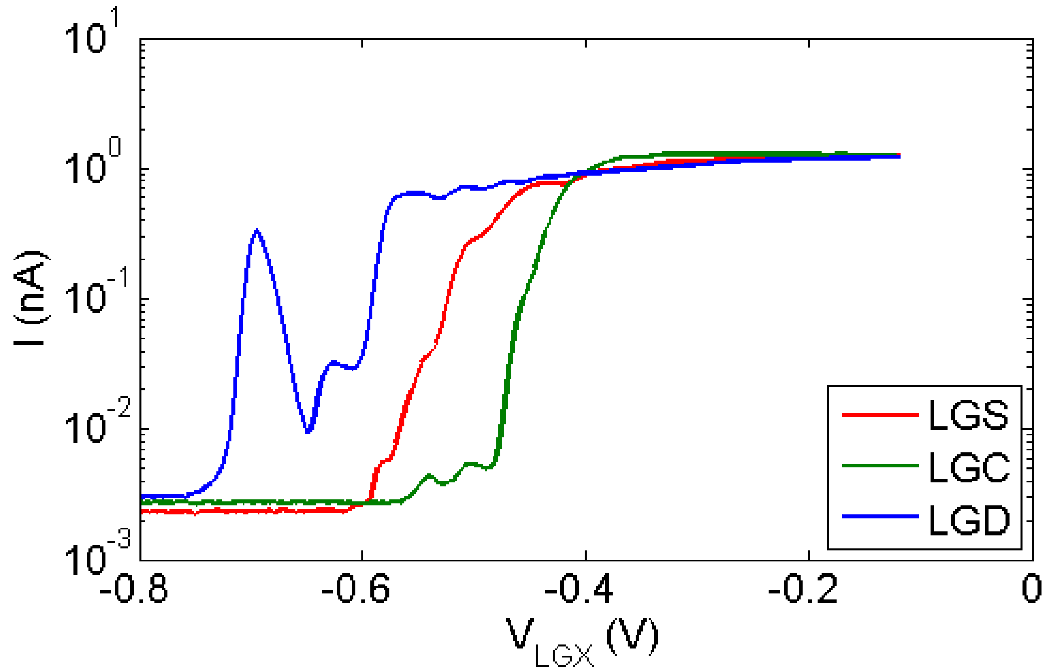


Figure 3.1. Current through the nanowire as a function of V_{LGS}, V_{LGC}, and V_{LGD}. Data from device AF-CA2U3D-3, run 2.56, Sep14_3 for LGS, Sep14_6 for LGC and Sep14_10 for LGD. Data taken at T = 400 mK, V_{UG} = 1 V and the other lower gates were set to $V_{LGX} = 0$.

To get a better look at these unintentional QDs, I show the current as a function of both LGD and UG in device 1 (AF-CA2U3D-3) in Figure 3.2(a). For comparison I show the current as a function of both UG and LGC in Figure 3.2(b). In Figure 3.2 (a) we see two sets of parallel peaks in the current, while we see a fairly smooth transition between no current and current in Figure 3.2(b). Because the lower gates should only be acting as a tunnel barrier, we expect to see something like 3.2(b). Notice that the LGD (blue) curve in Figure 3.1 corresponds to a horizontal slice through Figure 3.2(a).

In Figure 3.2(d) I show the current through another device, device 2 (AF-CA2C3B-2), which shows a very similar pattern when scanning both LGS and UG. In device 1 the unintentional QDs are coupled to LGD, and in device 2 the unintentional QDs are coupled to LGS. As can be seen in Figures 1.3 and 1.4, the only difference between LGS and LGD is the name, so that we can interpret 3.2 (a) and 3.2 (d) as being identical. (I could have relabeled both as LGD, but I chose not to so that the plot could easily be compared to the data in the published paper and the notebooks.) I was very surprised when I saw the data from device 2, because I thought that the pattern of current in device 1 was caused by random interface traps or dopants, so I did not expect to find the exact same pattern in the current in a second device. This was the first suggestion to me that the unintentional QDs might not be due to interface traps or dopants, but could be a systematic but unintended consequence of the fabrication. In this rest of this Chapter, unless I specifically refer to device 2, I will be discussing device 1.

There are several remarkable aspects of these data sets. If the device were working as intended, the data should just show a dark blue region in the lower left (where there is no current), a dark red region in the upper right (where there is a lot of current), and a smooth rainbow in between. Instead, we see two sets of parallel peaks. I showed in Chapter 1 that a single set of parallel peaks can be caused by Coulomb blockade through a single QD, so two sets of parallel peaks can be caused by a double QD (I will justify this in the next section). I will refer to the dot causing the more steeply sloped set of parallel peaks as dot A, and the dot causing the less steeply sloped set of parallel peaks as dot B. Dots A and B are highlighted in Figure 3.2(a). Also note that there are much

bigger current peaks where the lines due to current through dots A and B intersect, as shown in Figure 3-2(c). I will return to this observation in the next section.

The differences between the peaks corresponding to current dots A and B also stood out to us. Dot A has only a few peaks, whereas dot B has many peaks. Dot A is more strongly coupled to LGD, and dot B is more strongly coupled to UG. Notice that dot B has many charge states that are evenly spaced. If dot B were caused by an interface trap or dopant, it should not have this many evenly spaced charge states.

Also the dots appear to be interacting electrostatically; specifically, the current through dot B appears to be controlled by the charge state of dot A. In Figure 3.2(d) I added black bars where dot A changes charge states. To the left of the bars, there is no current through dot B. In between the bars, there is current through dot B. To the right of the bars there is an order of magnitude more current through dot B. Because the current through dot B changes suddenly when dot A changes, the current through dot B is dependent on the charge state of dot A.

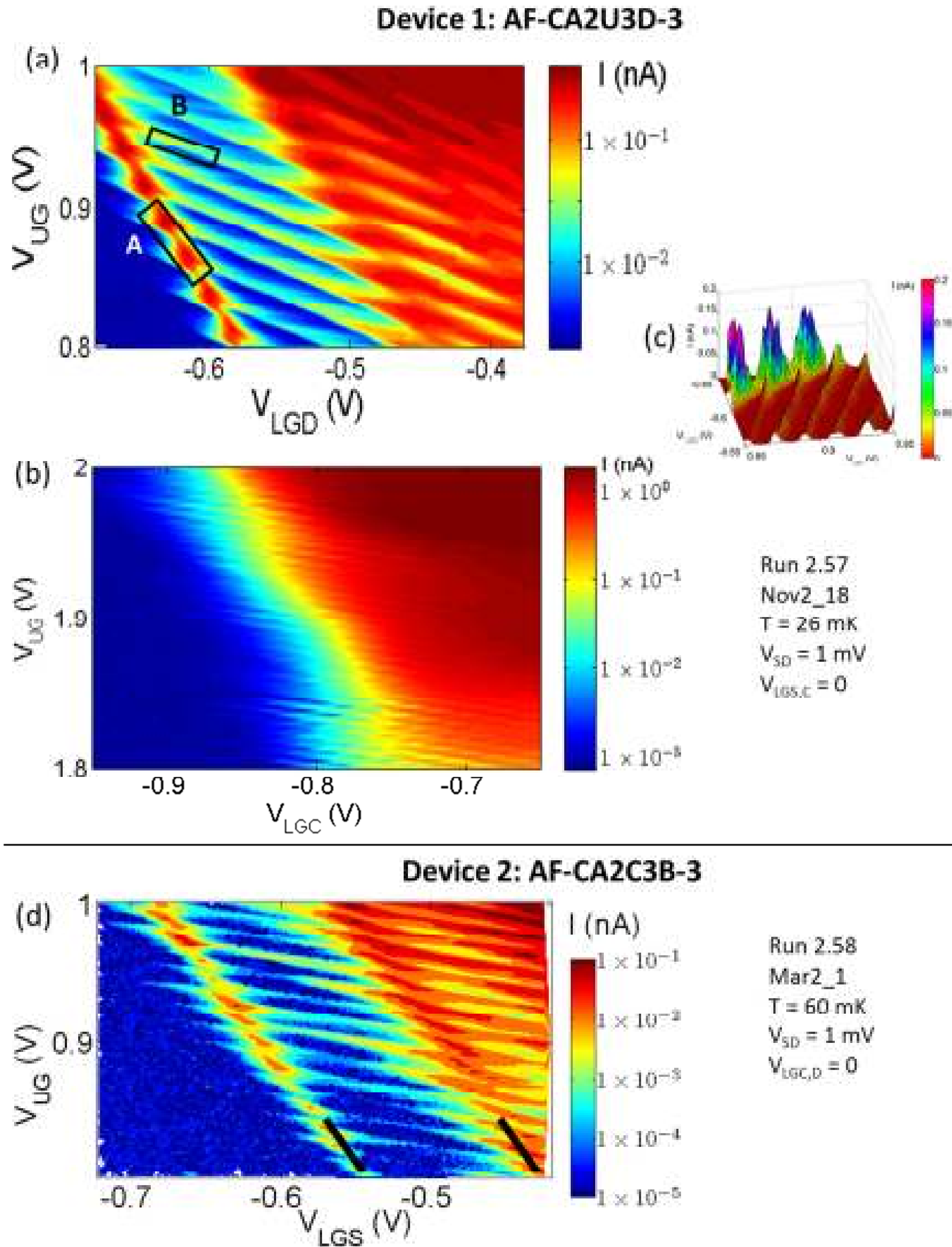


Figure 3.2. (a), (b) & (d) Current through the nanowire as a function of V_{UG} and V_{LGX} showing two unintentional QDs. The black boxes in (a) label dots A and B. The black bars in (b) indicate changes in the charge state of dot A. (c) Psuedo-3D views of data from (a), highlighting the larger current where the peaks due to dots A and B intersect.

2. Circuit Model

I mentioned that the data in Figure 3.2 is caused by two QDs. I talked about DQDs in Chapter 1, but this data violates one thing I said about DQDs in the first Chapter. I said that current should only flow when the chemical potentials of both QDs are in the bias window between the source and drain Fermi levels. Therefore, current should only occur at isolated points in a two gate scan, as shown by the black circles in Figures 1.14 and 1.16. Instead, in Figure 3.2 we see continuous lines of current with an extra-large peak where the lines intersect.

In Chapter 1 I discussed the series DQD. There is another kind of DQD called the parallel DQD [53–56]. The electrostatics of the series and parallel DQD are the same, but the arrangement of the QDs is different. In a series DQD the current path is source to dot A to dot B to drain, which is shown by the black circles in Figure 3.3. In contrast, in the parallel DQD, the source and drain are each tunnel coupled to both QDs, but the two QDs are capacitively rather than tunnel coupled to each other. Consequently, in a parallel DQD there are two current paths: source to dot A to drain and source to dot B to drain; the current can either flow through one QD or the other. Now, we can see line segments in the current of a two gate scan. In Figure 3.3, for the parallel DQD path A (which goes through dot A), current is allowed along the red lines. And in the parallel DQD path B, current is allowed along the blue lines. I find it helpful to think of the series DQD as an AND gate and the parallel DQD as an OR gate.

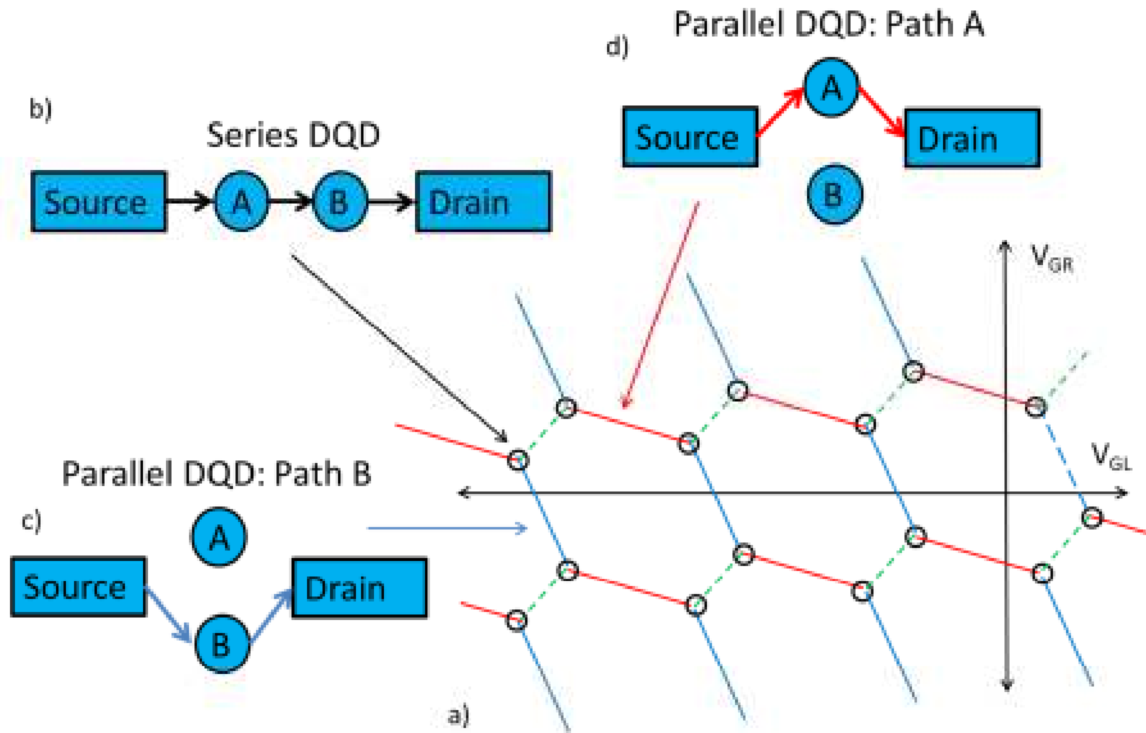


Figure 3.3. Current paths through series and parallel DQDs. (a) Charge stability diagram of DQD as shown in Chapter 1. In the series DQD (b), current must travel through both A and B, so current is only allowed at the black circles in (a). In path A of the parallel DQD (c), current only passes through dot A, so current is allowed along the red lines in (a). In path B of the parallel DQD (d), current only passes through dot B, so current is allowed along the blue lines in (a).

Which do we have: a series DQD or a parallel DQD? We cannot have a series DQD, because we observe continuous lines of current in Figure 3.2. But we also cannot have a parallel DQD. We see in Figures 3.2(c) that there is an order of magnitude more current where the lines corresponding to dots A and B intersect; this peak corresponds to where both dots A and B are in the bias window. At this peak we expect to see the sum

of the currents when dots A and B are individually in the bias window, but there is no reason to get an order of magnitude more current, which is what we see. This means that we cannot have either the series DQD or parallel DQD. If, however, we combine the series and parallel DQD, then we can explain the data. This hybrid series parallel model is shown in Figure 3.4(a).

The hybrid series-parallel model resolves the problems that both the series DQD and parallel DQD models had in explaining the data. The series path goes through both dots A and B: source-1-A-2-B-3-drain. This path has the least resistance, and it explains the peak in current in Figures 3.2 (c). There are also two parallel paths: path A is source-1-A-5-drain and path B is source-4-B-3-drain. These paths have higher resistances, but they explain why we see continous lines of current in Figure 3.2 (a) and (d).

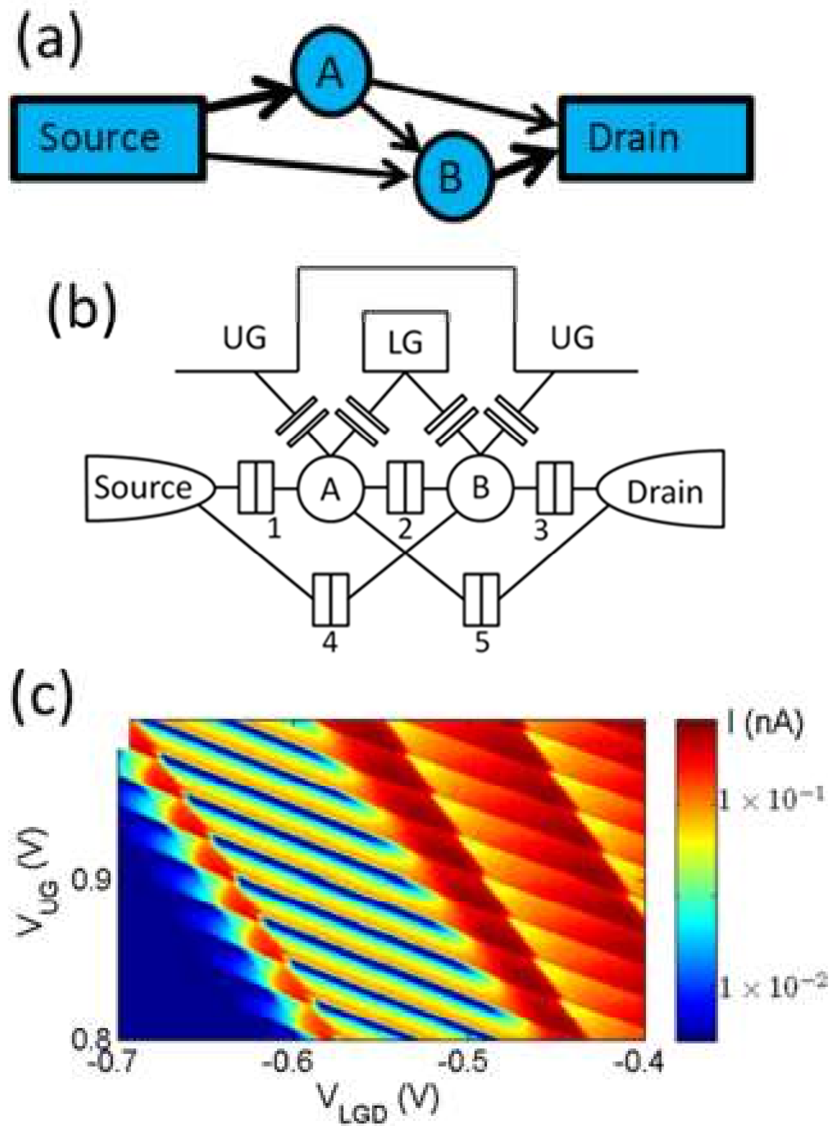


Figure 3.4 Hybrid series-parallel model. (a) Hybrid series-parallel model shown using the same representation as Figure 3.3. Thicker lines represent more current (not to scale). (b) The circuit diagram for hybrid series-parallel model, with tunnel junctions labelled 1-5. The series DQD path is source-1-A-2-B-3-drain. The parallel DQD path A is source-1-A-5-drain. The parallel DQD path B is source-4-B-3-drain. (c) Results from a simulation of the current through the circuit model shown in (b), with the parameters in Tables 3.1 and 3.2.

I was able to reject an alternative model to explain the data: cotunneling through a series DQD. Cotunneling, which is a higher order tunneling process, provides a current path through a DQD when only one of the QDs has a chemical potential in the bias window. But I was unable to find a set of resistances for the tunnel barriers could replicate the currents I measured, so this model was rejected.

3. Capacitances and Resistances

I will verify that the hybrid series-parallel model explanation is correct by simulating the circuit in Figure 3.4(a), but first I need all the capacitances and resistances in the circuit. The simplest parameters to measure are the gate capacitances. I showed how to measure gate capacitances and gate capacitance ratios in Chapter 1. The gate capacitances for both devices in Figure 3.2 are reported in Table 3.1.

Table 3.1. Gate capacitances for LG and UG to dots A and B and capacitance ratios.

Device 1	C_{LGD} (aF)	C_{UG} (aF)	Slope (C_{LGD}/C_{UG})	C_{LGC} (aF)
Dot A	2.3 + 0.3 – 1.3	1.3 + 0.2 – 0.6	-1.71 ± 0.02	< 0.1
Dot B	3.2 ± 0.2	7.9 ± 0.3	-0.41 ± 0.01	< 0.1
Device 2	C_{LGS} (aF)	C_{UG} (aF)	Slope (C_{LGS}/C_{UG})	C_{LGC} (aF)
Dot A	1.3 ± 0.2	0.9 ± 0.1	-1.46 ± 0.03	< 0.1
Dot B	2.2 ± 0.2	12.2 ± 0.6	-0.179 ± 0.007	< 0.1

The uncertainties in Table 3.1 represent the maximum and minimum values of periods and slopes that could be fitted to the width of the peaks. Notice that the relative uncertainties of the slopes are much smaller than the relative uncertainties of the gate capacitances. This is because I was able to take data similar to the data in Figure 3.2 from V_{UG} = 0.5 V to V_{UG} = 2.5 V (see Appendix D), so I could measure the slopes very well. The precision of the slope measurement will be important in determining the precision of the calculated locations of the unintentional QDs.

The asymmetric uncertainties for dot A in device 1 (C_{LGD} = 2.3 aF + 0.3 aF – 1.3 aF) comes from the aperiodic spacing of the more steeply sloped peaks in Figure 3.2(a). Aperiodic spacing between Coulomb blockade peaks is commonly observed in few electron QDs, and it can be explained by the addition of a size-quantized energy level to the charging energy. Therefore I took the smallest spacing between peaks and used that to calculate the best value (2.3 aF) and the smaller uncertainty, (+0.3 aF). Then I used the larger spacing to calculate the smallest possible value of capacitance, and I used that to determine the larger uncertainty (-1.3 aF).

Next, we would like to determine as much as we can about the barrier capacitances and resistances. The barrier capacitances are measured using the slopes of diamond diagrams as shown in Figure 3.5 (b) with dashed lines. I can identify which diamond corresponds to which QD by comparing the diamond diagrams to the V_{UG}-V_{LGD} plots in Figure 3.2 and Appendix D. The barrier capacitances and resistances are reported in Table 3.2. I used the slopes in Figure 3.5 corresponding to dot A to measure the capacitances of the 1st and 5th tunnel barriers. I used the slopes corresponding to dot B to

measure the capacitances of the 2nd and 4th tunnel barriers. I was unable to measure the capacitance of the 3rd barrier, so I choose a reasonable value (10 aF).

The barrier resistances were measured from the value of the current. I used the current through path A to calculate the resistance of the 5th tunnel barrier. I used the current through path B to calculate the resistance of the 4th tunnel barrier. I used the current through the series path to calculate the resistance of the 2nd barrier. I was unable to measure the resistances of the 1st and 3rd tunnel barrier because they have a smaller resistance than the other barriers, so I used a resistance of 100 kΩ for both barriers. See the next section for additional details about the measurement of the 4th tunnel barrier.

Table 3.2 Complete barrier resistance and capacitances for both devices with tunnel barriers as labeled in Figure 3.4. Uncertainties were not evaluated for this table.

Device 1				Device 2		
Barrier	R (Ω)	C (aF)		Barrier	R (Ω)	C (aF)
1	100 k	15		1	1 M	20
2	3 M	10		2	15 M	10
3	100 k	10		3	1 M	35
4	See below	See below		4	500 M	60
5	6 M	10		5	65 M	10

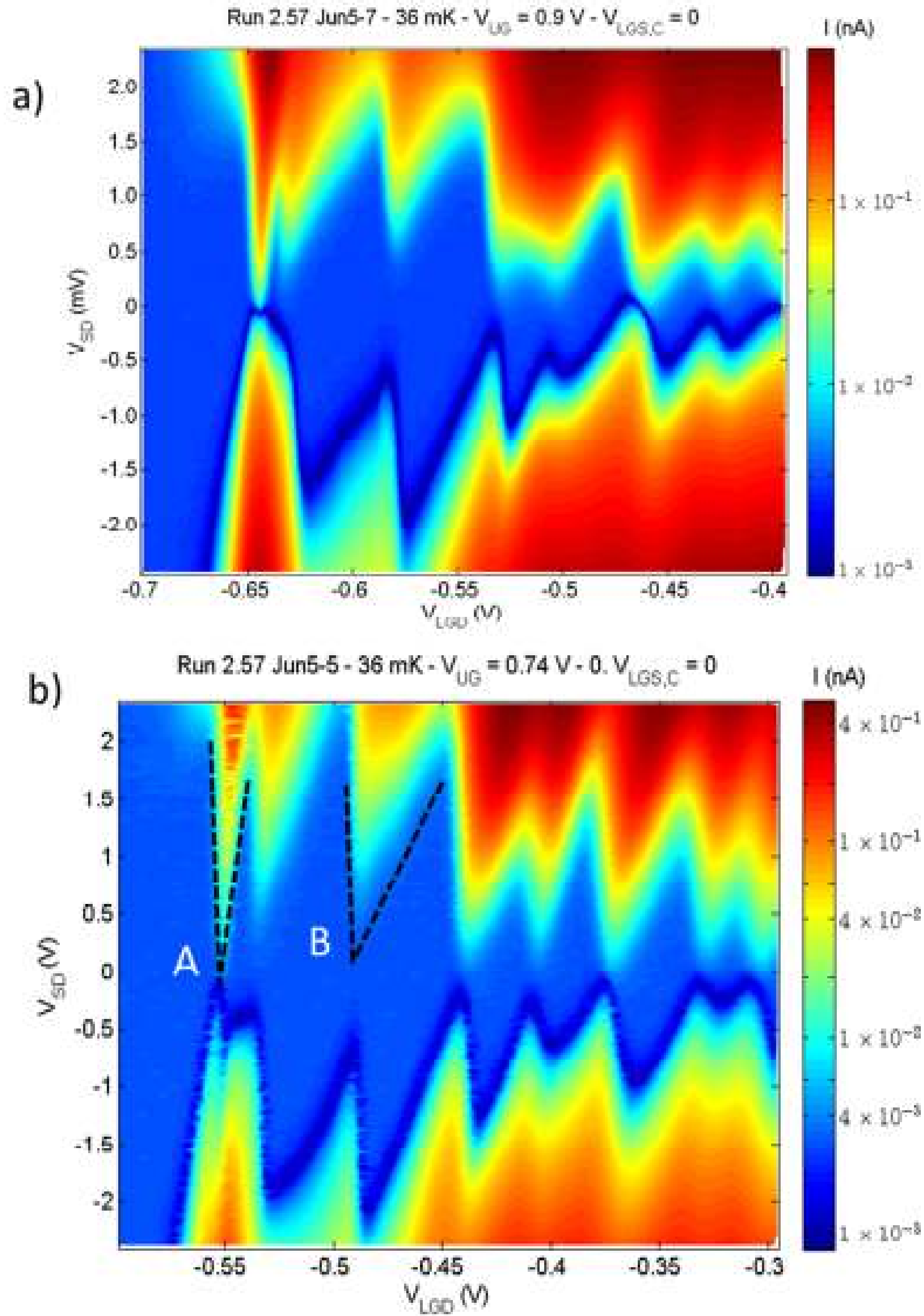


Figure 3.5 Diamond diagram. (a) Data from Run 2.57 Jun5_7, T = 36 mK, V_{UG} = 0.9 V, $V_{LGS,C}$ = 0. (b) Data from Run 2.57 Jun5_5, T = 36 mK, V_{UG} = 0.74 V, $V_{LGS,C}$ = 0. Dashed lines give the slopes used to calculate barrier capacitances.

4. Simulation of Hybrid Series-Parallel Model

I simulated the circuit in Figure 3.4(a) using SIMON (SIMulation Of Nanostructures)[Fig. 3.5] [57]. I used the parameters from Tables 3.1 and 3.2. The results of the simulation are shown in Figure 3.4(b).

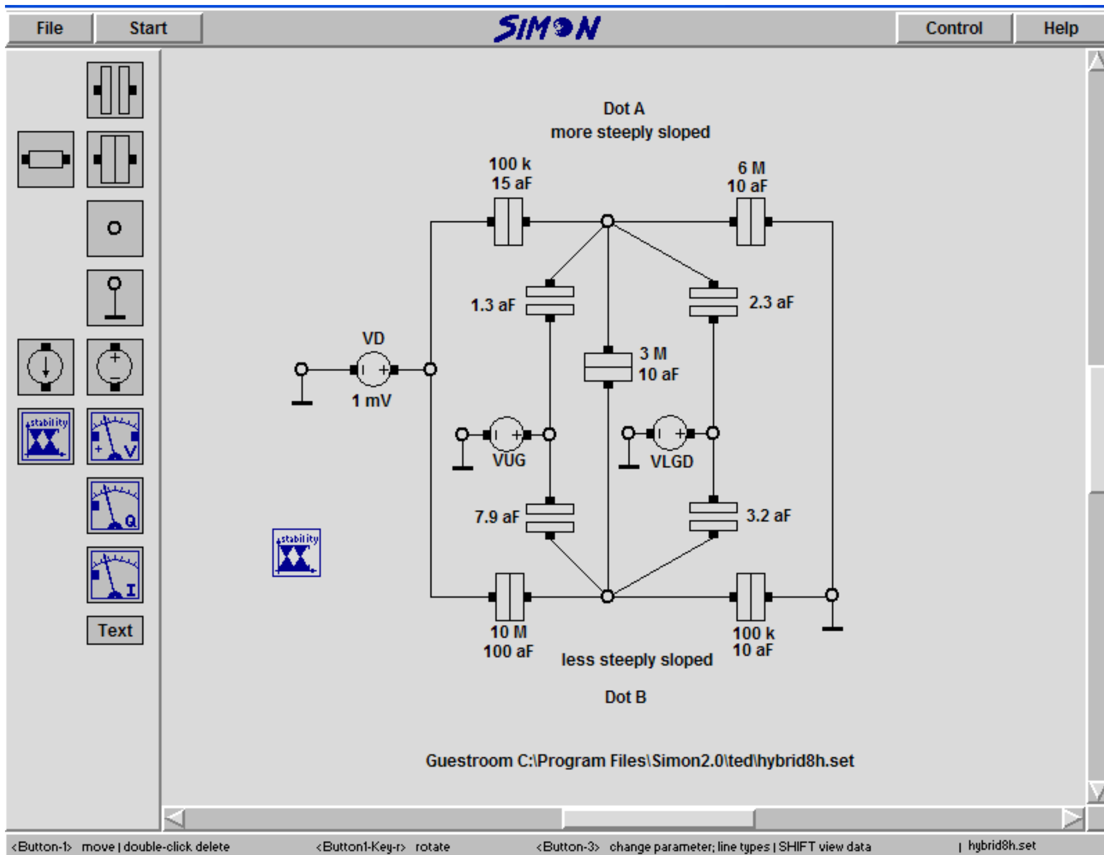


Figure 3.6 Screenshot of SIMON with parameters from Tables 3.1 and 3.2.

There was one aspect of the data I was unable to include in SIMON. As I mentioned before, the current through dot B depends on the charge state of dot A. This can be explained by a capacitive coupling between dot A and the fourth tunnel junction. To incorporate this, I did two separate SIMON simulations for two charge states of dot A,

and then combined the resulting simulations using an envelope function taken at high UG voltage (V_{UG} = 2.26 V). I will describe this envelope function in the next paragraph. In simulation 1, I used R = 10 MΩ and C = 100 aF for the resistance and capacitance of the fourth tunnel barrier (see Table 3.2). I call the current in this simulation $I_{sim,1}(V_{UG}, V_{LGD})$. In simulation 2 [$I_{sim,2}(V_{UG}, V_{LGD})$], I used R = 2 MΩ and C = 100 aF for the resistance and capacitance of the fourth tunnel barrier. Both $I_{sim,1}$ and $I_{sim,2}$ were simulated from V_{UG} = 0.8 V to 1 V and V_{LGD} = -0.7 V to -0.4 V.

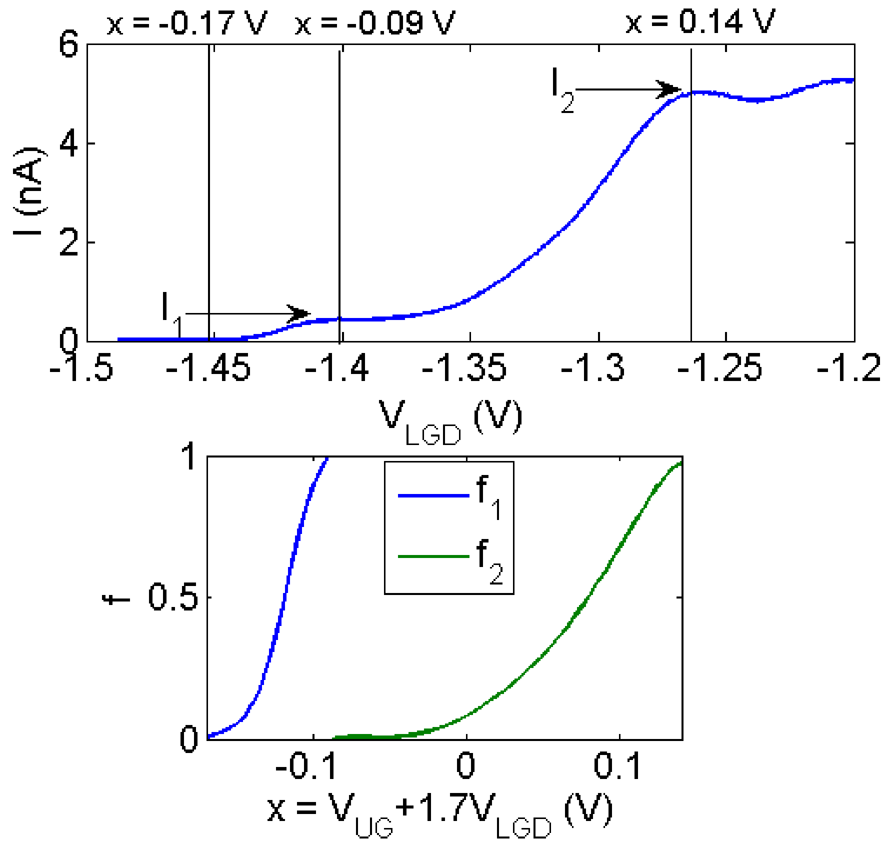


Figure 3.7 Details of the envelope function. (a) Current through the nanowire as a function of V_{LGD}, taken at V_{UG} = 2.26 V. (b) Dimensionless envelope functions f_1 and f_2.

In Figure 3.7(a) I show the I(V_{LGD}) taken at V_{UG} = 2.26. At higher V_{UG} there was no Coulomb blockade through dot B, but changes in the current due to the charge state of dot A can clearly be seen. Because this data was taken at a higher upper gate voltage, I use the variable $x = V_{UG} + 1.7V_{LGD}$, to compare V_{LGD} at lower V_{UG}. I used 1.7 because that is the measured ratio of C_{LGD}/C_{UG} (Table 3.1). I used this data, to construct the dimensionless envelope functions $f_1(x)$ and $f_2(x)$, shown in Figure 3.7(b). I can define $f_1(x)$ and $f_2(x)$ using the current shown in Figure 3.7(a).

$$f_1(x) = \frac{I(x)}{I_1} \text{ for } -0.17\,V < x < -0.09\,V$$

$$f_2(x) = \frac{I(x)-I_1}{I_2-I_1} \text{ for } -0.09\,V < x < -0.14\text{ V}, \qquad 3.1.$$

$$\text{where } x = V_{UG} + 1.7V_{LGD}.$$

where I_1 = 0.4 nA and I_2 = 4.95 nA.

To interpolate between the measured data points in Figure 3.6(a), I used a sixth order polynomial to the fit data. I will call these polynomial functions $f_1'(x)$ and $f_2'(x)$. I combined simulations 1 and 2 using the functions f_1 and f_2 with the equation:

$$I(V_{UG}, V_{LGD}) = \begin{cases} 0 & x < -0.17\,V \\ I_{sim,1}(V_{UG}, V_{LGD}) f_1'(x) & -0.17\,V < x < -0.09\,V \\ \begin{pmatrix} I_{sim,1}(V_{UG}, V_{LGD})\left(1 - f_2'(x)\right) \\ +I_{sim,2}(V_{UG}, V_{LGD}) f_2'(x) \end{pmatrix} & -0.09\,V < x < 0.14\,V \\ I_{sim,2} & 0.14\,V < x \end{cases} \qquad 3.2.$$

The results of Equation 3.2 are plotted in Figure 3.4 (c). Even though the simulation uses three free parameters, the agreement between Figure 3.4(b) and Figure 3.2(a) gives us confidence that the data in Figure 3.2 can be explained by two quantum dots, even though

it does not look like either a series or parallel DQD. The difficulty in measuring all of the tunnel barriers in Table 3.2 is an additional reason they are not used to determine the locations of the unintentional QDs.

D. Determining the Location of the Unintentional Quantum Dots

Now that we understand the circuit model of the DQDs and have extracted all the capacitances and resistances from the data, I will shift to describing the method I used to determine the locations of the unintentional QDs.

1. Qualitative Approach

I will begin by describing a qualitative approach to help us gain intuition about the locations of the unintentional QDs. Examining the gate capacitances in Table 3.1, we see that for both LGD and UG the capacitance to dot B is larger than the capacitance to dot A. Therefore, dot B is probably much larger than dot A. Because the capacitance from UG to dot B is similar to the capacitance from UG to an intentional short QD, dot B is likely similar in size to an intentional QD (between 40 and 100 nm long). The capacitances to dot A from UG and LGD are similar, but the capacitance to LGD is bigger, so dot A is probably is underneath the oxide in between UG and LGD.

2. Quantitative Approach

Now I describe a numerical method to determine the locations of the unintentional QDs. In the previous Chapter, I simulated the gate capacitances of the intentional QDs. In that case I knew the location and size of the QDs, so I could plug that into FASTCAP to calculate the capacitances. Now we have the inverse problem, we know the gate capacitances of an unintentional QD, and we want to determine the location. I solved this

problem by simulating the capacitances to 1 nm long slices of the nanowire. Each slice wraps around all four sides of the nanowire. The simulated capacitance to each slice can be treated as a differential gate capacitance, ΔC. The length of the slice is $\Delta x = 1$ nm. We can use this to approximate the derivative of the gate capacitance, $dC/dx \approx \Delta C/\Delta x$. The derivative of the gate capacitances to each of LGC, LGD, and UG are shown in Figure 3.7. As we intuitively expect, the differential capacitance to LGD is peaked underneath LGD, and the differential capacitance to the UG is peaked away from the LGs.

I can now numerically integrate the differential capacitances between any start (x_1) and end (x_2) positions within the nanowire. This allows me to compare the simulated and measured capacitances. We need to find a set of bounds, x_1 and x_2, between which all of the differential capacitances can be integrated to a value within the uncertainty of the measured gate capacitances as reported in Table 3.1. For device 1, dot B:

Dev. 1: Dot B: UG
$$3.0\text{ aF} < \int_{x_1}^{x_2} \frac{dC_{LGD}}{dx} dx < 3.4\text{ aF} \quad 3.3.$$

Dev. 1: Dot B: LGD
$$7.6\text{ aF} < \int_{x_1}^{x_2} \frac{dC_{UG}}{dx} dx < 8.2\text{ aF} \quad 3.4.$$

Dev. 1: Dot B: Slope
$$0.40 < \int_{x_1}^{x_2} \frac{dC_{LGD}}{dx} dx \Big/ \int_{x_1}^{x_2} \frac{dC_{UG}}{dx} dx < 0.42 \quad 3.5.$$

Dev. 1: Dot B: LGC
$$\int_{x_1}^{x_2} \frac{dC_{LGC}}{dx} dx < 0.1\text{ aF}. \quad 3.6.$$

We want to find all possible x_1 and x_2 that satisfy all four conditions. I assume that $x_1 < x_2$ to prevent double counting. Because we know that both dots A and B are near LGD, I only consider the drain half of the device, and I set the origin of the x-axis to be underneath the center of LGD. In Figure 3.8 I show all possible x_1 and x_2 which satisfy

the conditions in Eq. 3.3 through 3.6. In green are the solutions to condition 3.4 (C_{LGD}). In blue are the solutions to condition 3.3 (C_{UG}). In purple are the solutions to condition 3.5 (slope). In black are the solutions which satisfy all three conditions. There are two sets of solution, each of which I have highlighted by black rectangles in Figure 3.9. However, one set of those sets of solutions violate the condition in Eq. 3.6 (C_{LGC}), so it was eliminated.

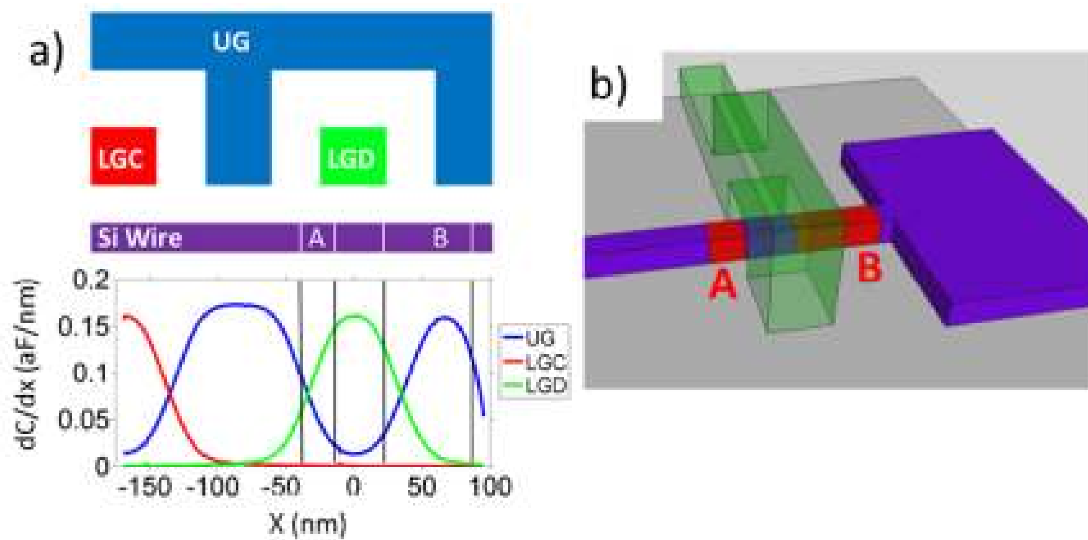


Figure 3.8 (a) Differential gate capacitances along the nanowire. The top half shows a cross section of the device with the Si nanowire, LGC, LGD and the UG shown to scale. The bottom half shows the differential gate capacitances for device 1. The origin of the x axis is the center of LGD. The vertical lines correspond to the deduced locations of dots A and B (see main text). (b) To scale pseudo 3-D view of the nanowire with dots A and B highlighted, LGD shown as translucent and the UG not shown.

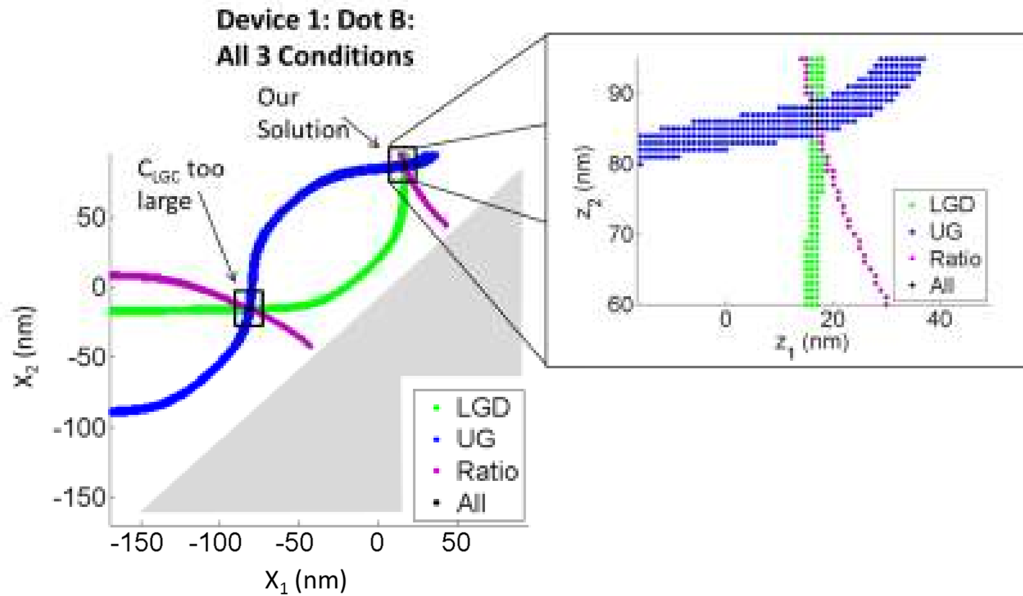


Figure 3.9 All x_1 and x_2 which satisfy conditions in Eq. 3.3 to 3.5. The lower-right half of the Figure is greyed out because in that region $x_2 < x_1$. In green are solutions to Eq. 3.4 (C_{LGD}), blue are solutions to Eq. 3.3 (C_{UG}), purple are solutions to Eq. 3.5 (slope), and black are solutions to all three.

In Figure 3.9 we see that all of the sets of x_1 and x_2 which satisfy all four of our criteria are contained within the bounds 16 nm $\leq x_1 \leq$ 18 nm and 85 nm $\leq x_2 \leq$ 89 nm. Therefore, dot B is between LGD (which is between x = -20 nm and x = 20 nm) and the end of the nanowire (which ends at +95 nm). This location is shown in Figures 3.8 and 3.11. This matches our expectation that dot B should be as long as one of the intentional QDs.

Next, we do the same analysis for dot A in Figure 3.10. The solutions for dot A are -43 nm ≤ x_1 ≤ 37 nm and -22 nm ≤ x_2 ≤ -16 nm. Dot A is on the other side of LGD, underneath the oxide in between LGD and UG, as we intuited earlier.

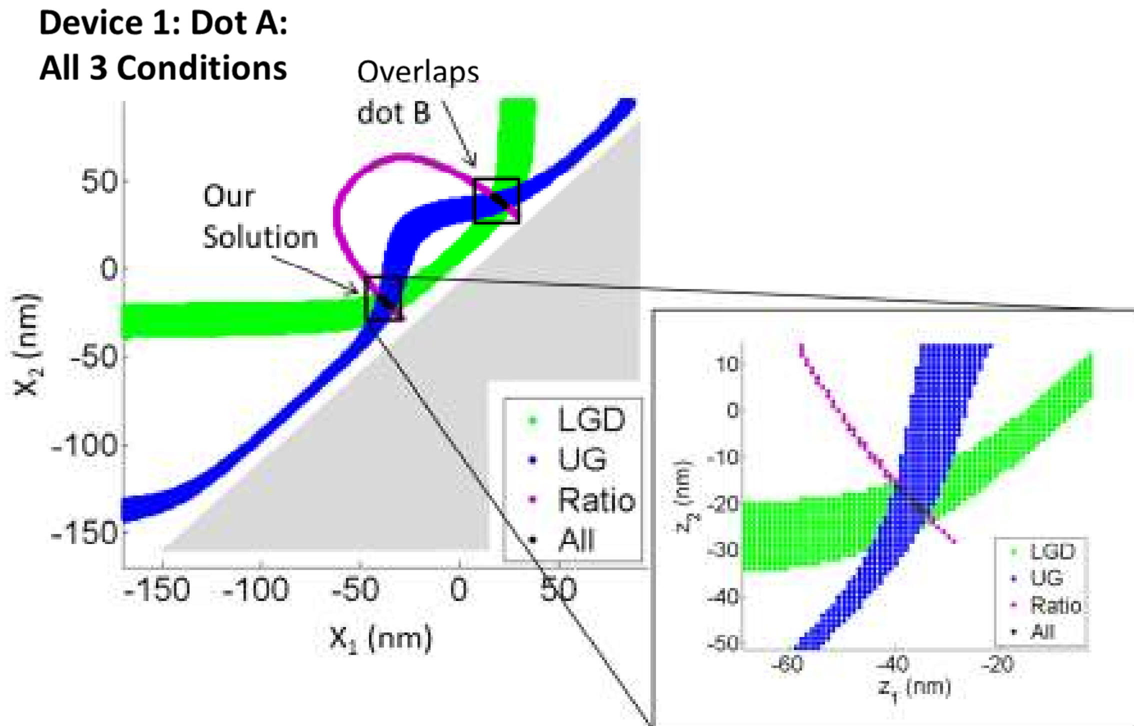


Figure 3.10. All x_1 and x_2 which satisfy gate capacitance criteria in Table 3.1 for dot A in device 1.

The locations of both dots A and B are shown in Figure 3.11. Table 3.3 contains the positions of both dots A and B in both devices 1 and 2.

Table 3.3. Positions, x_1 and x_2, of both dots A and B in both devices.

Device 1				Device 2		
	x_1 (nm)	x_2 (nm)			x_1 (nm)	x_2 (nm)
Dot A	-40 ± 3	-19 ± 3		Dot A	-37 ± 1	-22 ± 1
Dot B	17 ± 1	87 ± 2		Dot B	23 ± 2	117 ± 2

Notice the nanometer scale precision in Table 3.2. To determine if this is a reasonable precision, we can estimate what the precision of the position should be given the uncertainties in the gate capacitance measurements reported in Table 3.1.

$$\Delta x = \frac{\Delta C}{dC/dx}. \tag{3.7.}$$

Using the parameters for device 1, dot A: $\Delta C_{UG} = 0.3$ aF, $dC_{UG}/dx \approx 0.1$ aF/nm. This gives an estimate of $\Delta x = 3$ nm. So a precision of a few nanometers is a reasonable precision given the uncertainties in the gate capacitances.

Now take a look at the similarity in the location of both dot A and dot B in both devices. The only real difference is that x_2 is bigger for dot B in device 2, and this can be explained because the nanowires have different lengths in the two devices. The agreement in the locations of the unintentional QDs in the two devices is remarkable. It suggests that the cause of the unintentional QDs is not random, but rather it is an unintended consequence of the fabrication.

E. Implications and Conclusions

Now I have determined the locations of both dots A and B in both devices 1 and 2. In Chapter 5, I will use the locations to help determine the cause of the unintentional QDs. A schematic of the conduction band profile that is needed to create the unintentional QDs is drawn in Figure 3.11(c). We can use this Figure to start to develop intuition about the cause of the dots A and B.

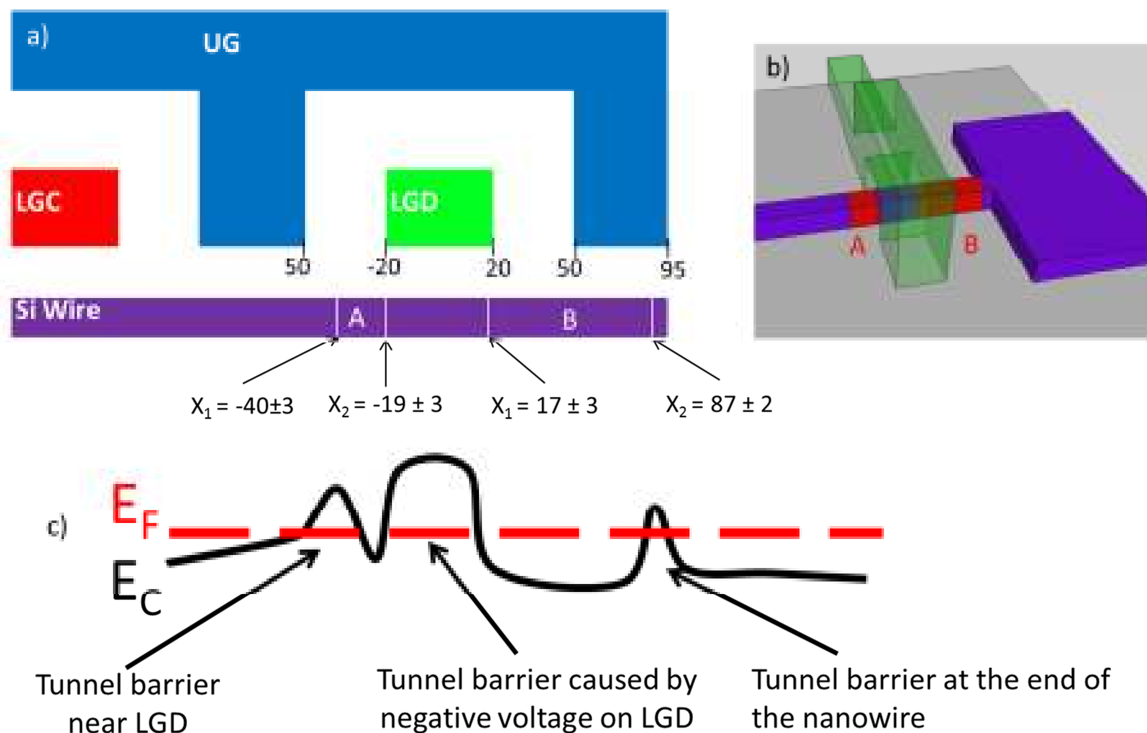


Figure 3.11 (a) Cross section of the device with the x-positions of dots A and B and key features of the device (all positions in nanometers). (b) Pseudo 3D view of dots A and B. (c) Schematic, aligned to (a), of the conduction band profile that could cause dots A and B.

Dots A and B are on either side of LGD, so there must be a tunnel barrier in between. This tunnel barrier is easy to explain because the negative voltage on LGD is supposed to create a tunnel barrier there. The other two tunnel barriers in Figure 3.11(c) cannot be explained by the electrostatics. There is a tunnel barrier at the end of the nanowire. In Chapter 5 I will show that strain from the thermal oxidation of the nanowire can create a tunnel barrier at the end of a mesa-etched nanowire. There is also a tunnel barrier located underneath the oxide in between the UG and LGD. In Chapter 5 I examine whether strain could also be the origin of this tunnel barrier. But before I discuss whether strain is the cause of these unintentional QDs, I need to review the basics of stress, strain and the silicon band structure. I will do this in the next Chapter.

Chapter 4: Reviews of Stress, Strain and the Band Structure of Silicon

A. Stress and Strain Review

At the end of the previous Chapter, I suggested that strain might be the cause of the unintentional QDs. The strain can come from fabricating the device or changing the temperature of the device away from its fabrication temperature. The strain alters the band structure of silicon, which changes the energies of the conduction band (CB) and valence band (VB). The strain-altered CB and VB can create tunnel barriers and QDs. In this Chapter I review the basics of stress and strain, as well as introduce the silicon band structure. I begin by providing some definitions and basic equations for stress and strain. For a more thorough treatment of stress and strain I recommend references [58,59].

1. Strain

A point in an object can be described by a position vector $\vec{x}$. When the object is deformed, that point will experience a displacement, $\vec{u}$. The new position vector is $\overrightarrow{x'} = \vec{x} + \vec{u}$ (Fig. 4.1).

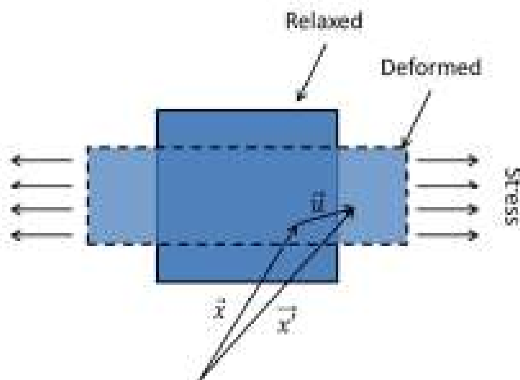


Figure 4.1 An object deformed by stress.

The total strain is the derivative of the displacement vector,

$$\epsilon_{tot,x} = \frac{\partial u_x}{\partial x} \quad \epsilon_{tot,xy} = \frac{1}{2}\left(\frac{\partial u_y}{\partial x} + \frac{\partial u_x}{\partial y}\right)$$
$$\epsilon_{tot,y} = \frac{\partial u_y}{\partial y} \quad \epsilon_{tot,yz} = \frac{1}{2}\left(\frac{\partial u_z}{\partial y} + \frac{\partial u_y}{\partial z}\right) \qquad 4.1.$$
$$\epsilon_{tot,z} = \frac{\partial u_z}{\partial z} \quad \epsilon_{tot,zx} = \frac{1}{2}\left(\frac{\partial u_x}{\partial z} + \frac{\partial u_z}{\partial x}\right).$$

where $\epsilon_{tot,x}$ is the x-component of the total normal strain and $\epsilon_{tot,xy}$ is the total shear strain in the *x*-*y* plane, and so on. Note that the definitions of the shear strains are symmetric, e.g., $\epsilon_{xy} = \epsilon_{yx}$. As a simple example, when a free-standing block of length l_x, undergoes a change in length of Δl_x, then $\epsilon_{tot,x} = \Delta l_x / l_x$. In subsequent examples, a typical order of magnitude of the strain is 0.1 %.

2. Stress

Stress describes the forces acting on a small volume within a solid from the neighboring particles. Figure 4.2 shows the sign and naming conventions used for stresses. Tensile (elongation) stress is positive and compressive stress is negative. Stress has units of pressure, and I will use both MPa and GPa.

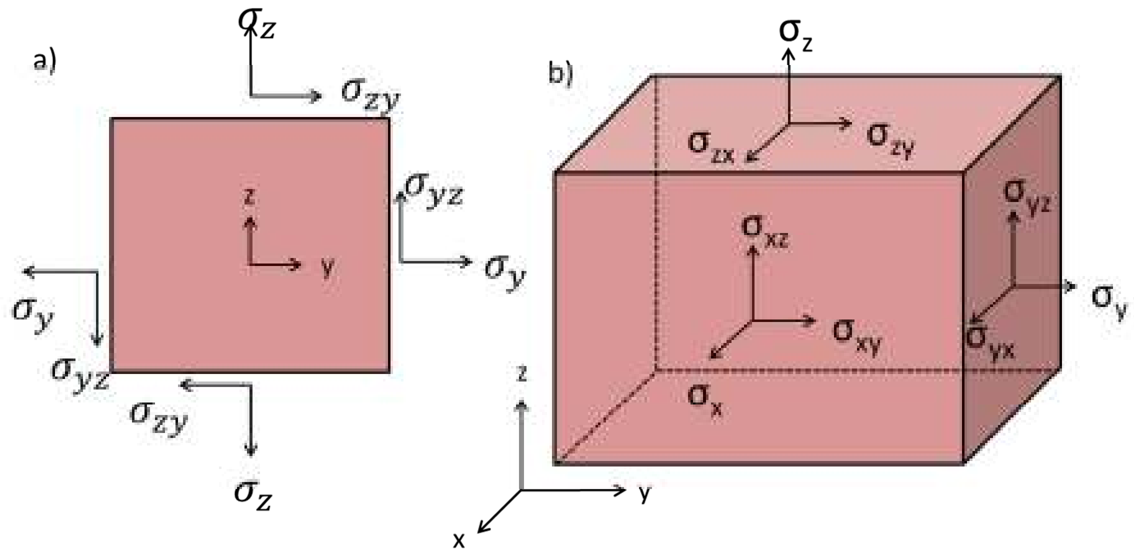


Figure 4.2 Schematic showing the stress components in (a) 2D and (b) 3D.

As can be seen in Figure 4.2(b) the forces acting on the right side of the cube are

$$\vec{f} = \sigma_{xy} A\hat{x} + \sigma_y A\hat{y} + \sigma_{yz} A\hat{z}. \qquad 4.2.$$

where A is the area of the side of the cube. Eq. 4.2 is just the force on one side of a cube, if we sum up the forces on all six sides of the cube, they should sum to zero, because the cube is not accelerating. This can be expressed mathematically in Eq. 4.3,

$$\frac{\partial \sigma_x}{\partial x} + \frac{\partial \sigma_{xy}}{\partial y} + \frac{\partial \sigma_{xz}}{\partial z} = 0$$

$$\frac{\partial \sigma_{yx}}{\partial x} + \frac{\partial \sigma_y}{\partial y} + \frac{\partial \sigma_{yz}}{\partial z} = 0 \qquad 4.3.$$

$$\frac{\partial \sigma_{zx}}{\partial x} + \frac{\partial \sigma_{zy}}{\partial y} + \frac{\partial \sigma_z}{\partial z} = 0.$$

These equations are called the equations of equilibrium. They are the first set of equations that must be solved when solving a continuum mechanics problem.

Both stress and strain are typically written as six element vectors.

$$\overrightarrow{\epsilon_{tot}} = \begin{pmatrix} \epsilon_{tot,x} \\ \epsilon_{tot,y} \\ \epsilon_{tot,z} \\ \epsilon_{tot,xy} \\ \epsilon_{tot,yz} \\ \epsilon_{tot,zx} \end{pmatrix} \quad \text{and} \quad \vec{\sigma} = \begin{pmatrix} \sigma_x \\ \sigma_y \\ \sigma_z \\ \sigma_{xy} \\ \sigma_{yz} \\ \sigma_{zx} \end{pmatrix}. \qquad 4.4.$$

3. Boundary Conditions

At the surfaces of an object we must specify the boundary conditions. For every point on the surface either the displacement or the external force applied to the surface, this is called the surface traction, must be specified.

I specify the displacements at some points on the object to prevent uniform displacements or rotations. In Figure 4.3(a) I specify that one point does not undergo any displacement to prevent a uniform displacement of the object (which would not result in strain). I specify some of the displacements at two other points to prevent rotations (which also would not result in strain). Alternatively, I would put a zero displacement condition on the entire bottom surface of the simulation, which was far from the area of the simulation I was interested in. Then I would verify that the stress caused by this displacement condition on the bottom surface relaxed, before it reached the area of the simulation I was interested in.

Everywhere else on the surface of the object I specify the boundary condition in terms of the external force applied to the object; these forces are called the surface tractions. In the simulations in the next Chapter the surface tractions are zero.

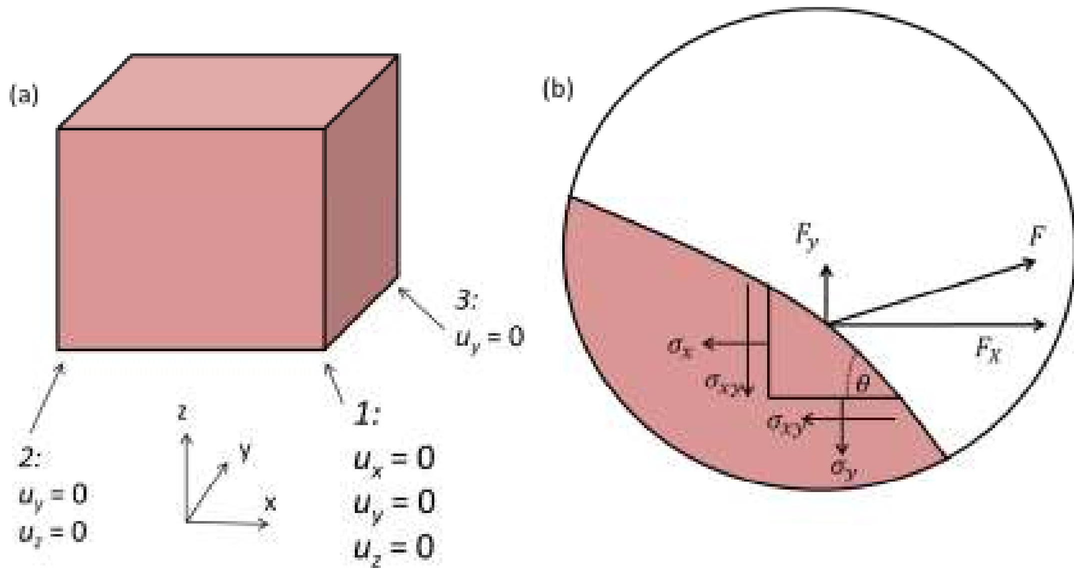


Figure 4.3 (a) Conditions for displacements (u_i) to prevent uniform displacements and rotations. (b) Balance of forces at the surface of an object.

We can derive the boundary condition by balancing the forces on a point of the object. The surface tractions, F, have units of pressure. For the 2D example shown in Figure 4.3(b),

$$\sigma_x \cos(\theta) + \sigma_{xy} \sin(\theta) = F_x,$$
$$y_x \sin(\theta) + \sigma_{xy} \cos(\theta) = F_y. \qquad 4.5.$$

where θ is the angle between the surface and the x-axis.

4. Hooke's Law

Stress and strain are related by Hooke's law, which is also called the constitutive equation. It is the second equation that must be satisfied when solving a stress-strain problem, and it can be written as

$$\boldsymbol{D}\vec{\sigma} = \overrightarrow{\epsilon_{el}} = (\overrightarrow{\epsilon_{tot}} - \overrightarrow{\epsilon_0} - \overrightarrow{\epsilon_{th}}). \quad 4.6.$$

I have introduced several new terms, including three new strain terms: $\overrightarrow{\epsilon_{el}}$, $\overrightarrow{\epsilon_0}$, and $\overrightarrow{\epsilon_{th}}$. First, the thermal strain, $\overrightarrow{\epsilon_{th}}$, is the product of α, the coefficient of thermal expansion (CTE), and the change in temperature (ΔT).

$$\overrightarrow{\epsilon_{th}} = \alpha \cdot \Delta T \begin{pmatrix} 1 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}. \quad 4.7.$$

I will only use isotropic CTEs i.e. (1,1,1,0,0,0) . To give a sense of magnitude, $\alpha \approx 3\text{x}10^{-6}$ / K for silicon and $\Delta T \approx 300$ K, so a typical thermal strain from room temperature to cryogenic temperature is ≈ 0.1 %.

Second, $\overrightarrow{\epsilon_{el}}$, is the elastic strain. This is the strain caused by stress. But, you might ask, isn't all strain due to stress? No. For example, thermal strain (for uniform α and ΔT) is a stress-less strain. The object has been deformed, but no stress occurs. Typically the elastic strain is what we care about, so unless I refer to a specific strain I am referring to the elastic strain.

Finally, $\overrightarrow{\epsilon_0} \equiv \boldsymbol{D}\overrightarrow{\sigma_0}$ is a term that will let us introduce other sources of stress into the problem. For example, when growing a thermal oxide, the silicon dioxide expands to incorporate the extra oxygen atoms. But the silicon dioxide cannot freely expand because

of the nearby silicon. This results in a compressive stress in the silicon dioxide of $\vec{\sigma_0} = -200$ MPa [60,61].

$\boldsymbol{D}$ is the compliance matrix, which allows us to convert between stress and strain. For an isotropic material the compliance matrix is

$$\boldsymbol{D} = \frac{1}{Y}\begin{bmatrix} 1 & -\nu & -\nu & & & \\ -\nu & 1 & -\nu & & 0 & \\ -\nu & -\nu & 1 & & & \\ & & & 1+\nu & 0 & 0 \\ & 0 & & 0 & 1+\nu & 0 \\ & & & 0 & 0 & 1+\nu \end{bmatrix}. \qquad 4.8.$$

where Y is the Young's modulus and ν is the Poisson's ratio. Using silicon as an example, Y = 130 GPa and $\nu = 0.27$. For the scenarios encountered later in this dissertation, the typical strain is between 0.1 % and 1 %, so the typical stress is between 100 MPa and 1 GPa.

B. Silicon Band Structure

I reviewed stress and strain because strain might be the cause of the unintentional QDs. Strain can do this by changing the band structure of silicon, so I will need to review the band structure of silicon. A good review of the band structure of silicon is given in ref. [62]. Silicon is an indirect band gap semiconductor. The conduction band (CB and E_C) and the valence band (VB and E_V) are shown in Figure 4.4. They are separated by the 1.12 eV band gap (E_G).

I note that band diagrams are typically drawn for electron energies, which have a negative charge. Electrons prefer the lower energy states, and the holes prefer higher energy states.

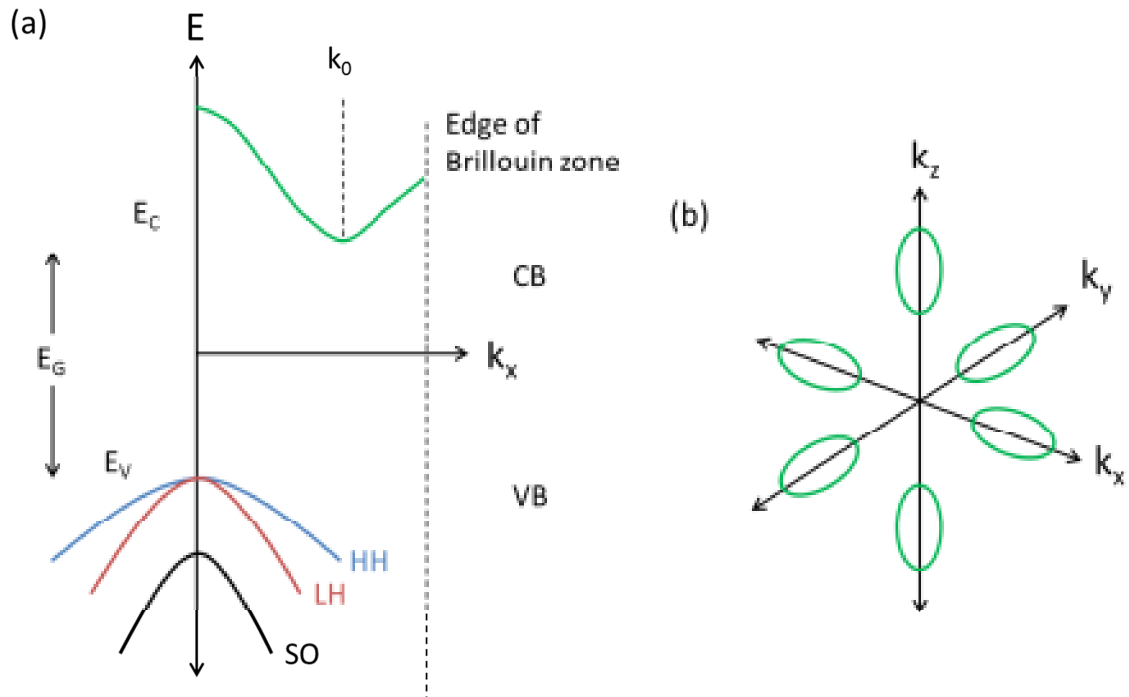


Figure 4.4. The band structure of silicon. (a) Schematic of the silicon bands for the top of the VB and the bottom of the CB. (b) Constant energy surfaces for the six valleys of the CB of silicon.

1. Conduction Band

a) Band Structure

There are six equivalent points in the Brillouin zone that are minima of the CB. They are located 85 % of the way to the edge of the Brillouin zone, at $\vec{k} = (\pm k_0, 0, 0)$, $(0, \pm k_0, 0)$ and $(0, 0, \pm k_0)$, where $k_0 = 0.85(2\pi/a_{si})$ and a_{si} is the lattice constant of silicon (a_{si} = 0.543 nm). Because each valley is a minimum in the CB, the energy near the center of the valley can be expanded to second order. This is the origin of the effective mass of the CB. In each valley the effective mass in the direction parallel to k_0 is different

than the effective mass in the transverse direction. The dispersion relations for each valley are

$$E_{C,\pm k_x}(\vec{k}) = \frac{\hbar^2}{2}\left(\frac{(k_x \mp k_0)^2}{m_l} + \frac{{k_y}^2}{m_t} + \frac{{k_z}^2}{m_t}\right)$$

$$E_{C,\pm k_y}(\vec{k}) = \frac{\hbar^2}{2}\left(\frac{{k_x}^2}{m_t} + \frac{\left(k_y \mp k_0\right)^2}{m_l} + \frac{{k_z}^2}{m_t}\right) \qquad 4.9.$$

$$E_{C,\pm k_z}(\vec{k}) = \frac{\hbar^2}{2}\left(\frac{{k_x}^2}{m_t} + \frac{{k_y}^2}{m_t} + \frac{(k_z \mp k_0)^2}{m_l}\right).$$

The longitudinal effective mass is $m_l = 0.98m_e$, and the transverse effective mass is $m_t = 0.19m_e$, where m_e is the free electron mass.

The six-fold valley degeneracy can be broken by both confinement and strain.

b) Confinement

To understand how confinement can break the six-fold valley degeneracy, we will examine the case of an inversion layer. An inversion layer forms in a metal-oxide-semiconductor (MOS) system, as shown in Figure 4.5(a). When a positive voltage is applied to the metal, positive charges appear at the metal-oxide interface, and negative charges appear at the oxide-semiconductor interface. This forces the silicon bands to bend with respect to the Fermi level [Fig. 4.5(b)].

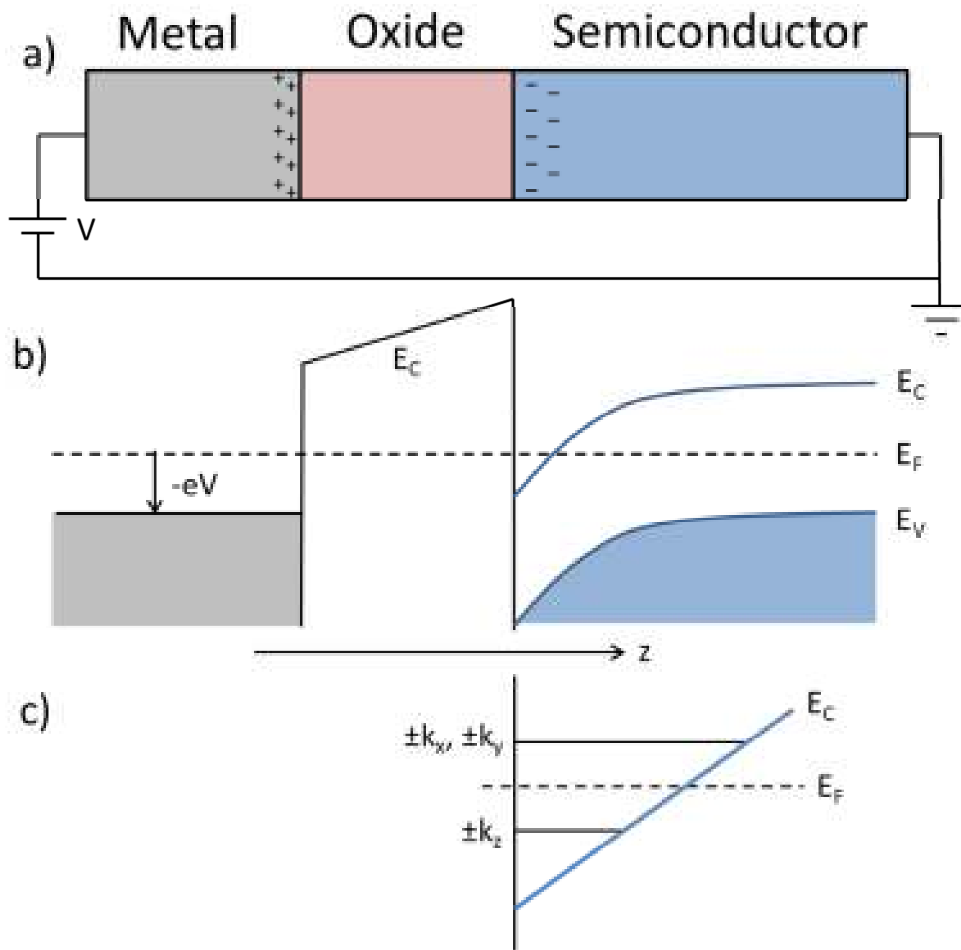


Figure 4.5. The inversion layer. (a) Circuit diagram for MOS capacitor. (b) Band diagram for the MOS capacitor in inversion. (c) Band diagram showing the semiconductor oxide interface and the six valley states. The black vertical lines in (b) and (c) represent the boundaries between the metal, the oxide and the semiconductor.

Figure 4.5(c) shows a zoomed in view of the CB near the interface. Here, the CB can be approximated as a triangular potential well. The oxide is treated as an infinitely high barrier, and the potential in the semiconductor is $e\mathcal{E}z$, where $\mathcal{E}$ is the electric field. The eigenenergies of this potential well are

$$E_n = \left(\frac{(e\mathcal{E})^2\hbar^2}{2m_z}\right)^{1/3} |a_n|, \qquad 4.10.$$

where a_n are the zeroes of the Airy function (a_0 = -2.33) [63]. Because the inversion layer is perpendicular the z-axis, $m_z = m_l$ for the $\pm k_z$ valleys, and $m_z = m_t$. for the $\pm k_x$ and $\pm k_y$ valleys. Therefore, the $\pm k_z$ valleys (E_0 = 37 meV using a conservative estimate of the strength of the electric field, 10^5 V/cm) are lower in energy than the $\pm k_x$ and $\pm k_y$ valleys (E_0 = 63 meV). Because this difference is many kT at cryogenic temperature, we can assume that only the $\pm k_z$ valleys are occupied. Therefore, confinement splits the six-fold degenerate CB into a two-fold degenerate $\pm k_z$ valley ground state and a four-fold degenerate $\pm k_x$ and $\pm k_y$ valley excited state.

c) Strain

Strain will also change the energy of the valleys. The change in energy for each pair of valleys is

$$\begin{aligned} \Delta E_{C,\pm k_x}(\epsilon_x, \epsilon_y, \epsilon_z) &= \Xi_u \epsilon_x + \Xi_d(\epsilon_x + \epsilon_y + \epsilon_z) \\ \Delta E_{C,\pm k_y}(\epsilon_x, \epsilon_y, \epsilon_z) &= \Xi_u \epsilon_y + \Xi_d(\epsilon_x + \epsilon_y + \epsilon_z) \\ \Delta E_{C,\pm k_z}(\epsilon_x, \epsilon_y, \epsilon_z) &= \Xi_u \epsilon_z + \Xi_d(\epsilon_x + \epsilon_y + \epsilon_z), \end{aligned} \qquad 4.11.$$

where for silicon Ξ_u = 10.5 eV and Ξ_d = 1.1 eV [62,64,65]. Because $\Xi_u >> \Xi_d$, the first term dominates these equations, and we approximate the change in energy of the $\pm k_z$ valleys as

$$\Delta E_{C,\pm k_z}(\epsilon_x, \epsilon_y, \epsilon_z) \approx \Xi_u \epsilon_z. \qquad 4.12.$$

Ξ_u is positive because the CB has a significant contribution from the atomic 3-d bonding orbitals. As a bonding orbital, if the atoms are pulled apart ($\epsilon > 0$), the band will increase in energy ($\Delta E_C > 0$), so the constant of proportionality should be positive ($\Xi_u > 0$).

2. Valence Band

a) Band Structure

The VB of silicon comes from bonding p-orbitals, so states in the VB have an orbital angular momentum $L = 1$, a spin $S = ½$. The differences in angular momentum give the VB three different hole states: the heavy hole, the light hole and the spin-orbit split-off band (Fig. 4.4). Heavy holes have a total angular momentum $J = 3/2$ and $J_z = \pm 3/2$. Light holes has a total angular momentum $J = 3/2$ and $J_z = \pm 1/2$. Holes in the spin-orbit spit-off band have $J = 1/2$ and $J_z = \pm 1/2$.

$$\begin{gathered}\text{Heavy Hole (HH)} \left|\tfrac{3}{2};\pm\tfrac{3}{2}\right\rangle \\ \text{Light Hole (LH)} \left|\tfrac{3}{2};\pm\tfrac{1}{2}\right\rangle \\ \text{Spin-Orbit (SO)} \left|\tfrac{1}{2};\pm\tfrac{1}{2}\right\rangle\end{gathered} \qquad 4.13.$$

Spin-orbit coupling splits off the spin-orbit band by 44 meV [62]. At low temperatures this is many kT, so in thermal equilibrium the spin-orbit band has no holes.

The Luttinger Hamiltonian, H_L, describes the VB of silicon [62,65].

$$\begin{gathered}H_L = \frac{1}{2m_e}\Big[\hbar^2\left(\gamma_1 + \tfrac{5}{2}\gamma_2\right)\boldsymbol{\nabla}^2 - \gamma_2(\boldsymbol{\nabla}\cdot\mathbf{J})^2 + \\ 2(\gamma_3 - \gamma_2)\left(\nabla_x^2 J_x^2 + \nabla_y^2 J_y^2 + \nabla_z^2 J_z^2\right)\Big].\end{gathered} \qquad 4.14.$$

where m_e is the free electron mass and $\boldsymbol{J}$ is the angular momentum operator and for silicon $\gamma_1 = 4.22$, $\gamma_2 = 0.39$ and $\gamma_3 = 1.44$ [66]. By applying the Luttinger Hamiltonian to the LH and HH states and ignoring the band mixing (the diagonal approximation), we get

$$H_{HH} = \frac{\hbar^2}{2m_e}\left[(\gamma_1 - 2\gamma_2)\mathrm{k}_x^2 + (\gamma_1 + \gamma_2)\left(\mathrm{k}_y^2 + \mathrm{k}_z^2\right)\right]$$
$$H_{LH} = \frac{\hbar^2}{2m_e}\left[(\gamma_1 + 2\gamma_2)\mathrm{k}_x^2 + (\gamma_1 - \gamma_2)\left(\mathrm{k}_y^2 + \mathrm{k}_z^2\right)\right]. \quad 4.15.$$

b) Confinement

Just like the CB, both confinement and strain will split the VB. Unfortunately, the VB is more complicated than the CB, so understanding the effect of confinement is more complicated. Because we will encounter a nanowire where holes are the charge carriers in Chapter 5, let us consider the effect of confinement in a nanowire parallel to the x-axis.

Recent theoretical work [67] has shown that the highest VB state is predominately LH in character with its spin aligned with the axis of the nanowire. I will go through a qualitative argument that shows that it is reasonable for the LH to be the ground state.

Given a nanowire parallel to the x-axis with a radius, R = 5 nm, the wave-vectors k_y and k_z are quantized, and the smallest wave-vectors should be of order $k_y, k_z \sim 2\pi/4R \approx 0.3\,\frac{1}{\mathrm{nm}}$. In contrast, along the x-direction the nanowire is very long, so the wave-vector k_x can be very small (limited by the thermal energy $\frac{\hbar^2 k_x^2}{2m_x} \sim kT$ so $k_x \sim 0.05\,\frac{1}{\mathrm{nm}}$). Using Eq. 4.15 and these estimates of the wave vectors to estimate the energy of both the LH and the HH,

$$H_{HH} = \frac{\hbar^2}{2m_e}\left[(\gamma_1 - 2\gamma_2)\mathrm{k}_x^2 + (\gamma_1 + \gamma_2)\left(\mathrm{k}_y^2 + \mathrm{k}_z^2\right)\right] \approx 31\text{ meV}$$

$$H_{LH} = \frac{\hbar^2}{2m_e}\left[(\gamma_1 + 2\gamma_2)\mathrm{k}_x^2 + (\gamma_1 - \gamma_2)\left(\mathrm{k}_y^2 + \mathrm{k}_z^2\right)\right] \approx 26\text{ meV}. \quad 4.16.$$

Thus it is reasonable that the highest energy VB state in a nanowire is the LH with its spin aligned with the nanowire.

c) Strain

Strain will also change the energy of the VB. Neglecting band mixing due to strain, for a hole with its spin aligned to the x-axis, the changes in the VB for the LH and the HH are

$$\Delta E_{LH} = a_v\left(\epsilon_x + \epsilon_y + \epsilon_z\right) + b_v\left(\epsilon_x - \left(\epsilon_y + \epsilon_z\right)/2\right)$$

$$\Delta E_{HH} = a_v\left(\epsilon_x + \epsilon_y + \epsilon_z\right) - b_v\left(\epsilon_x - \left(\epsilon_y + \epsilon_z\right)/2\right). \quad 4.17.$$

where $a_v = 2.1$ eV and $b_v = -2.33$ eV [62,64,65]. Because $a_v \approx - b_v$, the change in energy of the LH is approximately

$$\Delta E_{LH} \approx \frac{3}{2}a_v\left(\epsilon_y + \epsilon_z\right). \quad 4.18.$$

As I mentioned earlier, the VB comes from bonding p-orbitals, so a_v is positive.

C. Previous Work on Strain Effects in Nanostructures

At the beginning of this Chapter, I said that the progression I will follow in the next Chapter is that the stresses in a device create strain. The strain changes the CB and VB through the deformation potentials. QDs form in the strain-altered CB and VB. To conclude this Chapter, I now give some examples of semiconductor nanostructures that use the strain-altered band structure.

1. Strain-Induced QDs in III-Vs

Previous work has been reported on strain-induced QDs. In this work, carriers are confined within an InGaAs/GaAs quantum well due to strains from an InP stressor located above the quantum well. The strains, through the deformation potentials, create confinement in both the CB and VB. These strain-induced quantum dots are then studied optically (see the photoluminescence data shown in Figure 4.6). Local strain is needed to create confinement within a quantum well, because local electrostatic potentials (from surface gates) cannot create both a minimum in the CB and a maximum the VB. For good reviews of strain-induced QDs in III-Vs see references [68,69].

2. Intentional Strains in Silicon

Many silicon nanostructures *intentionally* incorporate strain to alter the silicon bandstructure. First, periodic strains have been used to make silicon super-lattices that have no materials interfaces [70]. Second, local strains are being used to address individual phosphorous donor qubits because strain changes the hyperfine coupling of the donor [31,71–73]. Third, transistors that are currently being fabricated use strain to increase the mobility of electrons in silicon [64,66,74].

3. Unintentional Strains in Silicon

There are also examples of the effect that *unintentional* strains can have on a device. First, in Si/SiGe nanowire resonant tunneling diodes, strain from the lattice mismatch is needed to understand the voltage dependence of the current [75]. These devices consist of a nanowire of $Si_{0.75}Ge_{0.25}$ with thin Si layers to create tunnel barriers for holes. The tunnel barrier comes from the band offset between Si and SiGe. The lattice mismatch between Si and SiGe creates strain, which creates confinement within

the nanowire. This confinement is needed to explain fine structure in current through the resonant tunneling diode.

Second, in Si/SiGe quantum wells, lattice defects can affect the formation of QDs [76]. In this work X-ray nanodiffraction was used to measure distortions in the silicon QW layer of a Si/SiGe heterostructure. They suggested that these distortions could affect the formation of QDs in these devices.

Third, in the next Chapter I will describe previous work at NTT on the PADOX (PAttern Dependent Oxidation) mechanism [77–79]. In this work, strain from the thermal oxidation of a silicon nanowire created tunnel barriers at the ends of the nanowire.

The work I will present in the next Chapter is different from this previous work in key ways. None of this previous work considered the strains from having electrostatic gates or contacts on a device. Because electrostatic gates and contacts are ubiquitous in the silicon QD community, I will suggest that strain-induced QDs may be very common. In the next Chapter, I will discuss the same strain-induced QD mechanism in three different device architectures. Thus, while previous work has shown some effect from strain in specific cases, I will make a more general argument, that strain can be as important as electrostatics to understanding a silicon QD device.

D. Conclusions

Now I have explained the basics of stress, strain and the band structure of silicon. In the next Chapter I will show how the strain altered band structure can cause QDs. We will see that the typical strains from either CTE mismatch or intrinsic strain are of order 0.1 %. Given the deformation potentials I discussed in this Chapter, these strains will

results in modulations of the CB and VB of order 10 meV. CB modulations of this magnitude are important to understand because 10 meV is the same magnitude as the change in the CB due to either an electrostatic tunnel barrier [80]or interface traps [32].

Chapter 5: Strain-Induced Quantum Dots

Based on “Determining the location and cause of unintentional quantum dots in a nanowire,” by Ted Thorbeck and Neil M. Zimmerman, published in Journal of Applied Physics **111**, 064309, (2012).

And “Formation of Strain-Induced Quantum Dots in Gated Semiconductor Nanostructures,” by Ted Thorbeck and Neil M. Zimmerman, to be published.

A. Preview

In the first Chapter I described how unintentional QDs are a common problem for us and for many other groups. In Chapter 3, I showed how we can use gate capacitances to determine the locations of the unintentional QDs in the nanowire. In this Chapter, I will simulate the strains in the device arising both from device fabrication and from cooling the device to operate at cryogenic temperatures. These strains are then converted to modulations in the conduction and valence band using the deformation potential as discussed in the previous Chapter.

I examine three device architectures. First, I examine the mesa-etched nanowire that I have discussed in the previous chapters. Second, I discuss bulk silicon devices, with both metal and poly-silicon gates. I showed a few examples of unintentional QDs in these devices in Chapter 1. Third, I discuss holes in a chemically-grown silicon nanowire with metal contacts. Tunnel barriers are often observed near the metal-nanowire interface, even in materials systems that should not form Schottky barriers. By discussing three different device architectures and seeing strain is important for each one,

we can see that strain can be as important as the electrostatics to the behavior of these devices.

B. Mesa-Etched Nanowire

In the previous Chapter I have suggested that strain might be the cause of dots A and B. In this section we will see that the strain from the nanowire from the lower gates is of order 0.1%. Knowing that the deformation potentials of the CB is about 10 eV from the previous Chapter, this will lead to modulations in the CB of order 10 meV. This is modulation of the CB is similar in magnitude to the CB modulation due to electrostatic tunnel barriers [80] or interface traps [32].

In this section I will show that dot B can be explained by a previous work on strain-induced tunnel barriers at the ends of mesa-etched nanowires. Dot A is more complicated. I will simulate the strain from a lower gate. Although the strains-induced peaks in the CB are large enough to create tunnel barriers, the peaks in the different valleys are not at the same locations. This means that while the strain-altered CB should affect the transport through the device, I cannot conclude that it should cause dot A.

In Figure 5.1 (a) and (b) I show the locations of the dots A and B, as determined in Chapter 3. A schematic of the CB profile needed to explain dots A and B is shown in Figure 5.1(c). I will begin this Chapter by examining the tunnel barrier at the end of the nanowire.

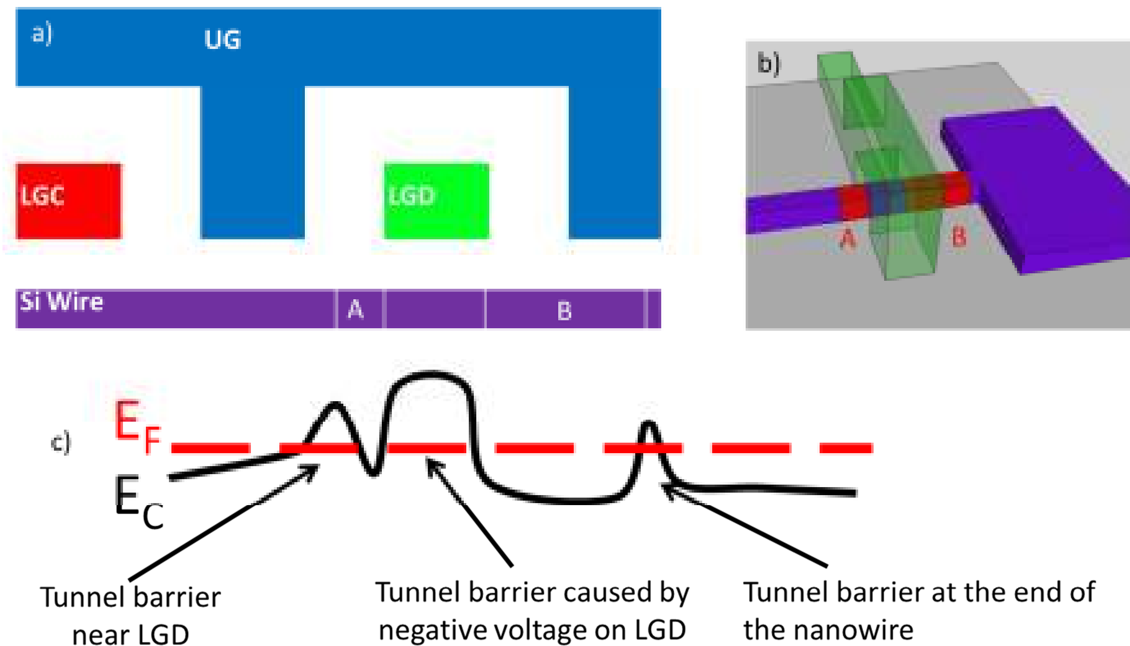


Figure 5.1. CB profile that could explain dots A and B. (a) Cross section of the device with the locations of dots A and B indicated. (b) Pseudo-3D view of dots A and B. (c) Schematic of CB profile that could cause dots A and B.

1. Dot B

To observe an unintentional QD at the location calculated for dot B, there must be a tunnel barrier underneath LGD and another TB at the end of the nanowire. The negative voltage on LGD will generate a tunnel barrier underneath it. The tunnel barrier at the end of the nanowire is harder to explain. According to the electrostatics, there should not be a tunnel barrier at the end of the nanowire.

An explanation for the tunnel barrier at the end of the nanowire comes from work at NTT in similar devices [43,77–79]. Those devices did not have lower gates, so there

were no electrostatic tunnel barriers. Nevertheless, Coulomb blockade was observed in those devices, and the capacitance to the (upper) gate scaled with the length of the nanowire [78]. They thus determined that there must be a mechanism generating tunnel barriers at the ends of the nanowire. The NTT group demonstrated that the combined effect of confinement and strain, from the thermal oxidation of the nanowire, was the cause of the tunnel barriers. This is referred to as the PADOX (PAttern Dependent OXidation) mechanism, and it is explained in Figure 5.2.

The NTT group attributed the tunnel barrier at the ends of the nanowire to the combined effects of confinement and strain. Confining electrons in the nanowire raises the CB. Strain from the thermal oxide lowers the CB in the center of the nanowire. A thermal oxide growing on a silicon nanowire must expand to incorporate the two extra oxygen atoms. The nanowire geometry prevents the oxide from fully expanding, causing compressive (negative) stress in the nanowire. As we saw in Chapter 4, compressive stress lowers the CB of silicon. The combined effect of confinement and strain on the CB will create peaks in the CB at the ends of the nanowire that will cause tunnel barriers.

The PADOX devices did not have any lower gates. Is it possible that the electrostatic effect of the lower gates could affect the tunnel barrier at the end of the nanowire? For these devices the end of the nanowire is 75 nm from the lower gate. As can be seen in Figure 3.7 the electrostatic effect of the lower gates falls off quickly along the nanowire. This is because the UG screens the end of the nanowire, preventing electric field lines from the lower gates from reaching the end of the nanowire. So there should be no electrostatic effect from the lower gates on the tunnel barriers at the ends of the nanowire.

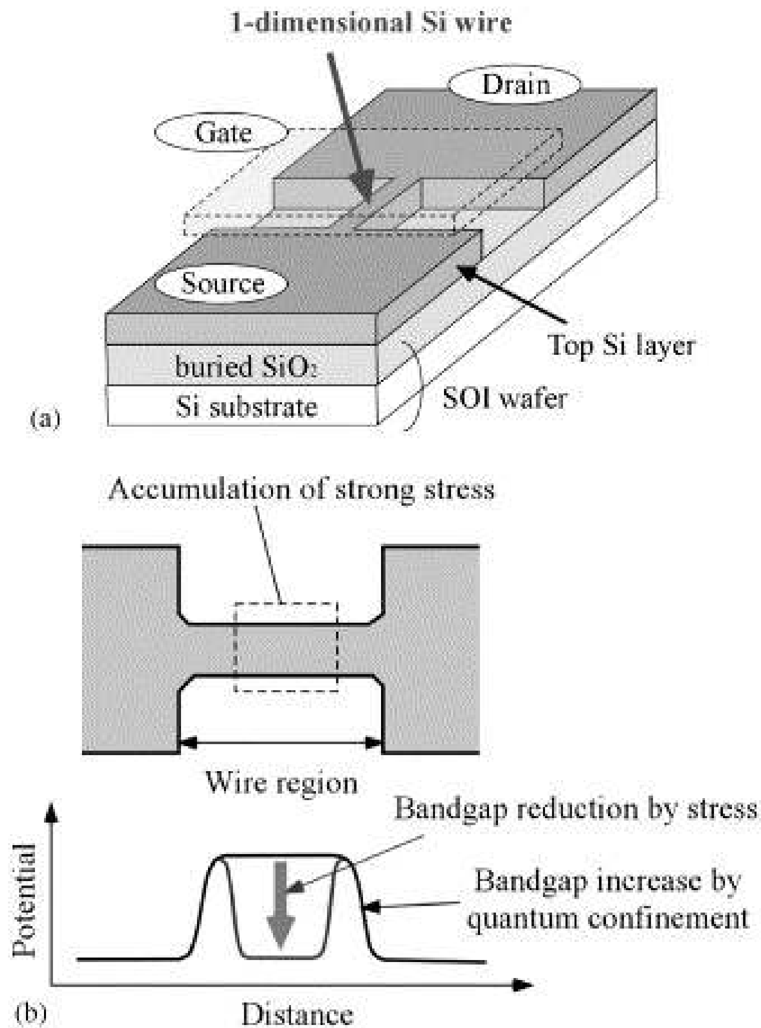


Figure 5.2 Schematic of the PADOX mechanism. (a) Pseudo 3-D view of the device. (b) Top view of nanowire showing region of high stress. (c) CB profile due to confinement and strain. Reprinted from Physica E, Vol. **19**, Y. Takahashi, Y. Ono, A. Fujiwara, and H. Inokawa, "Development of silicon single-electron devices", 95-101., Copyright (2003), with permission from Elsevier.

The tunnel barrier that I observed at the end of the nanowire is consistent with the PADOX mechanism. Now I can recommend methods to mitigate the unintentional QD. We could reduce the strain from thermal oxidation by changing the processing

conditions, such as temperature. But if we optimize the processing conditions to reduce strain, then we might have to sacrifice in the quality of the oxide in other ways, such as increasing the number of interface traps. Alternatively, we could reduce the impact of the unintentional QD, instead of reducing the strain. One method to do this is simply to increase the distance between LGD and the end of the nanowire. Making this distance longer would decrease the charging energy of the quantum dot. If the nanowire is long enough, then the charging energy will be too small to observe Coulomb blockade. Finally, we could increase the voltage on the UG to push the CB down so that any peaks due to strain would be below the Fermi level and would not cause tunnel barriers. The cost of this mitigation is a higher charge on the QD, because we also use the UG to control the charge on the QD. This cost could be overcome by splitting the UG into several gates, as shown in Figure 5.3. We are working on an architecture similar to this at NIST.

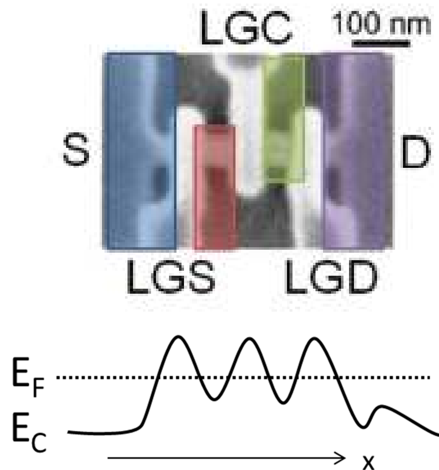


Figure 5.3. Split UG to separately gate the ends of the nanowire and the QDs. Increasing the voltage on the UG over the end of the nanowire lowers the CB so that the peak is below the Fermi level. Thus a tunnel barrier will not form.

2. Dot A

If dot B is due to a strain-induced tunnel barrier, then could dot A also be due to a strain-induced tunnel barrier? Nobody has previously considered the effect of strain from the gates in a QD device on the CB. I will address this topic as two separate questions. 1) Is the effect on the CB from the strain from the lower gates large enough to create tunnel barriers? The answer to this question is yes. 2) Can the strain-altered CB cause dot A? My attempts to answer this question were inconclusive.

To answer these questions, I simulated the strain in the device in COMSOL. Appendix C contains a simple tutorial on COMSOL. This simulated device was based on the dimensions of AF-CA2U3D-3 (device 1 from Chapter 3). The simulation includes both CTE mismatch and intrinsic stress, as discussed in the previous Chapter. The intrinsic stress in the SiO_2 is -200 MPa [60,61], and the intrinsic stress in the poly-silicon is -400 MPa [81,82]. I used the following device parameters: t_{si} = 20 nm, w_{si} = 20 nm, L_{LG} = 40 nm, L_{UG} = 70 nm, $t_{ox,1}$ = 20 nm, and $t_{ox,2}$ = 30 nm. The device is shown schematically in Figure 5.4(a). All three lower gates are simulated, but only one is shown for clarity. To see how sensitive the simulation is to variations in the oxide thickness, I made the oxide around LGD asymmetric. On one side of LGD the oxide is 35 nm thick. On the other side it is 25 nm. The file used in the COMSOL simulation is located on Guestroom PC: in C:\Ted\Stress\Save\poly gated nw final.mph.

In the top panel of Figure 5.4(c) I show a cross section of the device. In the middle panel of Figure 5.4(c) the strains calculated along the midline of the top of the nanowire (if the center of the nanowire is the origin then this line cut is along y = 0, z =

$t_{si}/2$– 1 nm). In Chapter 4 I showed that the change in CB from strain for the $\pm k_z$ valleys is

$$\Delta E_{C,\pm k_z} = \Xi_u \epsilon_z + \Xi_d(\epsilon_x + \epsilon_y + \epsilon_z), \quad 5.1.$$

where Ξ_u = 10.5 eV and Ξ_d = 1.1 eV [64]. I used this to calculate the bottom panel in Figure 5.4(c). We see that ΔE_C has the same shape as ϵ_z, as we expect from Eq. 4.10, $\Delta E_{C,\pm k_z} \approx \Xi_u \epsilon_z$. In the bottom panel of Figure 5.4(c) we see that the modulation of the CB due to the effect of the lower gate is about 10 meV. This is the same magnitude that the CB changes due to either an electrostatic tunnel barrier [80]or interface traps [32]. To answer the first question from the beginning of this section, the change in the CB due to the strain from the lower gates is large enough to create strain-induced tunnel barriers.

Now to consider the second question, can the strain-altered CB explain dot A? We need to consider electrons on all surfaces of the nanowire. In Figure 5.4(c), I consider the top of the nanowire, and in Figure 5.4(d) I consider the sides of the nanowire. The middle panels of 5.4(d) show the strains along the midline of the side of the nanowire. The strains in Figure 5.4 (d) were calculated along the middle of the side surface (y = $w_{si}/2$ – 1 nm, z = 0).

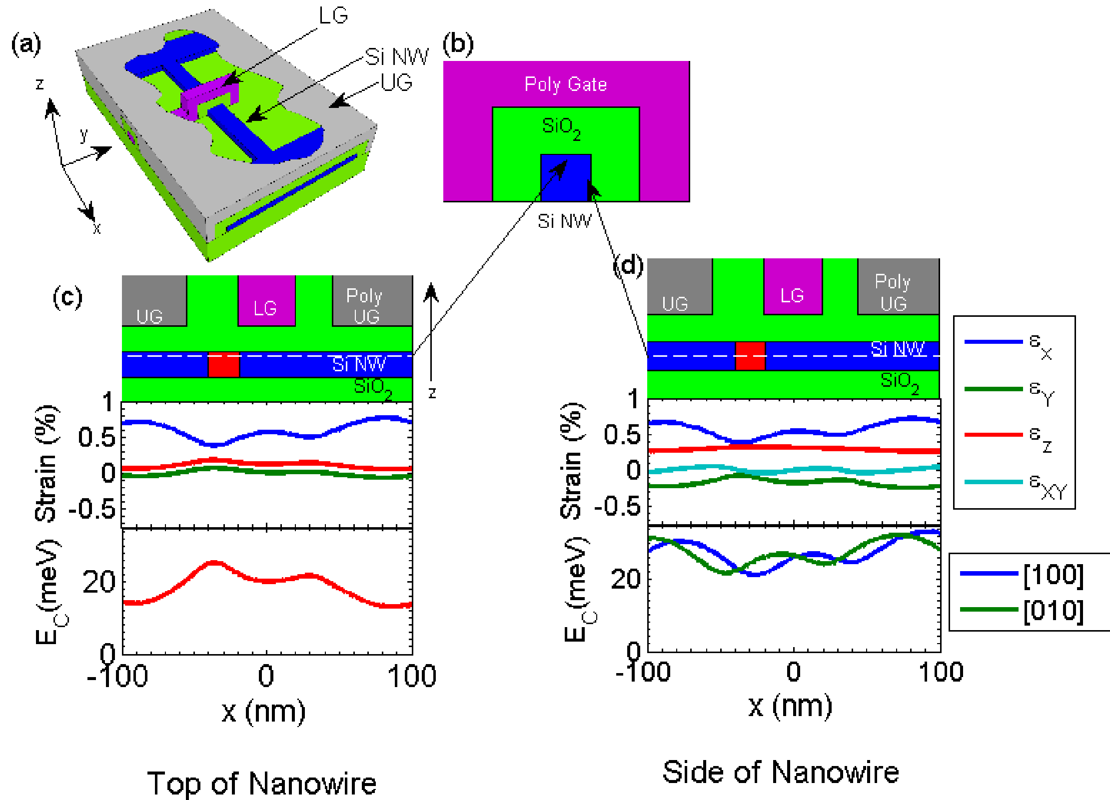


Figure 5.4 Strains and CB modulation in the mesa-etched nanowire. (a) A pseudo 3D image of the device, showing the nanowire (blue), SiO_2 (green), poly-Si LG (purple) and poly-Si UG (grey), partially cut away. (b) Schematic of the device perpendicular to the wire along the center of LG. (c) and (d) show the top and side surface of the nanowire. The upper panels are schematics of the device along the nanowire, with the location of dot A highlighted by a red square. The middle panels show the calculated strains along (c) a line in the center of the top of the nanowire and (d) the center of the side of the nanowire. The bottom panels are the calculated change in the CB along the nanowire in the CB due to strain. In (c) the calculation is for the $\pm k_z$ valleys. In (d) both the $k_{[010],[0\bar{1}0]}$ valleys, in blue and labeled [100], and $k_{[010],[0\bar{1}0]}$ valleys, in green and labeled [010], are calculated. All panels in (c) and (d) have the same horizontal axis scale.

We want to know how the strains change the CB along the top and side surfaces of the nanowire. The nanowire runs along the $[\bar{1}10]$ crystallographic direction and the wafer is perpendicular to the [001] direction. This is unfortunate because I have to either align my Cartesian coordinate system to the geometry of the nanowire or to the crystallographic directions; I cannot do both. As seen in Figure 5.4(a), I picked my coordinate system such that the nanowire is parallel to the x-axis. This means that the crystallographic directions are rotated 45° degrees from the nanowire axis. With this coordinate system there are no valleys in the k_x and k_y directions. Instead, those valleys have been rotated by 45°. I will label the valleys $k_{[100]}$, $k_{[010]}$, $k_{[\bar{1}00]}$, and $k_{[0\bar{1}0]}$. Because the rotation is in the x-y plane, there are still valleys in the $\pm k_z$ directions.

We will assume that the top surface can be treated as a bulk inversion layer. This means the $\pm k_z$ valleys are lowest in energy, as we saw in the last Chapter. However, we also have electron on all of the sides of the nanowire. Assuming that this can also be treated like a bulk inversion layer, then the $k_{[100]}$, $k_{[010]}$, $k_{[\bar{1}00]}$, and $k_{[0\bar{1}0]}$ valleys are degenerate and lower in energy than the $\pm k_z$ valleys. If this assumption is not correct then we need to consider all six valleys along on each side of the nanowire.

Along the sides of the nanowire we need to rotate the strains into the crystallographic basis.

$$\epsilon_{[100]} = \epsilon_x cos^2(\theta) + \epsilon_y sin^2(\theta) + \epsilon_{xy} sin(\theta) cos(\theta) = \frac{\epsilon_x + \epsilon_y + \epsilon_{xy}}{2}$$

$$\epsilon_{[010]} = \frac{\epsilon_x + \epsilon_y - \epsilon_{xy}}{2} \qquad 5.2.$$

where $\theta = \pi/4$ is the angle between the x-axis and the [100] crystallographic direction. The changes in energy of the $[100], [\bar{1}00], [010],$ and $[0\bar{1}0]$ valleys are

$$\Delta E_{[100],[\bar{1}00]} = \Xi_u \epsilon_{[100]} + \Xi_d(\epsilon_x + \epsilon_y + \epsilon_z)$$
$$\Delta E_{[010],[0\bar{1}0]} = \Xi_u \epsilon_{[010]} + \Xi_d(\epsilon_x + \epsilon_y + \epsilon_z). \quad 5.3.$$

Because $\epsilon_x + \epsilon_y + \epsilon_z$ is the trace of the strain tensor, it is the same in any basis, so it does not need to be written in the rotated basis. The lower panel of Figure 5.4(d) shows the change in energy of the $k_{[100],[\bar{1}00]}$ and $k_{[010],[0\bar{1}0]}$ valleys along the side of the nanowire.

We can compare the position of the tunnel barriers in the calculated CB profile to the location of dot A. The red boxes in the upper panels of Figures 5.4 (c) and (d) show the location of dot A. There must be barriers on both sides of the QD to create confinement. When we observe dot A, a negative voltage has been applied to LGD to create a tunnel barrier directly beneath LGD. This provides a confinement on the right-hand side of the QD. Can strain explain a tunnel barrier on the left-hand side of dot A? On the top of the nanowire we see that there is a peak in the CB of the $\pm k_z$ valleys due to strain at x = -40 nm. This barrier is located where a barrier must be to explain dot A. Unfortunately we do not see a similar barrier at x = -40 nm in either the $k_{[100],[\bar{1}00]}$ or $k_{[010],[0\bar{1}0]}$ valleys along the side wall. On both the top and side surfaces we see local modulations of the CB due to strain that are ≈ 5 meV. We will see in the next section that peaks of this height are large enough to create strain-induced QDs, but because these peaks are not in the same locations in the different valleys, this analysis fails to show that strain is the cause of dot A.

I have made two assumptions: 1) that the midpoint of the top of the nanowire is representative of the entire top of the nanowire and 2) only the k_z valleys are occupied on the top of the nanowire. In Figure 5.5 I show the change in energies of all six valleys at three different points along the top of the nanowire. We see that the energies of all of the valleys are similar at the three different positions I show, so assumption 1) is reasonable. If assumption number 2 is wrong, meaning that the other four valleys are also occupied, then we see that there are dips in the $k_{[100],[\bar{1}00]}$ or $k_{[010],[0\bar{1}0]}$ valleys at x = -40 nm, where there is a peak in k_z valleys. If these other four valleys are also occupied on the top of the nanowire, then we would not expect to observe a strain-induced tunnel barrier on the top of the nanowire at x = -40 nm.

The analysis that I discuss above incorporates the effect of strain in a simple way. Ultimately, I think a different approach is needed to determine if dot A is due to strain. First, I would calculate the strain as we have already done. Then I would use the strains to calculate the change in energy of each valley everywhere in the nanowire. Finally, I would self-consistently solve the electrostatics, the quantum mechanics and the semiconductor physics, to determine the charge density of the nanowire and the wavefunction of the electrons. If that predicts an isolated region of charge in the location of dot A, then we will have shown that dot A is due to strain. However, building a full simulation of strain, electrostatics and quantum mechanics is beyond the scope of this dissertation.

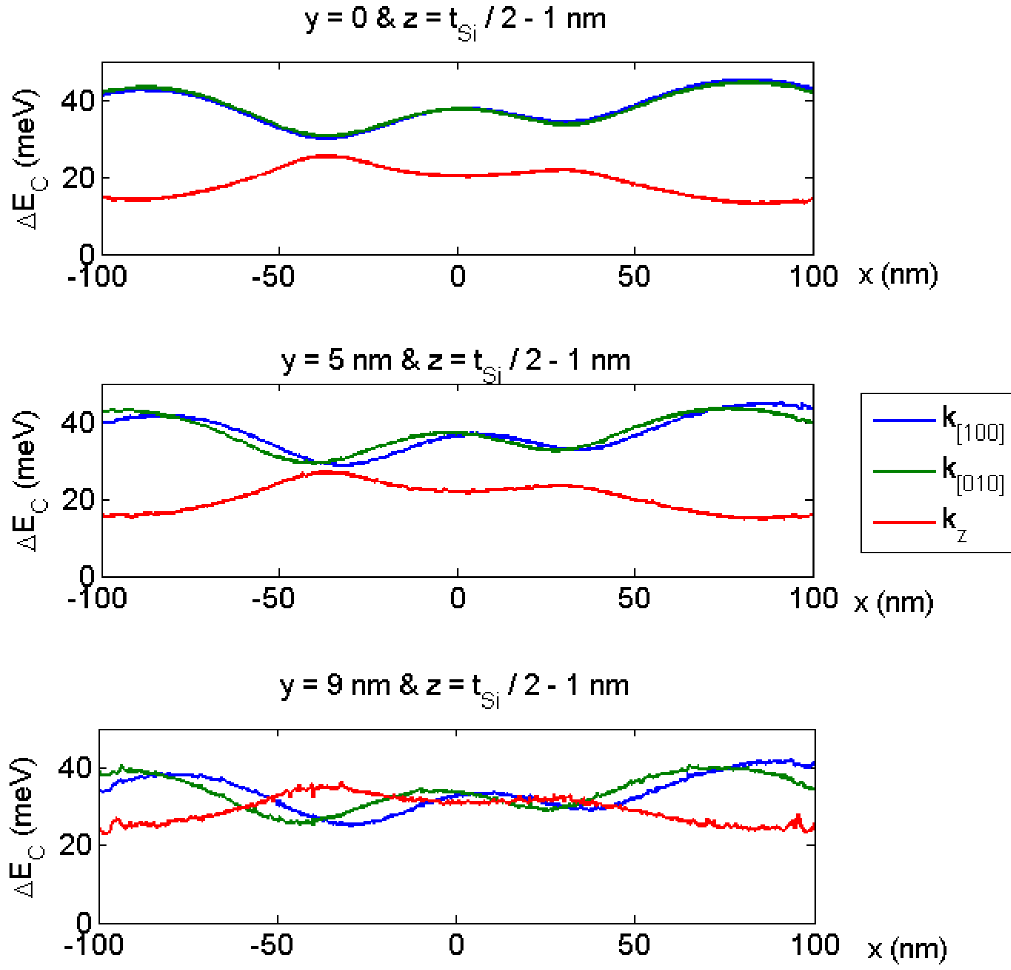


Figure 5.5 Change in energy of all six valleys at different points along the top of the nanowire. The blue curves represent the energy of the $k_{[100],[\bar{1}00]}$ valleys calculated using Eq. 5.3. The green curves represent $k_{[010],[0\bar{1}0]}$ valleys, calculated using Eq. 5.3. The red curve represents the energy of the k_z valleys calculated using Eq. 5.1. All three panels are at different positions along the top of the nanowire (z = $t_{si}/2$ - 1). The top panel corresponds to the center of the top of the nanowire (y = 0). The middle panel corresponds to the halfway between the midpoint of the top and the sidewall (y = $w_{si}/4$ = 5 nm). The bottom panel corresponds to the corner of the nanowire, 1 nm away from the sidewall (y = $w_{si}/2$ – 1 nm = 9 nm).

C. Metal-Gated Bulk Silicon

Having been unable to demonstrate that dot A was due to strain because of complications of the nanowire, I was curious to test the idea that strain could be the cause of the unintentional QDs in simpler geometries. I will first examine the strain in the metal-gated bulk silicon devices. I showed the effect of unintentional QDs in these devices in Chapter 1.

1. Metal Gates on Bulk Silicon

A metal-gated bulk-silicon device [28–30,32,39,83] is shown in Figure 5.6. This device consists of a lightly p-doped silicon wafer covered by 10 nm of SiO_2. The two aluminum gates, the UG and the LG, are isolated by 3 nm of AlO_X. The UG is 80 nm tall and 50 nm wide, and the LG has a 25 nm diameter. The UG and LG are perpendicular to each other. I only included one LG in the simulation even though multiple lower gates are needed to form an electrostatic QD. In these devices it is common to observe an unintentional QD below the metal lower gates, where there should only be a single tunnel barrier [22,29,30,32,45,84].

The electrical operation of this device is the same as the mesa-etched nanowires I have been discussing; the upper gate inverts the silicon and the lower gates are intended to deplete the silicon to form tunnel barriers. There is no silicon nanowire in this device, instead gate voltages create confinement in bulk silicon. Because this is a planar device, these devices only form an inversion layer perpendicular to the z-axis. Therefore, the electrons are only in the $\pm k_z$ valleys, and we do not have to worry about the other four valleys, as we did in the mesa-etched nanowire.

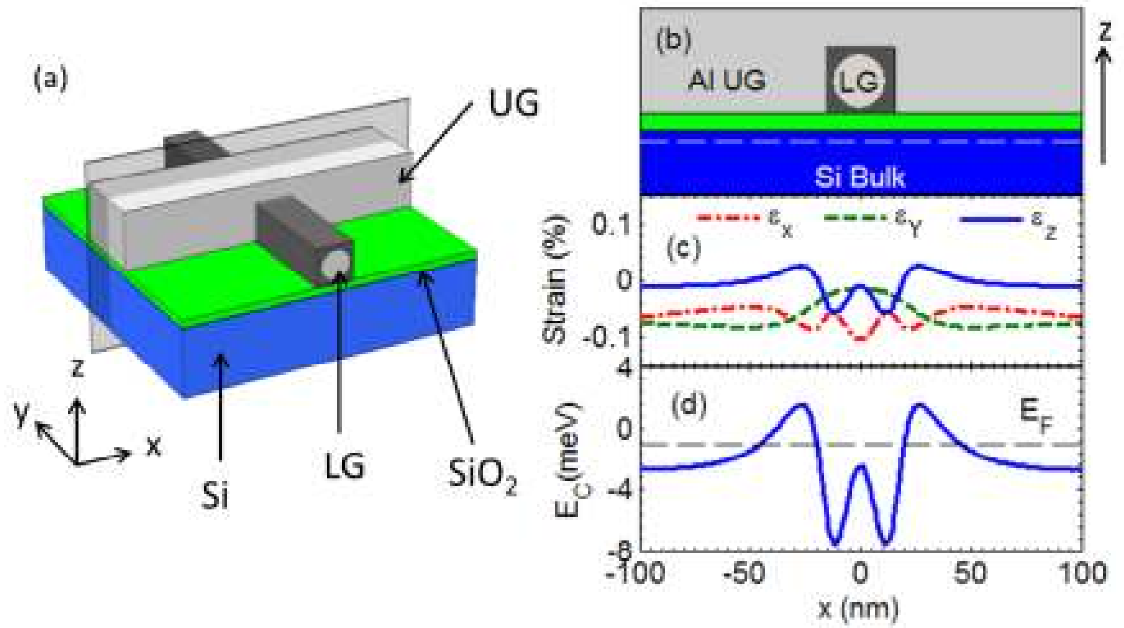


Figure 5.6 Strain-induced QD in metal-gated bulk silicon. (a) A pseudo-3D schematic of the device, showing the bulk silicon wafer (blue), SiO_2 (green), UG and LG (grey) and AlO_X (dark grey). (b) Schematic cross section through the semi-transparent plane in (a) with the same colors as (a). (c) Strains calculated using COMSOL in the inversion layer (white dashed line in (b)), showing the effects of the CTE mismatch between the Al and AlO_X. (d) Modulation of the CB from the strains in (c) showing tunnel barriers at $x = \pm 30$ and a QD in between. Dashed line represents the Fermi level. (b) - (d) All have the same horizontal scale.

The strains I calculated using COMSOL, as the device is cooled from 293 K to 1 K, are shown in Figure 5.6(c). The file used in the COMSOL simulation is located on Guestroom PC: in C:\Ted\Stress\Save\bulk si metal gates final.mph. The strains were

taken below the center of the UG, and 1 nm below the Si-SiO_2, because that is the center of the inversion layer. The strains elsewhere underneath the UG look similar.

The CTE mismatch is more important than the intrinsic stress in these devices. Al and AlO_X have small intrinsic stress when deposited [85]. There is intrinsic stress from the SiO_2, but from playing around with different values of the intrinsic stress in COMSOL, it does not appear to dominate the shape we see for the local strains. Now to consider the CTE mismatch, as the Al (α_{Al} = 23x10^{-6} / K) has a larger CTE than AlO_X (α_{AlOX} = 23x10^{-6} / K), the AlO_X is preventing the Al from contracting, so the Al is in tensile stress. Conversely, the AlO_X is in compressive stress. The stresses set up in the Al and AlOx propagate into the silicon, because of the equations of equilibrium. This explains why ϵ_z is negative below the AlO_X (x = ± 12 nm) and positive below the Al (x = ±30 nm).

The strain-altered CB is shown in Figure 5.6(d). Because the inversion layer is perpendicular to the z-axis, the electrons only occupy the $\pm k_z$ valleys. This makes the change in energy of the CB much easier to calculate than in the previous section, because we can simply use Eq. 5.1. We see that $\Delta E_C \approx \Xi_u \epsilon_z$, as we expect from Eq. 4.10.

Notice that ΔE_C has the right shape to form a strain-induced quantum dot, the peaks at ± 30 nm can form tunnel barriers and the dip between the tunnel barriers could form the QD. Now, I will determine if the peaks have the correct size to form tunnel barriers, by calculating the barrier resistance. The barrier resistance must be small enough that we can measure a tunneling current (R < 1 GΩ). But the barrier resistance must be much larger than the resistance quantum to observe discrete charging events on the QD

(R >> R_K = 26 kΩ). To estimate the resistance of the barriers, I use the WKB (Wentzel–Kramers–Brillouin) approximation,

$$1/R = G = N\frac{e^2}{\hbar}e^{-\frac{\pi}{\sqrt{2}}\frac{\sqrt{m^*}}{\hbar}\sqrt{\phi}L}. \quad 5.4.$$

where $m^* = 0.19m_0$ and the number of channels, N = 1 (because $N \approx wk_f \sim 1$ [86]). I calculate a tunneling resistance of 20 MΩ, given the height (φ = 4 meV) and length (L = 40 nm) of the barriers. This is large enough to quantize the charge on the QD without shutting current off.

I have shown that the strain-modulated CB in a metal-gated bulk silicon device has the correct shape and magnitude to induce a QD. Thus, unless the strain is mitigated during fabrication, a strain-induced QD would show up as an unintentional QD. The location of the strain-induced QDs match the observed location of the unintentional QDs in these metal-gated bulk silicon devices [22,29,30,32,45,84].

2. Poly-Silicon Gates on Bulk Silicon

Metals tend to have larger CTEs than semiconductors. Could we reduce the strain, and eliminate the strain-induced QDs, by replacing the metal gates with poly-silicon gates? Figure 5.7 shows a device identical to the device in Figure 5.6, except the Al has been replaced by poly-Si and the AlO_X has been replaced by SiO_2. Electrically, the poly-silicon gated devices operate just like the metal-gated device.

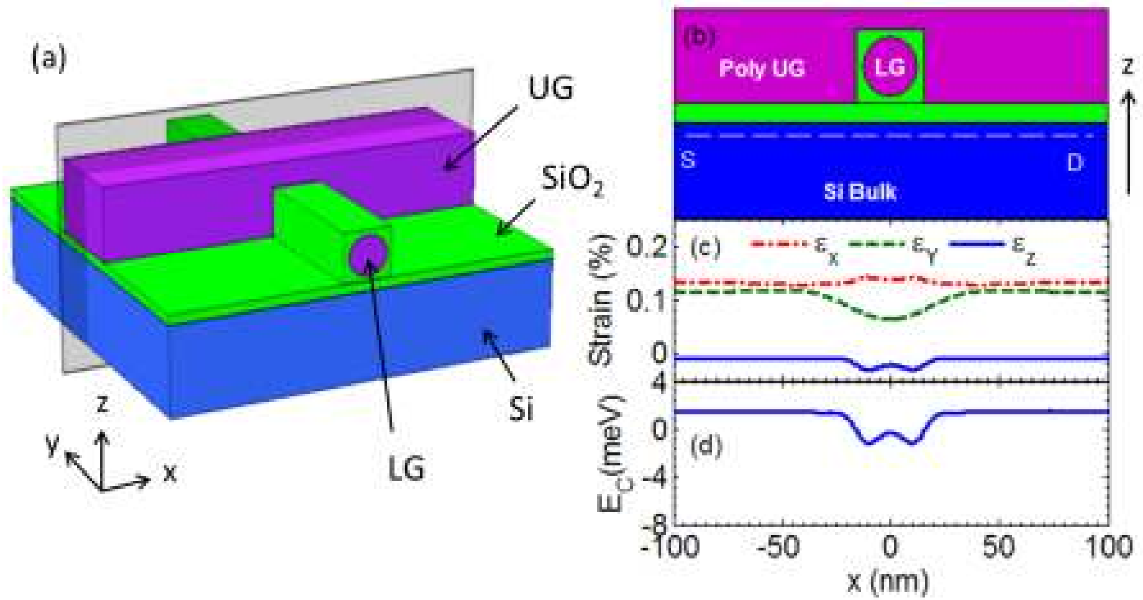


Figure 5.7 No Strain-induced QD in poly-Si gated bulk silicon. (a) A pseudo-3D view of the device, showing the bulk silicon wafer (blue), SiO_2 (green) and poly-silicon UG and LG (purple). (b) Schematic cross section through the semi-transparent plane in (a) with the same colors as (a). (c) Strain calculated by COMSOL in the inversion layer (white dashed line in (b)), showing smaller strains than in Figure 5.6. (d) Modulation of the CB from the strains in (c) showing no strain-induced tunnel barriers. (b) - (d) All have the same horizontal scale.

Switching the materials, from Al and AlO_X to poly-Si and SiO_2, significantly reduces the strain in the inversion layer [Fig. 5.7(c)]. The CB no longer has peaks to form tunnel barriers [Fig. 5.7(d)], so there is no strain-induced QD. Therefore, simply replacing the metal gates with poly-Si in a bulk QD device should reduce the frequency with which unintentional QDs are observed.

The file used in the COMSOL simulation is located on Guestroom PC: in C:\Ted\Stress\Save\bulk si poly gates final.mph.

D. Chemically-Grown Nanowire with Metal Contacts

So far I have consider whether strain could be the cause of tunnel barriers near gates in QD devices, now I will consider if strain from metal source and drain contacts can cause tunnel barriers near the metal contacts. Unlike the previous devices, there are no local gates to create electrostatic tunnel barriers, so there must be another mechanism generating tunnel barriers in the nanowire. Figure 5.8(a) shows a chemically-grown nanowire with metal contacts [87–90]. Unlike the previous device architectures I have discussed, this device architecture is for holes. Chemically-grown nanowires have some advantages over the mesa-etched nanowires that I discussed earlier; it is easy to grow a small diameter nanowire with little surface roughness [87].

In this device architecture, tunnel barriers must form near the nanowire-contact interface to confine holes within the nanowire. It is important to understand the cause of these tunnel barriers, because they are essential for forming the QDs. A Schottky barrier will often form at a metal-semiconductor interface, such as the contact-nanowire interface. But, in these devices, the semiconductor and the metal are often deliberately chosen to avoid a Schottky barrier. For example, at a bulk metal-InAs interface the Fermi level is pinned above the CB [91,92], thus preventing a Schottky barrier from forming. As another example, many semiconductor-metal interfaces that form Schottky barriers in bulk will not form Schottky barriers in nanostructures, because there are not enough interface states on the nanowire to pin the Fermi level [93,94]. Despite these efforts, tunnel barriers are often observed at nanowire-contact interfaces that should not

form Schottky barriers [95]. This is what led me to consider if strain-induced tunnel barriers could be forming at these interfaces.

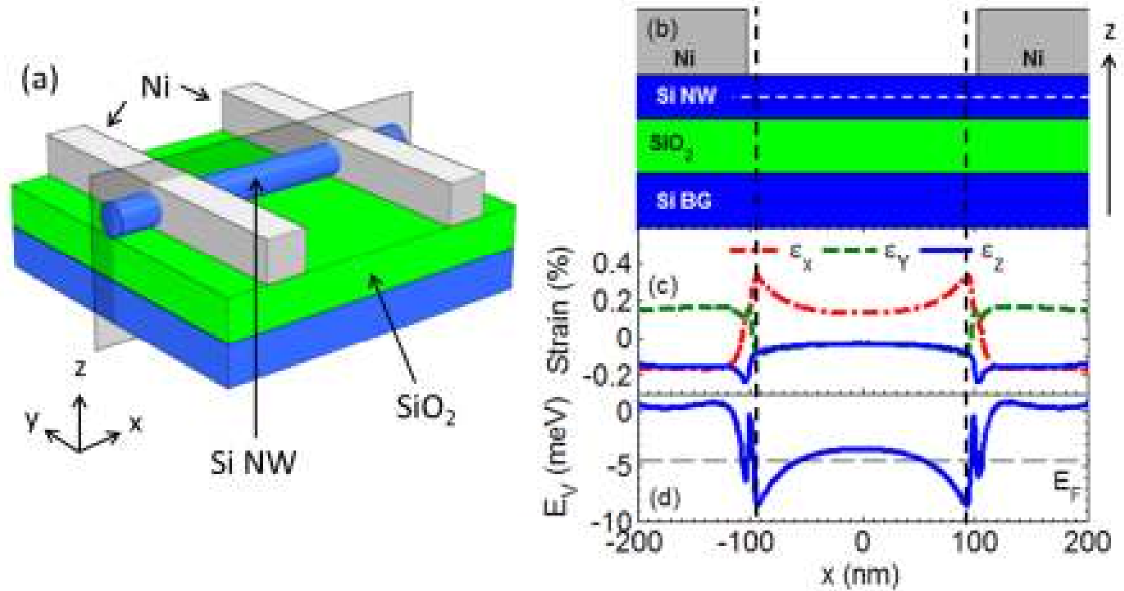


Figure 5.8 Strain-induced QD in chemically-grown nanowire with metallic contacts. (a) Pseudo-3D view of the device showing the bulk Si wafer (blue), SiO_2 (green), Si-nanowire (blue), nickel source and drain contacts (grey). (b) Cross section of the device, through the semi-transparent plane shown in (a), with same colors as (a). (c) Strains, due to CTE mismatch of Si-nanowire and nickel contact, calculated at center of the nanowire (white dashed line in (b)). (d) VB modulation for LH from strains in (c) showing tunnel barriers at x = ± 90 nm and a QD in between. Horizontal dashed line represents the Fermi level. (b)-(d) all have the same horizontal axis. Black vertical dashed line shows the location of the tunnel junctions in (b) to (d).

I simulated in COMSOL an undoped silicon nanowire with a 5 nm radius, sitting on top of a thick SiO_2 layer. Nickel contacts, 50 nm thick and separated by 200 nm, act as the source and drain. I assume that no Schottky barrier forms at the metal-nanowire interface, so holes can flow freely from the nickel into the silicon.

The calculated strains for a change in temperature of 293 K to 1 K are shown in Figure 5.8(c). The file used in the COMSOL simulation is located on Guestroom PC: in C:\Ted\Stress\Save\Si-NW final.mph. The strains in Figure 5.8(c) are for the center of the nanowire, but the strains elsewhere in the nanowire are similar. The largest source of strain in this device comes from the CTE mismatch of nickel ($\alpha_{Ni} = 13\text{x}10^{-6}$/K) and silicon ($\alpha_{Si} = 2.9\text{x}10^{-6}$/K). When the nickel contacts thermally contract, the nanowire is put into tension. Thus, the nanowire is in tension in ϵ_x and has compressive strains in ϵ_y and ϵ_z because of Poisson's ratio.

In the previous Chapter, I said that the topmost VB state in a nanowire is predominantly LH in character, with the spin of the hole aligned with the direction of the nanowire. Therefore, I will calculate the change in energy of the LH given the strains in Figure 5.8(c) using

$$\Delta E_{V,LH} = a_v(\epsilon_x + \epsilon_y + \epsilon_z) + b_v(\epsilon_x - (\epsilon_y + \epsilon_z)/2) \quad 5.5.$$

where a_v = 2.1 eV and b_v = -2.33 eV [64]. Figure 5.8(d) shows the calculated modulation of the VB due to strain using Eq. 5.5. Using Eq. 4.16, $\Delta E_{V,LH} \approx (3/2)a_v(\epsilon_y + \epsilon_z)$,we see that the change in the VB in Figure 5.8(d) is primarily due to the sum of the strains, $\epsilon_y + \epsilon_z$, in Figure 5.8(c).

Does the VB profile in Figure 5.8(d) have the right shape to form a QD for holes? We see two dips located at x = ± 90 nm, and a local maximum between the dips. The dips in the VB, located in between the metal contacts at x = ± 90 nm in Figure 5.8(d), form tunnel barriers for holes. The maximum in between can form a QD for holes. So this is the correct shape for a QD, but do the tunnel barriers have an appropriate tunneling resistance to measure a current through them? These tunnel barriers have a height of 5 meV and a length of 30 nm, which, using Eq. 5.4, gives an estimated tunneling resistance of 45 MΩ. This resistance is small enough to avoid shutting current off and large enough to quantize the charge on the QD. So the VB profile has both the correct shape and magnitude to form a QD for holes.

To sum up, I have shown that strain can create tunnel barriers at metal-semiconductor interfaces, which can explain why tunnel barriers are sometimes observed at nanowire-contact interfaces that should not form Schottky barriers [95].

E. Conclusions

I began this Chapter by considering the CB profile needed to explain the locations of dots A and B, as determined in Chapter 3. I was able to show that dot B is probably due to the combination of an electrostatic barrier below LGD and a strain-induced tunnel barrier at the end of the nanowire. Although the strain from the lower gates is large enough to create strain-induced tunnel barriers, the location of the peaks was different on different sides of the nanowire. So although the strain should have an effect on the transport, I cannot conclude that strain is the cause of dot A.

I was able to demonstrate strain-induced QDs in two other architectures. In a metal-gated bulk silicon device, I showed that strain due to CTE mismatch can induce a QD. This could explain the frequency with which unintentional QDs are observed in these devices. Replacing the metal gates with poly-silicon can eliminate the strain-induced QDs. I also showed that strain can create tunnel barriers for holes near the silicon-nanowire metal-contact interface in a chemically-grown nanowire.

There are advantages for strain-induced QDs as compared to electrostatically generated QDs. Making electrostatic QDs requires several gates, which sets the minimum size of the QD, whereas a strain-induced QD requires only a single gate. Because strain-induced QDs can be smaller, it should be easier to reach the few-electron limit, and they should have a larger size quantization energy.

There is a larger theme throughout this Chapter. We have seen that strain can dramatically alter the behavior of QD devices. This means that the strain can be as important to understanding a QD in silicon as the electrostatics. Therefore, the strain should be considered when either analyzing the results from a QD device or trying to design a new QD device.

Chapter 6: Conclusions and Future Work

A. Summary

In Chapter 1 I set myself the task of understanding the reproducibility of silicon QDs and why unintentional QDs are so common in silicon QD devices. In the subsequent chapter, I made a lot of progress to advance an understanding of these topics, but there is still much work to be done.

To advance understanding the reproducibility of silicon QDs, in Chapter 2 I discussed the reproducibility and predictability of the gate capacitances to intentional QDs. I showed that gate capacitances are reproducible to within 10%. I also showed that the gate capacitances scale with fabrication parameters and can be predicted by a capacitance simulator to within 20%.

To advance understanding why unintentional QDs are so common in silicon QD devices, in Chapter 3 I used the gate capacitances to determine the location of unintentional QDs with a precision of a few nanometers. Knowing the position of the QDs allowed me to compare the location to a calculation of the strain in Chapter 5. I was able to show that dot B is possible due to the combination of a strain-induced tunnel barrier at the end of the nanowire and an electrostatic barrier below LGD. I was unable to demonstrate that dot A was due to strain. But I was able to show strain induced QDs in two other device architectures: electrons in bulk silicon with metal gates and holes in a chemically-grown nanowire with metal contacts. This suggests that many of the

unintentional QDs that are observed, and attributed to interface traps or dopants, may actually be due to strain.

B. Future Work.

1. Barrier Capacitances

In Chapter 2 I discussed the challenges in simulating the barrier capacitances. Because the barrier capacitances are often larger than the gate capacitances, they can dominate the total capacitance of the QD, which is the capacitance that matters in determining the maximum operating temperature of a QD device. So the ability to predict barrier capacitances could help us to design QDs that operate at higher temperatures.

I also think there are more fundamental reasons to study the barrier capacitances. I speculated in Chapter 2 that the gate voltage dependence of the dielectric constant could be explained if the silicon tunnel barriers have not been fully depleted of charge carriers and therefore the dielectric constant of the silicon changes. If there are charge carriers in the silicon tunnel barriers, then it could have important implications for our devices. Having charge carriers in the tunnel barriers could affect the applications of the QDs. Could it increase the error rate of a charge pump or decrease the lifetime or coherence times of a qubit?

There could be other terms in the capacitance that we have been neglecting, such as quantum capacitance effects [96]. Capacitance is the change in charge as a voltage is changed. In a semiconductor nanostructure, it can take extra energy to add an electron, because of the density of states. This will change the voltage at which an electron is added to the QD, which will change the capacitance. This is called the quantum

capacitance, and it can be larger than the geometric capacitance in resonant tunneling diodes [97] and cooper pair boxes [98]. Because I calculate only the geometric capacitance and not the quantum capacitance, this might explain why the simulated and measured barrier capacitances are different.

2. Device Fabrication

I mentioned in Chapter 2 that we could compare the simulated and measured capacitances to determine if the devices are being fabricated as intended. If one dimension of the device is larger or smaller than intended, that should show up as a systematic difference between the simulated and measured gate capacitance. By doing this we could improve the fabrication of future devices.

3. Valley States in Nanowires

There is much work to be done to understand the valley physics of electrons in a nanowire in surface inversion. In Chapter 5 I suggested that, for a $[\bar{1}10]$ oriented nanowire, an electron on the top or bottom surfaces of the nanowire would likely be in the $\pm k_z$ valleys, and an electron on the side surfaces would likely be in the $\pm k_{[100]}$ or the $\pm k_{\{010]}$ valleys. But there is no evidence that this is correct. Also, assuming that multiple valley states are occupied, I do not know if the electron is in a mixed state or a superposition of the different valley states. Furthermore, I do not know if there is valley splitting, as there is in a 2DES (two dimensional electron system). Valley splitting is the splitting of a pair of valleys like the $+k_z$ and $-k_z$ valleys. Assuming there is a valley splitting, what is its magnitude?

As abstract as these questions seem, the answers are important for experiments we would like to do in these devices very soon. For example, we would like to do a charge or spin qubit experiment [99,100]. As Dimi Culcer has explained for a 2DES "in the presence of valley degeneracy a singlet-triplet qubit cannot be constructed, whereas for large valley splitting (>> k_BT) the experiment is similar to GaAs. [100]" It is clear that the valley physics of a nanowire is more complicated than that of a 2DES. Thus, I do not know the extent to which the valley physics affects our ability to do either a charge or spin qubit experiment.

What could be done to help us understand the valley physics of a nanowire in surface inversion? I do not expect these questions about valley physics in a nanowire to be analytically solved. Atomistic simulations have been done to understand the band structure and wavefunctions of electrons in smaller nanowires [101]. Because those simulations are atomistic, which are computationally expensive, our devices are too large to be simulated at present. But as computing power increases and our devices shrink, the length scales of what can be simulated and what we can make may soon meet.

4. Simulation Improvements

In Chapters 2 and 3, I used FASTCAP to simulate the electrostatics of a QD device, and in Chapter 5 I used COMSOL to simulate the strain in a QD device. I would like to understand how the strain affects the electrostatics. What I think is needed is a program that takes the strain-altered band structure, and then solves the electrostatic, semiconductor physics, and quantum mechanics simultaneously. If this program predicted a QD at the location of dot A, then we would at last know if dot A is a strain-induced QD.

5. Measurement of the Strain

I would like to see a measurement of the strain in the silicon below the gates. The measured strains could then be compared to the location of an unintentional QD to demonstrate that strain is the cause of a specific unintentional QD. We could also compare the measured strains to the simulated strain to help us improve our strain simulations. Either X-ray nanodiffraction or electron backscatter diffraction should be able to measure the strains [102]. One challenge is that the strain measurement would need to be performed at cryogenic temperatures. A second challenge is that the silicon nanowire, where the strain induced QDs are located, is below one or two poly-silicon layers. Therefore, when trying to measure the strain in the nanowire, the strain in the poly-silicon might be measured by accident.

6. Mitigation of Strain-Induced QDs

Understanding that strain is the cause of the unintentional QDs allowed me to suggest methods to reduce the strain and eliminate the unintentional QDs. In Chapter 5 I suggested that dot B in the poly-silicon gated nanowire could be eliminated by changing the oxidation conditions to reduce the strain or by splitting the upper gate into multiple gates. For the metal-gated bulk silicon device, I showed that the strain could be reduced by replacing the metal gates with heavily-doped poly-silicon gates.

A successful implementation of one of these mitigation techniques would be a nice confirmation of the results in this dissertation. More importantly, by reducing the frequency with which unintentional QDs are encountered, strain mitigation could help us to make a useful current standard from charge pumps or a useful quantum computer with either charge or spin qubits.

7. Strain-Induced QD architecture

Finally, I was able to show that strain-induced QDs have advantages over electrostatic QDs. But I was unable to devise a new architecture that uses the advantages of strain-induced QDs. In our current electrostatic architecture, the gates generate both the QDs and the tunnel barriers, as well as manipulate the electrons. Perhaps local stressors could generate the QDs and tunnel barriers leaving the gates to manipulate the charge on the QD. This type of design would give us more freedom in operating the gates.

Our devices look a lot like the finFETs that Intel is using in its current generation of transistors [17]. These transistors use strain engineering to increase the mobility of the silicon [74]. Repurposing the strain-engineering in the same finFETs to make QDs could let us design a new generation of devices. Because most of the techniques we use to fabricate our QD devices came from the semiconductor industry, it would be fitting if strain engineering became the next technique stolen.

Appendix A: The Measurement Circuit

A. The Old Circuit

1. Circuit Diagram

The electrical results presented in this dissertation were taken in an Oxford Kelvinox 100 dilution refrigerator. This appendix is to document the circuit that I used to obtain these results. The circuit diagram can be seen in Figure A.1. The circuit can be divided into two parts, room temperature electronics, and cryostat wiring.

2. The Room Temperature Electronics

SRS DS345 Function Generator: Our standard voltage source. Most of the DS345s are manually set to a DC voltage, and our LabVIEW programs sweep either one or two voltages to take the data.

Hewlett-Packard 3458A Digital Multi-Meter: Our standard voltage meter. We only have four of these, so we cannot monitor all of the voltages, even though the Figure is drawn as such.

DL Instruments Model 1211 Current Preamplifier: Our standard current preamplifier. We used the following standard setting

Sensitivity: 10^{-9} A/V

Rise Time: Min

Case ground defeated

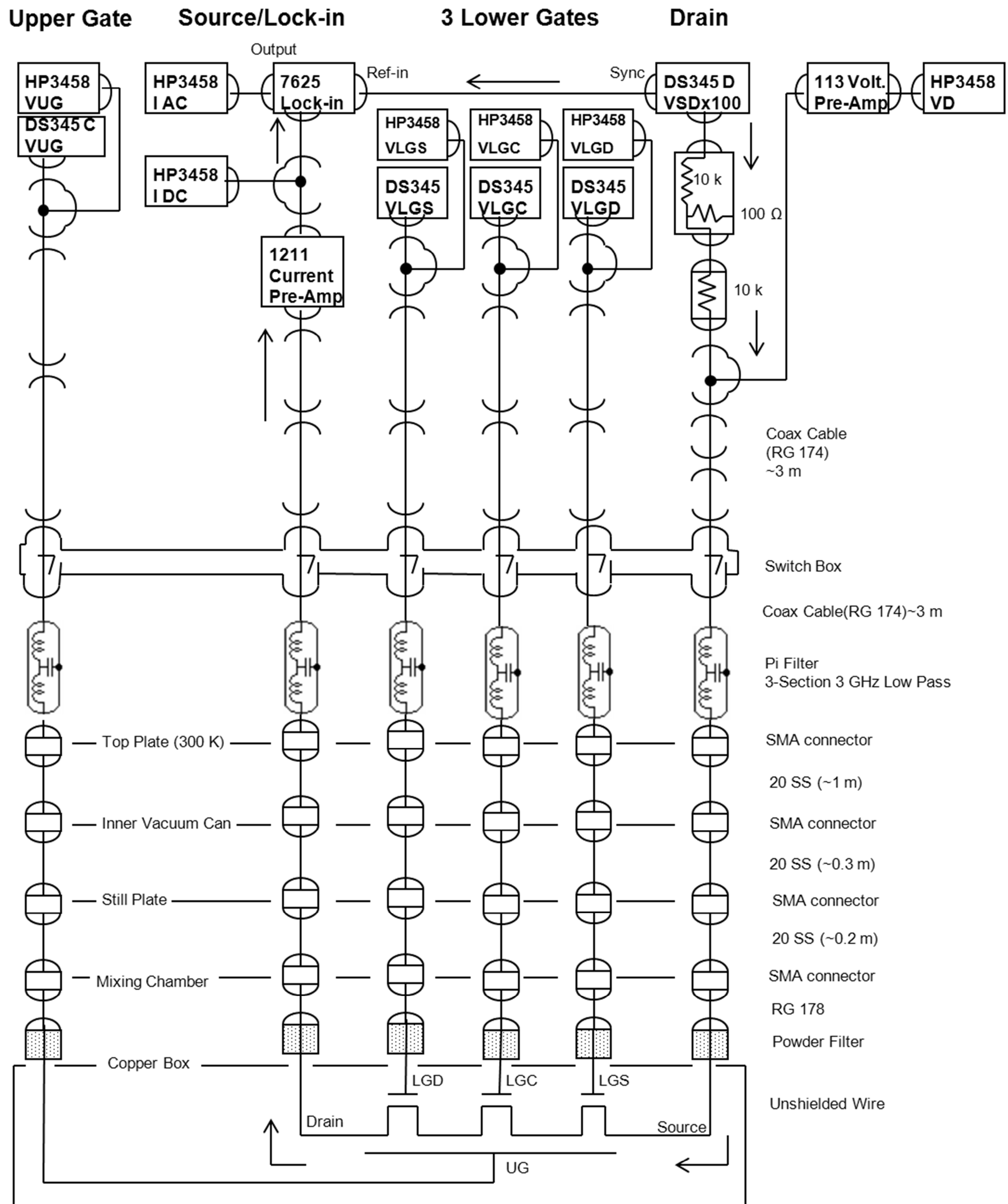


Figure A.1. Circuit diagram for the old circuit. Circuit recorded in SET-PC:C\Data\Measurement Diagnostics\Wiring10_1 circuit diagram.ppt. Not all of the voltmeters shown are used in any particular measurement.

EG&G 113 Voltage Preamplifier: We did not always use a voltmeter to record the drain voltage, but when we do we use this preamplifier to amplify the current by 100x. We used the following settings on the 113:

Single Ended

Gain = 100

LF Roll off = DC

HF Roll off = 10 kHz

Case Ground defeated

Perkin-Elmer 7265 DSP Lock-In Amplifier: Not all measurements were taken with a lock-in, so this was not always a part of the circuit. When it was we used the following settings:

Sine Wave Modulation

AC coupled

60 Hz Line Filter: ON

Anti-Aliasing Filter: ON

Output Filter: 12 dB/Oct LP

For each run we recorded: AC gain, Sensitivity, Phase, TC for output filter, Frequency of modulation

Voltage Divider: We use a 100:1 voltage divider to cut down the drain bias applied to the sample. We typically want a drain bias that is on the order of a millivolt and we might have an even smaller AC signal for the lock-in. These voltages are too small for the DS345. We fix this by put out a voltage that is 100x larger than we want, and then putting the signal through this voltage divider. The voltage divider consists of a 10 kΩ and a 100 Ω resistor in a resistive voltage divider geometry. Because our device resistance is always much larger than 100 Ω, we don't have to worry about the device resistance affecting the voltage divider. The additional 10 kΩ resistor after the voltage divider seems to reduce the noise.

3. Inside the Cryostat

Our cryostat has 14 voltage lines: 10 low frequency, 2 medium frequency and 2 high frequency. I primarily used the 10 low frequency lines. See appendix B of Emmanouel Hourdakis' dissertation for details about the cryostat wiring, including frequency dependence of the attenuation of the cables and filters as well as heat load from the cables. This dissertation can be found online at http://drum.lib.umd.edu/handle/1903/7619 or on Guestroom PC C:\manolis\manolis\thesis\committee corrections.pdf

B. The New Circuit

1. Motivation

Recently changes have been made to change the circuit to reduce the noise in the circuit. The new circuit is shown in Figure A.2. Several pieces of equipment have been

replaced. The AC and DC bias voltages have been separated. And ground loops were removed to reduce the 60 Hz noise.

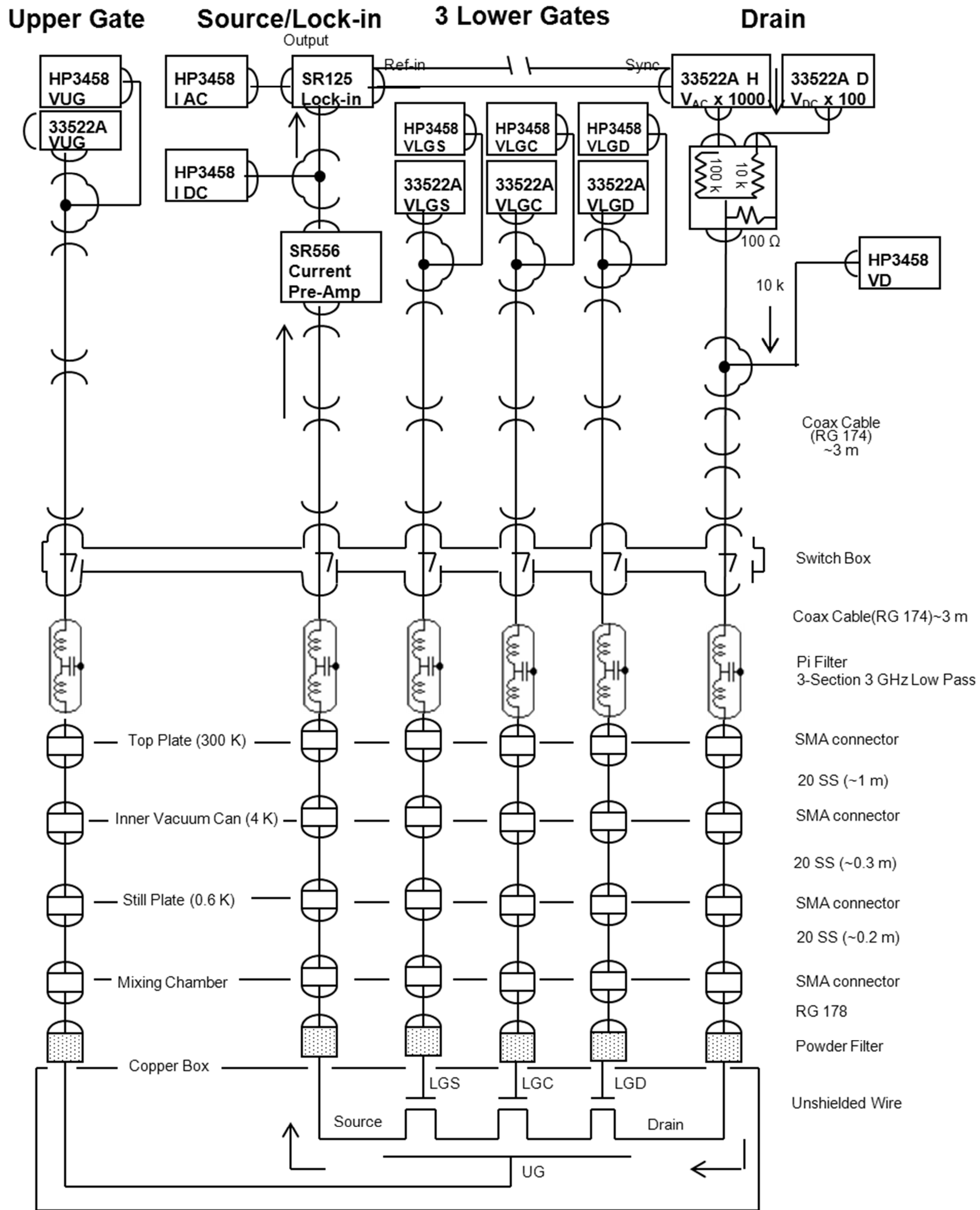


Figure A.2 The new circuit diagram showing new equipment.

2. The New Equipment

Agilent 33522A Voltage Source: The DS345 voltage sources have been replaced by Agilent 33522A. The biggest problem with the 345s was that signals would bleed through from the back-panel, either from the GPIB connection or the triggering. The new 33522As do not have this problem. Also the amplitude of voltage noise, as measured on an oscilloscope, is less than 10 μV.

SRS Model SR556 Current Preamplifier: The 1211 current preamp has been replaced by an SRS 556 because the 556 has a factor of three lower noise than the 1211.

SRS Model SR124 Lock-In Amplifier: This lock-in amplifier has similar performance to the Perkin-Elmer 7265 but the SR124 is analog.

Voltage Adder: Two voltage sources are now used for the drain voltage. One is used for the DC offset, the other provides the AC modulation for the lock-in amplifier. We made this change because it gave us a wider range of DC voltages that we could apply without limiting our ability to do a lock-in measurement.

Improvements to Reduce Noise: Improvements have been made to reduce the noise in the circuit. Several ground connections were removed from the insert, so the insert should now be floating. This reduced the amplitude of the 60 Hz noise in the current through a MOSFET from 2 nA to 10 pA. To remove ground connections from the insert: prevent the inner vacuum can pump line from touching the condenser line, unbolt the collar from the dewar, suspend the insert from rope, and prevent the transfer tube from touching the insert. Removing the EG&G 113 voltage preamplifier, which is used to amplify V_D, reduced the measured noise.

Also we observed that there were noise spikes 1 ms long with a 10 ms period that could be eliminated by 1) fully seating the 24-Pin Fischer connector and 2) not attaching the magnet ground wire to a broken power supply.

Appendix B: FASTCAP Tutorial

A. Motivation

I used FASTCAP [47], a part of the FASTFIELDSOLVERS package, to calculate the capacitance between the gates and the QDs as discussed in Chapters 2 and 3. The purpose of this section is to explain how FASTCAP works and how to use it. To explain how I performed these simulations, I will walk through a simple simulation. The device I will simulate is AF-CA2F3E-1, which was device 1 in Table 2.2.

FASTCAP 2.0 can be downloaded for free from fastfieldsolvers.com. Included in the FASTFIELDSOLVERS package are FASTMODEL and FASTCAP (as well as other programs which I don't use such as FASTHENRY). Buried in the package are also utilities such as cubegen.exe which I will use.

B. How FASTCAP Works

To calculate capacitances, FASTCAP numerically solves the integral form of Poisson's equations. This discussion is based on reference [47]. Let us say that we want to calculate the capacitances between m conductors. The surface of each conductor is divided into panels. There are a total of n panels. The potential and charge on conductor i are V_i and Q_i. The potential and charge on panel k are v_k and q_k. The charge on a conductor i can be found by summing up the charges on the panels of that conductor,

$$Q_i = \sum_{k \in i} q_k. \quad \text{B.1.}$$

For panel k, centered at x_k, the potential due to the charges on all of the other panels is

$$v_k = \frac{1}{4\pi\varepsilon}\sum_{l=1}^{n}\int_{panel,l}\frac{q_l/a_l}{|x' - x_k|}da', \qquad \text{B.2.}$$

where ε is the dielectric constant and a_l is the surface area of panel l. We have assumed that the surface charge on the panel is uniform, so we can pull q_l/a_l out of the integral. (This assumption is a potential source of error, so check your simulation by varying the number of panels to see if it affects the output.) Rewriting Eq. B2 as

$$v_k = \sum_{l=1}^{n}\left(\frac{1}{4\pi\varepsilon}\frac{1}{a_l}\int_{panel,l}\frac{1}{|x' - x_k|}da'\right)q_l. \qquad \text{B.3.}$$

We can define the quantity in the parenthesis, which has units of inverse capacitance, as p_{kl}.

$$p_{kl} = \frac{1}{4\pi\varepsilon}\frac{1}{a_l}\int_{panel,l}\frac{1}{|x' - x_k|}da' \qquad \text{B.4.}$$

p_{kl} is one element of an n by n matrix that I will call $\boldsymbol{p}$. If $\vec{v}$ is an n by 1 vector of the potential of each panel, and $\vec{q}$ is a n by 1 vector of the charge on each panel, these three can be related as

$$\vec{v} = \boldsymbol{p}\vec{q}. \qquad \text{B.5.}$$

If we invert the matrix $\boldsymbol{p}$, we see that it is the capacitance matrix for the panels.

$$\boldsymbol{c}\vec{v} = \vec{q} \qquad \text{B.6.}$$

where $\boldsymbol{c} = \boldsymbol{p}^{-1}$. The capacitance between two conductors, i and j, is the sum of the capacitances between all of the panels on the two conductors.

$$C_{ij} = \sum_{k \in i} \sum_{l \in j} c_{kl} \quad \text{B.7.}$$

This is how FASTCAP calculates capacitances. Note that FASTCAP assumes that all conductors are ideal and that all dielectrics are lossless.

C. Model and Meshing the Geometry

The basic steps of the simulation are first to model and mesh the geometry in an input file in FASTMODEL, and second to run the simulation in FASTCAP and finally interpret the results.

1. Two Level Hierarchy

The first step to doing a FASTCAP simulation is to define the model and the mesh in the input files in FASTMODEL. These files, which becomes the input for the FASTCAP simulation, both defines the geometry and defines the mesh. In order to change the mesh, different input files must be used.

The input files for FASTCAP have a two level hierarchy. The top level, which is edited in FASTMODEL, is a list file (.lst) The list file calls subfiles that are identified as .qui. Each line in the subfile is one quadrilateral panel, which is one element in the mesh of an object.

2. Batch File

In Chapter 2, I did FASTCAP simulations of many similar devices, so to speed up the process I wrote a batch file to generate the input files (both the list file and the .qui subfiles) for each device. For instructional purposes I have reproduced that batch file here. Running this batch file will generate all of the files that I will use in this tutorial.

Simply copy this text into notepad, save the file as maker.bat and put it in the same folder as cubegen.exe. I will describe cubegen.exe in the subfile section of this appendix, but it can be found in the utilities subfolder within the FASTCAP subfolder of the FastFieldSolver installation (Program Files\FastFieldSolvers\FastCap2\Utilities\cubegen.exe). Then run the batch file. I also recommend moving them to a subfolder of their own because maker.bat will generate 15 files. This program will generate two list files, one corresponding to the full island the other corresponding to the half island, as well as thirteen .qui files.

I will not go into how maker.bat works, because it is not essential for this tutorial. Playing around with the variables defined at the beginning of this file will change the dimensions of the device simulated. Note that for the program to run, the dimensions, except t_{si}, must be even.

3. List File

I will first describe the list file, which can be edited within FASTMODEL. The easiest way to follow this section is to open up FASTMODEL. Opening FASTMODEL will open FASTMODEL, FASTCAP and FASTHENRY. You can close FASTHENRY because we will never use it. For now you can minimize FASTCAP, because we will not use it for a little while. In FASTMODEL load CA2-3E1-full.lst, which was generated by the batch file in the previous section. The following is the text of CA2-3E1-full.lst.

```
*tsi=17
*wsi=30
*LLG=10
*LUG=40
*tox1=20
*tox2=30
```

```
*wire
C CA2-3E1-wire-left.qui 1 -185.E-9 -15.E-9 0 +
C CA2-3E1-wire-ox.qui 1 -145.E-9 -15.E-9 0 +
C CA2-3E1-wire-ox.qui 1 -130.E-9 -15.E-9 0 +
C CA2-3E1-wire-LLG.qui 1 -115.E-9 -15.E-9 0 +
C CA2-3E1-wire-ox.qui 1 -105.E-9 -15.E-9 0
C CA2-3E1-wire-ox.qui 1 -90.E-9 -15.E-9 0 +
C CA2-3E1-wire-LUG.qui 1 -75.E-9 -15.E-9 0 +
C CA2-3E1-wire-ox.qui 1 -35.E-9 -15.E-9 0 +
C CA2-3E1-wire-ox.qui 1 -20.E-9 -15.E-9 0 +
C CA2-3E1-wire-LLG.qui 1 -5.E-9 -15.E-9 0 +
C CA2-3E1-wire-ox.qui 1 5.E-9 -15.E-9 0 +
C CA2-3E1-wire-ox.qui 1 20.E-9 -15.E-9 0 +
C CA2-3E1-wire-LUG.qui 1 35.E-9 -15.E-9 0 +
C CA2-3E1-wire-ox.qui 1 75.E-9 -15.E-9 0
C CA2-3E1-wire-ox.qui 1 90.E-9 -15.E-9 0 +
C CA2-3E1-wire-LLG.qui 1 105.E-9 -15.E-9 0 +
C CA2-3E1-wire-ox.qui 1 115.E-9 -15.E-9 0 +
C CA2-3E1-wire-ox.qui 1 130.E-9 -15.E-9 0 +
C CA2-3E1-wire-right.qui 1 145.E-9 -15.E-9 0
*LGS
C CA2-3E1-LG-1.qui 1 -115.E-9 -65.E-9 0 +
C CA2-3E1-LG-2.qui 1 -115.E-9 -65.E-9 37.E-9 +
C CA2-3E1-LG-3.qui 1 -115.E-9 -35.E-9 37.E-9 +
C CA2-3E1-LG-4.qui 1 -115.E-9 35.E-9 37.E-9 +
C CA2-3E1-LG-1.qui 1 -115.E-9 35.E-9 0
*LGC
C CA2-3E1-LG-1.qui 1 -5.E-9 -65.E-9 0 +
C CA2-3E1-LG-2.qui 1 -5.E-9 -65.E-9 37.E-9 +
C CA2-3E1-LG-3.qui 1 -5.E-9 -35.E-9 37.E-9 +
C CA2-3E1-LG-4.qui 1 -5.E-9 35.E-9 37.E-9 +
C CA2-3E1-LG-1.qui 1 -5.E-9 35.E-9 0
*LGD
C CA2-3E1-LG-1.qui 1 105.E-9 -65.E-9 0 +
C CA2-3E1-LG-2.qui 1 105.E-9 -65.E-9 37.E-9 +
C CA2-3E1-LG-3.qui 1 105.E-9 -35.E-9 37.E-9 +
C CA2-3E1-LG-4.qui 1 105.E-9 35.E-9 37.E-9 +
C CA2-3E1-LG-1.qui 1 105.E-9 35.E-9 0
*UGLL
C CA2-3E1-UG-1.qui 1 -185.E-9 -65.E-9 0 +
C CA2-3E1-UG-2.qui 1 -185.E-9 -65.E-9 37.E-9 +
C CA2-3E1-UG-3.qui 1 -185.E-9 -35.E-9 37.E-9 +
C CA2-3E1-UG-4.qui 1 -185.E-9 35.E-9 37.E-9 +
C CA2-3E1-UG-1.qui 1 -185.E-9 35.E-9 0 +
*UGL
C CA2-3E1-UG-1.qui 1 -75.E-9 -65.E-9 0 +
```

```
C CA2-3E1-UG-2.qui 1 -75.E-9 -65.E-9 37.E-9 +
C CA2-3E1-UG-3.qui 1 -75.E-9 -35.E-9 37.E-9 +
C CA2-3E1-UG-4.qui 1 -75.E-9 35.E-9 37.E-9 +
C CA2-3E1-UG-1.qui 1 -75.E-9 35.E-9 0 +
*UGR
C CA2-3E1-UG-1.qui 1 35.E-9 -65.E-9 0 +
C CA2-3E1-UG-2.qui 1 35.E-9 -65.E-9 37.E-9 +
C CA2-3E1-UG-3.qui 1 35.E-9 -35.E-9 37.E-9 +
C CA2-3E1-UG-4.qui 1 35.E-9 35.E-9 37.E-9 +
C CA2-3E1-UG-1.qui 1 35.E-9 35.E-9 0 +
*UGRR
C CA2-3E1-UG-1.qui 1 145.E-9 -65.E-9 0 +
C CA2-3E1-UG-2.qui 1 145.E-9 -65.E-9 37.E-9 +
C CA2-3E1-UG-3.qui 1 145.E-9 -35.E-9 37.E-9 +
C CA2-3E1-UG-4.qui 1 145.E-9 35.E-9 37.E-9 +
C CA2-3E1-UG-1.qui 1 145.E-9 35.E-9 0 +
```

First, note that any file beginning with * is a comment, and does not alter the output. I begin this file with several lines that specify the dimensions of the device being simulated. Now let's take a look at a typical line from this file, the first non-commented line.

```
C CA2-3E1-wire-left.qui 1 -185.E-9 -15.E-9 0 +
```

The first letter "C" tells FASTCAP this line is a conductor. (Because SiO_2 is the only dielectric in the simulation, I only define the conductors in the list file.) "CA2-3E1-wire-left.qui" is a subfile that is called by the list file. The "1" sets the relative dielectric constant (I will change the global dielectric constant later). The next three numbers specify the x, y and z coordinates of the object defined in the subfile in units of meters. (This coordinate refers to the origin within the subfile.) The final character in the line is the "+" symbol, which means that this line and the following line are part of the same

conductor. So, for example, the five lines following the *LGS line are all a single conductor.

I have a few helpful notes. The conductors must not overlap. The .qui files and the list file must be in the same folder. Individual .qui subfiles can be called multiple times.

4. Subfiles

The .qui subfiles consist of only list of individual panels (and comment lines) which each line having the format…

Q <cond.name> <x1> <y1> <z1> <x2> <y2> <z2> <x3> <y3> <z3> <x4> <y4> <z4>

The "Q" specifies the panel is a quadrilateral. The <cond.name> is the name of the conductor, which is for me always named "1" because I find it much easier to never have multiple conductors in the same subfile. Four pairs of Cartesian coordinates follow to specify the four corners of the quadrilateral. Again the units are assumed to be meters. So the code for a square panel in the xy plane with 1 nm sides is centered at the origin is

Q 1 -0.5e-9 -0.5e-9 0.0 -0.5e-9 0.5e-9 0.0 0.5e-9 0.5e-9 0.0 0.5e-9 -0.5e-9 0.0

Note that I entered 0 as 0.0, because all numbers in FASTCAP need a decimal point.

Each conductor consists of hundreds or thousands of panels, so defining each panel by hand would be horrifying. Fortunately, FASTCAP includes utilities to write the .qui subfiles for you. To use these utilities, you must open up a DOS command prompt and go the utilities subfolder within the FASTCAP folder (Program Files\FastFieldSolvers\FastCap2\Utilities\cubegen.exe). To create the mesh I only use the

cubegen.exe program, which generates a box, and then I combine the boxes together, in the list file, to create gates and dots.

To create a box with 20 nm by 20 nm by 10 nm without a top or bottom with cubegen.exe, type into DOS command prompt

cubegen –xh20.E-8 –yh20.E-8 –zh10.E-8 –n10 –t –b >box.qui

I will break this line down. "cubegen" calls the utility. –xh20.E-8 specifies the dimension of the cube in the x direction (20 nm). –n10 specifies the size of the mesh, (the shortest side of the cube has 10 divisions). –t and –b remove the +z and –z sides of the cube respectively (-pbr removes the –x side, -pfl removes the +x side, -pfr removes the +y side, and –pbl removes the – y side). >box.qui specifies the file name as box.qui. Keep an eye on the size of the output file, if a subfile is more than a few hundred kb, FASTCAP will take a long time to run. If that is the case, change the –n10 to something smaller like –n5 to change the size of the mesh.

5. Looking at the Model

When the list file is opened in FASTMODEL, a separate window should show the geometry being simulated. In Figure B.1 I show the geometry as shown in FASTMODEL for CA2-3E1-full.lst.

I found that screening was very effective in minimizing the effect of conductors beyond about 50 nm from the quantum dots. Therefore I did not have to include the handle wafer or connect the different pieces of the upper gate.

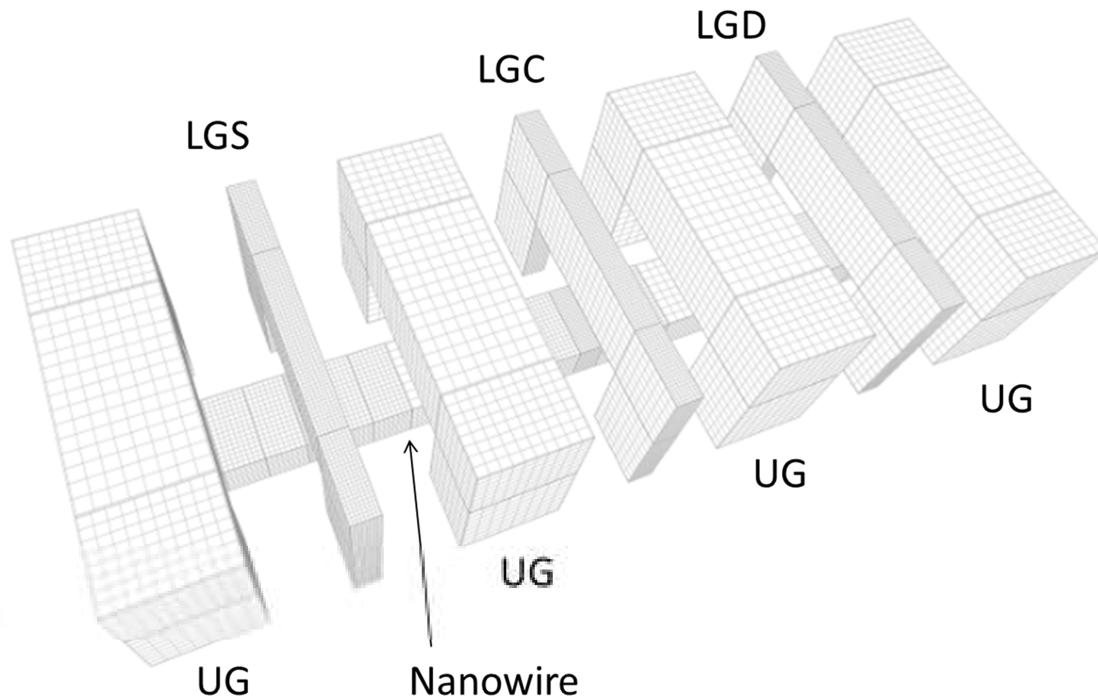


Figure B.1 The geometry as shown in Fast Model for CA2-3E1-full.lst

D. Running the Simulation

Once the input files have been created, go to FASTCAP2 and under file select open. A dialog box will open. Select browse, and find the file CA2-3E1-full.lst. There are several options regarding the simulation. First, define the global relative dielectric constant, which is 3.9 for SiO_2, in the "Global Permittivity Constant" section. The other two parameters that I use in this dialog box are the "Order of the Multipole Expansion" and "Iteration Tolerance". The higher the "Order of the Multipole Expansion" the better the simulation will be, but the longer it will take. Likewise, the "Iteration tolerance," which by default is set to 1% of the total capacitance, can both make the simulation more accurate and take longer. I typically will start with the default values of 2 and 0.01 and do

a quick simulation. I will later increase order the expansion and decrease the tolerance to do slower simulations. I stop when changing these parameters does not change the results, to the accuracy that I care about.

Now press run. With the default settings, this simulation takes less than a minute on my computer. The output is a capacitance matrix…

CAPACITANCE MATRIX, every unit is 1e-021 farads

1 2 3 4 5 6 7

1%GROUP1 1 2.992e+004 -1.248e+004 -21.17 -3840 -91.85 -46.81 -
1.19e+004

1%GROUP2 2 -1.248e+004 6.357e+004 -1.248e+004 -2661 -6389 -2654 -
2.485e+004

1%GROUP3 3 -21.17 -1.248e+004 2.991e+004 -47.9 -93.46 -3841 -
1.189e+004

1%GROUP4 4 -3840 -2661 -47.9 3.422e+004 -466.3 -178.1 -2.37e+004

1%GROUP5 5 -91.85 -6389 -93.46 -466.3 3.431e+004 -479.2 -2.373e+004

1%GROUP6 6 -46.81 -2654 -3841 -178.1 -479.2 3.423e+004 -2.371e+004

1%GROUP7 7 -1.19e+004 -2.485e+004 -1.189e+004 -2.37e+004 -2.373e+004 -
2.371e+004 1.531e+005

This output does not look very pretty, so I always copy this to a spreadsheet, as shown in Table B.1.

Table B.1 Capacitance matrix in units of zF (one thousandths of an aF)

	1	2	3	4	5	6	7
1	2.99E+04	-1.25E+04	-21.17	-3840	-91.85	-46.81	-1.19E+04
2	-1.25E+04	6.36E+04	-1.25E+04	-2661	-6389	-2654	-2.49E+04
3	-21.17	-1.25E+04	2.99E+04	-47.9	-93.46	-3841	-1.19E+04
4	-3840	-2661	-47.9	3.42E+04	-466.3	-178.1	-2.37E+04
5	-91.85	-6389	-93.46	-466.3	3.43E+04	-479.2	-2.37E+04
6	-46.81	-2654	-3841	-178.1	-479.2	3.42E+04	-2.37E+04
7	-1.19E+04	-2.49E+04	-1.19E+04	-2.37E+04	-2.37E+04	-2.37E+04	1.53E+05

The capacitance matrix is a symmetric matrix which contains both the self and mutual capacitances between a set of conductors. The off diagonal elements, the mutual capacitances, are the negative of the capacitances between any two conductors. The diagonals are the total capacitance to any node (including both the mutual and self-capacitances). The sum of any row or column is the self-capacitance, which should be much smaller than the mutual capacitances. If this is not true (some conductor has a large self-capacitance), then many electric field lines are going off to infinity. Typically, a large self-capacitance means that more of the geometry needs to be simulated to capture

those electric field lines. For a full discussion of the capacitance matrix see references [18,21]

The conductors are listed in the order they are defined in the .lst file. I used:

1. The part of the wire to the left of the QD (can be ignored)
2. The quantum dot
3. The part of the wire to the right of the QD (can be ignored)
4. LGS
5. LGC
6. LGD
7. UG

Therefore the capacitance to the QD are from the simulation are in Table B.2

Table B.2 Capacitance to the full quantum dot in aF with default parameters (2nd order expansion and 1% tolerance).

	LGS	LGC	LGD	UG
Full-QD	2.661	6.389	2.654	24.9

The self-capacitance for the quantum dot is only 3.2% of the total capacitances, which is pretty good (I got this taking the sum of column 2 and dividing by the diagonal element of that column). You can see a small difference between the capacitance from LGS and LGD, which should the same by symmetry.

Now we will increase the order of the expansion to 4 and the iteration tolerance to 0.001. I show the results in Table B.3

Table B.3 Capacitance to the full quantum dot in aF with refined parameters (4th order expansion and 0.1% tolerance).

	LGS	LGC	LGD	UG
Full-QD	2.6487	6.3789	2.6488	24.838

Notice that the discrepancy between the capacitances to LGS and LGD has decreased, but the capacitances have not changed very much, which means this is a good solution. Therefore these are the solutions I used in Table 2.2.

E. Finished

Congratulations, you have finished your first FASTCAP simulation. To begin to do your own FASTCAP simulations, I suggest to first play around with the device parameters at the beginning of maker.bat. Try changing the size of the mesh in maker.bat to see the effect on the capacitances. Then I suggest making your own input file using by hand using cubegen.exe.

Appendix C: COMSOL Multiphysics Tutorial

A. Preview

In Chapter 5 I used COMSOL Multiphysics to simulate the strains in a QD device due to both CTE mismatch and intrinsic strain. As a tutorial in COMSOL, in this Chapter I show how to simulate the bimetallic strip. Because the bimetallic strip has been solved analytically, I will compare the analytical solution to the simulation results.

B. What COMSOL is Doing – One Dimension

COMSOL is a finite element simulator. The basic idea of a finite element analysis is that an object can be broken up into discrete elements, and the strain and stress can be determined from the displacement of the elements [103]. In Figure C.1(a) I draw a thin rod, immobile at one end and with tension T applied to the other end. I show a few of these elements in Figure C.1(b). Each element is connected to its neighbors by a spring.

We can rewrite the equations that govern stress and strain in terms of the forces in the springs and the displacement of the elements. Element i, which has an initial position x_i, will undergoes deformation u_i.

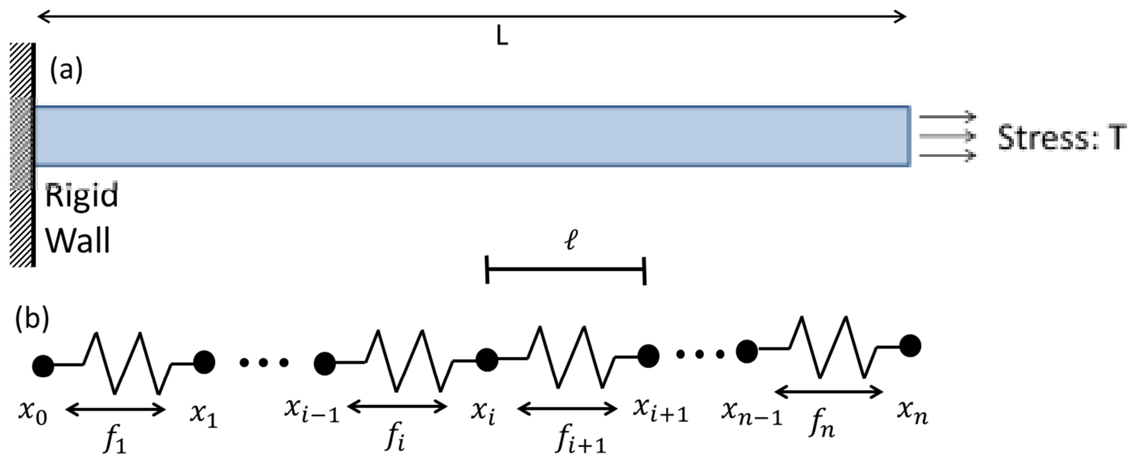


Figure C.1 (a) Rod under tension. (b) Finite elements connected by springs.

The spring constant, k, comes from the Young modulus of the rod Y, and the cross sectional area of the rod A, and rod is divided into segments of length ℓ,

$$k = \frac{YA}{\ell} \tag{C.1.}$$

The strain can be obtained from the displacement of the elements

$$\epsilon_i = \frac{u_{i+1} - u_i}{\ell} \tag{C.2.}$$

We can rewrite the equations from Chapter 4 in terms of these elements. First, Hooke's law (eq. 4.1) can be written as

$$f_i = k(u_i - u_{i-1}) \tag{C.3.}$$

Second, the equilibrium equations, which say the forces on an element in equilibrium (Eq. 4.2) must be zero, can be rewritten as

$$f_i - f_{i-1} = 0 \tag{C.4.}$$

We also know the boundary equations for this problem

$$u_0 = 0$$

$$\text{and } f_n = TA \quad \text{C.5.}$$

where T is the tension applied to one end of the rod and A is the cross sectional area of the rod. This problem is trivial to solve, but in general solving these simultaneous equations requires the use of a computer program.

$$f_i = TA$$

$$u_i = \frac{f_i}{k} i \quad \text{C.6.}$$

$$\epsilon_i = \frac{u_i}{\ell} = \frac{f_i i}{k\ell} = \frac{TAiL}{YA\ell} = \frac{T}{Y}$$

These results for the strain of a rod under tension are what we intuitively expect. This was a trivial example, but it outlines the basic steps that any finite element simulator must apply to solve a stress-strain problem.

C. What COMSOL is Doing – Two Dimensions

Now I will show how COMSOL solves a 2D problem [103]. I assume the plane stress condition, in which out of plane stresses are set to zero, i.e. $\sigma_z = \sigma_{xz} = \sigma_{yz} = 0$. I will go through the case of a constant strain triangle. For a good reference on the finite element method in 2D see reference [103].

At vertices of the triangle are three nodes, labeled 1, 2 and 3. A node, i, has a position (x_i, y_i), undergoes a displacement ($u_{x,i}$, $u_{y,i}$), and experiences forces ($F_{x,i}$, $F_{y,i}$).

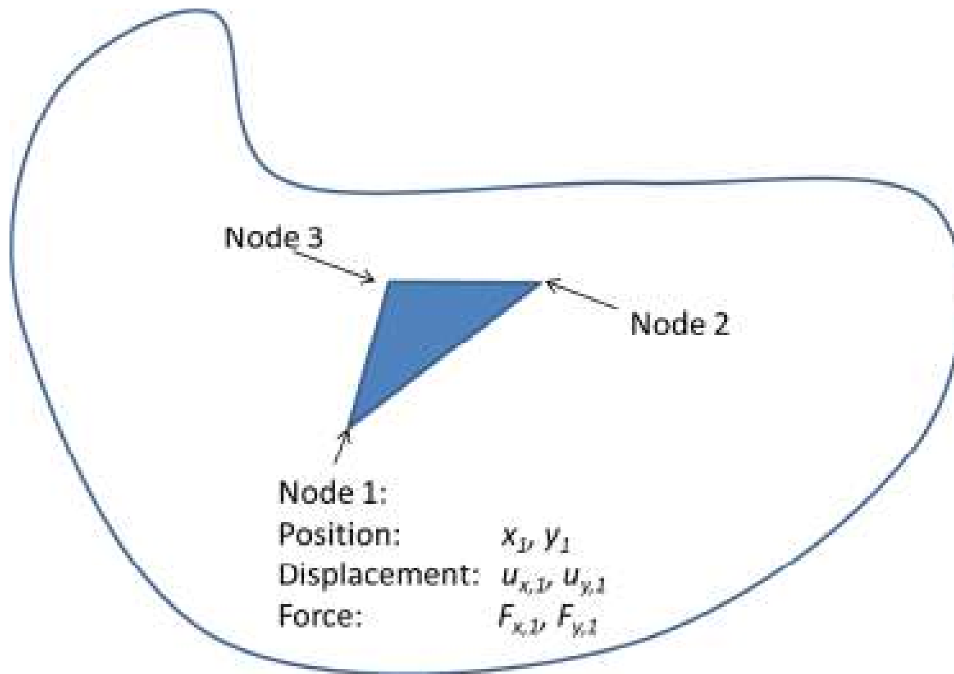


Figure C.2 A triangular mesh element with nodes 1, 2 and 3.

First, I will go through the equations that COMSOL is solving.

1) In the plane stress case we can relate the stress and strain in the triangle with

$$\vec{\epsilon} = \boldsymbol{D}^{-1}\vec{\sigma}$$

$$\vec{\epsilon} = \begin{bmatrix} \epsilon_x \\ \epsilon_y \\ \epsilon_{xy} \end{bmatrix} \text{ and } \vec{\sigma} = \begin{bmatrix} \sigma_x \\ \sigma_y \\ \sigma_{xy} \end{bmatrix} \tag{C.7.}$$

$$\boldsymbol{D}^{-1} = \frac{Y}{1-\upsilon^2}\begin{bmatrix} 1 & \upsilon & 0 \\ \upsilon & 1 & 0 \\ 0 & 0 & 1-\upsilon \end{bmatrix}$$

2) For the constant strain triangle, the strain in each triangular mesh element is constant and is a function of the displacements of the nodes of the triangle,

$$\begin{bmatrix} \epsilon_x \\ \epsilon_y \\ \epsilon_{xy} \end{bmatrix} = \boldsymbol{B}\vec{d}$$

$$\boldsymbol{B} = \frac{1}{2A}\begin{bmatrix} y_2 - y_3 & 0 & y_3 - y_1 & 0 & y_1 - y_2 & 0 \\ 0 & x_3 - x_2 & 0 & x_1 - x_3 & 0 & x_2 - x_1 \\ x_3 - x_2 & y_2 - y_3 & x_1 - x_3 & y_3 - y_1 & x_2 - x_1 & y_1 - y_2 \end{bmatrix} \quad \text{C.8.}$$

$$\vec{d} = \begin{bmatrix} u_{x,i} \\ u_{y,i} \\ u_{x,j} \\ u_{y,j} \\ u_{x,m} \\ u_{x,m} \end{bmatrix}$$

where A is the area of the triangular element ($2A = x_1(y_2 - y_3) + x_2(y_3 - y_1) + x_3(x_1 - x_2)$). The matrix $\boldsymbol{B}$ is a function of only the positions of the nodes.

3) The boundary stresses from a specific element e, can expressed the nodal equivalent forces

$$\overrightarrow{f^e} = \boldsymbol{k^e}\vec{d}$$
$$\boldsymbol{k^e} = tA\boldsymbol{B^T}\boldsymbol{D^{-1}}\boldsymbol{B}\vec{d} \quad \text{C.9.}$$

The first two elements of the nodal force vector can be written as

$$f_{x,1} = t\frac{y_2 - y_1}{2}\sigma_x + t\frac{y_1 - y_3}{2}\sigma_x + t\frac{x_3 - x_1}{2}\sigma_{xy} + t\frac{x_1 - x_2}{2}\sigma_{xy}$$
$$f_{y,i} = t\frac{x_3 - x_1}{2}\sigma_y + t\frac{x_1 - x_2}{2}\sigma_y + t\frac{y_2 - y_1}{2}\sigma_{xy} + t\frac{y_1 - y_3}{2}\sigma_{xy} \quad \text{C.10.}$$

In Figure C.3, we see the set of boundary stresses that are being are concentrated as the element nodal force vector.

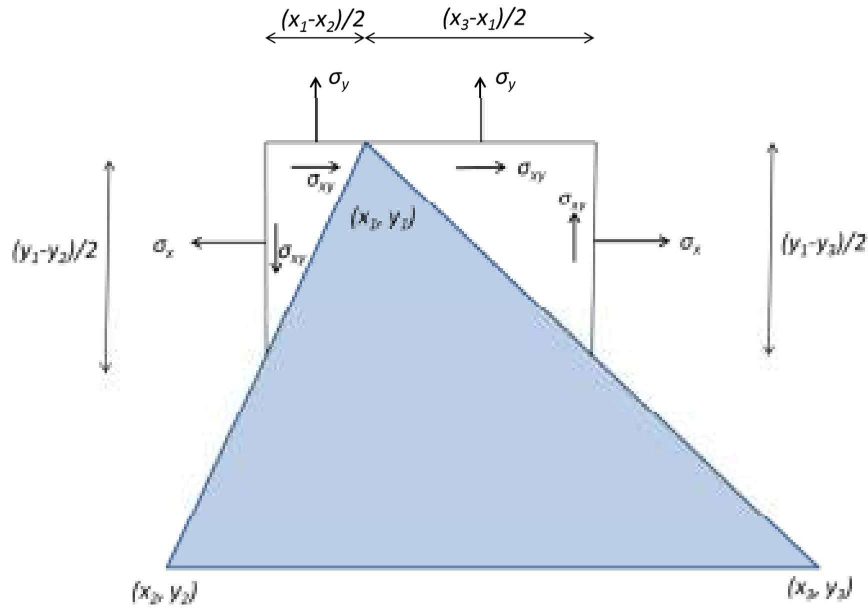


Figure C.3 Element nodal force

4) The nodal force vector is a the sum the elemental nodal force vectors over all of the elements adjacent to a node

$$F_{x,i} = \sum_{\substack{elements \\ e}} f_{x,i}^{e} \qquad \text{C.11.}$$

The nodal displacement vector contains all of the nodal displacements

5) The nodal force vector and the nodal displacement are related by

$$\begin{bmatrix} F_{x,1} \\ F_{y,1} \\ \cdot \\ \cdot \\ \cdot \\ F_{x,N} \\ F_{y,N} \end{bmatrix} = \boldsymbol{K} \begin{bmatrix} u_{x,1} \\ u_{y,1} \\ \cdot \\ \cdot \\ \cdot \\ u_{x,N} \\ u_{y,N} \end{bmatrix} \qquad \text{C.12.}$$

where N is the total number of nodes and

$$K_{xi,xi} = \sum_{\substack{elements \\ e}} k^e_{xi,xi} \quad \text{C.13.}$$

Now I will describe the algorithm that COMSOL is using to compute stress and strain.

1) COMSOL meshes the object. Each element of the mesh must be a single material

2) External forces and displacement constraints are specified. At the surface of the objet the either the force applied to a node or the displacement of a node is specified.

$$F_{x,i} = -F_{x,i,ext}$$
$$F_{y,i} = -F_{y,i,ext}. \quad \text{C.14.}$$

3) COMSOL calculates $\boldsymbol{k^e}$ for every element in the mesh, and then sums them up, as shown in Eq. C.13 to calculate $\boldsymbol{K}$.

4) COMSOL solves Eq. C.12, which is a set of 2N equations with 2N unknowns.

5) COMSOL uses Eq. C.8 to calculate the strains, and Eq. C.7 to calculate the stress.

D. The Bimetallic Strip

The bimetallic strip (or thermostat) is a classic elasticity problem; it was first solved by S. Timoshenko in 1925 [104]. The bimetallic strip consists of two materials, with different CTEs, that are glued together. Both materials experience the same change in temperature (ΔT) [Fig. C.3(a)], thus causing the bimetallic strip to bend [Fig. C.3(b)]. I will use Si and SiO_2 as the two materials; this combination was previously studied in

Ref. [105]. The Si has a thickness, d_{Si}, CTE, α_{Si} ,and Young's modulus, Y_{Si}. The SiO_2 has the same variables: d_{Ox}, α_{Ox}, and Y_{Ox}. It is assumed that the length of the bimetallic strip is much greater than the thickness.

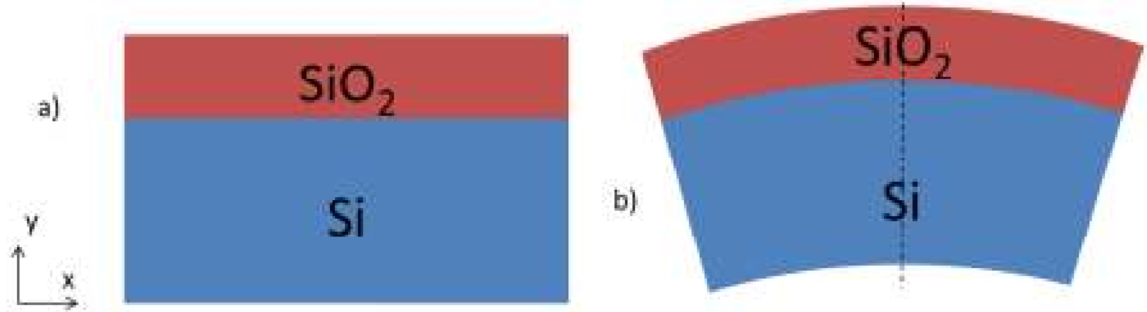


Figure C.4 (a) The bimetallic strip at the initial temperature. (b) The bimetallic strip at the final temperature. The stress will be calculated along the black dashed line.

I will not go through the derivation of the analytical solution to the bimetallic strip because it is a lot of algebra, but I recommend the derivation in references [104–106]. Upon deformation the radius of curvature, R, for the bimetallic strip is

$$\frac{1}{R} = \frac{6(1+m)^2}{\left[3(1+m)^2 + (1+mn)\left(m^2 + \frac{1}{mn}\right)\right]} \frac{(\alpha_{Ox} - \alpha_{Si})\Delta T}{h}$$

$$\text{where } h = d_{Si} + d_{ox} \qquad \text{C.15.}$$

$$m = d_{Si}/d_{ox}$$

$$n = E_{Si}/E_{ox}$$

The stresses along the black dashed line in Figure C.3(b) are given by

$$\sigma_y = \sigma_{xy} = 0$$

$$\sigma_{x,Si}(y) = \frac{1}{R}\left[E_{Si}\left(y + \frac{d_{Si}}{2}\right) + \frac{E_{Si}d_{Si}^3 + E_{ox}d_{ox}^3}{6d_{Si}(d_{Si} + d_{ox})}\right] \quad \text{C.16.}$$

$$\sigma_{x,ox}(y) = \frac{1}{R}\left[E_{Ox}\left(y - \frac{d_{ox}}{2}\right) - \frac{E_{Si}d_{Si}^3 + E_{Ox}d_{ox}^3}{6d_{ox}(d_{Si} + d_{ox})}\right]$$

where y = 0 is set to the Si-SiO_2 interface.

Table C.1 Properties in the following simulation.

Silicon Properties			Silicon Dioxide Properties	
d_{Si}	50		d_{Ox}	10 nm
Y_{Si}	130 GPa		Y_{Ox}	73 GPa
α_{Si}	$2.6\text{x}10^{-6}$ / K		α_{Ox}	$0.5\text{x}10^{-6}$ / K
R = 0.16 mm				

E. COMSOL Simulation

Start by opening up COMSOL Multiphysics (Fig. C.5). I am using COMSOL 4.1.

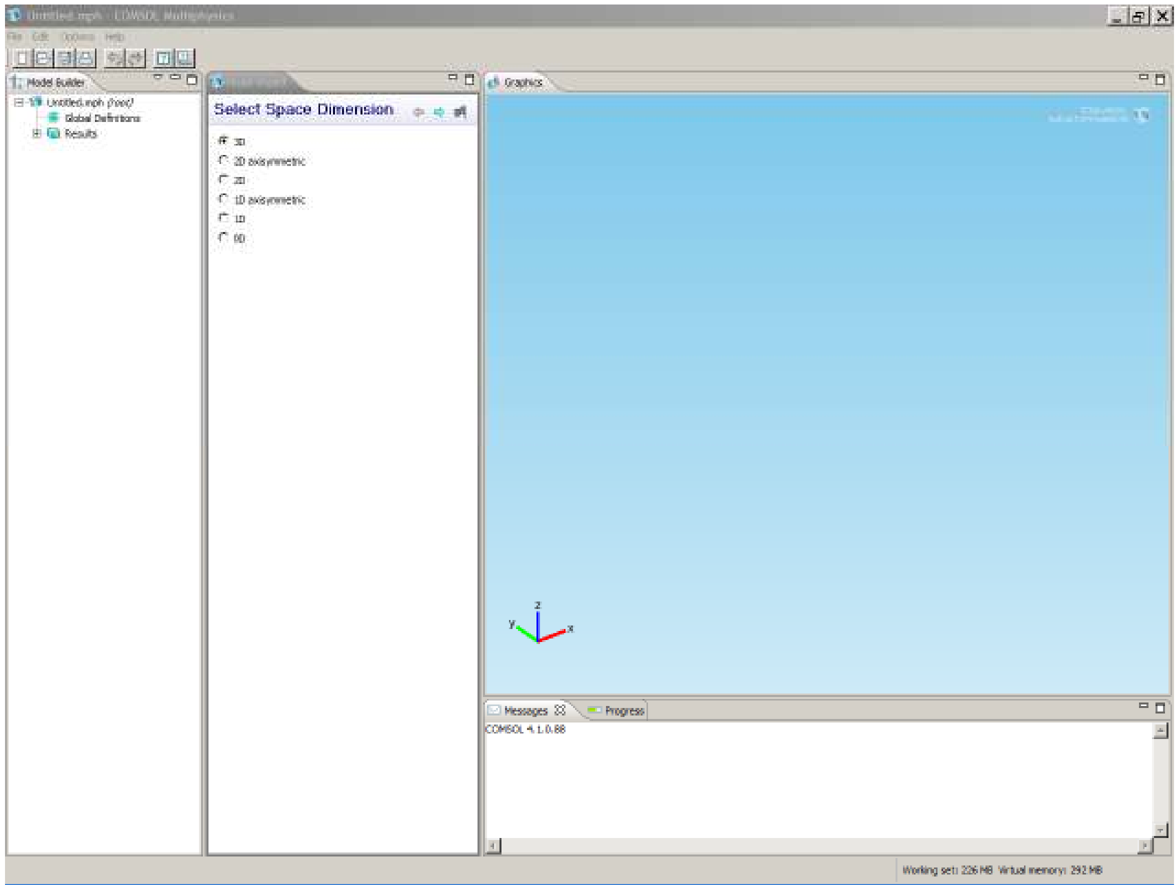


Figure C.5 The COMSOL startup screen.

The model wizard should have started up. Select “2D” and press the blue arrow.

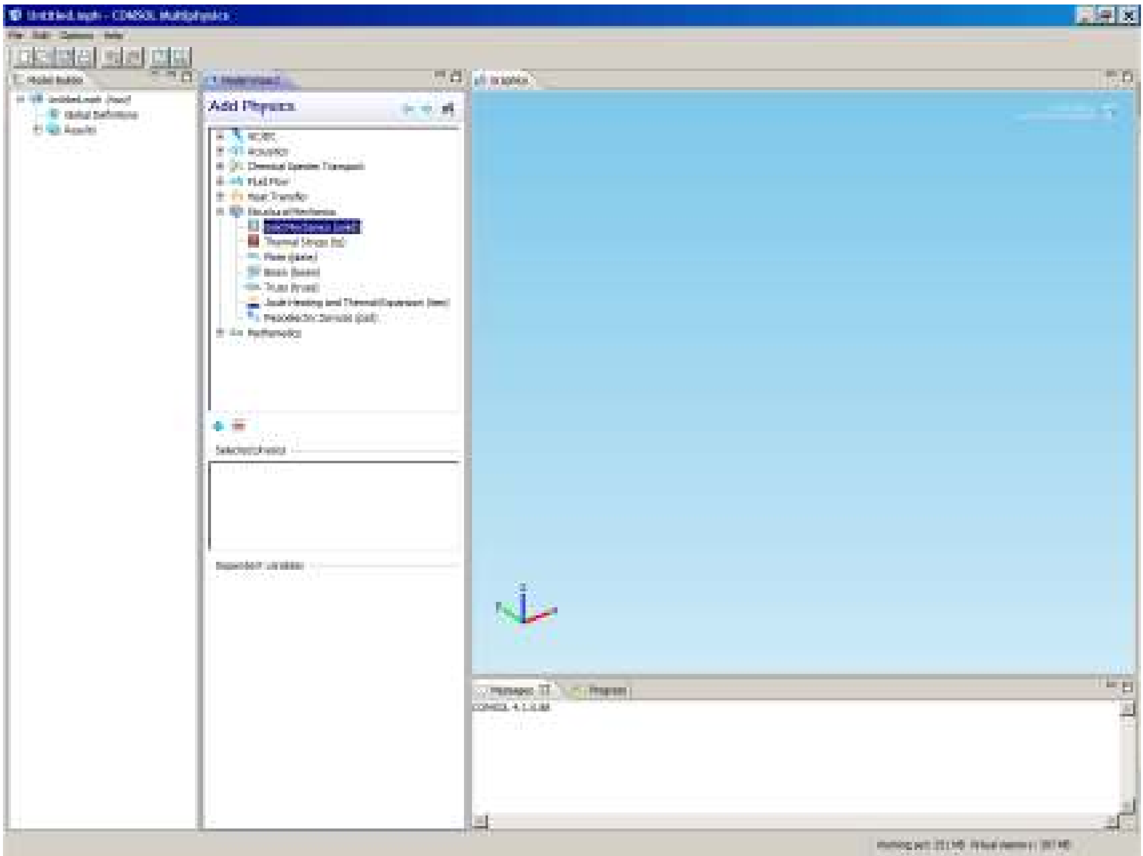


Figure C.6 Add Physics: Solid Mechanics

Next we need to add some physics to the model. We want to solve a solid mechanics problem, so under "Structural Mechanics" select "Solid Mechanics." Press the blue "+" sign to add the physics to the model (Fig. C.6). Now press the blue right arrow again.

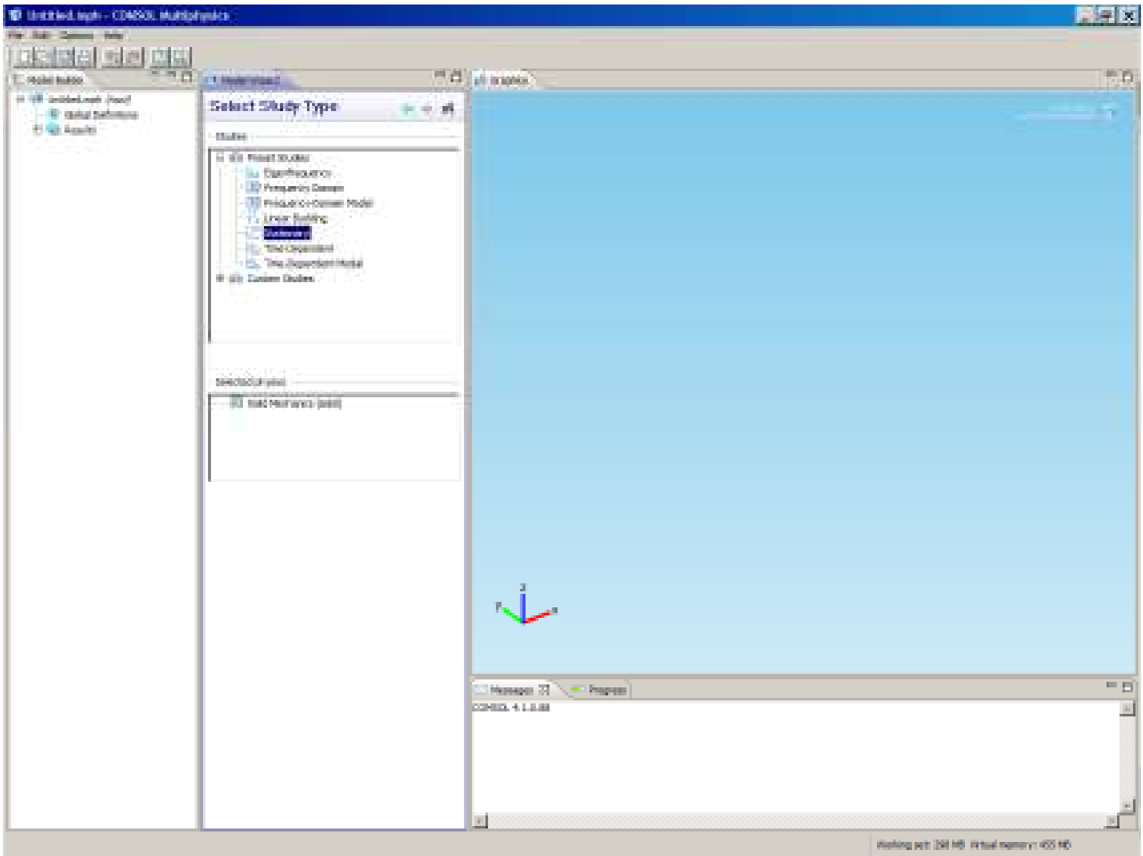


Figure C.7. Stationary Study

Now select “Stationary” and then press the checkered flag (Fig. C.7). We want to do a stationary simulation because we do not need any time dependence.

Next, we want to put in some variables, so under “Model Builder” right-click on “Global Definitions” and select “Parameters” (Fig. C.8).

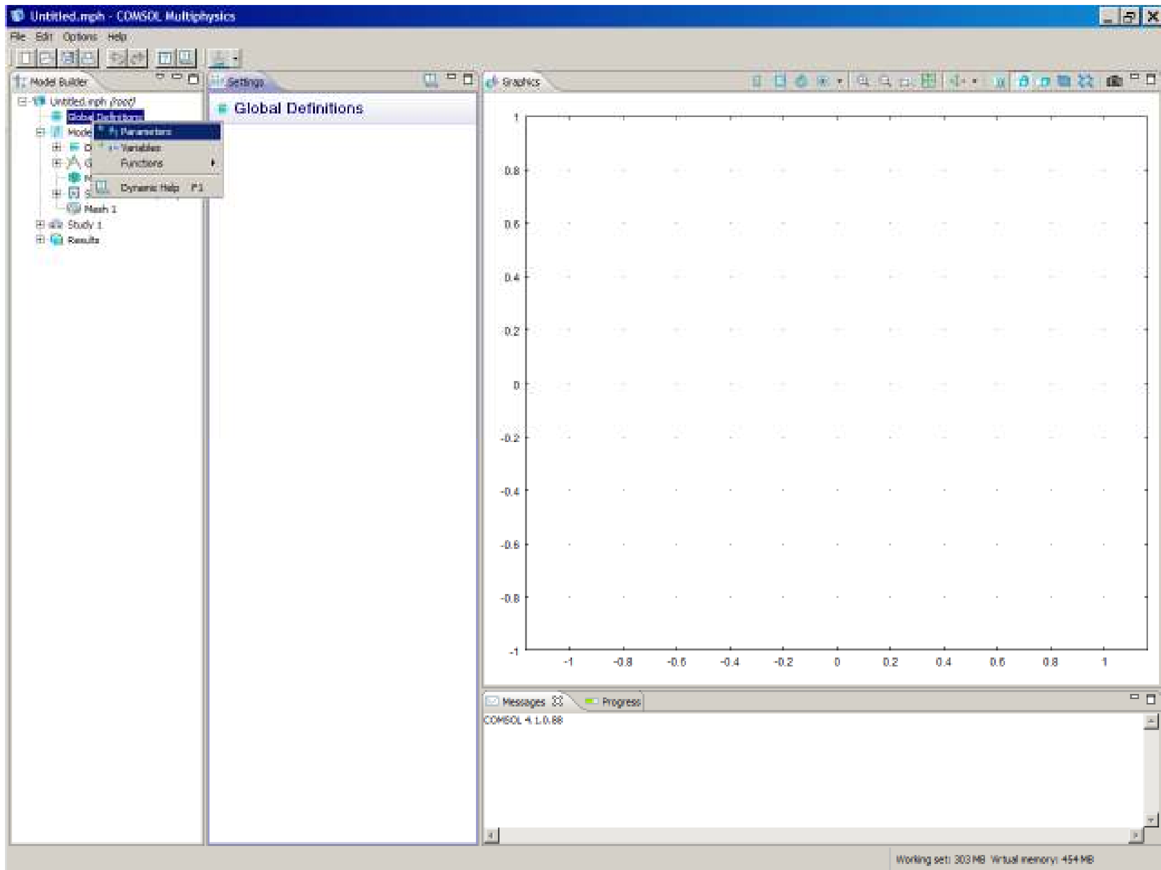


Figure C.8 Add parameters

Now we want to define d_{Si} and d_{ox} as variables. Under “Name” enter “dsi”, and under “Expression” enter ”50[nm]” (Fig. C.9). Next add d_{Ox} = 10 nm and L = 100 nm.

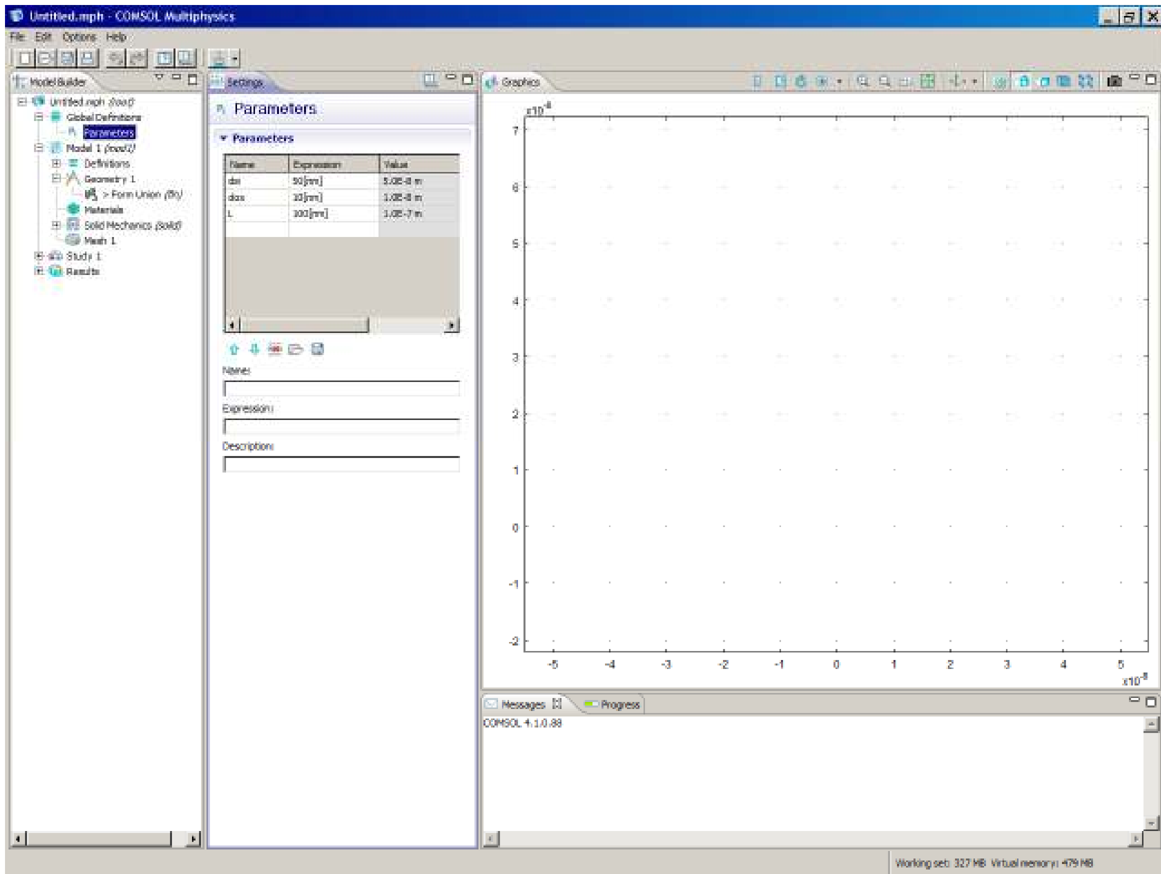


Figure C.9 Finished adding parameters.

Next, we want to make the geometry shown in Figure C.4 (a). We start by making a rectangle that will correspond to the oxide, so right-click on "Geometry 1" and select "Rectangle." In the rectangle settings box change under "Size", change "Width" to "L" and "Height" to "dox". Next, under the "Position" change the "x:" setting to "-L/2". Finally, click the button that looks like a building to make this rectangle, this is the "Build-All" button (Fig. C.10).

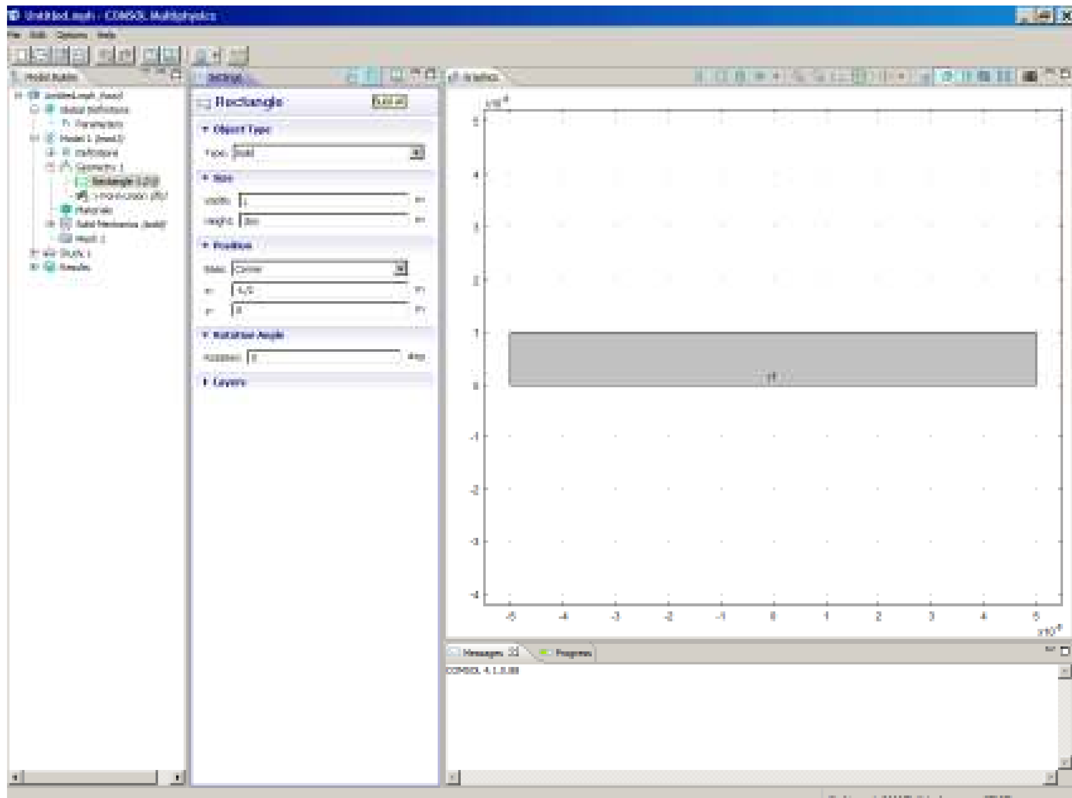


Figure C.10 Making the oxide rectangle.

Next, we make the silicon rectangle. Again, under “Geometry” select “Rectangle.” Give this rectangle a width of L and height of d_{Si}. Under “Position”, type “-L/2” for “x:” and “–dsi” for “y:”, and press the “Build-All” button to make the second rectangle (Fig. C.11).

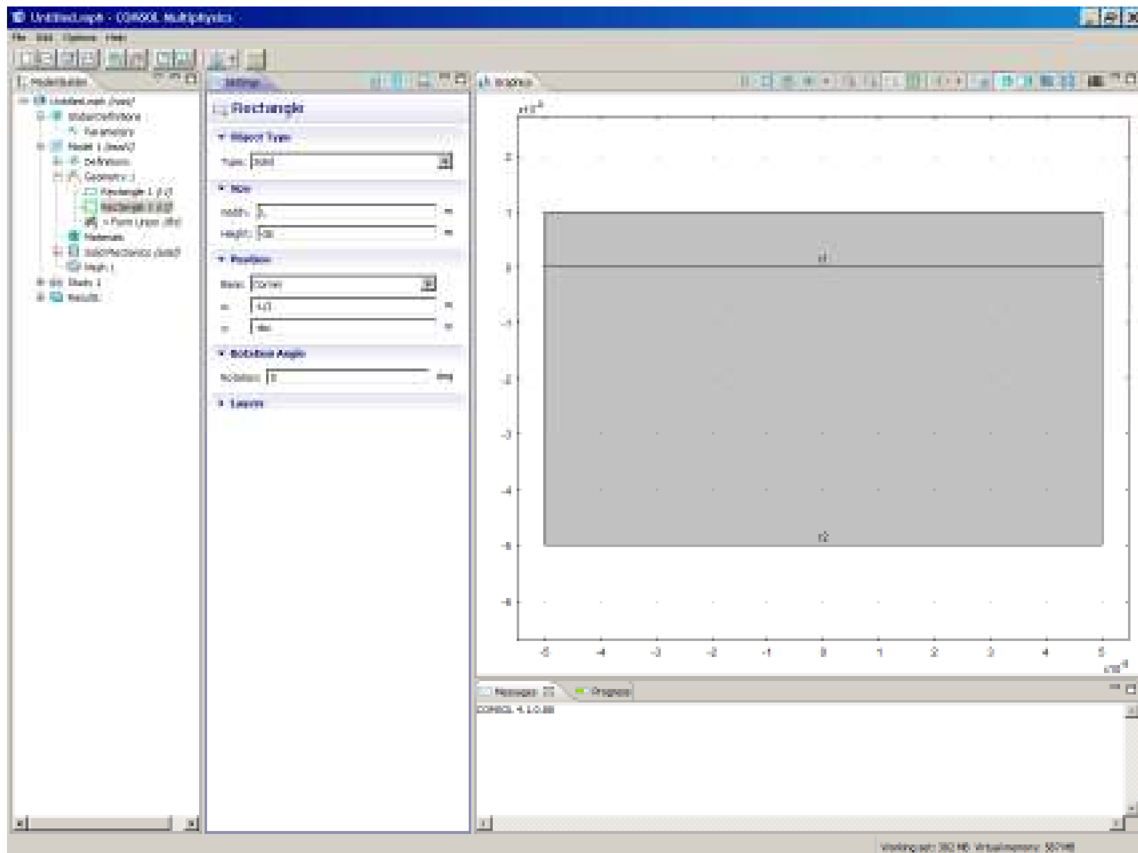


Figure C.11 Adding the second rectangle.

We have the two rectangles, but COMSOL does not yet know what material the two rectangles are made of. In the “Model Builder” right-click on materials and select “Open Material Browser.” Under the “Built-in” materials list, select “Silicon” and press the blue “+” sign to add the material to the model. Now click on silicon and change the coefficient of thermal expansion to “2.6E-6[1/K]". Now click on the “Material Browser” again, and under “MEMS” and “Insulators” find SiO_2 and add it to the model (Fig. C.12).

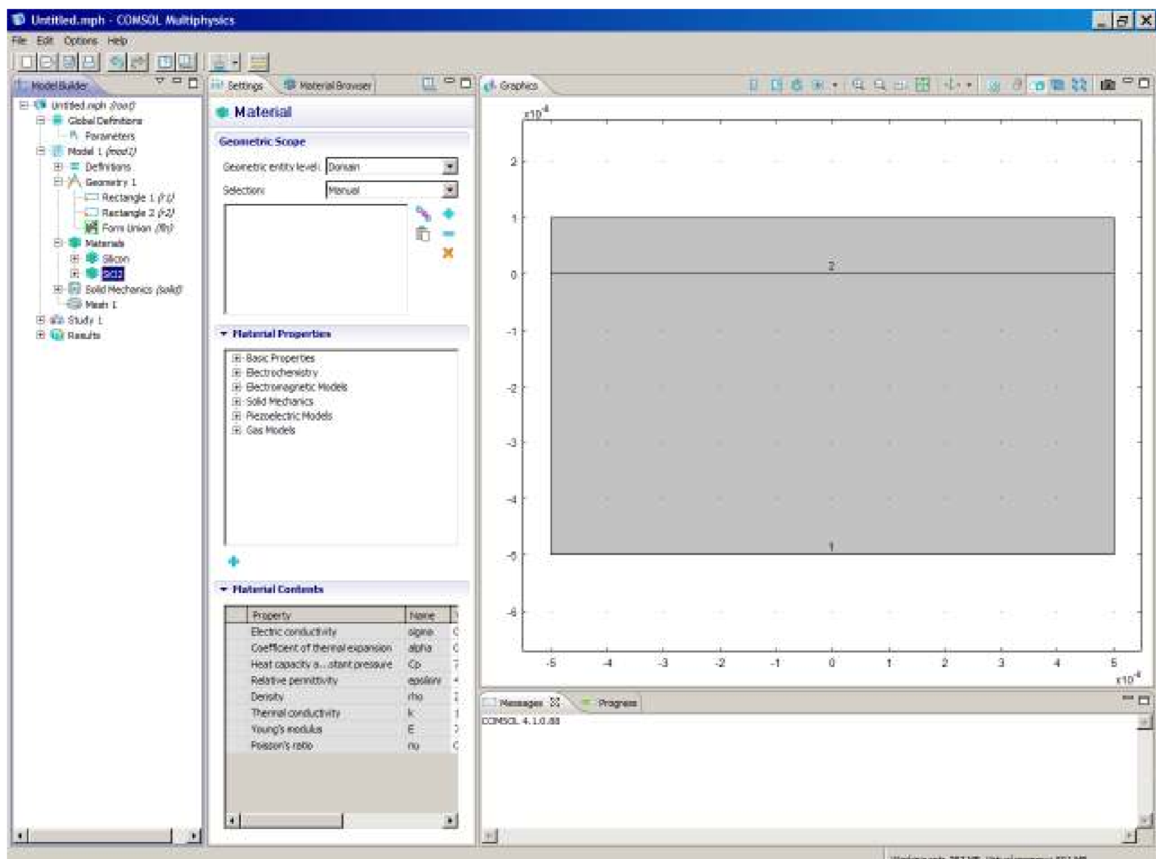


Figure C.12 Adding Si and SiO_2 to the model.

Now we have added the materials properties for Si and SiO_2 to the model. But we added Si first, so COMSOL assumed that both of the rectangles are made out of Si. We need to tell COMSOL that the upper rectangle is made of SiO_2. To do this, left-click on "SIO2" under the "Model Builder." In the graphics window, first left-click on the upper rectangle to highlight it, which will turn the rectangle pink. Then right-click the upper rectangle to add it to the selection list under SiO2, which will turn the rectangle blue (Fig. C.13).

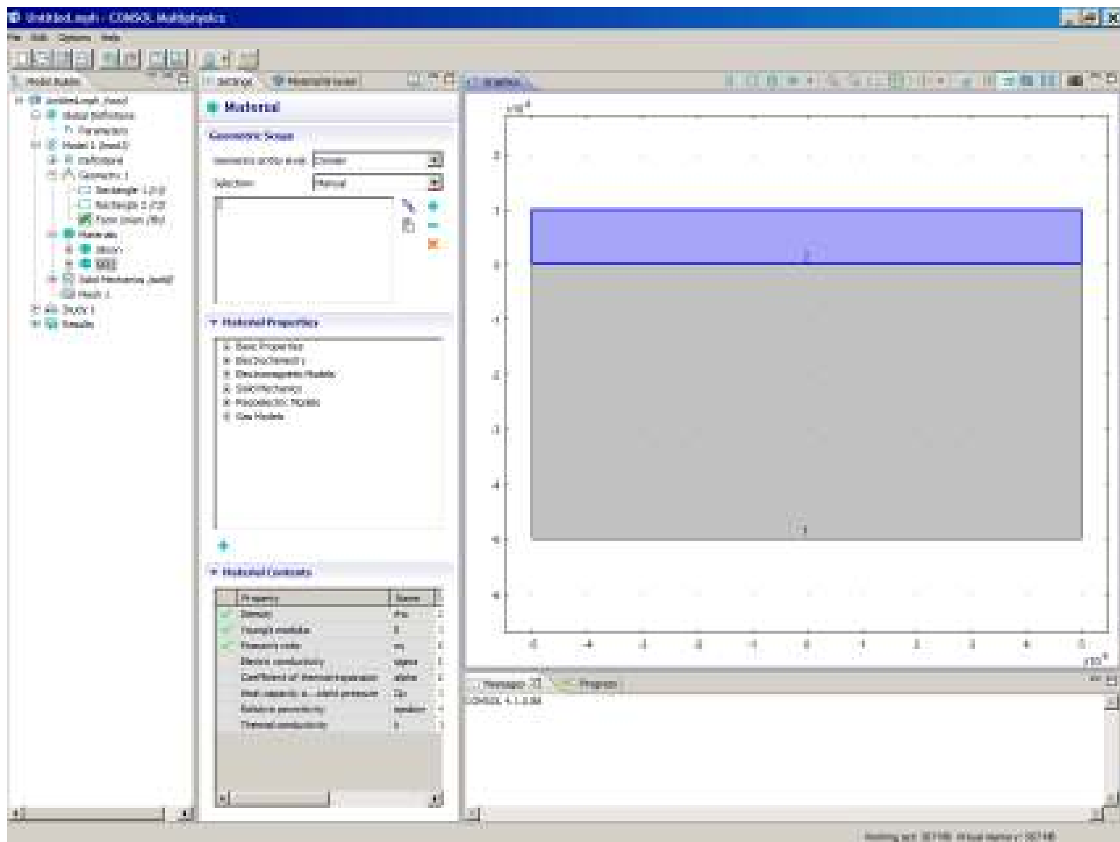


Figure C.13. Making the upper rectangle SIO_2.

COMSOL now knows the geometry and materials we want, but it does not yet know what we want to simulate. Under "Model Builder", expand the list below "Solid Mechanics." Right-click on "Linear Elastic Material Model 1" and select "Thermal Expansion" (Fig. C.14).

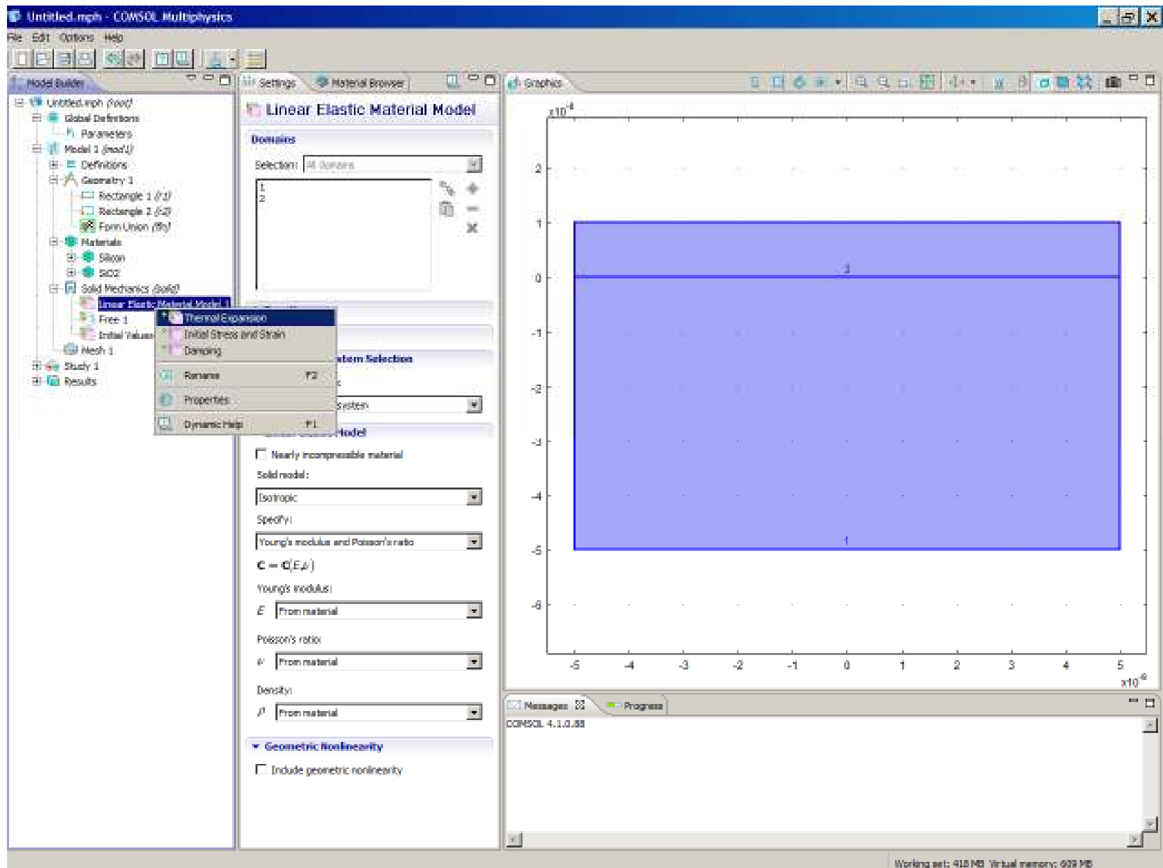


Figure C.14 Adding thermal expansion.

By default, when we add the thermal expansion, COMSOL will apply the thermal expansion to the entire geometry (Fig. C.15). We need to change the temperature "T" to "1[K]" (I think of this as T_{final}). By default the strain reference temperature is 293 K (this is what I think of as $T_{initial}$).

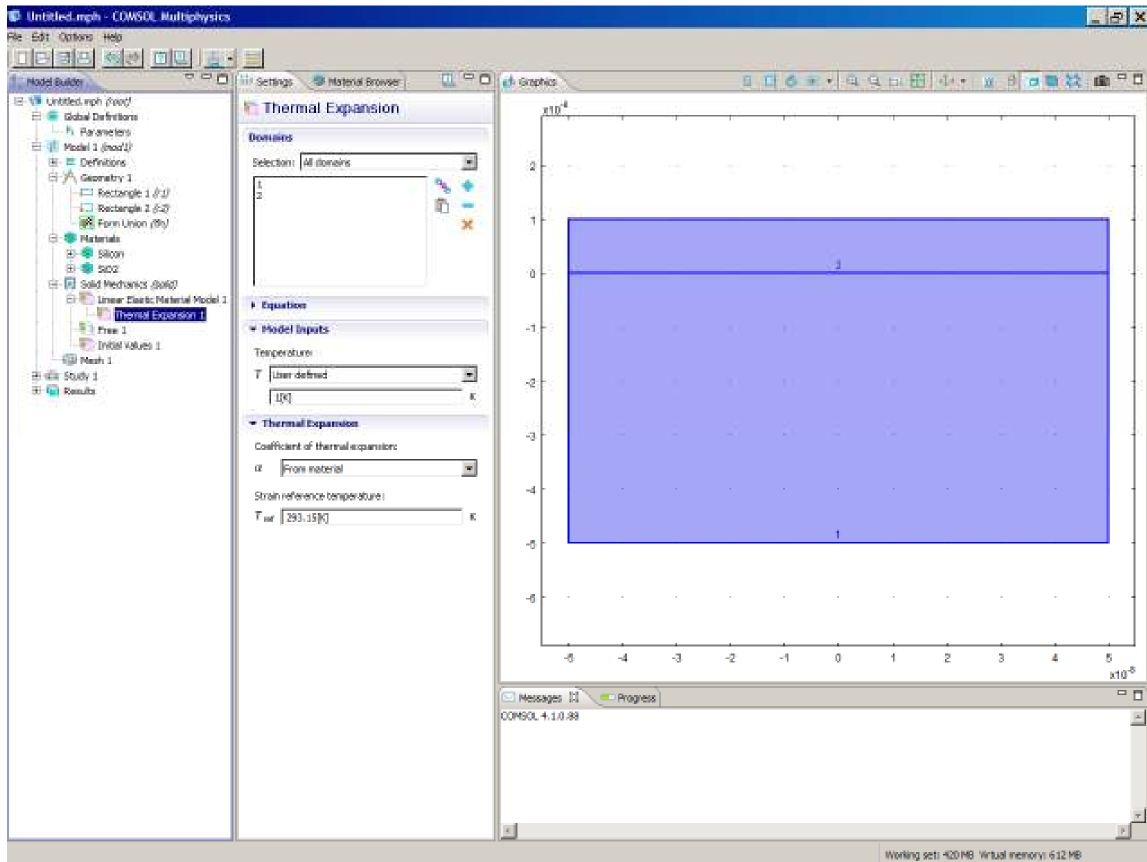


Figure C.15 Setting the change in temperature from 293 K to 1 K.

Now COMSOL knows that we want to simulate a change in temperature from $T_{initial}$ to 293 K to T_{final} = 1 K. We also need to tell COMAOL that the object is stationary and not rotating. Right-click on "Solid Mechanics" and under "Points" select "Fixed Constraint" (Fig. C.16). We want to tell COMSOL to not displace the lower left corner of the silicon rectangle, so left-click on the point to select it, then right-click on it to add it to the list of fixed constraints (this is what I think of as $T_{initial}$).

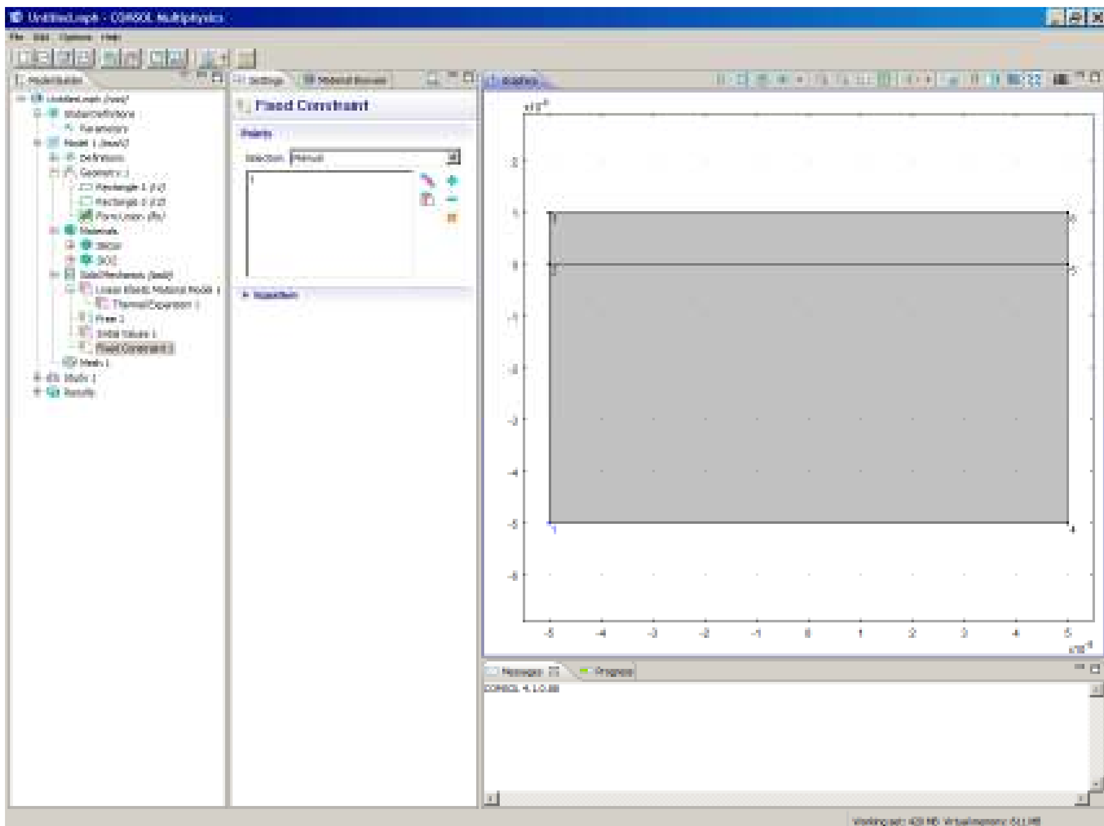


Figure C.16 Adding the lower left corner to the list of fixed constraints.

COMSOL now knows that the whole object cannot undergo a global displacement, but it does not yet know not to rotate the object. To add this, right-click on "Solid Mechanics" and under "Points" select "Prescribed Displacement." Add the lower left corner to the list of points of prescribed displacement the same way we did in the previous step. To prevent rotation, we do not want this corner to experience a displacement in the y-direction, but we do not care about displacement in the x-direction. Check the box next to "prescribed in y-direction." Zero is entered in the box below by default (Fig. C.17).

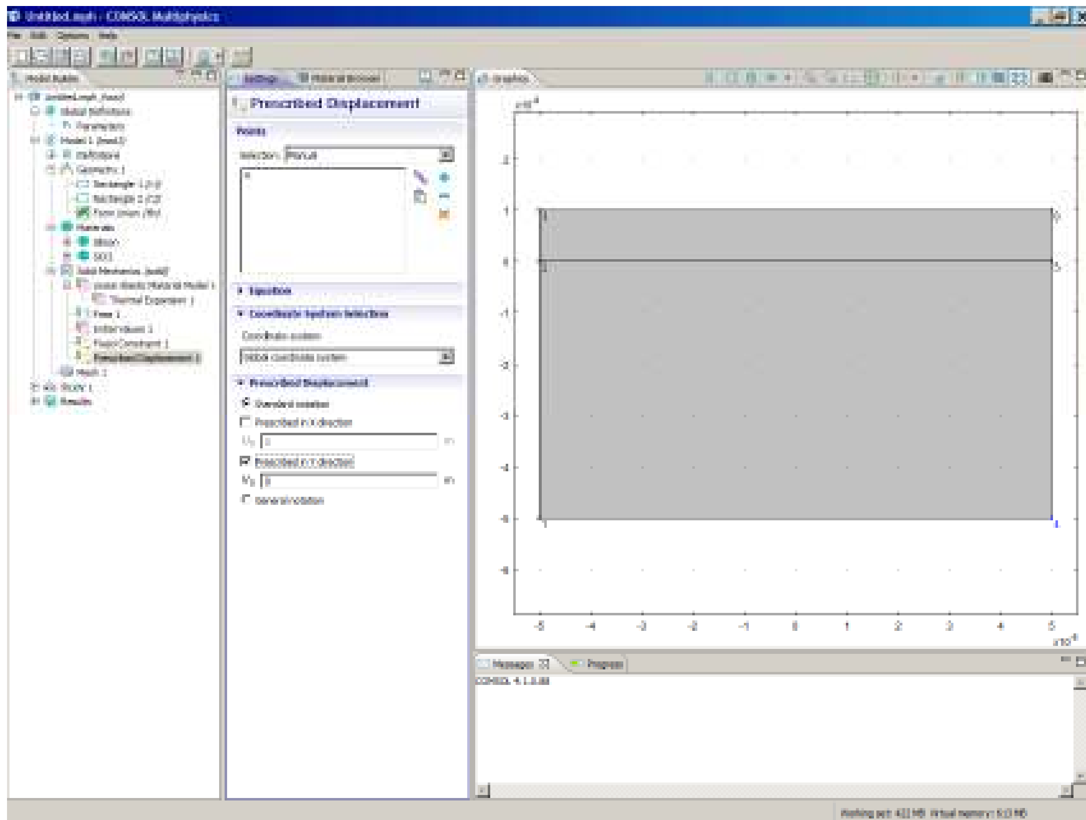


Figure C.17 Prescribed Displacement for the lower right corner.

Now under “Model Builder” left-click on ”Solid Mechanics.” Then under “2D approximation” change “Plane Strain” to “Plane Stress.” This will set the out of plane stress components, σ_z, σ_{xz} and σ_{yz}, to zero. This corresponds to the case of the case of very thin (in the out-of-plane direction) Si and SiO_2. I did not make this choice for a physics reason; rather I am making this choice because it is the case we have an analytical solution for.

Next we need to mesh the rectangles, so click on “Mesh 1” under the “Model Builder.” Change the element size from “Normal” to “Extra Fine.” Then select “Build All” to mesh the object. COMSOL will display the resulting mesh (Fig. C.18).

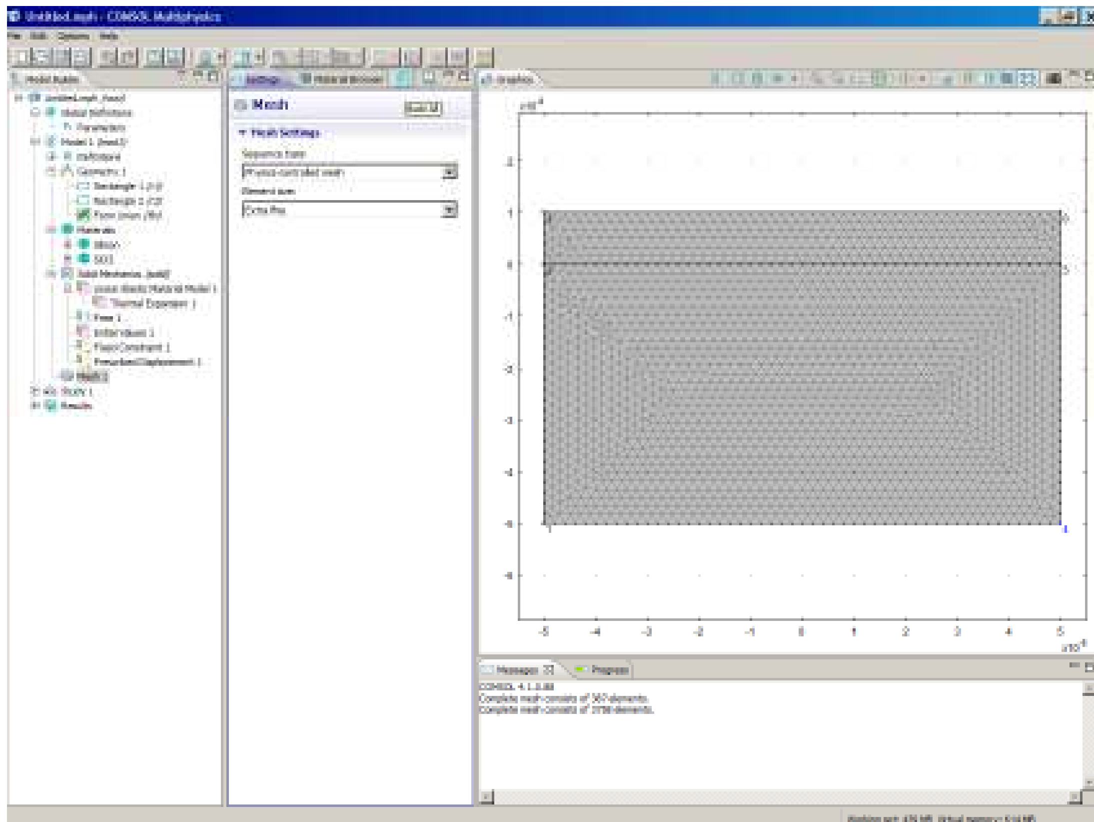


Figure C.18 Meshing the object.

Now we are done making the object, we just need to tell COMSOL to simulate it. Under “Model Builder” right-click on “Study 1” and then select “Compute”. For a geometry this simple, it should only take a couple of seconds. By default a surface plot of the displacement will appear (Fig. C.19). Notice that the displacement is zero in the lower-left corner, where we set a fixed constraint.

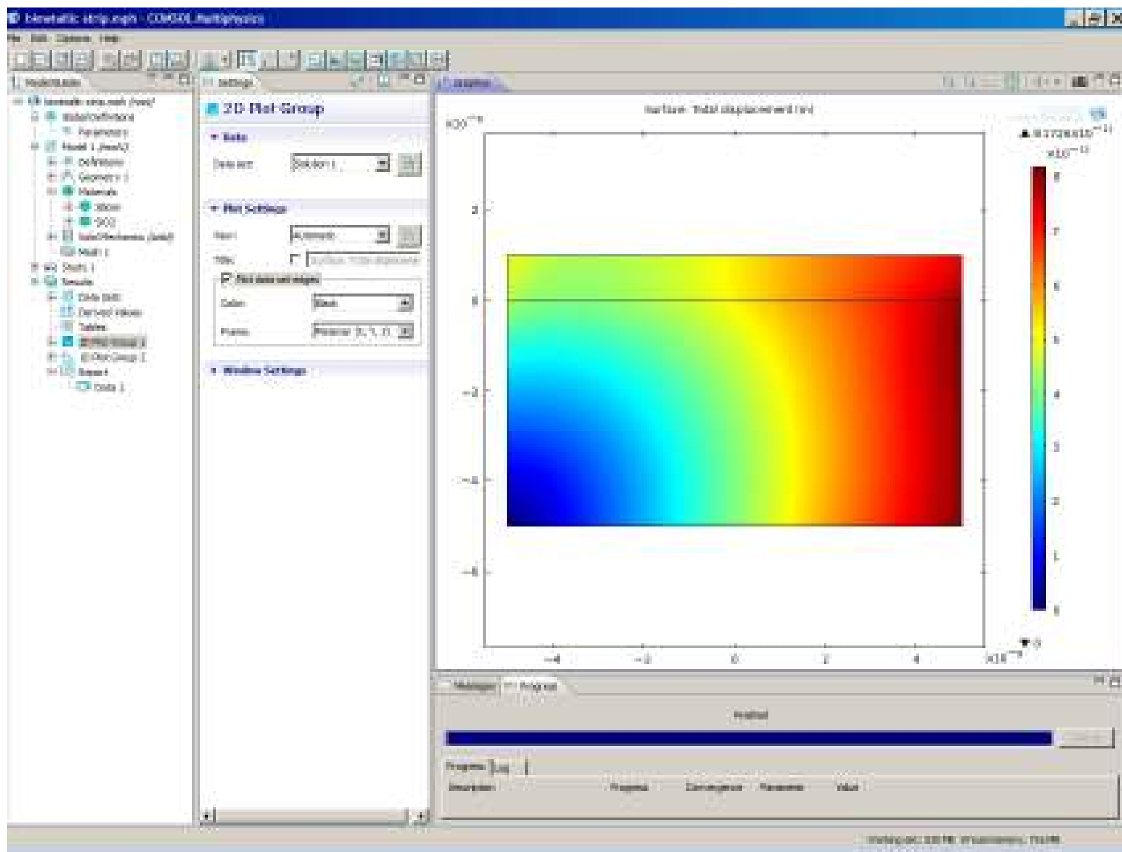


Figure C.19 The displacement.

We want to compare the stress in the simulation to the stress calculated from Eq. C.8. We want to plot the stresses along the black dashed line in Figure C.4(b). To tell COMSOL to take a cut along the center of the object, right-click on “Data Sets” under “Results” in the “Model Builder,” and select “Cut Line 2D.” Under “Line Data” for “Point 1:” enter “0” for “X:” and “-dsi” for “Y:”. And for “Point 2:” enter “0” for “X:” and “dox” for “Y:”. Then click on the “Plot” button, which is where the “Build All” button used to be. The line cut should appear in the graphics window (Fig. C.20).

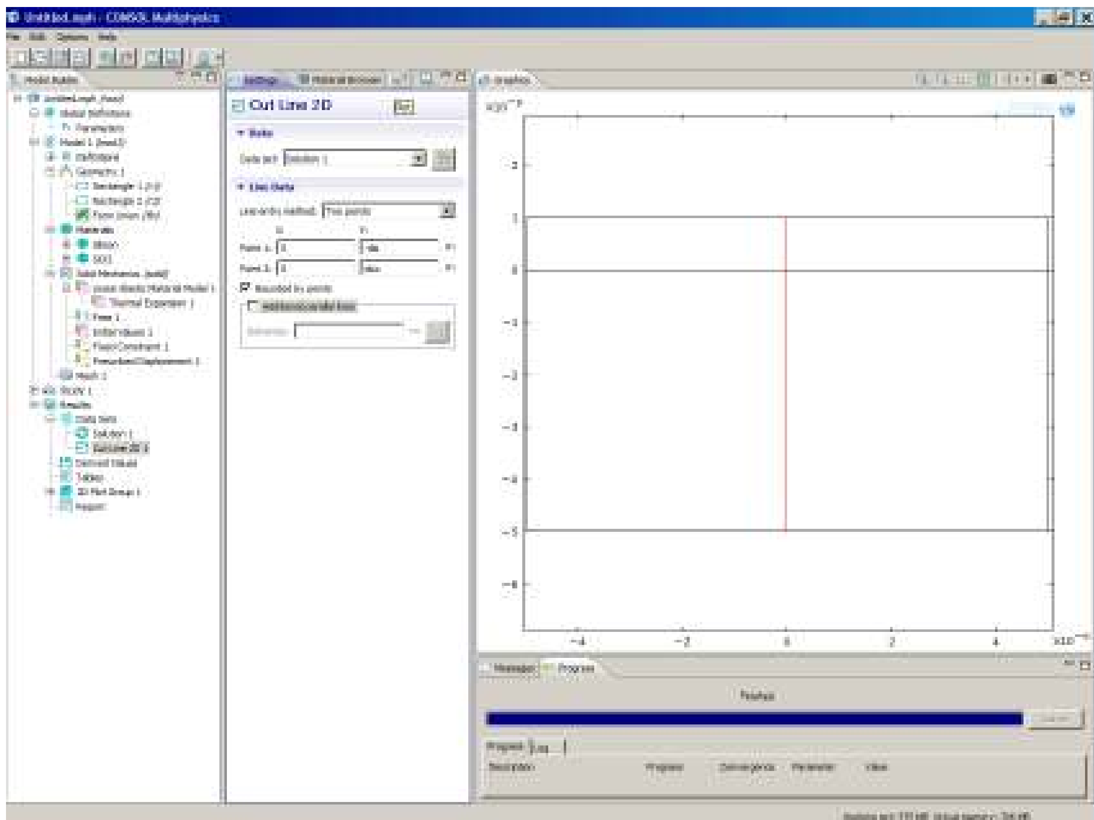
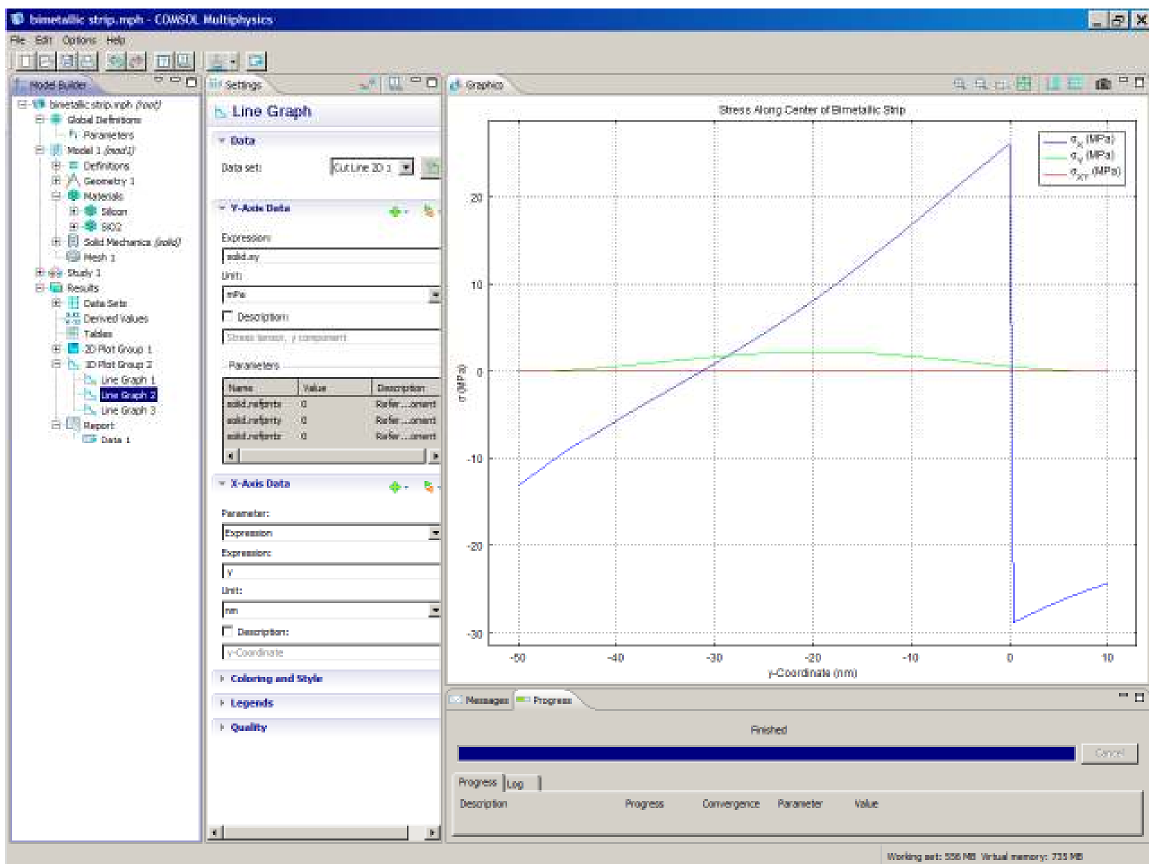


Figure C.20 The line cut.

Next we want to create a plot of the stress along the line cut we just made, so right-click on “Results” and select “1D Plot Group.” Then under “Results” right-click on “1D Plot Group 2” and select “Line Graph.” Under “Data set” select “Cut Line 2D 1.” For “Y-Axis Data” under “Expression:” enter “solid.sx”, and change the unit to “MPa.” Under X-Axis Data” change the “Parameter” to “Expression” then enter “y” under “Expression:” and “nm” under “Unit:”. Finally, click “Plot.” A plot of σ_x versus y should appear. I will go ahead and add the stresses σ_y and σ_{xy}, and a legend. We see that σ_x has the form expected from expected from Eq. C.8: separate linear functions in the silicon and the oxide.



C.20 The stresses σ_x, σ_y and σ_{xy}.

In Figure C.22, we see that the stress σ_y in the y-direction is not zero. This is because I did not make the length of the bimetallic strips much greater than the thicknesses of the silicon and oxide. We change L to 200 nm, and recalculate the stresses (Figure C.22). Now we see that the stress in the y-direction has gone to zero.

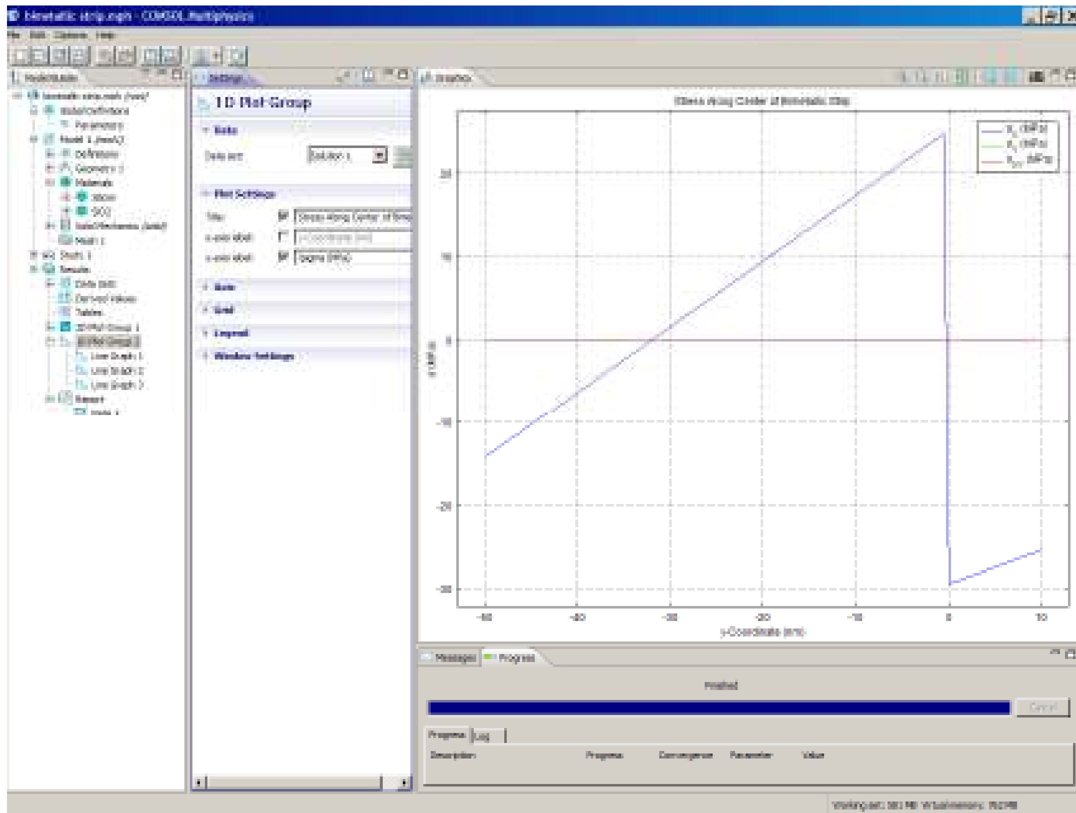


Figure C.22 The stresses σ_x, σ_y and σ_{xy}, for L = 200 nm.

Congratulations, if you have followed along, you have completed your first COMSOL simulation.

Just in case it would be helpful to you, I have saved this example on GuestroomPC in C:\Ted\Programs\comsol\bimetallic\bimetallic strip.mph.

F. Comparison

In Figure C.23 I compare the numerical solution from COMSOL to the analytical expression in Eq. C.14. We see that the agreement between the two methods is very good. This gives us confidence that COMSOL is calculating the stress as we expect.

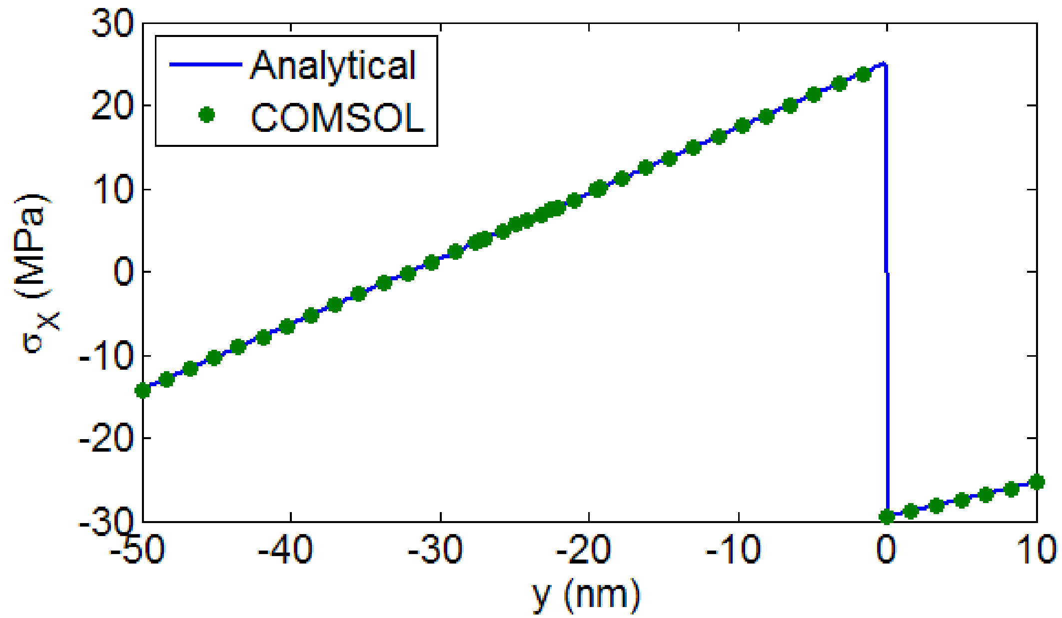


Figure C.23. Comparison of analytical expression for stress in the y-direction from Eq. C.14 (blue line) to the COMSOL calculation (green squares).

G. Material Properties

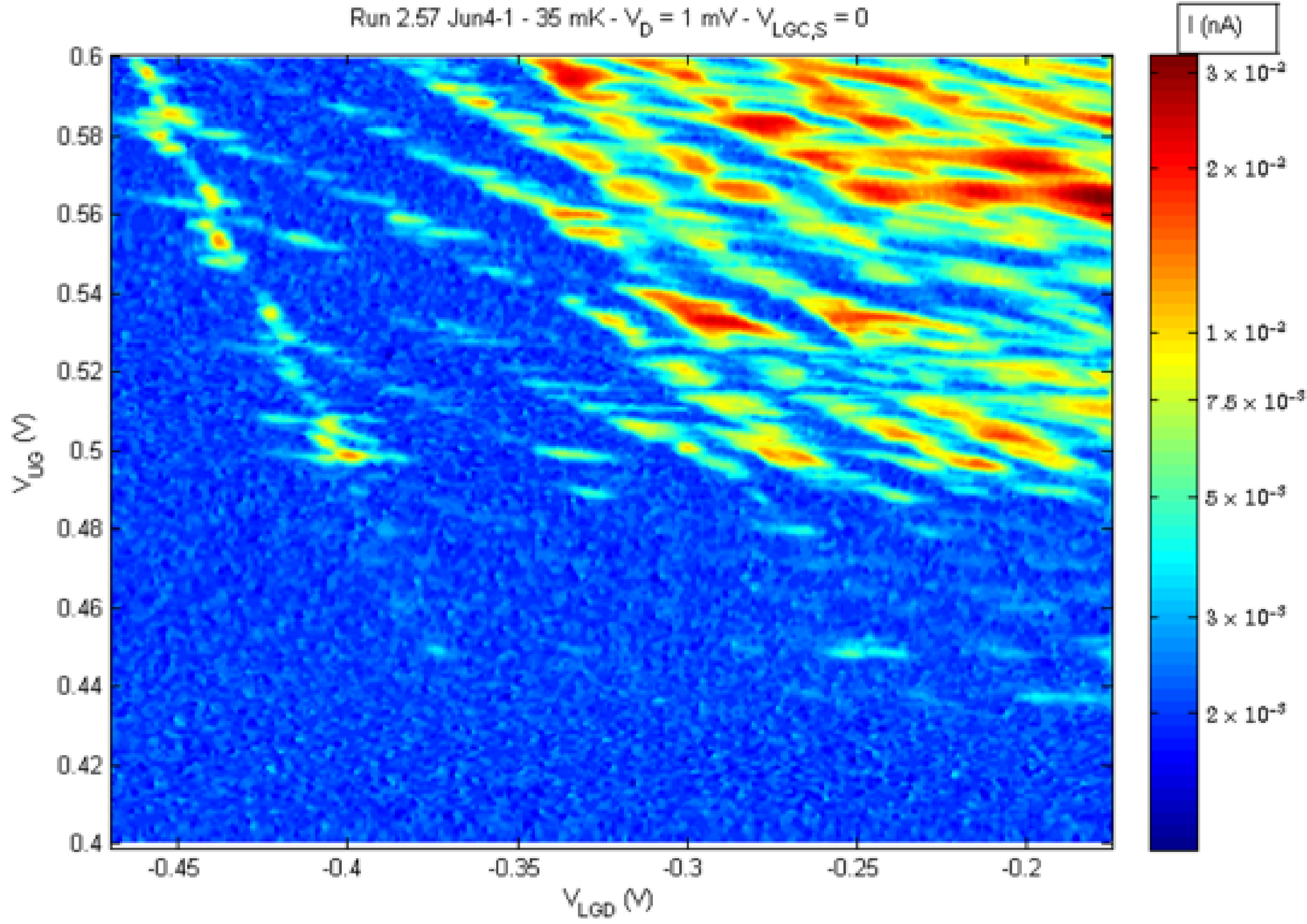
In Table C.2 I show the material properties I used in the COMSOL simulations in Chapter 5.

Table C.2 List of parameters used in the COMSOL simulations, including Young's modulus, Poisson's ratio, density and CTE.

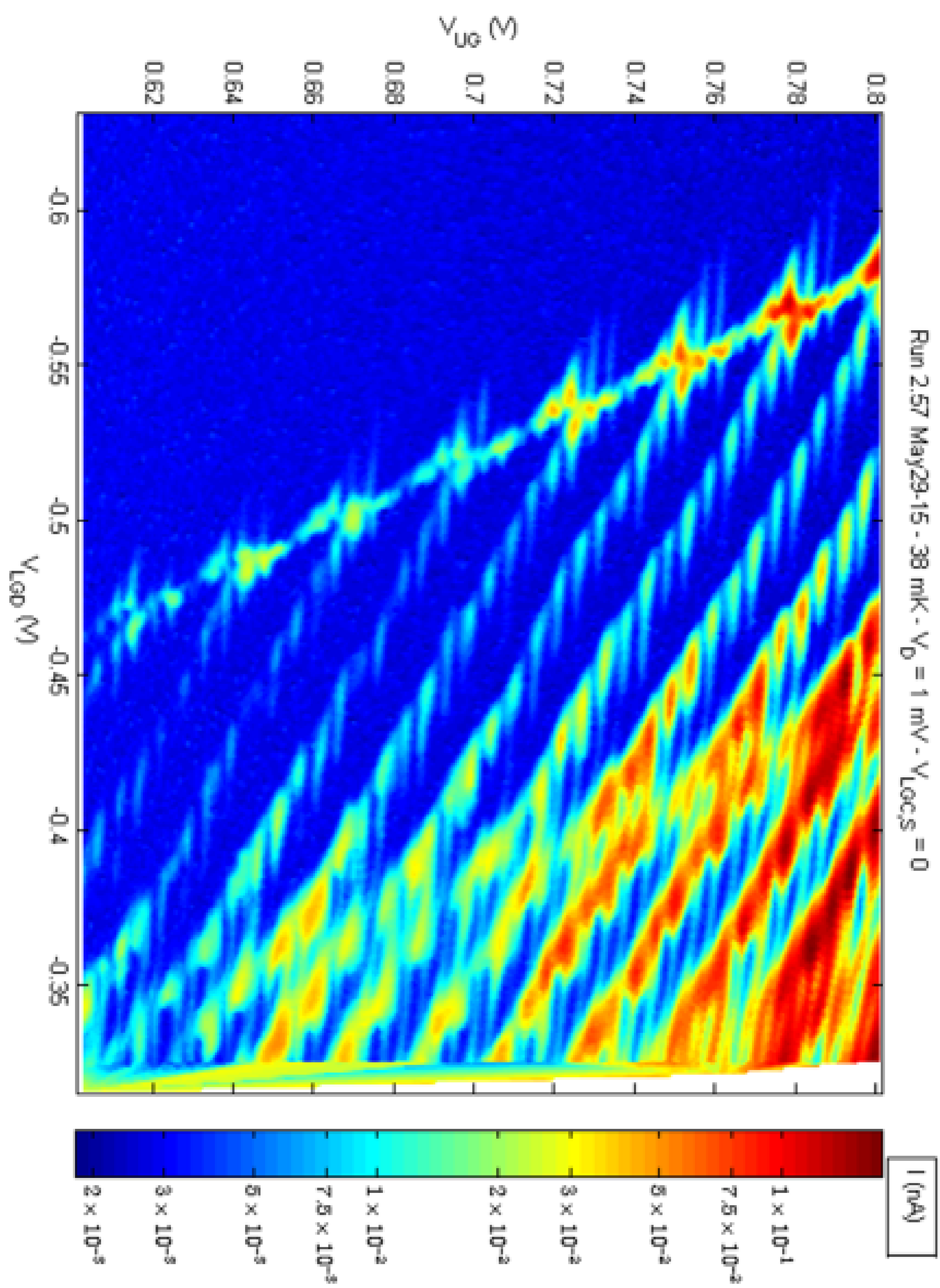
	Young's Modulus [GPa]	Poisson's ratio	Density [kg/m^3]	CTE (x10^{-6}/K)
Silicon	130	0.27	2300	2.6
SiO_2	73	0.17	2200	0.49
Aluminum	70	0.35	2700	23
AlO_X	300	0.22	3900	5.4
Nickel	220	0.31	8900	13
Poly-Si	170	0.22	2300	2.9

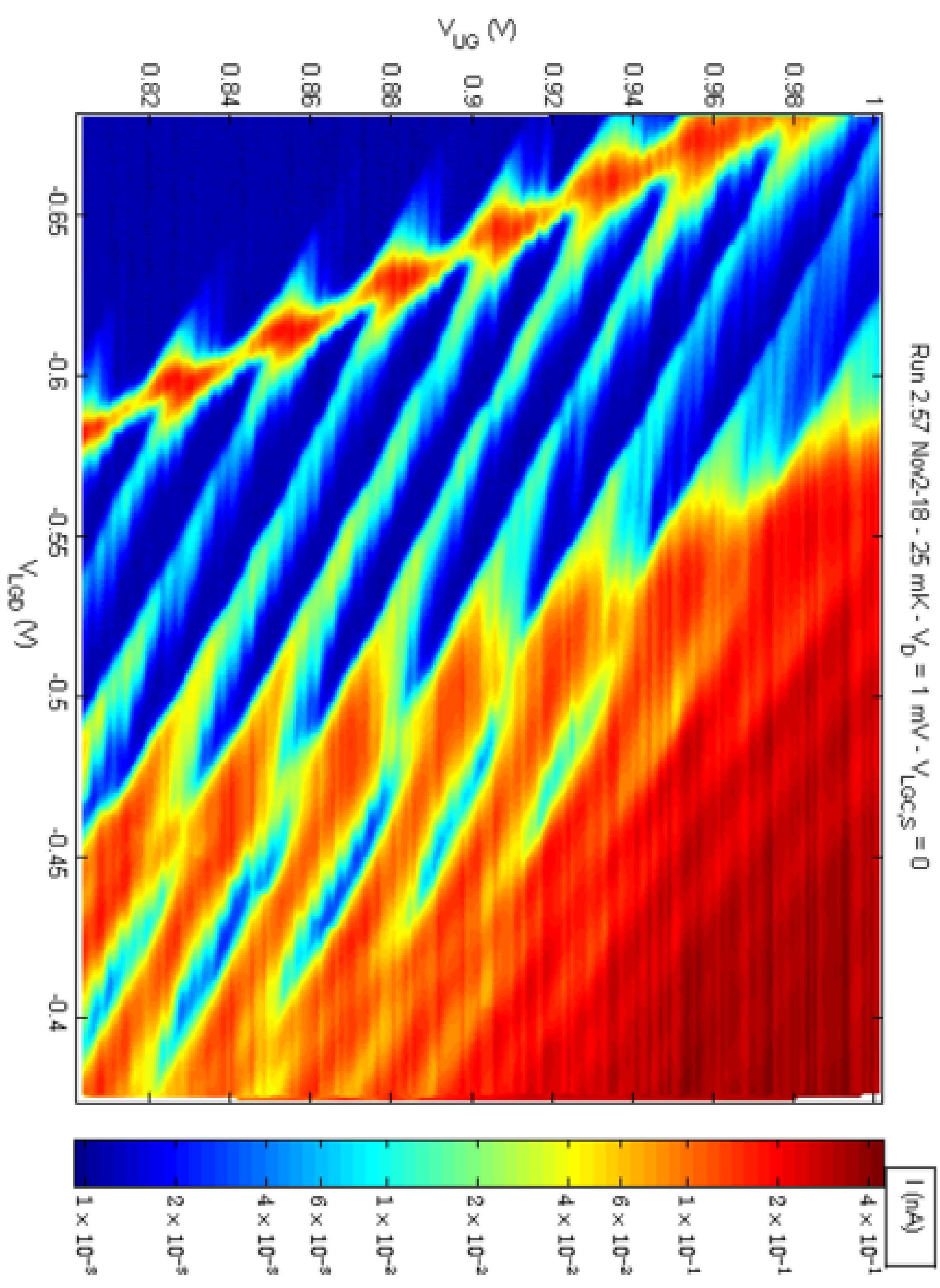
Appendix D: Additional Unintentional QD Data

In this appendix I feature current as a function of V_{UG} and V_{LGD}, showing both dot A and B, from V_{UG} = 0.5 V to 2.4 V in device AF-CA2U3D-3. Having data over such a wide voltage range was helpful in measuring the capacitances as I discussed in Chapter 3. Each plot shows a different range of V_{UG}. It is remarkable that the same two unintentional QDS are observable over such a range of gate voltages.

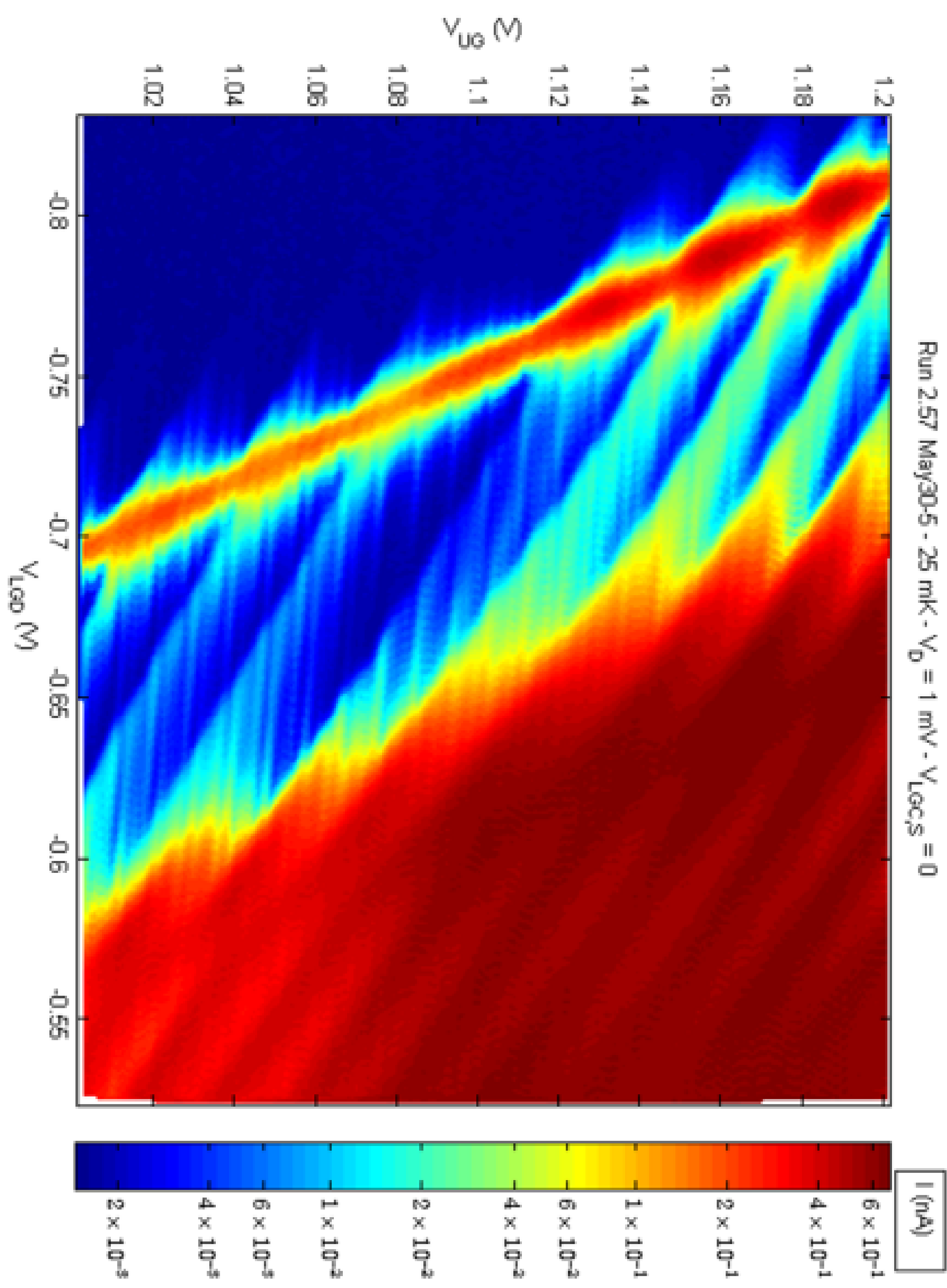


Run 2.57 Jun4-1 - 35 mK - V_D = 1 mV - V_LGC,S = 0
V_UG (V)
V_LGD (V)
I (nA)
0.4
0.42
0.44
0.46
0.48
0.5
0.52
0.54
0.56
0.58
0.6
-0.45
-0.4
-0.35
-0.3
-0.25
-0.2
2 × 10^-3
3 × 10^-3
5 × 10^-3
7.5 × 10^-3
1 × 10^-2
2 × 10^-2
3 × 10^-2

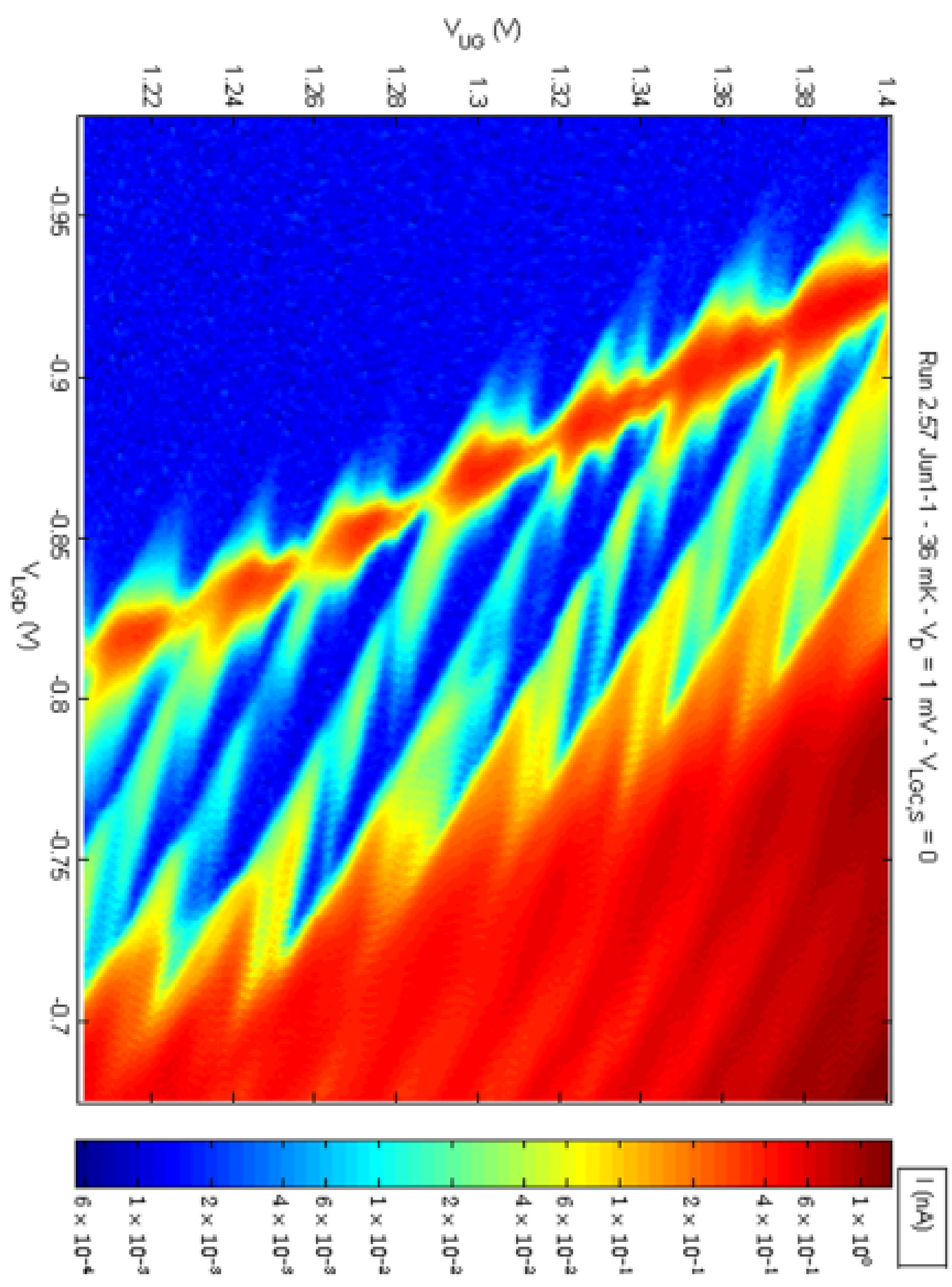

Run 2.57 May29-15 - 38 mK - V_D = 1 mV - V_LGC,S = 0
V_UG (V)
0.62
0.64
0.66
0.68
0.7
0.72
0.74
0.76
0.78
0.8
V_LGD (V)
-0.6
-0.55
-0.5
-0.45
-0.4
-0.35
I (nA)
2 × 10⁻³
3 × 10⁻³
5 × 10⁻³
7.5 × 10⁻³
1 × 10⁻²
2 × 10⁻²
3 × 10⁻²
5 × 10⁻²
7.5 × 10⁻²
1 × 10⁻¹

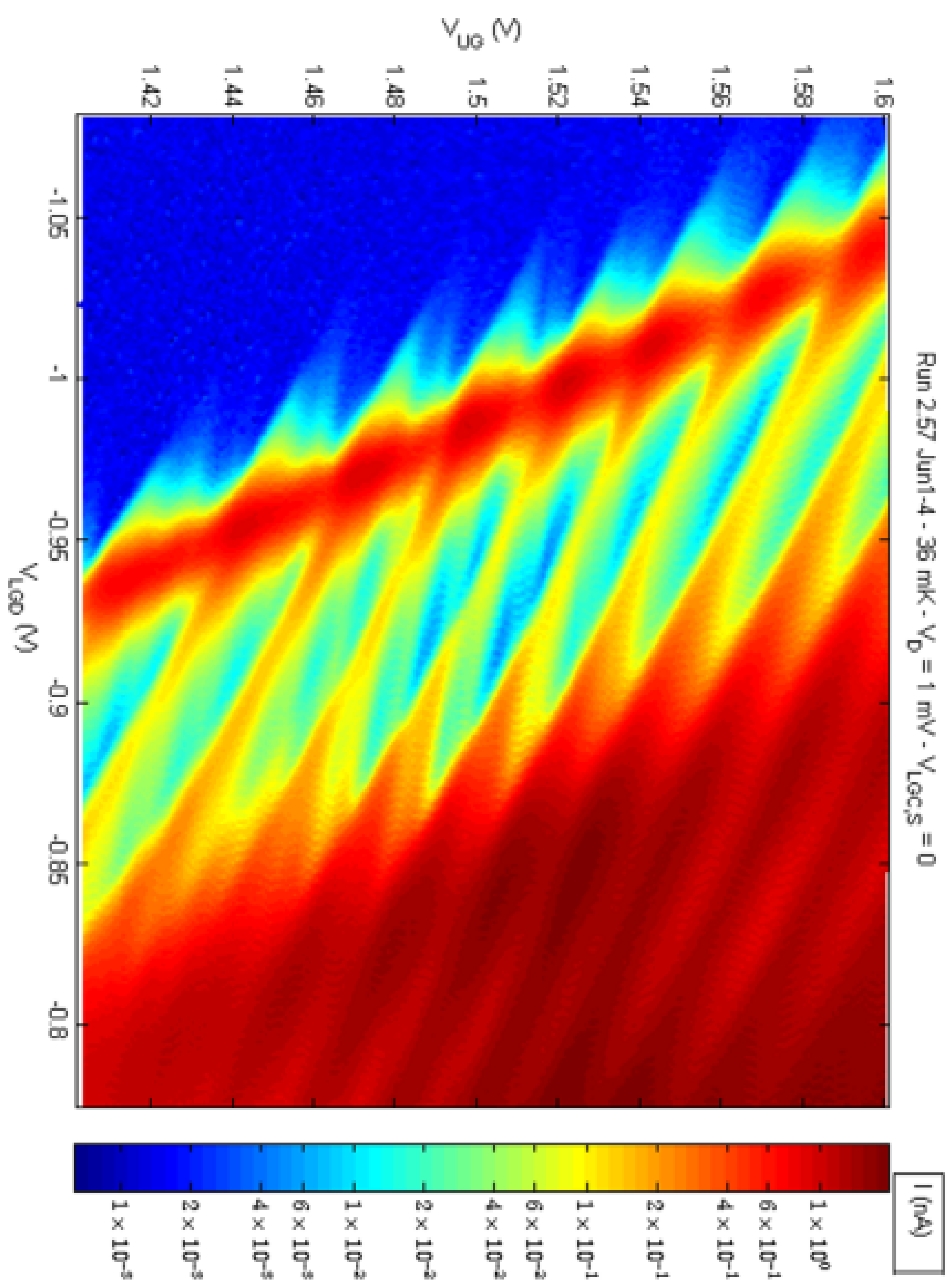


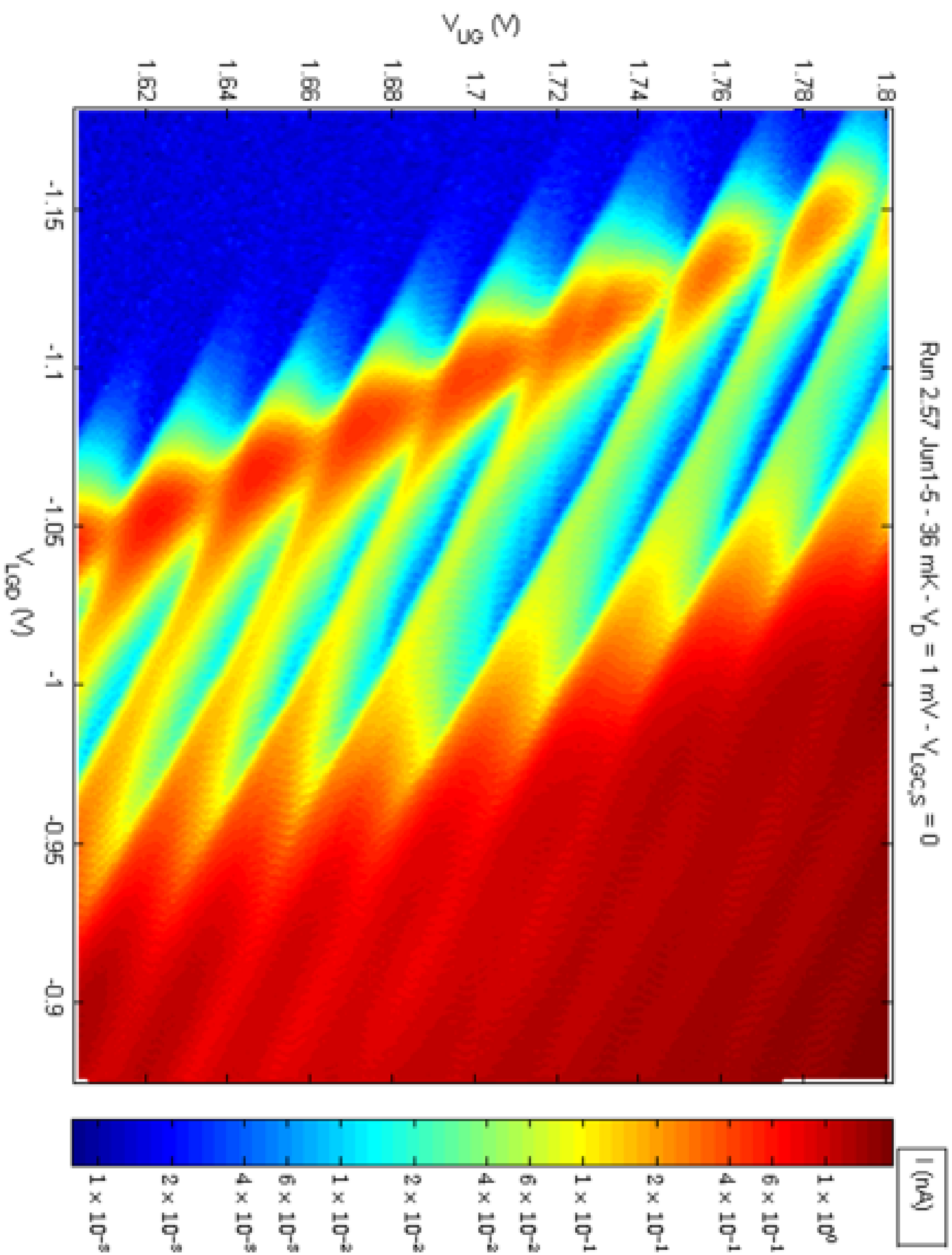
Run 2.57 Nov2-18 - 25 mK - V_D = 1 mV - V_LOC,S = 0
V_UG (V)
V_LGD (V)
I (nA)



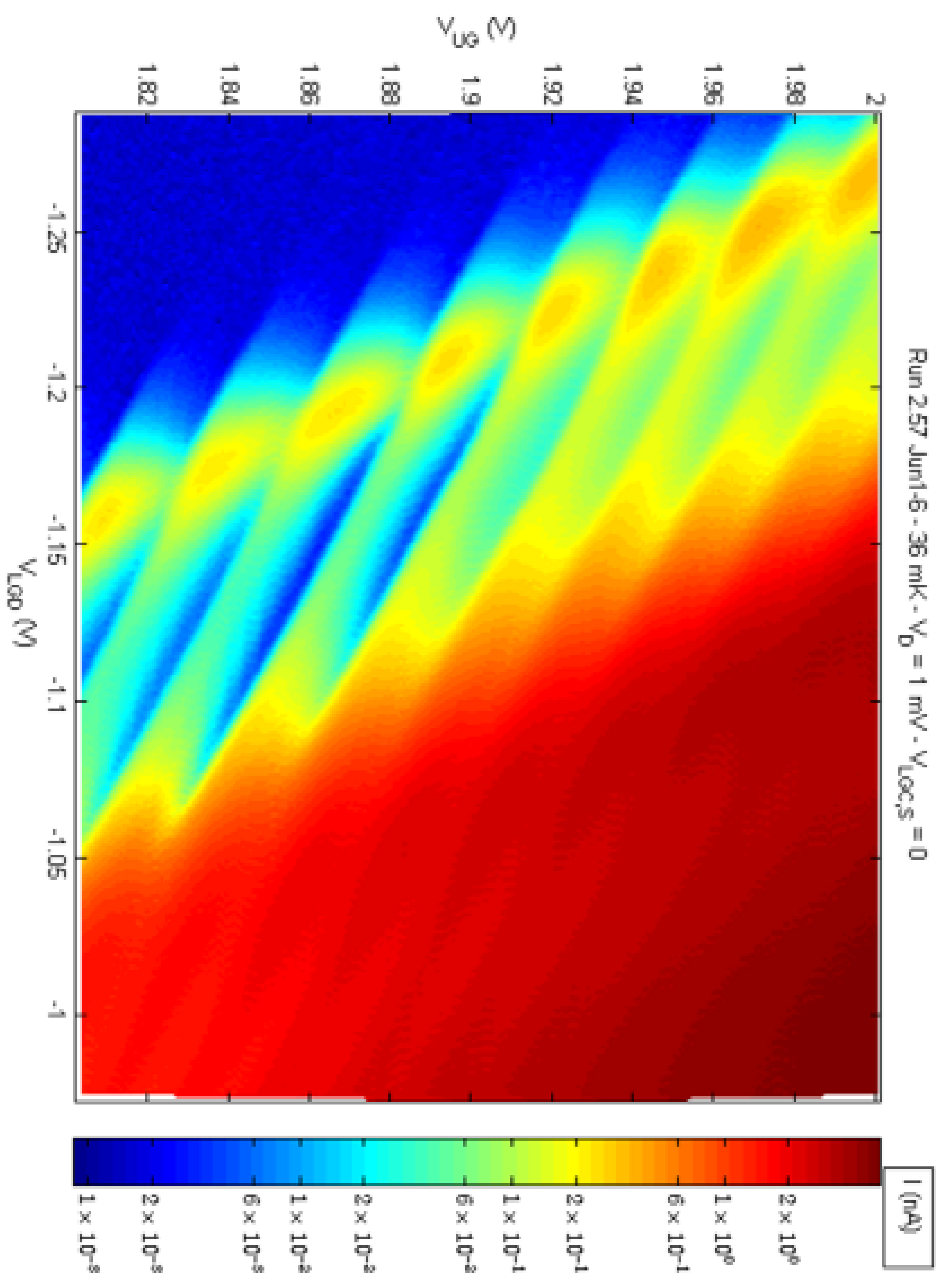
Run 2.57 May30-5 - 25 mK - V_D = 1 mV - $V_{LGC,S}$ = 0
V_{UG} (V)
1.02
1.04
1.06
1.08
1.1
1.12
1.14
1.16
1.18
1.2
V_{LGD} (V)
-0.8
-0.75
-0.7
-0.65
-0.6
-0.55
I (nA)
2×10^{-3}
4×10^{-3}
6×10^{-3}
1×10^{-2}
2×10^{-2}
4×10^{-2}
6×10^{-2}
1×10^{-1}
2×10^{-1}
4×10^{-1}
6×10^{-1}

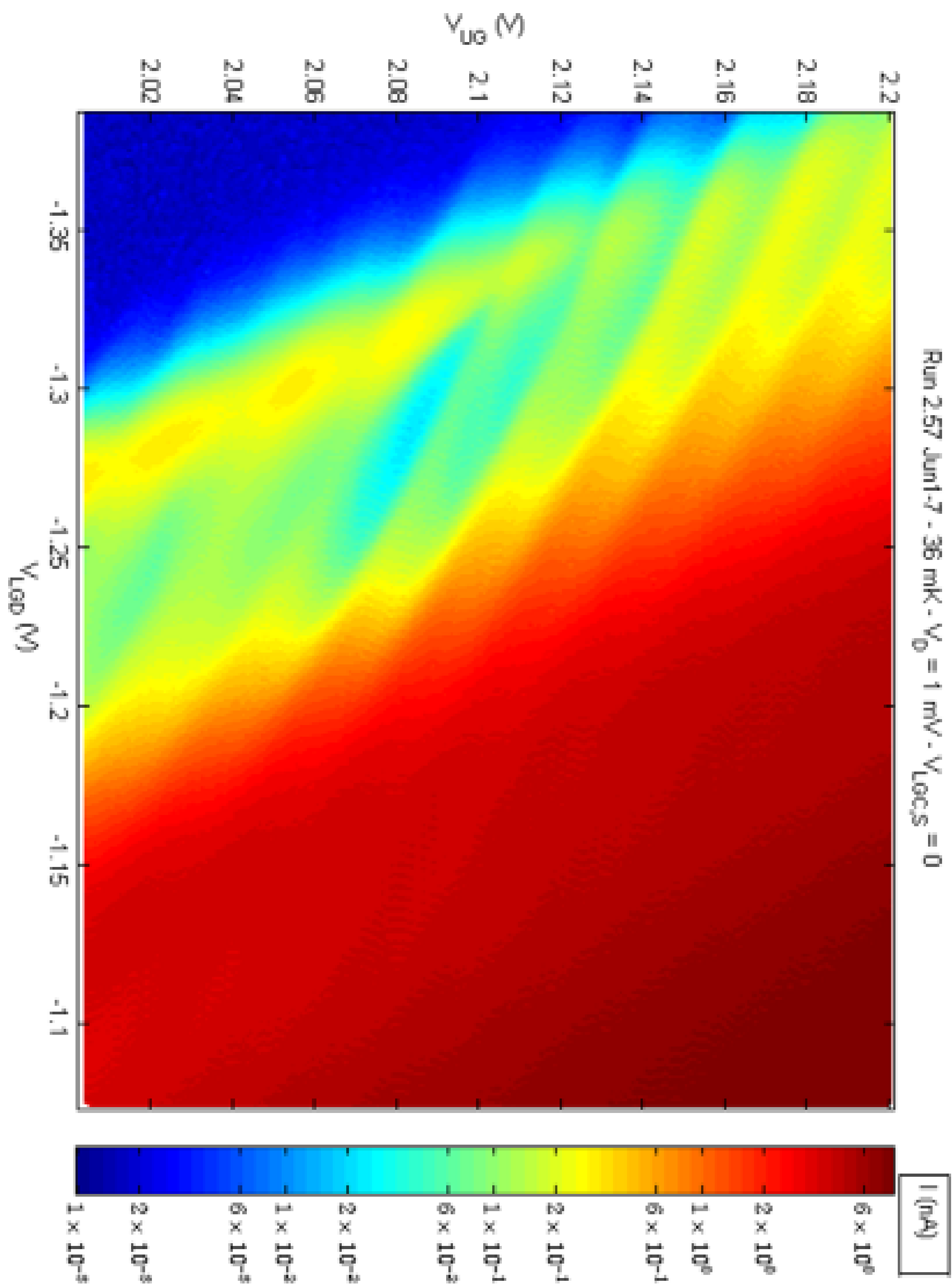

Run 2.57 Jun1-1 - 36 mK - V_D = 1 mV - $V_{LGC,S}$ = 0
V_{UG} (V)
V_{LGD} (V)
I (nA)
1.22
1.24
1.26
1.28
1.3
1.32
1.34
1.36
1.38
1.4
-0.95
-0.9
-0.85
-0.8
-0.75
-0.7
6×10^{-4}
1×10^{-3}
2×10^{-3}
4×10^{-3}
6×10^{-3}
1×10^{-2}
2×10^{-2}
4×10^{-2}
6×10^{-2}
1×10^{-1}
2×10^{-1}
4×10^{-1}
6×10^{-1}
1×10^{0}

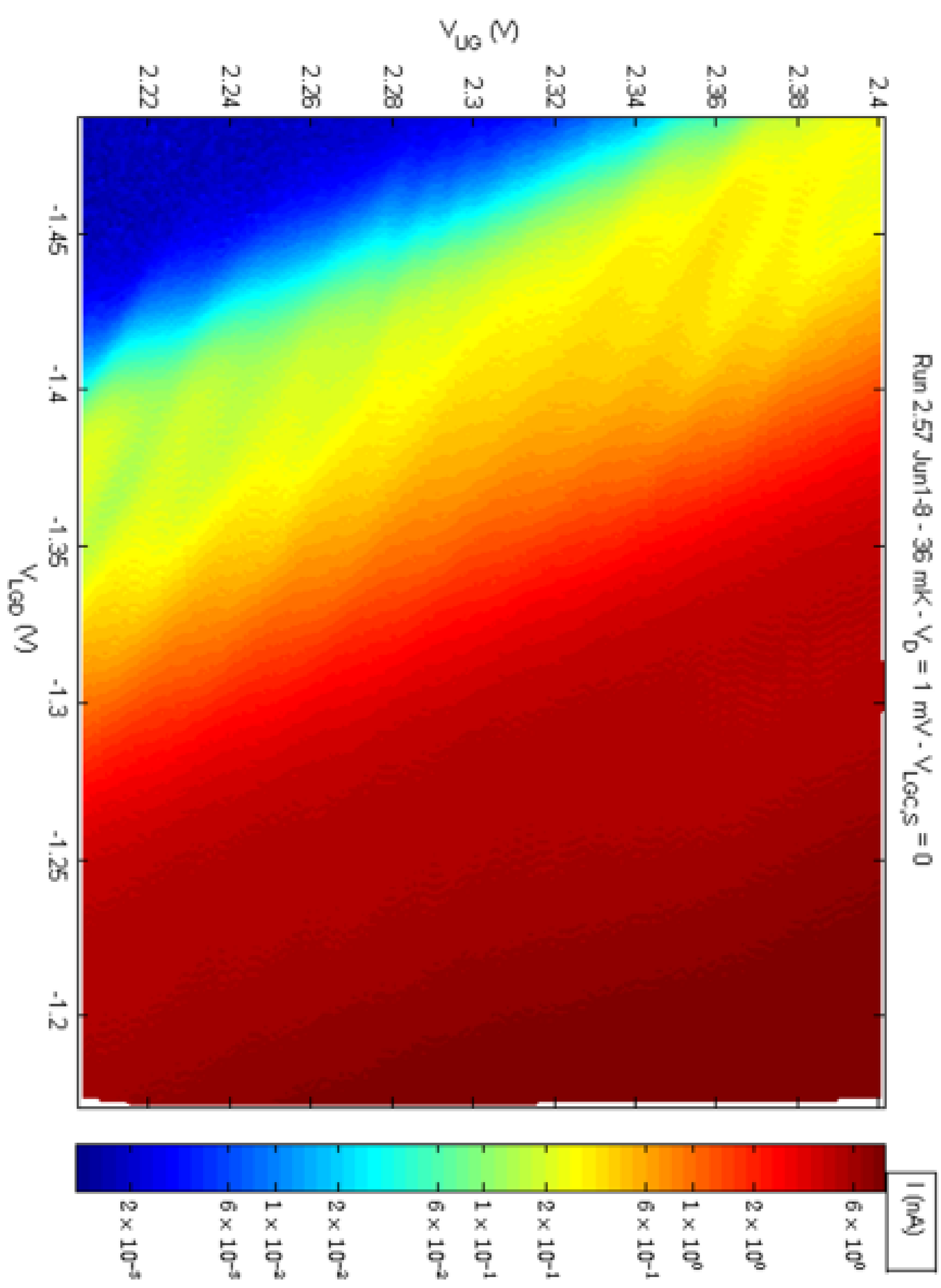

Run 2.57 Jun1-4 - 36 mK - V_D = 1 mV - $V_{LGC,S}$ = 0
V_{UG} (V)
1.42
1.44
1.46
1.48
1.5
1.52
1.54
1.56
1.58
1.6
-1.05
-1
-0.95
-0.9
-0.85
-0.8
V_{LGD} (V)
I (nA)
1×10^{-3}
2×10^{-3}
4×10^{-3}
6×10^{-3}
1×10^{-2}
2×10^{-2}
4×10^{-2}
6×10^{-2}
1×10^{-1}
2×10^{-1}
4×10^{-1}
6×10^{-1}
1×10^{0}



Run 2.57 Jun1-5 - 36 mK - V_D = 1 mV - V_LGC,S = 0
V_UG (V)
1.62
1.64
1.66
1.68
1.7
1.72
1.74
1.76
1.78
1.8
V_LGD (V)
-1.15
-1.1
-1.05
-1
-0.95
-0.9
I (nA)
1 × 10^-3
2 × 10^-3
4 × 10^-3
6 × 10^-3
1 × 10^-2
2 × 10^-2
4 × 10^-2
6 × 10^-2
1 × 10^-1
2 × 10^-1
4 × 10^-1
6 × 10^-1
1 × 10^0


Run 2.57 Jun1-6 - 36 mK - V_D = 1 mV - V_LGC,S = 0
V_UG (V)
2
1.98
1.96
1.94
1.92
1.9
1.88
1.86
1.84
1.82
-1.25
-1.2
-1.15
-1.1
-1.05
-1
V_LGD (V)
I (nA)
2 × 10^0
1 × 10^0
6 × 10^-1
2 × 10^-1
1 × 10^-1
6 × 10^-2
2 × 10^-2
1 × 10^-2
6 × 10^-3
2 × 10^-3
1 × 10^-3


Run 2.57 Jun1-7 - 36 mK - V_D = 1 mV - $V_{LGC,S}$ = 0
V_{UG} (V)
V_{LGD} (V)
I (nA)
2.02
2.04
2.06
2.08
2.1
2.12
2.14
2.16
2.18
2.2
-1.35
-1.3
-1.25
-1.2
-1.15
-1.1



Run 2.57 Jun1-8 - 36 mK - V_D = 1 mV - $V_{LGC,S}$ = 0
V_{UG} (V)
2.22
2.24
2.26
2.28
2.3
2.32
2.34
2.36
2.38
2.4
V_{LGD} (V)
-1.45
-1.4
-1.35
-1.3
-1.25
-1.2
I (nA)
2×10^{-3}
6×10^{-3}
1×10^{-2}
2×10^{-2}
6×10^{-2}
1×10^{-1}
2×10^{-1}
6×10^{-1}
1×10^{0}
2×10^{0}
6×10^{0}

References

[1] N. M. Zimmerman, Am. J. Phys. **66**, 324 (1998).

[2] J. P. Pekola, O.-P. Saira, V. F. Maisi, A. Kemppinen, M. Möttönen, Y. A. Pashkin, and D. V. Averin, arXiv:1208.4030 (2012).

[3] F. Piquemal and G. Genev s, Metrologia **37**, 207 (2000).

[4] M. W. Keller, Metrologia **45**, 102 (2008).

[5] M. A. Nielsen and I. L. Chuang, *Quantum Computation and Quantum Information*, 1st ed. (Cambridge University Press, 2000).

[6] J. J. L. Morton, D. R. McCamey, M. A. Eriksson, and S. A. Lyon, Nature **479**, 345 (2011).

[7] F. A. Zwanenburg, A. S. Dzurak, A. Morello, M. Y. Simmons, L. C. L. Hollenberg, G. Klimeck, S. Rogge, S. N. Coppersmith, and M. A. Eriksson, Arxiv12065202 Appear Rev Mod Phys (2013).

[8] Y. Nakamura, Y. A. Pashkin, and J. S. Tsai, Nature **398**, 786 (1999).

[9] T. Hayashi, T. Fujisawa, H. D. Cheong, Y. H. Jeong, and Y. Hirayama, Phys. Rev. Lett. **91**, 226804 (2003).

[10] J. R. Petta, A. C. Johnson, C. M. Marcus, M. P. Hanson, and A. C. Gossard, Phys. Rev. Lett. **93**, 186802 (2004).

[11] J. Gorman, D. G. Hasko, and D. A. Williams, Phys. Rev. Lett. **95**, 090502 (2005).

[12] A. Fujiwara, K. Nishiguchi, and Y. Ono, Appl. Phys. Lett. **92**, 042102 (2008).

[13] Z. Shi, C. B. Simmons, D. R. Ward, J. R. Prance, T. S. Koh, J. K. Gamble, X. Wu, D. E. Savage, M. G. Lagally, M. Friesen, S. N. Coppersmith, and M. A. Eriksson, arXiv:1208.0519 (2012).

[14] A. Fujiwara, H. Inokawa, K. Yamazaki, H. Namatsu, Y. Takahashi, N. M. Zimmerman, and S. B. Martin, Appl. Phys. Lett. **88**, 053121 (2006).

[15] P. J. Koppinen, M. D. Stewart, and N. M. Zimmerman, Ieee Trans. Electron Devices **60**, 78 (2013).

[16] A. Fujiwara, Pers. Commun. (2012).

[17] J.-P. Colinge, editor, *FinFETs and Other Multi-Gate Transistors*, Softcover reprint of hardcover 1st ed. 2008 (Springer, 2010).

[18] T. Heinzel, *Mesoscopic Electronics in Solid State Nanostructures*, 3rd ed. (Wiley-VCH, 2010).

[19] H. Grabert and M. H. Devoret, editors, *Single Charge Tunneling: Coulomb Blockade Phenomena in Nanostructures*, 1st ed. (Springer, 1992).

[20] L. P. Kouwenhoven, C. M. Marcus, P. L. McEuen, S. Tarucha, R. M. Westervelt, and N. S. Wingreen, in *Mesoscopic Electron Transp.* (Springer, 1997).

[21] W. G. van der Wiel, S. De Franceschi, J. M. Elzerman, T. Fujisawa, S. Tarucha, and L. P. Kouwenhoven, Rev. Mod. Phys. **75**, 1 (2002).

[22] K. W. Chan, M. Möttönen, A. Kemppinen, N. S. Lai, K. Y. Tan, W. H. Lim, and A. S. Dzurak, Appl. Phys. Lett. **98**, 212103 (2011).

[23] S. B. Field, M. A. Kastner, U. Meirav, J. H. F. Scott-Thomas, D. A. Antoniadis, H. I. Smith, and S. J. Wind, Phys. Rev. B **42**, 3523 (1990).

[24] J. H. F. Scott-Thomas, S. B. Field, M. A. Kastner, H. I. Smith, and D. A. Antoniadis, Phys. Rev. Lett. **62**, 583 (1989).

[25] H. Ishikuro and T. Hiramoto, Appl. Phys. Lett. **71**, 3691 (1997).

[26] C. de Graaf, J. Caro, S. Radelaar, V. Lauer, and K. Heyers, Phys. Rev. B **44**, 9072 (1991).

[27] H. W. Liu, T. Fujisawa, Y. Ono, H. Inokawa, A. Fujiwara, K. Takashina, and Y. Hirayama, Phys. Rev. B **77**, 073310 (2008).

[28] N. Shaji, C. B. Simmons, M. Thalakulam, L. J. Klein, H. Qin, H. Luo, D. E. Savage, M. G. Lagally, A. J. Rimberg, R. Joynt, M. Friesen, R. H. Blick, S. N. Coppersmith, and M. A. Eriksson, Nat. Phys. **4**, 540 (2008).

[29] B. Hu and C. H. Yang, Phys. Rev. B **80**, 075310 (2009).

[30] S. J. Angus, A. J. Ferguson, A. S. Dzurak, and R. G. Clark, Nano Lett. **7**, 2051 (2007).

[31] J. J. Pla, K. Y. Tan, J. P. Dehollain, W. H. Lim, J. J. L. Morton, D. N. Jamieson, A. S. Dzurak, and A. Morello, Nature **489**, 541 (2012).

[32] E. P. Nordberg, G. A. T. Eyck, H. L. Stalford, R. P. Muller, R. W. Young, K. Eng, L. A. Tracy, K. D. Childs, J. R. Wendt, R. K. Grubbs, J. Stevens, M. P. Lilly, M. A. Eriksson, and M. S. Carroll, Phys. Rev. B **80**, 115331 (2009).

[33] E. P. Nordberg, H. L. Stalford, R. Young, G. A. Ten Eyck, K. Eng, L. A. Tracy, K. D. Childs, J. R. Wendt, R. K. Grubbs, J. Stevens, M. P. Lilly, M. A. Eriksson, and M. S. Carroll, Appl. Phys. Lett. **95**, 202102 (2009).

[34] N. M. Zimmerman, W. H. Huber, B. Simonds, E. Hourdakis, A. Fujiwara, Y. Ono, Y. Takahashi, H. Inokawa, M. Furlan, and M. W. Keller, J. Appl. Phys. **104**, 033710 (2008).

[35] N. M. Zimmerman, B. J. Simonds, A. Fujiwara, Y. Ono, Y. Takahashi, and H. Inokawa, Appl. Phys. Lett. **90**, 033507 (2007).

[36] N. M. Zimmerman, W. H. Huber, A. Fujiwara, and Y. Takahashi, Appl. Phys. Lett. **79**, 3188 (2001).

[37] E. Hourdakis, J. A. Wahl, and N. M. Zimmerman, Appl. Phys. Lett. **92**, 062102 (2008).

[38] C. Tahan and R. Joynt, arXiv:1301.0260 (2013).

[39] B. M. Maune, M. G. Borselli, B. Huang, T. D. Ladd, P. W. Deelman, K. S. Holabird, A. A. Kiselev, I. Alvarado-Rodriguez, R. S. Ross, A. E. Schmitz, M. Sokolich, C. A. Watson, M. F. Gyure, and A. T. Hunter, Nature **481**, 344 (2012).

[40] N. M. Zimmerman, A. Fujiwara, H. Inokawa, and Y. Takahashi, Appl. Phys. Lett. **89**, 052102 (2006).

[41] M. Boehm, M. Hofheinz, X. Jehl, M. Sanquer, M. Vinet, B. Previtali, D. Fraboulet, D. Mariolle, and S. Deleonibus, Phys. Rev. B **71**, 033305 (2005).

[42] M. Hofheinz, X. Jehl, M. Sanquer, G. Molas, M. Vinet, and S. Deleonibus, Appl. Phys. Lett. **89**, 143504 (2006).

[43] Y. Takahashi, Y. Ono, A. Fujiwara, and H. Inokawa, J. Phys. Condens. Matter **14**, R995 (2002).

[44] H. Stalford, R. W. Young, E. P. Nordberg, C. B. Pinilla, J. E. Levy, and M. S. Carroll, Nanotechnol. Ieee Trans. **10**, 855 (2011).

[45] W. H. Lim, H. Huebl, L. H. Willems van Beveren, S. Rubanov, P. G. Spizzirri, S. J. Angus, R. G. Clark, and A. S. Dzurak, Appl. Phys. Lett. **94**, 173502 (2009).

[46] G. J. Podd, S. J. Angus, D. A. Williams, and A. J. Ferguson, Appl. Phys. Lett. **96**, 082104 (2010).

[47] K. Nabors and J. White, Comput.-Aided Des. Integr. Circuits Syst. Ieee Trans. **10**, 1447 (1991).

[48] M. Hofheinz, X. Jehl, M. Sanquer, G. Molas, M. Vinet, and S. Deleonibus, Phys. Rev. B **75**, 235301 (2007).

[49] M. Miranda, Ieee Spectr. **49**, 32 (2012).

[50] G. P. Lansbergen, R. Rahman, C. J. Wellard, I. Woo, J. Caro, N. Collaert, S. Biesemans, G. Klimeck, L. C. L. Hollenberg, and S. Rogge, Nat. Phys. **4**, 656 (2008).

[51] N. C. Bishop, R. W. Young, G. A. T. Eyck, J. R. Wend, E. S. Bielejec, K. Eng, L. A. Tracy, M. P. Lilly, M. S. Carroll, C. B. Pinilla, and H. L. Stalford, arXiv:1107.5104 (2011).

[52] G. Leti, E. Prati, M. Belli, G. Petretto, M. Fanciulli, M. Vinet, R. Wacquez, and M. Sanquer, Appl. Phys. Lett. **99**, 242102 (2011).

[53] F. Hofmann, T. Heinzel, D. A. Wharam, J. P. Kotthaus, G. Böhm, W. Klein, G. Tränkle, and G. Weimann, Phys. Rev. B **51**, 13872 (1995).

[54] A. S. Adourian, C. Livermore, R. M. Westervelt, K. L. Campman, and A. C. Gossard, Appl. Phys. Lett. **75**, 424 (1999).

[55] I. H. Chan, R. M. Westervelt, K. D. Maranowski, and A. C. Gossard, Appl. Phys. Lett. **80**, 1818 (2002).

[56] A. Hübel, K. Held, J. Weis, and K. v. Klitzing, Phys. Rev. Lett. **101**, 186804 (2008).

[57] *Computational Single-Electronics*, 2001st ed. (Springer, 2001).

[58] S. Timoshenko, *Theory of Elasticity*, 3rd ed. (McGraw-Hill Publishing Company, 1970).

[59] R. W. Soutas-Little and Physics, *Elasticity* (Dover Publications, 1999).

[60] E. Kobeda and E. A. Irene, J. Vac. Sci. Technol. B Microelectron. Nanometer Struct. **4**, 720 (1986).

[61] E. Kobeda and E. A. Irene, J. Vac. Sci. Technol. B Microelectron. Nanometer Struct. **5**, 15 (1987).

[62] P. Y. Yu and M. Cardona, *Fundamentals of Semiconductors: Physics and Materials Properties*, 2nd ed. (Springer-Verlag Telos, 1999).

[63] T. Ando, A. B. Fowler, and F. Stern, Rev. Mod. Phys. **54**, 437 (1982).

[64] M. V. Fischetti and S. E. Laux, J. Appl. Phys. **80**, 2234 (1996).

[65] J. Singh, *Electronic and Optoelectronic Properties of Semiconductor Structures*, 1st ed. (Cambridge University Press, 2007).

[66] S. E. Thompson, M. Armstrong, C. Auth, M. Alavi, M. Buehler, R. Chau, S. Cea, T. Ghani, G. Glass, T. Hoffman, C.-H. Jan, C. Kenyon, J. Klaus, K. Kuhn, Z. Ma, B. McIntyre, K. Mistry, A. Murthy, B. Obradovic, R. Nagisetty, P. Nguyen, S. Sivakumar, R. Shaheed, L. Shifren, B. Tufts, S. Tyagi, M. Bohr, and Y. El-Mansy, Ieee Trans. Electron Devices **51**, 1790 (2004).

[67] D. Csontos, P. Brusheim, U. Zülicke, and H. Q. Xu, Phys. Rev. B **79**, 155323 (2009).

[68] F. Boxberg and J. Tulkki, Reports Prog. Phys. **70**, 1425 (2007).

[69] H. Lipsanen and M. Sopanen, in *Opt. Quantum Dots Wires*, edited by G. W. Bryant and G. S. Solomon (Artech House, 2004), p. 547.

[70] C. Euaruksakul, Z. W. Li, F. Zheng, F. J. Himpsel, C. S. Ritz, B. Tanto, D. E. Savage, X. S. Liu, and M. G. Lagally, Phys. Rev. Lett. **101**, 147403 (2008).

[71] B. E. Kane, Nature **393**, 133 (1998).

[72] H. Huebl, A. R. Stegner, M. Stutzmann, M. S. Brandt, G. Vogg, F. Bensch, E. Rauls, and U. Gerstmann, Phys. Rev. Lett. **97**, 166402 (2006).

[73] L. Dreher, T. A. Hilker, A. Brandlmaier, S. T. B. Goennenwein, H. Huebl, M. Stutzmann, and M. S. Brandt, Phys. Rev. Lett. **106**, 037601 (2011).

[74] Y. Sun, S. E. Thompson, and T. Nishida, J. Appl. Phys. **101**, 104503 (2007).

[75] C. D. Akyüz, A. Zaslavsky, L. B. Freund, D. A. Syphers, and T. O. Sedgwick, Appl. Phys. Lett. **72**, 1739 (1998).

[76] P. G. Evans, D. E. Savage, J. R. Prance, C. B. Simmons, M. G. Lagally, S. N. Coppersmith, M. A. Eriksson, and T. U. Schülli, Adv. Mater. **24**, 5217 (2012).

[77] S. Horiguchi, M. Nagase, K. Shiraishi, H. Kageshima, Y. Takahashi, and K. Murase, Jpn. J. Appl. Phys. **40**, L29 (2001).

[78] Y. Ono, K. Yamazaki, M. Nagase, S. Horiguchi, K. Shiraishi, and Y. Takahashi, Solid-State Electron. **46**, 1723 (2002).

[79] K. Shiraishi, M. Nagase, S. Horiguchi, H. Kageshima, M. Uematsu, Y. Takahashi, and K. Murase, Phys. E Low-Dimens. Syst. Nanostructures **7**, 337 (2000).

[80] H. Sellier, G. P. Lansbergen, J. Caro, S. Rogge, N. Collaert, I. Ferain, M. Jurczak, and S. Biesemans, Appl. Phys. Lett. **90**, 073502 (2007).

[81] D. Maier-Schneider, A. Köprülülü, S. B. Holm, and E. Obermeier, J. Micromechanics Microengineering **6**, 436 (1996).

[82] A. Shintani and H. Nakashima, Appl. Phys. Lett. **36**, 983 (1980).

[83] M. Xiao, M. G. House, and H. W. Jiang, Phys. Rev. Lett. **104**, 096801 (2010).

[84] M. F. Gonzalez-Zalba, D. Heiss, G. Podd, and A. J. Ferguson, Appl. Phys. Lett. **101**, 103504 (2012).

[85] M. J. Madou, *Fundamentals of Microfabrication and Nanotechnology, Third Edition, Three-Volume Set*, 3rd ed. (CRC Press, 2011).

[86] C. W. J. Beenakker and H. van Houten, in *Solid State Phys.*, edited by Henry Ehrenreich and David Turnbull (Academic Press, 1991), pp. 1–228.

[87] W. Lu and C. M. Lieber, J. Phys. Appl. Phys. **39**, R387 (2006).

[88] J. Salfi, S. Roddaro, D. Ercolani, L. Sorba, I. Savelyev, M. Blumin, H. E. Ruda, and F. Beltram, Semicond. Sci. Technol. **25**, 024007 (2010).

[89] Z. Zhong, Y. Fang, W. Lu, and C. M. Lieber, Nano Lett. **5**, 1143 (2005).

[90] F. A. Zwanenburg, C. E. W. M. van Rijmenam, Y. Fang, C. M. Lieber, and L. P. Kouwenhoven, Nano Lett. **9**, 1071 (2009).

[91] S. Bhargava, H.-R. Blank, V. Narayanamurti, and H. Kroemer, Appl. Phys. Lett. **70**, 759 (1997).

[92] M. Scheffler, S. Nadj-Perge, L. P. Kouwenhoven, M. T. Borgström, and E. P. A. M. Bakkers, Phys. E Low-Dimens. Syst. Nanostructures **40**, 1202 (2008).

[93] F. Léonard and A. A. Talin, Nat. Nanotechnol. **6**, 773 (2011).

[94] F. Léonard and A. A. Talin, Phys. Rev. Lett. **97**, 026804 (2006).

[95] P. Dahl Nissen, T. Sand Jespersen, K. Grove-Rasmussen, A. Márton, S. Upadhyay, M. Hannibal Madsen, S. Csonka, and J. Nygård, J. Appl. Phys. **112**, 084323 (2012).

[96] S. Luryi, Appl. Phys. Lett. **52**, 501 (1988).

[97] Y. Hu and S. Stapleton, Appl. Phys. Lett. **58**, 167 (1991).

[98] T. Duty, G. Johansson, K. Bladh, D. Gunnarsson, C. Wilson, and P. Delsing, Phys. Rev. Lett. **95**, 206807 (2005).

[99] D. Culcer, Ł. Cywiński, Q. Li, X. Hu, and S. Das Sarma, Phys. Rev. B **82**, 155312 (2010).

[100] D. Culcer, Ł. Cywiński, Q. Li, X. Hu, and S. Das Sarma, Phys. Rev. B **80**, 205302 (2009).

[101] N. Neophytou, A. Paul, M. S. Lundstrom, and G. Klimeck, Ieee Trans. Electron Devices **55**, 1286 (2008).

[102] G. Agostini and C. Lamberti, editors, *Characterization of Semiconductor Heterostructures and Nanostructures*, 1st ed. (Elsevier Science, 2008).

[103] D. L. Logan, *A First Course in the Finite Element Method: 3rd (Third) Edition* (CL Engineering, 2001).

[104] S. Timoshenko, J. Opt. Soc. Am. **11**, 233 (1925).

[105] S. D. Brotherton, T. G. Read, D. R. Lamb, and A. F. W. Willoughby, Solid-State Electron. **16**, 1367 (1973).

[106] S. P. Timoshenko and J. M. Gere, *Theory of Elastic Stability*, 2nd ed. (Dover Publications, 2009).