The Dissertation Committee for Dharmendar Reddy Palle
certifies that this is the approved version of the following dissertation:

Modeling of Graphene-based FETs for Low Power Digital Logic and Radio Frequency Applications

Committee:

Leonard F. Register, Supervisor

Sanjay K. Banerjee

Emanuel Tutuc

Arjang Hassibi

Allan H. McDonald

Gary D. Carpenter

Modeling of Graphene-based FETs for Low Power Digital Logic and Radio Frequency Applications

by

Dharmendar Reddy Palle, B.Tech., M.S.E.

DISSERTATION

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT AUSTIN

August 2013

Dedicated to my mother Shyamala.

Acknowledgments

I began my long journey aspiring for a doctorate degree in the second year of my undergraduate at IIT Kanpur. I have had many gurus who helped me hone my ability to think critically. Unfortunately, I have interacted with most of them only indirectly via their excellent books. Among others, there are three very important gurus who I worked with very closely. I must thank Dr. Shyam Sunder who was my physics tutor for the pre-undergraduate classes and has been very influential in my developing a strong interest in physics and my IIT selection. My baby steps in research took place in Dr. Samares Kar's lab at IIT Kanpur. He in my view is a perfectionist who expected a similar work from the small tasks given to me. In the two years of working with him, I must thank him for helping me learn to pay attention to finer details and a liking towards research in solid state electronics.

The next guru is Dr. Frank Register, my PhD Adviser at UT. I have learnt many subjects from reading books but sometimes the learning and the knowledge transfer that occurs from a direct interaction with a teacher is impossible by other means. I must thank Frank for being such a great adviser and teacher, perhaps the best I have had to date. In the five years of working with him I have learnt a lot in terms of technical aspects of subjects. While I may forget the learnt subjects with age but the happy experience in pursuing

my PhD research with frank is definitely unforgettable for me. Finally I thank my parents and sisters for all their support and of course my many roommates who deserve a special thanks for keeping up with me. I believe that a simple thank you is not at all sufficient for all the people mentioned above. In fact, I deliberately keept it simple so that I can never forget that I have not thanked enough and I will always be indebted to them.

I must also specially thank Dr. Banerjee, Dr. Tuutc, Dr. Allan MacDonald, Dr. Hassibi and my mentor at IBM Gary Carpenter who all are my PhD Committee members and I have had opportunity to work and interact with all of them. I also thank Jean Toll for the excellent support she provides to all off the MER students.

I thank NRI, SWAN, DARPA CERA program and Intel for financial support for my research work. Also, I thank IBM for the PhD Scholarship for the year 2010. Finally, I thank the Texas Advanced Computing Center for providing the computational facilities.

Modeling of Graphene-based FETs for Low Power Digital Logic and Radio Frequency Applications

Publication No. _____

Dharmendar Reddy Palle, Ph.D.
The University of Texas at Austin, 2013

Supervisor: Leonard F. Register

There are many semiconductors with nominally superior electronic properties compared to silicon. However, silicon became the material of choice for MOSFETs due to its robust native oxide. With Moore's observation as a guiding principle, the semiconductor industry has come a long way in scaling the silicon MOSFETs to smaller dimensions every generation with engineering ingenuity and technological innovation. As per the 2012 International Technology Roadmap for Semiconductors (ITRS), the MOSFET is expected to be scaled to near 6 nm gate length by 2025. However, materials, design and fabrication capabilities aside, basic physical considerations such as source to drain quantum mechanical tunneling, channel to gate tunneling, and thermionic emission over the channel barrier suggest an end to the roadmap for CMOS is on the horizon. The semiconductor industry is already aggressively looking for the next switch which can replace the silicon FET in the long term. My Ph.D. research is part of the quest for the next switch.

The promises of process compatibility with existing CMOS technologies, fast carriers with high mobilities, and symmetric conduction and valence bands have led to graphene being considered as a possible alternative to silicon. This work looks at three devices based on graphene using first principles atomistic transport simulations and compact models capturing essential physics: the large-area graphene RF FET, the Bilayer pseudoSpin FET, and the double electron layer resonant tunneling transistor. The characteristics and performance of each device is explored with a combination of SPICE simulations and atomistic quasi static transport simulations. The BiSFET device was found to be a promising alternative to CMOS due to extremely low power dissipation. Finally, I have presented formalism for efficient simulation of time dependent transport in graphene for beyond quasi static performance analysis of the graphene based devices explored in this work.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction and background

1.1 Introduction

Creating, processing, storing and communicating information has been an integral part of intelligent human life since time immemorial. The representation of our expressions and thoughts which naturally occur as gestures and sounds has moved to higher levels of abstraction. For example, from encoding information into a set of alphabets in English to 0s and 1s used in computers. Our never ending quest in understanding how nature works led to orders of magnitude increase in sophistication and abstraction in processing information. At the heart of such remarkable progress is our current understanding of the quantum aspects of nature. The innovation of semiconductor-based transistor nearly 50 years ago enabled the information technology revolution that has a significant positive impact on many aspects of our lives.

The silicon based Metal Oxide Semiconductor Field Effect Transistor (MOSFET) has been and it still is the workhorse of semiconductor industry. A field effect transistor (FET) is an electronic device with at least three terminals where the conductivity between two terminals, the source and drain, is modulated by the electric field created by the third (etc.) gate terminal(s).

This gate control enables the use of a FET as switch. The switch is on and conducts high current from drain to source when the gate to source voltage, V_{gs}, is greater than a threshold voltage, V_{th}. The switch is considered off in the sub-threshold region when V_{gs} is less than V_{th}, and the current is sufficiently low. Physically, this control is achieved by gate modulation of the free carrier density in the channel between the source and drain contacts from a high value in the ON state to a low value in OFF state. The switches can be used to implement Boolean logic gates such as NOT, OR and AND etc. These basic gates can be composed to form functional blocks to implement various electronic products used for information storage and processing.

To promote market growth for electronic products it is desirable to have more functionality per chip at lower cost. The path followed by the industry to date is roughly governed by the Moore's observation in 1975 that the number of transistors on an integrated chip doubles every two years. Transistors must be scaled to smaller sizes to increase the transistor density which helps in adding more functionality per chip at lower cost. Dennard provided the formal basis for defining the scaling laws(constant field) for the early n-MOS technology [15]. For constant field scaling the geometrical variables such as gate length, oxide thickness and junction depths are decreased by a factor of α. To keep the electric fields in the channel and across the gate oxide constant, the supply voltage is also decreased by the same factor of α. Such scaling laws provide guidelines for defining a smaller transistor with mostly improved or at least similar performance for the next technology node.

The ratio of current in the ON state and that in OFF sate, referred to as a ON/OFF ratio, is a very important performance metric of digital switches, and the larger the better. A large on current, I_{on}, allows for quick charging of a capacitive load, which typically includes the gates of one or more subsequent transistors as well as the metallic interconnects. Considering only gate capacitance as the load, the switching time is approximately C_gV/I_{on}, where C_g is the gate capacitance and V is the supply voltage. Under constant field scaling, for a long channel MOSFET and assuming gradual channel approximation is valid, the on current decreases by a factor of α. Since the oxide thickness and supply voltage is also scaled down by α, the switching time decrease by a factor of α. In other words, the intrinsic circuit speed is expected to increase by a factor of α at constant switching power density at every generation. However, the current in the sub-threshold region of the device varies as $I_{sub} \propto exp(-(V_{gs} - V_{th})/k_BT)$ where k_B is the Boltzmann's constant and T is the temperature. The off current, I_{sub} at $V_{gs} = 0$ should be kept low for lower static or leakage power. The leakage current increase exponentially under constant field scaling as threshold voltage must be scaled by same actor as the supply voltage. Furthermore, gate leakage current which has an exponential dependance on the oxide thickness also can contribute significantly to the leakage power when the physical gate oxide thickness is scaled to sub 2nm. A better representation of the sub-threshold current, valid for short channel device as well, is $I_{sub} \propto 10^{-V_{th}/S}$ where S is the sub-threshold slope defined as the gate voltage change required for a decade of increase in

drain current. The sub-threshold slope $S = 2.3\frac{k_B T}{q}\frac{\partial \phi_{barrier}}{\partial V_g}$, where $\phi_{barrier}$ is the potential barrier between source and drain. In the short channel regime, the drain current is due to carrier flow over the gate controlled barrier near source. For an FET with current due to over the barrier carrier flow, the best possible sub-threshold slope is about 60 mV/decade at 300 K achievable under a very strong electrostatic control of the barrier height by the gate voltage. Due to the exponential dependance of the sub-threshold current on V_{th} and the lower bound on sub-threshold slope, $2.3k_BT/q$, which in-turn results from the Boltzmann distribution of the carrier occupation, the threshold voltage cannot be scaled to an arbitrarily small value. Historically, the industry used both constant field and constant voltage scaling and currently a scaling where both the supply voltage and electric fields are scaled every generation. The scaling rules were adjusted to achieve near doubling of transistor on chip and improved circuit performance indicated by up to a 30 % increase in intrinsic circuit speed at every node transition (roughly 18 months). Due to the limitations of threshold voltage scaling, the supply voltage scaling has slowed, for e.g., by a factor of 0.7 to 0.9 for the sub 100 nm node transistors. While the scaling of the transistor geometry continued, the leakage power contribution to the total chip power began to increase.

To address the increasing power per performance gain, clock frequency of the chips were kept low and performance was boosted by exploiting mutlti-thread/multi-core architecture for chips. Where as at the device level, performance improvement is becoming increasingly difficult to achieve due to var-

ious short channel effects such as mobility degradation, drain induce barrier lowering, band to band tunneling etc. Furthermore, the source/drain parasitic resistance has also become a bottleneck to achieve higher drive current by gate length scaling. Despite various challenges posed by the physics of short channel devices, engineering ingenuity and technological improvements have kept the device scaling going with products using 22nm node transistors already in market this year. For example, the industry has used high-k metal gate stacks to reduce the leakage currents in sub 2nm effective oxide thickness devices and structural changes such as using silicon on insulator, multigate (e.g., FinFETs) for tighter electrostatic control. Larger on currents were achieved by improving carrier velocities via carrier mobility and thermal velocity improvements via a range of methods such as channel material choice, straining the channel material, modulation doping, etc. Salicided contacts are used to improve contact resistance.

As per the 2012 ITRS roadmap device scaling is targeted to 6 nm physical gate length by 2025 [16]. What is the path beyond this point ? Introduction of novel device concepts for beyond 5 nm nodes in commercial products may requires at least 15 to 20 years of research and development. To continue the power and performance gains, in the near term, the possible solutions can be based on evolutionary approaches to further the performance scaling of MOSFETs to its physical limits by replacing silicon with alternate channel materials such as III-V or Germanium as channel material of FinFETs or nanowire FETs. Graphene is another channel material which is attractive

due to very high carrier mobility, high current carrying capacity, its planar geometry and process compatibility with conventional silicon technology.

However, materials, design and fabrication capabilities aside, basic physical considerations such as source to drain quantum mechanical tunneling, channel to gate tunneling, and thermionic emission over the channel barrier suggest an end to the roadmap for CMOS is on the horizon. The semiconductor industry is already aggressively looking for the next switch which can replace the silicon FET [17] in the long term. To overcome the threshold voltage scaling limitations due to thermal barrier, tunnel FETs may be used for sub 0.5 V supply voltage devices. While the sub 60 mV/decade sub-threshold slope theoretically achievable in tunnel FETs is attractive, tunnel FETs have low drive current (best values of the order of less than 0.5E-2 mA/micron). It is not clear yet that both sufficiently low power and high performance will be possible in a combination to challenge CMOS. Going further, revolutionary solutions could be provided by devices based on alternate state variables, such as spin, material phase and pseudospin etc, using novel switching mechanisms. For example, spin of the electron is used as a state variable of logic in so called spintronics. Due to long spin precession times possible in graphene, graphene could prove valuable for spintronics as a channel material as well. Yet another example is the exotic device concept of the Bilayer pseudoSpin Field Effect Transistor (BiSFET) [18] which is based on the theoretically predicted possibility of room temperature superfluidity in two adjacent layers of graphene, where "pseudospin" refers to the which layer degree of freedom here

My Ph.D. research is part of the quest for the next switch. This work looks at three devices based, at least initially, on graphene using first principles atomistic transport simulations and compact models capturing essential physics. The high current capacity of the graphene is attractive for analog/radio frequency applications. In Chapter 2, I discuss a hardware correlated compact model for a large area graphene FET intended for RF applications. In Chapter 3, I address the potential performance of the BiSFET through Spice-level circuit simulation. The essential physics of the device is captured in compact models, and plausible benefits of the BiSFET are accessed via figures of merits of basic Boolean logic gates such as inverter and NAND. In Chapter 4, I re-evaluate a device concept which has been called Double Electron Layer Tunneling Transistor [19] and which we refer to as an Interlayer Tunnel Field Effect Transistor (ITFET), with graphene and III-V quantum wells as the electron layers with sub 100 nm channel length. I use both analytic models and non-equilibrium Green's function (NEGF) calculations to identify the essential physics of these short channel devices. In Chapter 5, I provide a framework for time dependent quantum transport in graphene based devices that will be useful for understanding the intrinsic speed limitations of e.g., Graphene RF FETs, ITFETs and even the BiSFET, which beyond its critical voltage, also could act as a terahertz source . Finally, in Chapter 6, I provide the conclusion of my research work and possible future directions. In the rest of this chapter I discuss the relevant electronic properties of Graphene.

1.2 Introduction to Graphene[1]

Carbon has various crystalline allotropes such as diamond, graphite, graphene, nanotubes and Buckminsterfullerenes. The carbon atoms are sp^3 hybridized in diamond and sp^2 hybridized in rest of these materials. Graphene is a two dimensional network of carbon atoms arranged on a honeycomb lattice. A stack of graphene sheets bound by weak van der Waals forces forms graphite. A graphene sheet rolled into a tube is carbon nanotube. A graphene sheet with at least 12 pentagonal defects results in fullerene. Graphene was not known to exist in isolated form until 2004 when it was first isolated by Novoselov's group in Manchester [20]. Ever since 2004 graphene has become a play ground for researchers in observing interesting physical phenomenon such as anomalous Integer Quantum Hall Effect [21, 22], Klein tunneling [23], and even in predictions of possibility of exotic states such as Bose condensates at room temperature in bilayer graphene [24]. All of these phenomena are a result of graphene's 2D lattice structure and consequent electronic properties: a zero band gap semiconductor, i.e., a semi-metal, whose low energy quasi particles are chiral, massless Dirac fermions with large Fermi velocities [21–23, 25]. Semiconducting materials with such fast carriers are very attractive for use in high frequency circuit applications [26–28].

Graphene with its exceptional electronic properties represents a possible alternative to silicon. Although carbon nanotubes(CNTs) which offer

[1]Text reproduced with permission from IOP Publishing: Journal of Physics. D [1], copyright (2011).

similar electronic properties have been considered as a replacement for silicon, graphene with its planar geometry can be processed with more conventional complementary metal oxide semiconductor (CMOS) technology giving it a significant advantage over CNTs. However, a major drawback for such use of bulk graphene is its zero band gap, which makes the graphene-based FETs hard to switch off. Alternatives to circumvent this problem are to use nanoribbons of graphene or bilayer graphene, both of which can have a band gap under the right conditions.The solutions can also be revolutionary where alternative switching mechanisms in graphene based devices can be used. For example, due to long spin relaxation times for electrons in graphene ribbons, it can be used in spintronics-based devices, where the spin of the electron is used as a state variable for logic [29, 30].

Owing to the large mobility of carriers and high current carrying capability (about 10^8 A/cm2) [31, 32], graphene may also be useful for application in radio frequency (RF) circuits. Today, the best performing circuits in this field employ III-V based high electron mobility transistors (HEMTs). However these technologies are expensive and not as scalable as Si, whereas graphene processing appears compatible with traditional CMOS technologies. Also, because of its planar geometry, it is the ultimate thin body channel material and is less likely to suffer from performance degradation due to scaling. Finally, a combination of graphene's unique properties and process compatibility with existing CMOS technologies gives hope for the possibility of all-graphene-based or hybrid graphene/silicon high performance and high density system on chips

in the future.

1.2.1 Electronic Properties : Monolayer, Bilayer and Nanoribbons

In this section we will discuss the electronic structure of graphene and the transport properties relevant to FETs. An extensive review of various electronic properties of graphene is given in refs [4, 33]. Carbon has a ground state electronic configuration $1s^2 2s^2 2p^2$ with 4 valence electrons (2 in 2s sub shell and 2 in 2p sub shell). When forming bonds, the valence shell orbitals hybridize by promoting an electron from 2s sub shell to 2p. The resulting hybrid orbitals are named sp, sp^2 or sp^3 orbitals depending on the number of p-orbitals involved in hybridization. These hybrid orbitals are strongly directional and form sigma (σ) bonds forming the backbone of covalently bonded materials. For example, in graphene, 3 of the 4 valence electrons are present in the three sp^2 hybridized orbitals due to mixing of 2s, $2p_x$ and $2p_y$ atomic orbitals and the 4^{th} electron is present in the unhybridized $2p_z$ orbital. Since, the three sp^2 hybrid orbitals are 120° apart and lie in xy plane, graphene has a planar geometry. Graphene's primitive Bravais lattice is hexagonal, and the primitive unit cell contains two atoms, A and B, as shown in Figure 1.1(a). The carbon atoms on one sub-lattice are connected to three atoms from other sub-lattice via equal length sp^2 hybridized sigma bonds shown by vectors δ_i,$i = 1, 2, 3$ which are 120° apart as, shown in Figure 1.1(a). Graphene's crystal lattice is also referred to as a honeycomb lattice for obvious reasons.

The strong sigma bonds formed by the carbon atoms result in deep

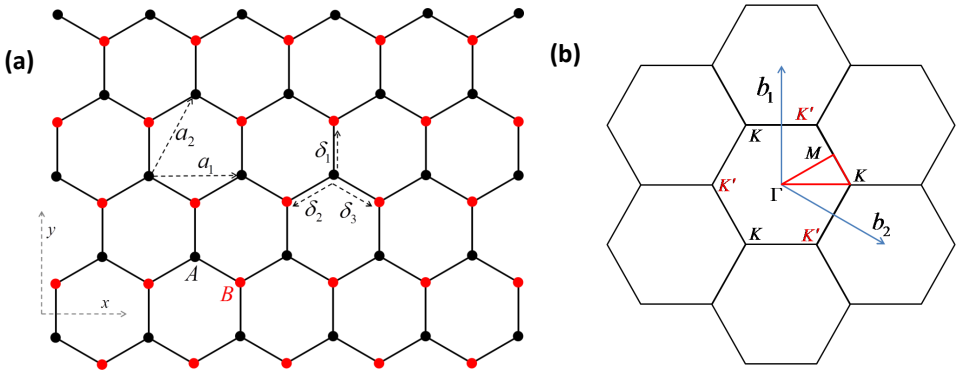

Figure 1.1: Real space lattice of graphene showing the primitive lattice vectors a_1 and a_2 of triangular lattice with lattice constant $a = \sqrt{3}c$, where c=0.142nm is the carbon carbon bond length. (b) Reciprocal lattice of graphene showing lattice vectors b_1 and b_2 and first Brillouin zone with high symmetry points , M and Dirac points K and K'. (Reprinted with permission from IOP Publishing: Journal of Physics. D [1], copyright (2011).)

lying valence bands of graphene's electronic structure, and are responsible for the excellent mechanical properties of this material. The electrons in the unhybridized $2p_z$ orbitals, which are responsible for the optical and electronic properties of graphene, delocalize along the plane of the graphene surface to form π(bonding) and π^* (anti-bonding) bands. Graphene's electronic properties can be reasonably described by the low energy Hamiltonian (see Appendix D for details)

$$h\left(\mathbf{k}\right) = -\begin{pmatrix} 0 & f\left(\mathbf{k}\right) \\ f^*\left(\mathbf{k}\right) & 0 \end{pmatrix} \tag{1.1}$$

where

$$f\left(\mathbf{k}\right) = t\left(e^{i\mathbf{k}\cdot\delta_1} + e^{i\mathbf{k}\cdot\delta_2} + e^{i\mathbf{k}\cdot\delta_3}\right) \tag{1.2}$$

The Hamiltonian in Eq. (1.1) can be diagonalized to obtain the spectrum $E(\mathbf{k}) = \pm|f(\mathbf{k})|$ which is shown in Figure 1.2. The momentum vector $\mathbf{k}$

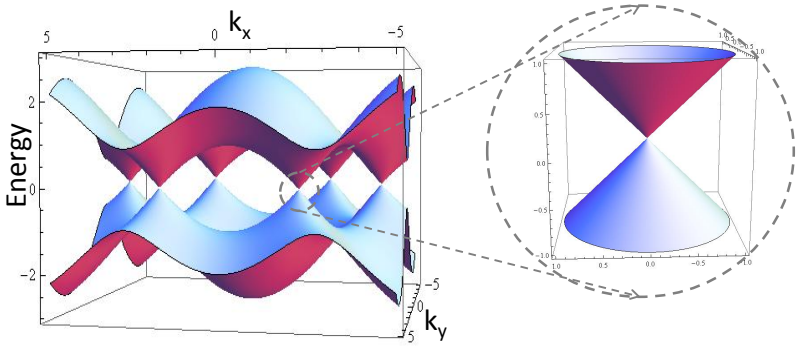

Figure 1.2: left: Tight binding based band structure of graphene (axes have arbitrary units), right: Low energy band structure near Dirac point. (Adapted with permission from IOP Publishing: Journal of Physics. D [1], copyright (2011).)

belongs to first Brillouin zone of graphene reciprocal lattice shown in Figure 1.1(b). The high symmetry points $\mathbf{K}$ and $\mathbf{K}'$ are referred to as Dirac points because of electrons near the points resemble Dirac Fermions, although much slower. Most of the transport properties of graphene are determined by the nature of spectrum near these points where the energy spectrum is given by $E = \pm v_F \hbar \mathbf{q}$, where v_F is the Fermi velocity, the wave vector $\mathbf{q}$ is referenced to the Dirac points, and the negative energy branch is the valence band and the positive energy band is the conduction band, which has been confirmed based on angle-resolved photo emission spectroscopy [25]. The velocity is $v_F = 3tc/(2\hbar) \approx 1.1 \times 10^8$ cm/s consistent with a nearest neighbor hopping potential of t =-2.7 eV. Since the $2p_z$ orbital of each carbon atom has one electron, the valence band is completely filled and the conduction band empty. That is, graphene has half filled bands. Also, because the the bands touch at the Dirac points, graphene is a zero band gap semiconductor, i.e, a semimetal.

From Eq. (1.1), one obtains the low energy Hamiltonian

$$H_{LE}(\mathbf{q}) = \frac{3tc}{2}\sigma \cdot \mathbf{q} \tag{1.3}$$

near the Dirac point at $\mathbf{K} = \left(\frac{4\pi}{3a}, 0\right)$ with $\sigma = (\sigma_x, \sigma_y)$ where $\sigma_{x/y}$ are the Pauli matrices. Near the other Dirac point at $\mathbf{K}'$, $\sigma = (\sigma_x, -\sigma_y)$. One can obtain the continuum approximation for H_{LE} by substituting $k_x = -i\partial_x$ and $k_y = -i\partial_y$,

$$H = -i\hbar v_F \sigma \cdot \nabla \tag{1.4}$$

We see that the above equation is similar to 2D Dirac equation with zero mass but with the Fermi velocity instead of velocity of light. The Pauli matrices in the Eq. (1.4) represent one form of what can be termed as "pseudospin" which does not correspond to real spin but to the two sub-lattice degrees of freedom present in graphene. (Note that for the BiSFET name, "pseudospin" refers to the "which layer"' degree of freedom, and not the sub-lattice pseudospin, although the latter also affects BiSFET properties.) The pseudospin does not merely represent a formal mapping of a system with two degrees of freedom into the language of spin operators; the pseudospin has observable consequences. For example, for low energies, as long as the approximate Hamiltonian in Eq. (1.3) is valid, pseudospin is a good quantum number. The eigenstates of the low energy Hamiltonian are also the eigenstates of the helicity operator defined as $h = \sigma \cdot \mathbf{q}/q$ which has only two eigenvalues ± 1. The $+1$ and -1 eigenvalue corresponds to a state whose pseudospin is parallel and anti-parallel to its momentum, respectively. Pseudospin in graphene is illustrated in Figure 1.3. One of the major consequences of the pseudospin of

the carriers in graphene is perfect "Klein tunneling" between conduction and valence band, which can be described as due to the suppression of intra-valley back scattering by pseudospin conservation requirements. An electron with normal incidence on a potential step, if it undergoes back scattering, must go from State 1 to State 2 as shown in Figure 1.3. However, since the pseudospin of State 1 is anti-parallel to the pseudospin of State 2 such scattering events are forbidden unless the scattering with the potential barrier can induce pseudospin flip. Such spin flip can happen only in case of potential steps that are abrupt on scale of the primitive unit cell. Interaction of an electron with short range potential which can transfer large momentum, such as deformation potential phonon scattering, can result in allowed inter-valley back scattering from State 2 to State 3 as shown in Figure 1.3.

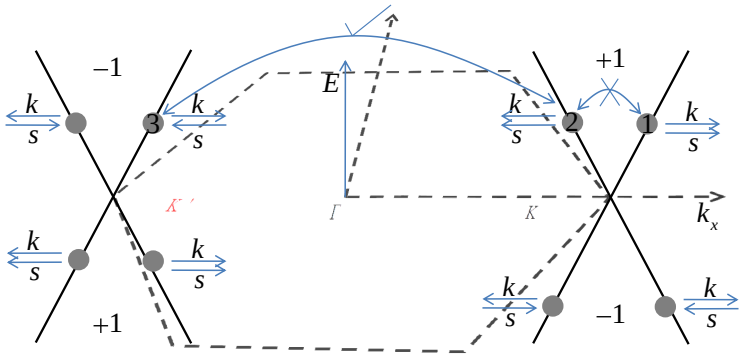

Figure 1.3: Schematic showing the low energy linear dispersion at the two inequivalent valleys in the first Brillouin zone of graphene. Each valley has two bands which are marked by the chirality of the carriers in that band. The small arrows at the dots show the direction of the momentum k and pseudospin s. Also shown are the forbidden intra-valley backscattering from 1 to 2 and allowed inter valley back scattering from 2 to 3. (Reprinted with permission from IOP Publishing: Journal of Physics. D [1], copyright (2011).)

As we have seen above, a simple tight binding description of monolayer graphene gives a good insight into the electronic properties of graphene. This was studied by Wallace to understand the electronic properties of graphite which is a 3D allotrope of carbon consisting of a large number of stacked graphene layers [34]. The graphene layers in graphite are bound by the van der Waals like interaction between the graphene layers. Such stacking can have a significant effect on the electronic structure of the resulting material. In graphite, the stacks are usually oriented such that only half of the carbon atoms of one layer overlap with carbon atoms of next layer. For example, consider first two layers such that the sub-lattices are named A1 and B1 in Layer 1, and A2 and B2 in Layer 2. Now consider that Layer 2 stacked on Layer 1 has the position shown in Figure 1.4(a) (adapted from ref [2]) where the A2 atoms are exactly above B1 atoms in, and B2 atoms are located above the centers of the hexagons in Layer 1. We may call say that Layer 1 has Positioning 1 here, and that Layer 2 has Positioning 2. A third layer could be added that precisely aligns with the first, that is with Positioning 1 such that A3 atoms are directly above A1 and B3 atoms are directly above B1 atoms. However the third layer could also be added such that the the A3 atoms are directly above B2 atoms instead, which we may refer to as Positioning 3. Stacking order where Layer one has Positioning 1 and Layer 2 has Positioning 2 is called Bernal stacking. When a large number of Bernal stacked bilayers are stacked in the pattern 121212... we get the most naturally occurring bulk graphite. Other types of graphite with a stacking order 123123123 ... (rhombohedral stacking) and

graphite with no discernible stacking order, known as turbostatic graphite, have also been observed. Bilayer graphene can also have another configuration, usually referred to as hexagonal or AA stacking, where both layer have same Positioning, i.e., all atoms of one layer are exactly above atoms of another layer.

Like monolayer graphene, bilayer graphene also has interesting electronic properties. For example, from device application point of view, Bernal stacked bilayer graphene is a zero band gap semiconductor but with parabolic bands, at least with a zero interlayer field/potential energy difference, as illustrated by solid lines in Figure 1.4(b). Interestingly, with application of an interlayer field, a bandgap can be opened up in bilayer graphene producing a semiconductor rather than the semimetal of bulk mono-layer graphene as illustrated by dash-dot lines in Figure 1.4(b). However opening a band gap in bilayer graphene requires large displacement fields of the order of 2 to 3 V/nm to open band gaps up to 200 meV as shown by the experimental data in Figure 1.4(c). The band gap in Figure 1.4(c) is obtained using infrared microscopy measurements on graphene bilayer FET [3].

Although graphene has atomic dimensions perpendicular to its plane, it can have large dimensions in plane and, thus, can be considered as a bulk material in plane. Much as for conventional semiconductors, one can then prepare nanostructures of graphene by reducing the extent in one or both of the in-plane dimensions with electronic and optical properties that are different from the bulk. Graphene nanostructures with confinement in one dimension

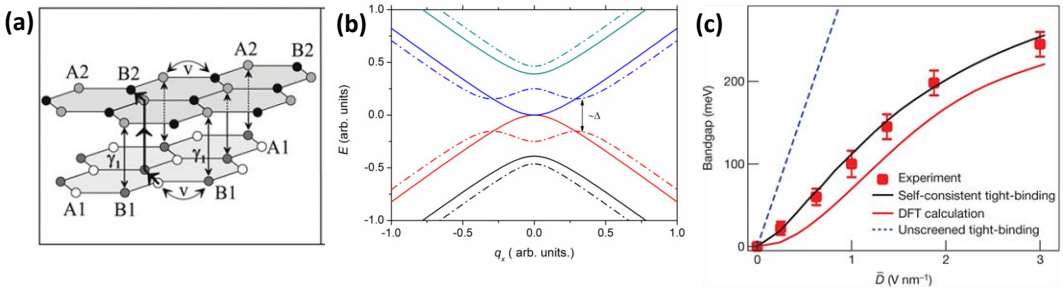

Figure 1.4: (a) Schematic of the bilayer lattice containing four atoms per unit cell: A1 (white circles) and B1 (grey circles) in the bottom layer, and A2 (grey circles) and B2 (black circles) in the top layer (figure reprinted from [2] with permission from Elsevier, copyright (2007)). (b) Unscreened tight-binding based bandstructure of a bilayer graphene near Dirac point along $\mathbf{q}_x$-direction with zero potential difference between layers (solid lines) and non zero potential difference between layers (dash-dot lines). (c) Electric-field dependence of tunable energy bandgap in graphene bilayer. Experimental data(red squares) are compared to theoretical predictions based on self-consistent tight-binding (black trace), ab-initio density functional (red trace),and unscreened tight-binding calculations (blue dashed trace).The error bar is estimated from the uncertainty in determining the absorption peaks in the spectra (reprinted by permission from Macmillan Publishers Ltd: Nature [3] copyright(2009)).

are known as Graphene nanoribbons (GNRs), whereas confinement in both directions results in graphene quantum dots. Depending on the configuration of the carbon atoms on the graphene edge parallel to the large spatial extent, graphene nanoribbons are classified as arm-chair (ac) or zig-zag (zz) nanoribbons. The two types of nanoribbons are shown in Figure 1.1(a). In Figure 1.1(a), if the transport direction is along the x-direction and confinement along the y-direction, then it is a zig-zag nanoribbon, named after the zig-zag arrangement of edge carbon atoms along x direction. Similarly if the transport direction is along y and confinement along x, it is called an arm chair nanoribbon. Theoretically, the energy gap of GNR depends on the width and crystallographic orientation of the graphene nanoribbon [35, 36]. As one can see from the band structure in Figure 1.5 zig-zag nanoribbons are metallic and have edges states (at least in the tight-binding model with no edge reconfiguration). Where as armchair GNRs can be metallic or semiconducting depending on the width of GNR. This implies, a precise control on atomic scale of the edge roughness is required to realize a semiconducting GNR. Moreover as can be seen from Figure 1.6(a), to obtain band gaps of reasonable values ($>$ 100 meV) requires patterning of graphene ribbons of widths less than 10 nm [5]. The band gaps were obtained by the measurement of conductance across lithographically patterned graphene nanoribbons of varying widths and crystallographic orientations. Note that the empirical scaling shown by the dotted line has the form $E_g \propto (W - W_o)^{-1}$, with W_o about 16 nm suggesting smaller effective width of the actual graphene ribbon. Also, the lack of direc-

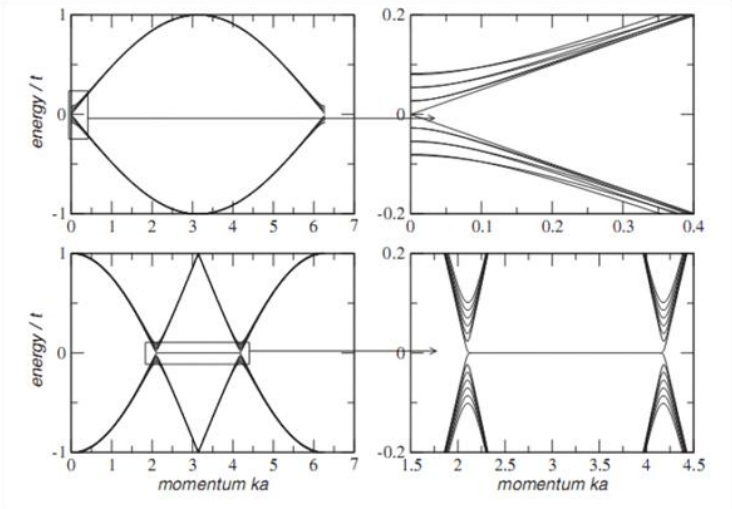

Figure 1.5: Electronic dispersion of graphene nanoribbons. Left: energy spectrum as calculated from the tight-binding equations, for a armchair-nanoribbon (top) and zig-zag nanoribbon (bottom). The width of nanoribbon is N=200 unit cells. Only 14 eigenstates are depicted. Right: zoom of the low energy states shown on the right (figure reprinted with permission from [4] Copyright (2009) by the American Physical Society).

tional dependence of the gap indicates that the GNR edges are not atomically precise. Also note that there are considerable error bars on these results. This variations is consistent with theoretical predictions that indicate that edge roughness can significantly degrade the performance of graphene nanoribbon based FETs [6, 37] and induce device to device variation [6], as illustrated by Figure 1.6(b).

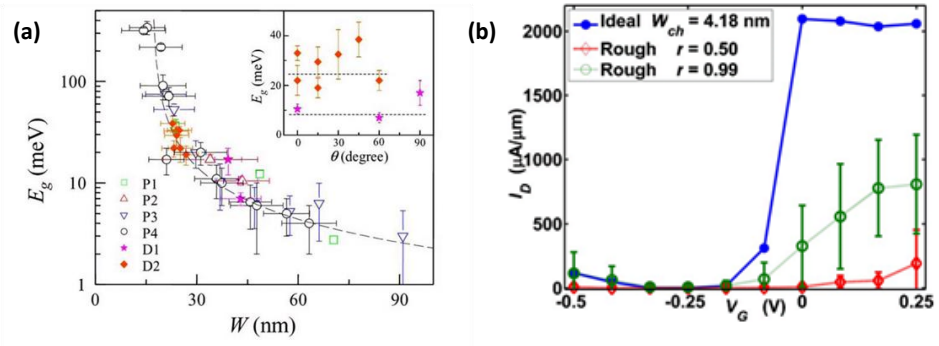

Figure 1.6: (a) Conductance measurement based band gap vs width for 6 devices sets (P1 P4 and D1-D2). The devices P1-P4 have parallel GNRs with W from 15 nm to 90 nm and devices D1,D2 have similar width but different crystallographic directions. The inset shows E_g vs relative angle θ for the device sets D1 and D2. Dashed lines in the inset show the value of E_g as predicted by the empirical scaling of E_g vs W (figure reprinted with permission from [5], copyright (2007) by the American Physical Society). (b) $I_D - V_G$ of a dual-gate MOSFET with $W_{ch} = 4.18$ nm, and using different values of edge roughness parameter r. Error bars indicate standard deviation in I_D for ten devices having randomly different edges. $V_D = 0.3$ for all these simulations (figure reprinted with permission from [6], copyright (2008) by the American Institute of Physics).

Chapter 2

Large Area Graphene Field Effect Transistor: Compact Model

2.1 Introduction

Large thermal velocities and an ultrathin body make graphene a promising choice for radio frequency (RF) circuit applications [28,38–40]. In fact, RF devices with projected cutoff frequency as high as 210 GHz have already been demonstrated [41]. Graphene FETs are ambipolar due to a zero band gap, which, while problematic for conventional switching, can be used to make interesting circuits, as for the example of the demonstration of a single graphene FET based Mixer [38]. While a technological effort is underway to make better graphene FETs, we believe that having a SPICE model can help in predictive analysis of performance metrics and design tradeoffs of larger graphene based circuits. Also, a hardware correlated model will help us understand the impact of process changes on device and circuit level metrics. To address these needs, this work introduces a surface potential based compact model for single gate Graphene FET in the diffusive transport regime considering both the drift and diffusion contributions to the current. The model describes both n and p regions of device operation continuously.

One can obtain the surface potential which is defined as the conduction band referenced Fermi energy by solving the one dimensional Poisson equation perpendicular to the gate. The approach results in an implicit equation which has no closed form solution. To obtain an approximate expression for the surface potential, I solve the implicit equation numerically and fit the resulting solution with an analytic expression. The analytic expression satisfies the implicit equation in the asymptotic limits. Details of this procedure and C-V model are discussed in Section 2.3. The expression for drain current is obtained using the terminal voltage dependent surface potential assuming a gradual channel approximation and a diffusive transport regime. The approximation seems to be reasonable for Graphene FET devices up to 500 nm given that most of the Graphene FET devices reported are not clean enough to observe the long mean free paths expected. The I-V model is discussed in more detail in Section 2.4. Finally in Section 2.5 I provide the analysis and verification of the model.

2.2 Graphene Electrostatics

Graphene is a single 2D network of carbon atoms arranged on a honeycomb lattice. As a consequence of this specific arrangement of carbon atoms, graphene is a zero band gap material whose low energy quasi particles are chiral, massless Dirac fermions. The low energy dispersion is given by, $E(k) = \hbar v_F k$, where the constant $v_F \approx 10^8$ cm/s is the Fermi velocity. Because most graphene FETs are gated only up to 250 meV of Fermi energy,

it is reasonable to use the linear dispersion relation which is expected to be valid for $E < 1$eV. The free carrier density in graphene at a given Fermi level, μ, can be obtained by integrating over density of states,

$$g(E) = \frac{g_s g_v}{(2\pi)} \frac{|E - E_D|}{(\hbar v_f)^2} \tag{2.1}$$

where $g_s = 2$ is the spin degeneracy, $g_v = 2$ is the valley degeneracy and E_D is the energy at the Dirac point. Because of the zero band gap, the conduction band edge E_c and the valence band edge E_v are the same as E_D. In addition, the zero band gap makes graphene a degenerate semiconductor, and the carrier density is given by the Fermi integral of order 1 due to linear dependence of density of states on energy.(In conventional semiconductors like silicon within the parabolic regime, the carrier density is given by Fermi integral of order 1/2). The total charge density as function of Dirac-point-referenced Fermi energy or the surface potential $\zeta = \mu - E_D$ in graphene is given by

$$Q_g(\zeta) = en_i \left(F_1\left(-\frac{\zeta}{k_B T}\right) - F_1\left(\frac{\zeta}{k_B T}\right)\right) \tag{2.2}$$

where $n_i = \frac{g_s g_v}{(2\pi)} \left(\frac{k_B T}{\hbar v_f}\right)^2$ is the intrinsic carrier density at temperature T(at room temperature $n_i = 9.81 \times 10^{10} cm^{-2}$), and F_1 is the Fermi integral (of index 1). See appendix A for details.

2.3 Capacitance Voltage (C-V) Model

Consider a metal-insulator-graphene system for which the band energy diagram at an applied gate bias, V_G, is shown in Figure 2.1(b), where W_m

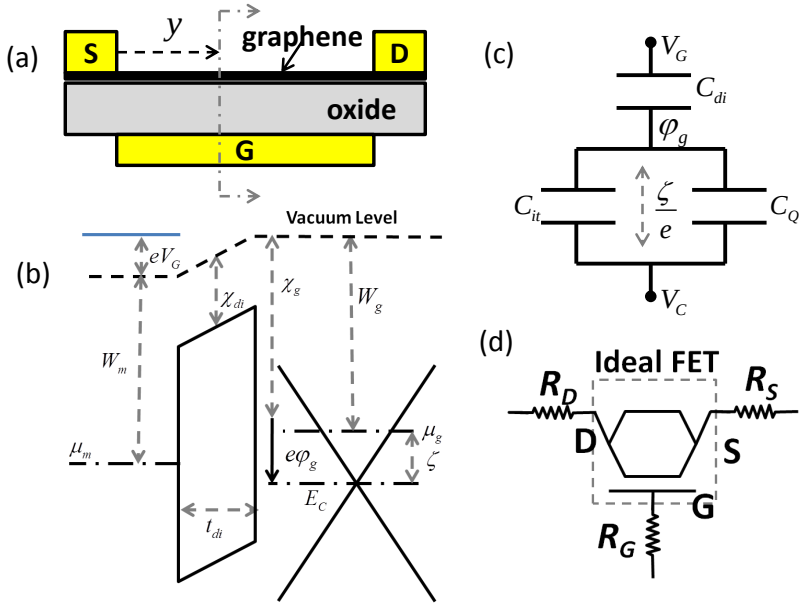

Figure 2.1: (a) Two dimensional schematic of a graphene field Effect transistor, (b) band energy diagram of the graphene FET along the vertical cross section at a distance y from the source shown in (a),(c) Equivalent capacitor circuit for a metal-insulator graphene system, and (d) FET circuit model with ideal FET shown in dashed box and contact resistances lumped into $R_{S/D/G}$

is the metal work function, and χ_g is the electron affinity of graphene. The Fermi level in the graphene is $\mu_g = \zeta + E_c = \zeta - \chi_g - e\varphi_g$ and in the metal is $\mu_m = -W_m - e\varphi_m$, where, again, ζ is the conduction band referenced Fermi energy in graphene and φ_g and φ_m are the electrostatic potentials in graphene and metal respectively. Under an applied gate bias, V_G, the Fermi level in graphene, μ_g, and the Fermi level in metal, μ_m, are related as $\mu_m - \mu_g = -e(V_G - V_C)$ which can be written as

$$-e(V_G - V_C) = -W_m - \zeta + \chi_g + e\left(\varphi_g - \varphi_m\right) \tag{2.3}$$

where V_C is the bias in the channel at location y shown in Figure 2.1(a). The electrostatic potential difference between graphene and metal in the above equation can be obtained by solving a 1D Poisson equation (see appendix B for details). Assuming a fixed charge density Q_F at the interface and an interface trap density Q_{it}, the potential difference is given by,

$$\varphi_g - \varphi_m = \frac{Q_g(\zeta) + Q_F + Q_{it}(\zeta)}{C_{di}} \tag{2.4}$$

where C_{di} is the dielectric capacitance and Q_g is the charge in the graphene given by Eq. (2.2). We can combine Eqs. (2.3) and (2.4) to obtain,

$$e(V_G - V_C) = W_m - \chi_g + \zeta - e\frac{Q_g(\zeta) + Q_F + Q_{it}(\zeta)}{C_{di}} \tag{2.5}$$

The gate voltage at which the graphene is charge neutral can be defined as the flat band voltage V_{FB} which is given by,

$$V_{FB} = \frac{W_m - \chi_g}{e} - \frac{Q_F + Q_{it}(\zeta = 0)}{C_{di}} \tag{2.6}$$

Assuming a constant interface trap density D_{it}, we have an interface trap charge density $Q_{it} = -eD_{it}\zeta$ and an interface trap capacitance $C_{it} = e^2 D_{it}$. Finally we can write the charge voltage relation for a metal-insulator-graphene structure as:

$$C_{di}\left(V_G - V_C - V_{FB} - \frac{\zeta}{e}\right) = C_{it}\frac{\zeta}{e} - Q_g(\zeta) \tag{2.7}$$

The above equation is partitioned such that it reflects the equivalent capacitance circuit for the graphene MIS structure as shown in the Figure 2.1(c). The term on the left hand side of Eq. (2.7) is the charge on the metal gate,

i.e., the charge on the dielectric capacitor in the equivalent circuit shown in Figure 2.1(c). The first term on the right hand side is the net trapped charge and the second term is the net charge available for conduction. In other words, the right hand side terms account for the net charge on the parallel capacitor branch due to an effective potential drop of ζ/e across it. By scaling the variables in Eq. (2.7), we can write it in dimensionless form as,

$$x = z\left(1 + C_1\right) - C_0\left(F_1\left(-z\right) - F_1\left(z\right)\right) \tag{2.8}$$

where $x = \left(V_G - V_{FB} - V_C\right)/V_{th}$, $z = \zeta/k_BT$, $C_1 = C_{it}/C_{di}$, $C_0 = en_i/C_{di}V_{th}$ and $V_{th} = k_BT/e$. The solution z of the above equation in the asymptotic limit of small x is $\propto x$ and in the limit of large x is $\propto \sqrt{x}$. Based on these limits, I fit the numerical solution of Eq. (2.8) with the form $z = ax\left(b + \sqrt{|x|}\right)^{-1}$, where a and b are fitting parameters which are fixed for a given combination of C_{di}, C_{it} and T. The total charge on the gate node Q_G is obtained by integrating the charge density at a position y in the channel given by the left hand side term in Eq. (2.7). For the purpose of Verilog-A model implementation, the gate charge is partitioned as $Q_S = -2Q_G/3$ for the source node and $Q_D = -Q_G/3$ for the drain node. Total charge on the gate node is given by (see appendix B for details),

$$Q_G = -W\frac{L}{V_{ds}}C_{di}V_{th}^2\frac{\left(x_d^2 - x_s^2\right)}{2} + W\frac{L}{V_{ds}}C_{di}V_{th}^2\int_{x_s}^{x_d} z dx \tag{2.9}$$

where $x_{s/d} = \left(V_G - V_{FB} - V_{s/d}\right)/V_{th}$. Also, for the purpose of this calculation I have assumed that V_C varies linearly from V_s at source to V_d at

drain i.e., a gradual channel approximation. In the limit of $V_{ds} \to 0$, $Q_G = WLC_{di}\left(V_{gs} - V_{FB} - \zeta\left(x_s\right)/e\right)$.

2.4 Current Voltage (I-V) Model

Current in graphene in the diffusive transport regime can be modeled using the drift-diffusion equation. Because graphene is degenerate, the ratio of the mobility to the diffusion constant is carrier density dependent. The total current density taking into account both the drift and diffusion terms can be written as [42],

$$J = J_n + J_p = \mu_n n \nabla\mu + \mu_p p \nabla\mu \tag{2.10}$$

where $\mu_{n/p}$ represent the electron mobility and the hole mobility respectively. We can write the electron density (n) and hole density (p) using Eq. (2.8) and the identity $F_1\left(-z\right) + F_1\left(z\right) = z^2/2 + 2F_1\left(0\right)$ as (see appendix C for details),

$$n/n_i = \frac{z^2}{4} + F_1\left(0\right) + \frac{x - z\left(1 + C_1\right)}{2C_0} \tag{2.11}$$

$$p/n_i = \frac{z^2}{4} + F_1\left(0\right) - \frac{x - z\left(1 + C_1\right)}{2C_0} \tag{2.12}$$

The electrochemical potential is expressed as $\mu = \mu_{eq} - eV_C$, where μ_{eq} is a equilibrium chemical potential in the channel. We can now integrate Eq. (2.10) using $d\mu = -edV_C$ to obtain the drain to source current,

$$I_{ds} = \frac{W}{L} e n_i V_{th}\left(\mu_{on} g\left(x_s, x_d, 1\right) + \mu_{op} g\left(x_s, x_d, -1\right)\right) \tag{2.13}$$

for a graphene FET of width W and length L. The function g is defined as

$$g\left(x_s, x_d, s\right) = \int_{x_s}^{x_d} \frac{z^2}{4} + F_1\left(0\right) dx + s \int_{x_s}^{x_d} \frac{x - z\left(1 + C_1\right)}{2C_0} dx \tag{2.14}$$

Eq. (2.13) is simply the result of the Pao-Sah modeling approach [42] applied to graphene. However, there is no double integral in this case because the channel is confined to a single atomic plane of the graphene. The expressions for the integral of z and z^2 appearing in Eq. (2.14) are given in appendix C

2.4.1 High Field Mobility

The high field mobility model is assumed to be of the form,

$$\mu_{n/p} = \frac{\mu_{o,n/p}}{1 + \frac{\mu_o |E|}{v_{sat}}} \tag{2.15}$$

where $\mu_{o,n/p}$ is the low field mobility for electrons/holes, which is used as a fitting parameter. Note that in the current model the low field mobility in the denominator is assumed to be same for both holes and electrons to simplify the equations. Saturation velocity in graphene is modeled as $v_{sat} = \frac{\hbar v_F \Omega}{E_F}$, where Ω is the frequency of substrate optical modes (fitting parameter), v_F is the Fermi velocity of carriers in graphene and $E_F = k_B T z$ is the Fermi energy [43]. This model can be implemented by dividing Eq. (2.13) by,

$$\left(L + \frac{\mu_o k_B T}{\hbar v_F \Omega} \int_0^L |z| \left| \frac{d\varphi}{dx} \right| dx \right) / L \tag{2.16}$$

which is nothing but the integration of the denominator of the high field mobility Eq. (2.15). The integral in the above equation has a closed form expression using the fact that the gradient of potential is $d\varphi = dV_C + V_{th} dz$ in the electric field term.

2.4.2 Contact Resistance

I address the source and drain contact resistance using the equivalent FET model shown in Figure 2.1(d). The contact resistance is not expected to be different at the drain and source ends from a device fabrication point of view. However, in the present model we allow them to vary independently to account for the often seen asymmetry in transfer characteristics. The current approach to address the contact resistance is not accurate. Ideally, one should include the effect of contact resistance using gate voltage dependent contact barrier height. In fact, our current understanding is that the work function difference between the gate and graphene will be a function of local bias across graphene and metal, i.e., the voltage drop due to current flowing across the contact. Consequently, one must solve a nonlinear equation self consistently to address the contact resistance.

2.5 Model Verification and Analysis

I have fit measured C-V and I-V data to verify our model. The data used for verification of this model is based on embedded gate graphene transistors with CVD graphene and HfO_2 or Al_2O_3 dielectrics. Device fabrication details are given in [44]. We consider three different devices referred to as A, B (from the same wafer) which have about 4.4 nm HfO_2 as gate dielectric and C which has about 3 nm Al_2O_3 as gate dielectric. Figure 2.2(a) shows the fit of C-V model to the measured low frequency (100 KHz) capacitance of device A. The extracted dielectric capacitance agrees well with the dielectric capacitance

expected from the physical thickness. Also, I have estimated the interface trap density to be $2.5 \times 10^{13} cm^{-2} eV^{-1}$. For verifying the I-V model I have fit the model to measured transfer (Device B) and output (Device C) characteristics. As it can be seen from the fit to the transfer characteristics in Figure 2.2(b), the model tracks the data quite well and it addresses both n-type and p-type conduction regimes typical of zero band gap graphene transistors. The gate voltage at which the minimum occurs in the transfer characteristics is referred to as the Dirac voltage. At low drain to source voltage (V_{ds}), the Dirac voltage corresponds to the flat band voltage or the neutrality point in graphene which is an indicator of the default doping of the graphene due to impurities or other such sources. While it is reasonable to expect the non zero minimum current at finite temperatures, there is also a possible contribution from the minimum conductivity (σ_{min}) of the graphene. We account for this by adding a term $\sigma_{min} V_{ds}$ to Eq. (2.13) with σ_{min} as a fitting parameter. When measured at high V_{ds}, the Dirac point shifts typically in the direction of V_{ds} and may shift in the opposite direction in short channel devices [45]. Figure 2.2(c) shows the fit of the current model to the measured output characteristics of Device C.

Table 2.1 shows the model parameters extracted from fitting the data for Device B and Device C. The fitting parameters used in approximating the solution of Eq. (2.8) are $a = 2.85$ and $b = 4.52$ for Device B and $a = 3.02$ and $b = 8.18$ for device C. For Rows 3 and 4 I have allowed the R_s and R_d to vary independently. In another approach I set R_s and R_d to be same to obtain the model parameters shown in Rows 5 and 6. The contact resistances seem

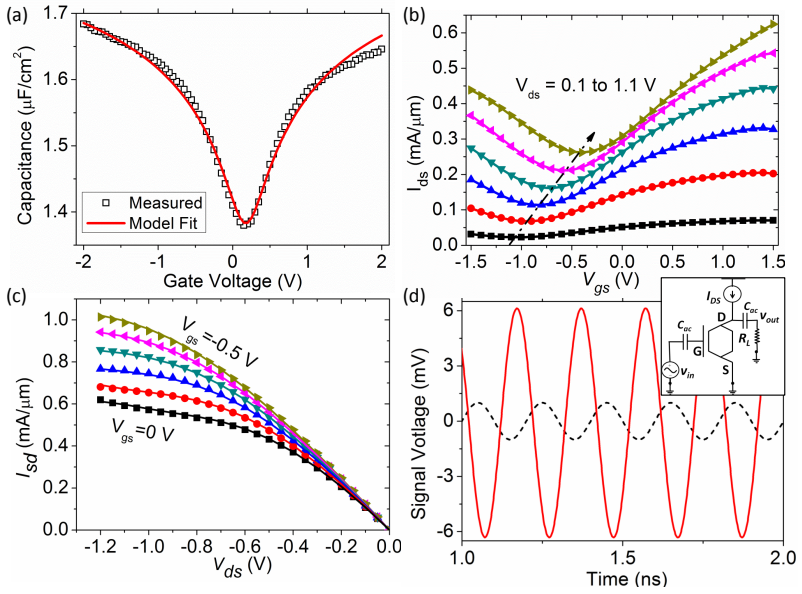

Figure 2.2: (a) Fit (line) of the CV model to measured data (symbol), (b) Fit (line) of the IV model to measured transfer characteristics (symbol) for V_{ds} = 0.1 to 1.1 V as shown by the dashed arrow, (c) Simulated output characteristics (line) vs measured output characteristics (symbol) and (d) Simulated output signal (solid line) and input signal (dashed line) for common source amplifier using the device whose output characteristics is shown in (c).

consistent with the typically seen 500 to 1000 $\Omega \cdot \mu m$ values in our samples. For the parameter value given in column 7, Ω is scaled by 55 meV. For Device C, using $\mu_o = \mu_p$, we estimate the Ω to be about 24.6 meV. The lowest energy surface optical phonon frequency for Al_2O_3 is 55 meV [46].

If the device does not show full current saturation, there will be large error bar in the value of Ω. In fact, this is the case for Device C, as the number of data points near saturation(i.e., near Vds=-1 Volt) is less than the number of data points where the device is not in saturation. For Device C, omega

Table 2.1: Extracted model parameters with 95% confidence interval for Device B and Device C both with 10 μm width and 500 nm length.

	Rs	Rd	μ_n	μ_p	V_{FB}	μ_o/Ω	σ_{min}
	Ω	Ω	$\frac{m^2}{Vs}$	$\frac{m^2}{Vs}$	V	$\frac{m^2}{V}$	$\frac{4e^2}{h}$
C	14.96 ± 62%	55.50 ± 17%	0.41 ± 58%	0.78 ± 11%	0.69 ± 5%	1.74 ± 67%	11.64 ± 22%
B	55.45 ± 31%	61.14 ± 26%	0.09 ± 16%	0.08 ± 17%	-1.02 ± 4%	3.32 ± 9%	12.94 ± 5%
C	34.70	34.70	0.26	0.67	0.76	1.63	11.64
B	60.10	60.10	0.10	0.08	-1.02	3.22	12.94

has a near 50% variation in the 95% confidence interval, whereas for Device B, which shows better saturation behavior compared to Device C, omega has about 10% variation in 95 % confidence interval. I believe that when more data is available where the device shows stronger velocity saturation related current saturation, omega can be extracted with smaller error bar. I also see large error bars in the extracted contact resistance values which needs to be corrected with better contact resistance model.

I have implemented the model in Verilog-A using the model equations and the extracted physical parameters shown in third row of Table 2.1. Figure 2.2(d) is a transient analysis result for a common source amplifier shown in the inset and simulated in Spectre® [47] using the model parameters of Device C. The device is biased at $V_{GS} = -0.3$ V and $V_{DS} = -1.0$ V and has a voltage gain of 15.8 dB with a 5 GHz input signal.

2.6 Conclusions and Future Work

While graphene FET fabrication is at a stage where only ICs with one or two devices have been demonstrated [40], the hardware correlated graphene FET model presented in this work can be used to analyze the performance met-

rics and design tradeoffs for large graphene based circuits and hybrid circuits involving conventional silicon based FETs. The model captures the essential physics of the long channel graphene FETs and agrees well with the measured data. Given that most devices fabricated at present still operate in diffusive transport limit, we believe that the presented model does address the essential physics of these devices. However, this model will be inadequate for nanoscale graphene devices where band-to-band tunneling and quasi-ballistic regime of transport become significant. When the pinch off point moves into the channel, it may be necessary to address the band-to-band tunneling in channel in future work. Contact resistance is another issue that is not addressed accurately in the current model. As was mentioned earlier, a constant contact resistance seems inadequate resulting in large error bars in extracted parameters. Due to the presence of Schottky barriers at the metal graphene junctions source and drain contact resistances can be different with different combinations of terminal voltages. For example in a p-type device, when the pinch off point is in the middle of the channel, the source side has metal to p-type graphene junction, and the drain side has metal to n-type graphene junction. In the current model such asymmetry is addressed using constant but different source and drain resistance. A more accurate model should capture this asymmetry based on the local barrier height at the source and drain ends resulting in non-linear gate voltage dependent contact resistance. This limitation also should be addressed in future work.

Chapter 3

Bilayer pseudoSpin Field Effect Transistor

3.1 Introduction

The **Bi**-layer pseudo**S**pin **F**ield **E**ffect **T**ransistor, BiSFET, represents a proposal for a novel graphene based transistor intended to enable much lower voltage and power operation than possible with Complementary Metal Oxide Semiconductor (CMOS) Field-Effect Transistor (FET) based logic [12, 18]. The BiSFET is based on predicted room temperature superfluidity in dielectrically separated bilayer graphene [24, 48]. The BiSFET is currently only a concept and I recognize the limitations of theory, particularly initial efforts, and the technological challenges to its realization. However, if realizable, in principle BiSFETs could operate at voltage comparable or even smaller than the thermal energy leading to drastic power reductions. Furthermore, output characteristics would exhibit negative differential, also quite unlike those of MOSFETs. In this work, therefore, I discuss the physics underlying the BiSFET, evolving BiSFET design, and possible BiSFET-based Boolean logic and memory circuits. While it might seem premature to worry about circuits for a still hypothetical device, the circuit level work that I performed is necessary to help measure the potential payoff of continuing device work (and circuit work) and to inform that work through identification of critical device physics

and technological challenges

In Section 3.2, I discuss the essential physics, materials requirements and describe the associated SPICE models of the BiSFETs used for the circuit simulations of this work in Section 3.3. Implementation of various Boolean logic gates and their operating principles are discussed in Section 3.4. A BiSFET-based memory cell and associated preliminary SPICE simulations are presented in Section 3.5.

3.2 Underlying Physics

The BiSFET is a conceptually radical device concept based on novel physics in a novel material system. Understanding the proposed BiSFET designs and expected I-V characteristics requires understanding the underlying physics. Indeed, as our understanding has improved the design has evolved. The phenomenon of superfluidity in electron-hole systems was theoretically first predicted by Lozovik and Yudson [49]. Past experimental investigations have focused on closely spaced electron-hole bilayers in GaSb-InAs [50], Si [51], and GaAs [52, 53]. However, the Bose condensate in III-V systems is observed only at very low temperatures and high magnetic fields. The phenomenon of exciton condensation is also expected to occur in two dielectrically separated layers of graphene. However, due to unique electronic properties of graphene, it is theoretically predicted that the superfluid state due to exciton condensation might occur above room temperature in this latter system [10, 24, 48]. In this section I will review the essential physics of the condensate formation in bilayer

graphene and various factors that govern the room temperature transition from a device application point of view.

3.2.1 Condensate Formation in Bilayer Graphene

The possibility of room temperature superfluid condensation in graphene bilayers is a consequence of a synergy of multiple properties of graphene. In a double quantum well system, under certain conditions, electrons in one semiconductor layer can pair with holes in an adjacent layer (both Fermions) resulting in electron-hole-pairs/excitons (Bosons) which can then condense. A perfect or strong nesting of the electron and hole Fermi surfaces, achieved by using strong magnetic fields in III-V systems, is required to observe the coherent many body state. In graphene, because of the nearly symmetric electron-hole band structure of graphene, a dielectrically separated bilayer graphene system allows such nesting absent any magnetic field; it can be achieved simply with equal electron and hole densities in opposite layers. The predicted maximum temperature for coherence is slightly above $0.1E_F/k_B$ where E_F is the magnitude of the Fermi energy relative to the band edge [24,48]. The zero band gap and low density of states makes it relatively easy to move the Fermi level well into the bands, as needed for condensation, with limited carrier concentrations, and, thus, manageable gating fields. For graphene, 300 K = 0.1EF/kB translates to electron and hole densities of only $n_o \approx p_o \approx 5 \times 10^{12} cm^{-2}$ for condensation at 300 K [18]. The carrier density in graphene layers has been electrostatically modulated to as high as 10^{13} cm^{-2} using independent

gates [54]. Of course, being nearly perfect two-dimensionality, allows even dielectrically separated graphene layers to be brought much closer together than is conceivable for quantum wells in III-V systems. The effective dielectric permittivity for the required electron-hole exchange potentials is largely an extrinsic property, governed more by the surrounding dielectrics than the graphene itself.

Condensate formation is taken to be the result of the Fock many-body exchange interactions between layers, V_F, which can be approximated for graphene layers, in mean-field theory on a p_z/π-orbital based atomistic tight binding lattice, as

$$V_F\left(\mathbf{R}_T,\mathbf{R}_B\right) \approx \frac{e^2}{4\pi\varepsilon_{eff}\sqrt{\left|\mathbf{R}_T-\mathbf{R}_B\right|^2+d^2}} \sum_{\alpha,\mathbf{k},s} n_{\alpha,\mathbf{k},s}\varphi_{\alpha,k,s}\left(\mathbf{R}_T\right)\varphi^*_{\alpha,\mathbf{k},s}\left(\mathbf{R}_B\right) \tag{3.1}$$

assuming a two-dimensional "bulk" region [7,9]. In addition, the authors of [7, 9] also assumed a constant effective dielectric permittivity ϵ_{eff} to simplify the discussion. See ref [10] for the mean field calculations with detailed screening models. $\mathbf{R}_T$ and $\mathbf{R}_B$ are the two-dimensional (2D) in-plane location vectors for the atoms in the top and bottom graphene layers, respectively, and d is the separation between the two layers. The $\varphi_{\alpha,\mathbf{k},s}$ are the tight-binding electron energy eigenfunctions, and the $n_{\alpha,\mathbf{k},s}$ are the occupancy factors. The eigenstate labels α, $\mathbf{k}$, and s denote the band index of the many body system, the wave-vector and the (real) spin state, respectively.

Now consider the n and p-type graphene layers with their symmetric

conduction and valence bands with equal charge densities under equilibrium conditions, such that the conduction band Dirac cone and the valence band Dirac cone intersect at the common Fermi level. Assuming some coupling between the layers, there will be an anti-crossing/band gap formed about the Fermi level with interlayer coherent states such that $\varphi_{\alpha,k,s}\left(\mathbf{R}_T\right)\varphi^*_{\alpha,\mathbf{k},s}\left(\mathbf{R}_B\right) \neq 0$ near the anti-crossing, with those below the band gap preferentially occupied according to Fermi statistics. Therefore, the Fock interaction potential $V_F\left(\mathbf{R}_T, \mathbf{R}_B\right)$ of Eq. (3.1) will be nonzero. So there will be coupling between the layers, which will produce a band gap and interlayer coherence, which will produce a non-zero Fock interation, and so forth and so on. Moreover, the formation of the band gap will lower the energy of the preferentially occupied states near but below the band gap, which will reduce the overall energy of the system, making it the energetically favored state. While some single-particle/bare coupling can strengthen the condensate, it is in fact possible to find self-consistent solutions for the many-body condensate without any bare coupling, so-called "spontaneous" condensates.

Illustrative calculations using the Fock interaction of Eq. (3.1) within and otherwise nearest neighbor atomistic tight-binding calculations—the Fock interactions is far from nearest neighbor–are provided without and with bare interlayer coupling are in references [7] and [9], respectively. In [7], the dielectrically separated bilayer system was assumed to be weakly coupled if at all. (The nominal stacking, Bernal or otherwise, is essentially irrelevant for spontaneous condensates.) The solid lines in Figure 3.1(a) shows the low energy

band structure of the weakly coupled bilayer graphene layers in the absence of interlayer exchange coupling. Black solid lines indicate the band structure of electron layer with Fermi level, and red solid lines is the band structure of the hole layer again with Fermi level at . The dotted lines show the zero temperature low-energy dispersion for the bilayer graphene in presence of exchange interactions for different values of interlayer separation. It can be observed that the zero temperature band gap E_{g0} opens due to the Fock interaction and the band gap increases with decreasing interlayer separation i.e., due to strengthening Coulomb interactions between layers, and, thus, strengthening Fock exchange corrections to the Coulomb interactions. Figure 3.1(b) shows the temperature dependence of the band gap due to interactions for three values of dielectric constant. The critical temperature T_c for the transition from normal state to superfluid state occurs when the curve collapses, rather abruptly, to the temperature axis. Contributions to the exchange interaction from above the band gap are opposite in phase to those from below. (Think about formation of symmetric and antisymmetric states when two otherwise isolated degenerate states couple.). The abruptness is due to a positive feedback loop in which with increasing temperature and the associated occupation of states above the gap and emptying of those below, the exchange interaction and the condensate and associated band gap are weakened, leading to more occupation of states above the band gap and more emptying below. As a result, the form of the temperature dependencies when normalized to the zero temperature band gap E_{g0} are quite similar independent of the dielec-

tric constant, with $k_B T_c \approx 0.25 E_{g0}$. However, T_c increases proportional to E_{g0} that, in turn, increase even more strongly with decreases in the dielectric constant than with layer separation. Notably, and wherein lies the Moniker

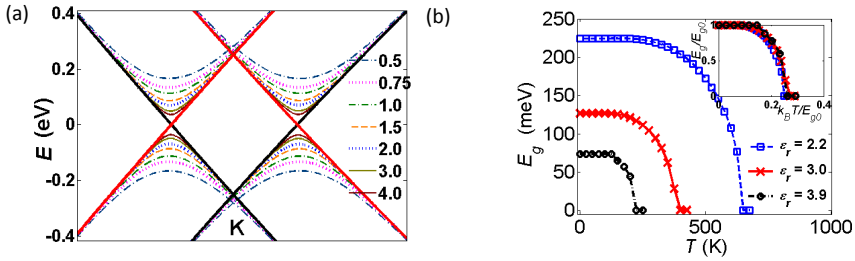

Figure 3.1: (a) Low-energy dispersion of the graphene bilayer system with a potential energy difference between layers of $\Delta = 0.5$ eV, a relative dielectric constant $\epsilon_r = 2.2$ at 0 K, and balanced charge distributions, as a function of layer separation d shown in the legend in units of nm. The solid black and red lines are the band structures of the top and bottom layer graphene, respectively, in absence of interlayer exchange coupling. (b) Temperature dependence of the band gap for three different dielectric constants with $\Delta = 0.5$ eV, d = 1 nm and balanced charge distributions. Lower ϵ_r result in a Coulomb interaction and, thus, larger Fock correction potential, which, in turns, leads to larger 0 K band gaps that are, therefore, also more robust at higher temperatures. The top right insert shows the same data illustrates the similarity of the T dependence of band gap for different ϵ_r when normalized by the 0 K band gap, E_{g0} . (figure a and b reprinted with permission from [7], copyright (2010) by the American Physical Society)

for the BiSFET, if we described the "which layer" degree of freedom for the electrons as a (another) two-level pseudospin state, then the condensate can be described as a coherent pseudospin state between the layers, with an associated pseudo-spin strength—the stronger the condensate, the larger the pseudospin strength—and pseudo-spin phase.

3.2.2 Effects of biasing; Enhanced Interlayer Tunneling and Critical Current

The BiSFET is based on the low bias-voltage transport properties of the condensate. The low-voltage tunneling current between independently contacted bilayers is significantly enhanced by the condensate formation in bilayer systems, as has been confirmed experimentally in III-V systems [55,56].

Here I consider two possible configurations for contacts to independent layers. First, as shown in Figure 3.2(a) with contacts C1 and C2 to layer-1 and contacts C3 and C4 to layer-2, when a current is driven through layer-1 due to an applied bias across C1 and C2, $V_{1,2}$, in the presence of a condensate, and equal amount of current flows in layer-2. In other words, when a hole is injected into layer-1 via contact C1 moving towards C2, it drags along with it an electron from C3 to C4. At low bias one expects perfect Coulomb drag and an almost dissipation less exciton transport from one end to another, and the current then will be mainly contact resistance limited. Recent experimental results provides a clear illustration of such near perfect coulomb drag [56] in a III-V double quantum well system in presence of such a condensate, although again at very low temperatures and high magnetic fields, of course [57]. The drag current should ultimately drop at voltages $V_{1,2}$ on the scale of the equilibrium condensate band gap via destruction of the condensate, resulting in negative differential resistance (NDR) in the device characteristics, as qualitatively shown in Figure 3.2(c); much like with temperature increase, pushing the Fermi levels in C1 and C2 toward opposite band gap edges will weaken

the exchange interaction by altering the occupation probabilities above and below the condensate band gap, leading to the self-consistent destruction of the condensate.

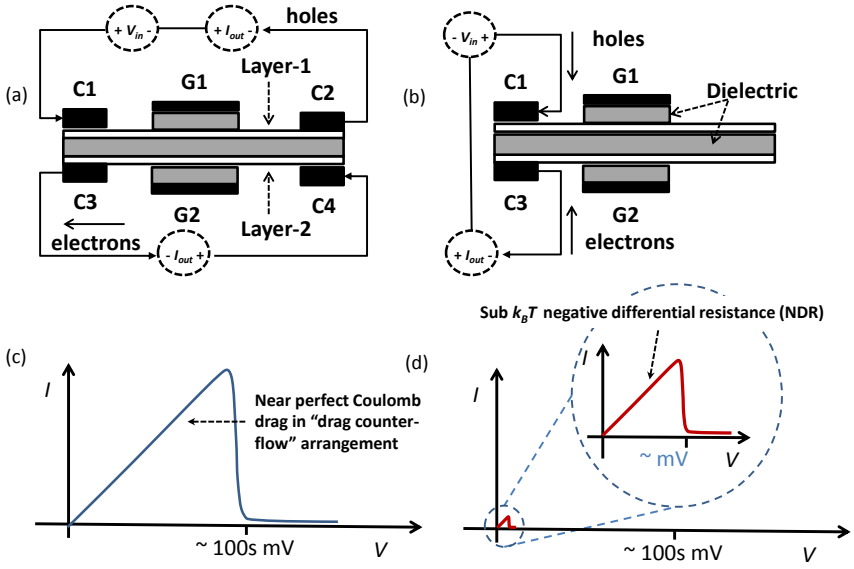

Figure 3.2: Schematic showing possible BiSFET device geometry with gates G1 and G2 to layer-1 and layer-2 respectively and (a) four independent contacts C1 to C4, (b) two independent contacts C1 and C3, (c) illustration of current voltage characteristic for the device in drag counter flow arrangement shown in (a) and (d) current voltage characteristics for the device geometry shown in (b).

Figure 3.2(b) shows another possible geometry, the one relevant for the BiSFET, where the current flow is due to condensate enhanced resonant tunneling at very low bias $V_{1,3}$ between contacts C1 and C3, and contacts C2 and C3 are simply removed. An NDR current-voltage characteristic for such a device geometry is illustrated schematically in Figure 3.2(d). In comparison to the Coulomb drag counterflow geometry shown in Figure 3.2(c), the maximum

or "critical" current for tunneling can occur at much lower bias voltages $V_{1,3}$, and via a very different process as discussed more below. The same experimental work that showed the near perfect Coulomb drag also clearly demonstrated this behavior when biased in this configuration [56].

Part of the essential physics of interlayer current flow in the presence of the condensate is that while the associated exchange interaction alone can self-consistently support a spontaneous condensate, it cannot by itself support an interlayer current. (As in any tight-binding calculation of current flow between two sites coupled by some hopping potential) The current flow between the layers depends on the imaginary part of product of the complex conjugate of wave-function on the first layer, the wave-function on the second layer, and the interlayer hopping potential. Here, however, in the absence of bare tunneling, the interlayer coupling is provided by the exchange interaction that also depends on the wave functions consistent with Eq. (3.1). The self-consistent result is that the phase of the wave-functions are uniformly opposite that of the exchange interaction—i.e., the collective pseudo-spin phase is opposite that of the exchange interaction—and the above product has no imaginary part and the current vanishes. (Notably, however, in the absence of bare/single particle coupling, the phase relationship between the layers is also arbitrary.) However, if—real-valued for the sake of argument—bare interlayer coupling is added, then there is an "additional" current term associated current flow between the layers that depends on the product of the wave-function on the first layer, the complex conjugate of the wave-function on the second site, and, this time, the

real interlayer hopping potential. Thus, bare coupling can provide a nonzero current (as well as establish a non-arbitrary phase relationship between the layers). Moreover, because the product of the wave-function on the first layer and the complex conjugate of the wave-function on the second is greatly enhanced over a wide range of energies by condensate formation, the tunneling current is greatly enhanced compared to the bare current. Indeed, for small interlayer voltages the layers are expected to be essentially shorted together (as will be illustrated through simulations reviewed below) independent of the bare coupling; variations in single particle coupling are simply compensated for by increases in the pseudospin phase.

With increases in interlayer voltage, the current flow increases essentially linear with small voltages—which are the goal here—consistent with Landauer-Büttiker theory (where the transverse density of states changes little with small voltage changes), at least up to a point. However, the steady-state (DC) current that can be supported in this way is maximized when the pseudospin phase reaches $\pm\pi/2$, which corresponds to critical interlayer current and is analogous to the critical current in a superconducting Josephson junction. However, with the pseudo-spin phase set to $\pm\pi/2$ at the critical current, variations in the single particle coupling cannot be compensated for. As a result, the critical current, itself, is governed by the single-particle coupling and varies approximately linearly with its strength [9, 55, 57, 58]. If the interlayer voltage is increased further, the phase relationship for the condensate between the layers cannot be stabilized and the DC current would vanish, although

the condensate itself does not breakdown. Rather the interlayer phase and, thus, the condensateenhanced current is expected to oscillate with time with an oscillation frequency of roughly of $qV_{1,3}/h$ where h is Plank's constant, analogous to the AC Josephson effect. Accordingly, in principle an interlayer voltage $V_{1,3} = 25$ mV, as commonly used for BiSFET simulations, would produce oscillation frequencies greater that 6 THz! On the scale of a 10 Ghz clock frequency, this signal would essentially vanish. On the other hand, the BiSFET could also be used as linear voltage-to-frequency Thz generator, again much like a Josephson junction, at still lower voltages if the signal could be effectively coupled out. Note that so long as the condensate exists, there is nothing in this discussion that is sensitive to temperature; in principle the condensate can have critical interlayer voltages arbitrarily comparable to or smaller than k_BT/q, allowing the proposed BiSFET to potentially operate at or below k_BT/q voltages.

To illustrate the above transport physics, colleagues have performed quantum transport simulation of interlayer current flow with the exchange interaction self-consistently calculated via an iterative method at a temperature of 300 K (at no small computational cost, as the exchange interaction is non-local in the plane of the layers [8]). These simulations considered a biasing arrangement like that of Figure 3.2(b) with a 15 nm long condensate region, except that the right contacts were grounded rather than left floating in these initial calculations. (Note that the 250 mV band-gap formed in the channel by the condensate actually eliminates most current flow between left and right

contacts through the region of the condensate in this latter configuration, such that the device is effectively contacted much the same as Figure 3.2.) Noting the primitive units cells of each layer contain two atoms, an A atom and a B atom, they considered only the A atom to A atom component of coupling through the presumed dielectric tunnel barrier, and took the coupling strength to be 0.5 mV. (Any A atom to B atom component of coupling is essentially orthogonal to the nature of the condensate [7], eventually forcing a change in the condensate itself for A atom to B atom coupling only before significant critical current is expected [9].) The pseudospin phase as a function of iteration step after applying an interlayer voltage V_{il} across, and equally divided between, the left contacts, is shown in Figure 3.3(a). For a 5 meV interlayer potential, a stable phase of approximately $\pi/8$ is approached, with an interlayer current density per unit area in the center of the condensate region of approximately 6 nA/nm^2. This current, approaching 70 to 80% of the limit of Landauer Büttiker theory, shows the layers to be essentially shorted together despite the very weak bare coupling. As the interlayer current flow is proportional to both the interlayer voltage and the sine of the phase angle up to the critical current, this results would suggest a critical current density and voltage of voltage of, respectively, $6\text{nA/nm}^2 \times \sin(\pi/2)/\sin(\pi/8) \cong 16\text{nA}$and $5\text{mV} \times \sin(\pi/2)/\sin(\pi/8) \cong 13\text{mV}$. Moreover, the calculated critical/switching voltage is approximately only one-half of kBT at the 300 K simulation temperature, supporting the possibility of sub-kBT switching in BiSFETs. Notably, these results from quantum transport calculations are perhaps two

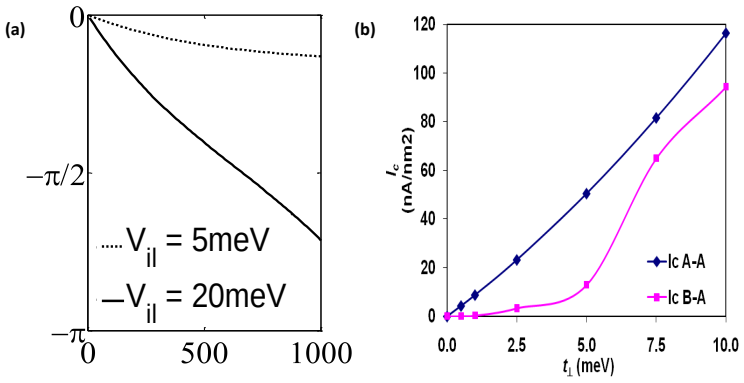

Figure 3.3: (a) Variations in pseudo-spin phase difference between layers (it's roughly independent of position) during iterative calculations of the condensate and associated steady state current flow between layers, as obtained through quantum transport calculations at a simulation temperature of 300 K(figure adapted from [8]). The largest possible current flow between layers occurs when the interlayer phase angle is $\pm\pi/2$. For a 5 meV interlayer potential, a stable phase of $\approx \pi/8$ is approached. As the interlayer current flow is proportional to both the interlayer voltage and the sine of the phase angle, below the critical current, this results would suggest a critical current density and voltage of voltage of, respectively, $6\text{nA/nm}^2 \times \sin(\pi/2)/\sin(\pi/8) \cong 16\text{nA}$and $5\text{mV} \times \sin(\pi/2)/\sin(\pi/8) \cong 13\text{mV}$. For a 20 meV interlayer voltage, no stable solution is found, consistent with the expectation of being above the critical current. Notably, these results from quantum transport calculations are perhaps two to three times the 0.5 meV bare coupling results shown on the critical current vs. hopping potential curve of (b) for the same typeA atom to A atomof interlayer coupling, which was obtained earlier from more "back of the envelope" calculations [9], but the condensate here (as measured by the band gap) is somewhat stronger that for which the prior results were obtained as well, so that the agreement is as good as could be reasonably expected (figure reprinted with permission from [9], copyright (2011) by the American Physical Society)

to three times the 0.5 meV bare coupling results shown on the critical current vs. hopping potential curve of Figure 3.3(b) for the same type of interlayer coupling, which was obtained earlier from more "back of the evelope" calculations [9]. However, the results of Figure 3.3(b) were obtained with a weaker (smaller band gap) condensate as well, so the agreement is quite good. For a 20 meV interlayer voltage in these quantum transport calculations, no stable solution is found, consistent with the expectation of being above the critical current. Rather, the phase just keeps increasing with iterations, going beyond $\pi/2$ limit for a stable result and continuing with no end in sight, or expected.

3.2.3 Effects of Charge Imbalance

The independent contacts to the layers drive the tunneling current between the layers enhanced by the condensate. The gates are used to electrostatically induce the required electron and hole density in the layers. When the carrier densities in the layers are equal, there is a maximum overlap of the Fermi surfaces and the band gap opening due to interactions is at its maximum value. However, when the carrier density in the layers is not equal, the Fermi level shifts toward one of the condensate band gap edges which can weaken the condensate in much the same way that increasing the temperature can, and resulting in collapse of the condensate for strong imbalance. Figure 3.4 illustrate the effect of the carrier density imbalance on the band gap [59]. Observe that the band gap decreases with increasing imbalance ultimately collapsing when the imbalance is more than about 15 % of the total carrier density. A

similar effect occurs for the critical current that is maximized when the carrier density in the layer is equal but decrease with increasing imbalance. As shall be seen, in the initial BiSFET designs gate-controlled charge imbalance was proposed as the switching mechanism, although concerns over screening (see below) have led us to consider alternate gating mechanisms.

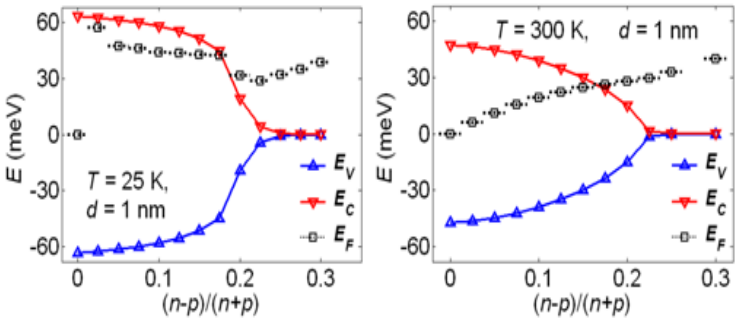

Figure 3.4: Energy band edges and Fermi level as a function of carrier imbalance between top layer electron density and bottom layer hole density for graphene bilayers separated by 1 nm at 25 K and 300 K with $\epsilon_r = 3$ and $\Delta = 0.5$ eV (figure reprinted with permission from [7], copyright (2010) by the American Physical Society).

3.2.4 Effects of Screening

The consideration of screening has led to an active debate in the literature over feasible critical temperatures T_c. The exchange interaction depends on the Coulombic interaction between the carriers in opposite layers. As already seen, larger permittivities ϵ and larger interlayer spacings d reduce the interlayer Coulombic interactions and, thus the exchange interaction and T_c. The strength of Coulomb interactions can also decrease due to screening from surrounding charges. There is no perfect theory of either superfluidity or

screening. The calculations are often done within a random phase approximation (RPA). Above room temperature condensation was obtained in [24] based on calculations with un-screened interactions (although the effects of screening were discussed in some detail), while calculations based on static screening predict extremely low transition temperatures [60]. Recent calculations accounting for the dynamic nature of interactions and self-consistent reduction in the screening with condensate band gap formation still predict the possibility of a room temperature T_c, although requiring smaller dielectric permittivity than originally estimated [10]. Figure 3.5 (a) and (b) adapted from the work of [10] show the ratio of estimated band gap to Fermi level as function of effective fine structure constant for graphene, $\alpha = e^2(4\pi\epsilon\hbar v_D)$ in SI units where v_D is the fixed group velocity magnitude for carriers near the Dirac point in graphene. Figure 3.5 (a) shows the band gap (left to right) unscreened interactions and N = 1,2,3,4 degenerate Dirac cones including spin (where N = 1 for the topological insulator Bi2Se3, for example, while N = 4 for graphene) as a function of the fine structure constant which is inversely proportional to the dielectric constant. The vertical dashed line corresponds to the fine structure constant of vacuum. The horizontal gray highlighted area is an approximate range of spontaneous condensate gaps required for room temperature transition. For graphene with N = 4, the fine structure constant has to be greater than 1.5 for spontaneous coherence at or above room temperature. For robust condensate formation at or above room temperature, the dielectric constant and the Fermi level in graphene should fall in the region

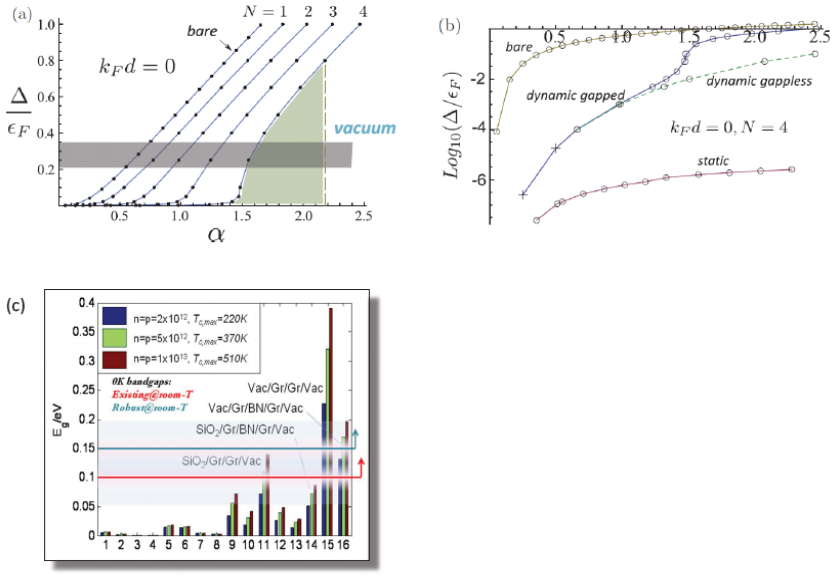

Figure 3.5: (a) Spontaneous gap as a function of effective fine structure constant $\alpha = e^2/(4\pi\epsilon\hbar v_D)$ in SI units where v_D is the fixed group velocity magnitude for carriers near the Dirac point in graphene, for (left to right) unscreened interactions and N = 1,2,3,4 degenerate Dirac cones (N = 1 for Bi2Se3, N = 4 for graphene). The vertical dashed line corresponds to fine effective fine structured constant of vacuum (b) Comparison between several approaches to screening. When the screening is treated by using static screening, an extremely small gap is predicted. If the reduction in screening in the coherent state is neglected the sudden rise at $\alpha \approx 1.5$ is absent (for the dashed line we fixed the gap to be $10^{-4}E_F$ inside the polarization functions, thus virtually computing them in the gapless phase.). The full dynamical and gaped screening result approaches the result obtained with bare interactions at strong coupling but differs from it by several orders of magnitude at weak coupling (figure a and b reprinted with permission from [10], copyright (2012) by the American Physical Society) and (c) Band gap for various dielectric stacks for the double gated dielectrically separated graphene bilayer system; only plausible stacks according to this calculation are labeled. (The band gaps in this latter calculation were obtained within the nominally unscreened self-consistent Fock approximation, but the dielectric permittivities were uniformly scaled up such that in the region of the grey bars in Figure3.5(a), they reproduced approximately the same band gaps as the screened interaction with the actual dielectric constants. High frequency dielectric constants were used representing, in classical terms, the high velocity of the carriers making ion-based screening of individual carriers difficult [11].)

of intersection of the green and gray shaded areas. Figure 3.5(b) shows the comparison between several approaches to screening. When the screening is treated by using static screening, an extremely small gap is predicted and associated critical temperature is predicted. There is improvement by considering dynamic screening. However, only when the self-consistent reduction in screening in the sudden rise in the predicted condensate bandgap coherent state is neglected, the sudden rise at $\alpha \approx 1.5$ is absent. The full dynamical and gapped screening result approaches the result obtained with bare interactions at strong coupling but differs from it by several orders of magnitude at weak coupling. Figure 3.5(c) shows the estimated band gap for different combinations of dielectric stacks for the double-gated dielectrically separated graphene bilayer system [11] based of the results of [10]. To the extent these estimates are accurate—which again should not be overestimate given the complexity of the system—dielectric environments including partial vacuums would clearly be beneficial. (The band gap for this latter calculations in this calculation was calculated with the nominally unscreened self-consistent Fock approximation, but the dielectric permittivities were uniformly scaled up such that in the region of the grey bars in Figure 3.5(a), they reproduced approximately the same band gaps as the screened interaction with the actual dielectric constants. High frequency dielectric constants were used representing, in classical terms, the high velocity of the carriers making ion-based screening of individual carriers difficult. We note that graphene device with air-gap gates have been demonstrated [61], and air gaps are even being considered for reducing

gate to source and drain capacitance in CMOS [62, 63]

3.3 BiSFET Design and Compact Modeling

The initial BiSFET design was based on the calculations for estimating transition temperature without free-carrier screening. The required carrier densities for condensate formation were assumed to be induced by gate work function engineering, and switching via small gate voltage-induced variations in the carrier densities. I here refer to this device as BiSFET 1. With a refined understanding of the effects of such screening and the required dielectric environment based on the work of [10], we developed an alternative design where the carrier densities are created by relative large but fixed gate voltage on gate further removed from the region of condensation, as they could provide a significant source of screening too. In this way, these gates do not enter into the switching or steady-state power consumption. Switching control is provided by gate induced current crowding in parallel conduction paths via limited changes in the input resistance. I here refer to this second design as BiSFET 2. In this section, I present a discussion of device structure and compact model for BiSFET I and BiSFET 2 shown in Figure 3.6(a) and (b), respectively. In the next section, I will consider SPICE-level circuits simulations based on these compact device models. To that extent, it's important to understand that, despite the physical differences between the BiSFET 1 and BiSFET 2 designs, their anticipated $I - V$ characteristics are much the same.

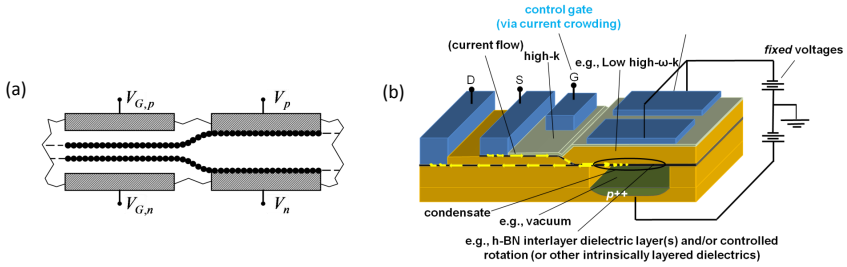

Figure 3.6: Illustration of different gating schemes: (a) Device schematic of BiSFET-1. (Note the region of condensation could be defined by layer proximity as shown, or changes in dielectric constant or equilibrium carrier densities, etc.) (b) Device schematic of BiSFET-2

3.3.1 BiSFET 1

The essential device elements of BiSFET 1 are shown schematically in Figure 3.6(a). For strong Coulomb interaction between carriers, the p-type and the n-type layers should be encapsulated in a low dielectric constant (low-k) environment. The tunnel barrier should be sufficiently thick to limit, but not eliminate, single particle transport between layers. As illustrated, each graphene layer has a metallic contact and is electrostatically coupled to a gate electrode through a gate dielectric. In the absence of applied gate biases, carriers would be induced in the graphene layers by use of differing work functions for the gates, or ferroelectric oxides as dielectrics, and/or back-gating. Applied gate voltage signals are intended only to balance or slightly unbalance the charge concentrations between layers to improve or degrade the electron-hole pairing and, thus, the condensate and interlayer critical tunneling current as discussed in Section 3.2.3. Note, that the gates are not intended to destroy the condensate, merely weaken it and reduce the associated critical current.

In the proposed circuits, which ever device reaches it critical current first will fall into its NDR regime and be strongly turned off; the gates are there to establish which device reaches its critical current first. Furthermore, for most circuits considered here, a switchable input signal to only one gate is required, leaving a good deal of flexibility in what constitutes the other gate. A fixed input voltage signal to these other gates could be effectively achieved via small changes to gate work function, etc.

As discussed in Section 3.2.2, in the presence of condensate formation, at very low biases the layers can be effectively shorted together, but once the critical current is reached the interlayer resistance increases drastically, producing a pronounce negative differential resistance characteristic that we expect to qualitatively much like that exhibited in Figure 3.2(d). And, in principle, the critical current can be varied arbitrarily, at least conceptually, via variations in the single particle coupling (although, achieving the desired single particle coupling is a significant technological issue.). Moreover, gating the charge imbalance would weaken the condensate (pseudospin) and, thus, the critical current.

For the purpose of compact modeling, the device geometry and NDR behavior discussed in the preceding two paragraphs was modeled as illustrated in Figure 3.7(a) and (b). As taken from [18], all gate lengths have been chosen to be = 10 nm, slightly larger than then estimated Josephson length. Unless otherwise specified, the gate length L to channel width W ratio for the BiSFETs was taken to be 2; larger specified values of W/L indicate wider

gates. Effective oxide thicknesses (EOT) of 1 nm were assumed for gate and interlayer dielectrics correspond to capacitances $C_{G,p}$, $C_{G,n}$and, C_{il} =all of 3.5×10^{-6} F/cm^2. The nominal carrier densities in graphene layers were taken to be $n_o \approx p_o \approx 5 \times 10^{12}$ cm^{-2} to allow room temperature operation as discussed in Section 3.2.1. These densities could be provided by, e.g., opposing gate work-functions of approximately ± 1 eV, respectively, or by one ± 1 eV gate work function and an opposing fixed "'back gate"' bias.

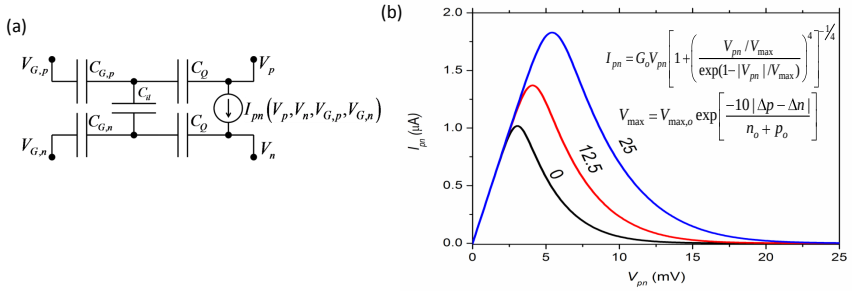

Figure 3.7: (a) Equivalent circuit model of BiSFET and (b) I-V characteristics of BiSFET for three different gate voltages $V_{G,n}$ with $V_{G,p} = -25$ mV. (Reprinted with permission from [12], copyright (2010) by the IEEE)

The NDR behaviors were modeled via the interlayer voltage relation,

$$I_{pn} = G_o V_{pn}\left[1 + \left(\frac{V_{pn}/V_{\max}}{\exp\left(1 - |V_{pn}|/V_{\max}\right)}\right)^4\right]^{-1/4} \tag{3.2}$$

with gate voltage dependent $V_{\max}$,

$$V_{\max} = V_{\max,o}\exp\left[\frac{-10\,|\Delta p - \Delta n|}{(p_o + n_o)}\right] \tag{3.3}$$

where Δn and Δp are the variations in charge densities with all four terminal voltages consistent with the model of Figure 3.7(a). A comparison of the

above modeled dependence of the critical current on percentage imbalance, and that later obtained from self-consistent numerical calculations for a bulk double layer graphene system for the same interlayer separation, tunnel and gate dielectrics, is shown in Figure 3.8. At least in principle, however, the differences are not critical. As long as the critical current decreases with increasing imbalance in a controllable manner, the BiSFET logic circuits should work. Again, the gates are there only to establish which device reaches its critical current first in circuits. The onset of NDR we expect to be more like that of Figure 3.2(d) than that of Eq. (3.2) as illustrated in Figure 3.7(b), but the latter represents a much more conservative "leaky" approximation to the OFF state beyond the critical voltage. Indeed, I have used still slower, leakier models of the OFF state with relatively little effect on circuit simulation results. The essential requirement is only there is an NDR and that it can have an onset at critical voltages below the $k_BT/q \approx 26$ mV at room temperature.

3.3.2 BiSFET 2

With screening making it more difficult to achieve a condensate than initially estimated, the proximity of the gates—i.e., metallic field screening layers—in BiSFET 1 to the condensate became conceptually problematic, requiring the gates to be moved further away. With the gates further away, work function engineering would be insufficient to create the necessary carrier densities. The carrier densities would have to be provided by substantial gate voltages or by substantial doping with the gates removed. In either case, us-

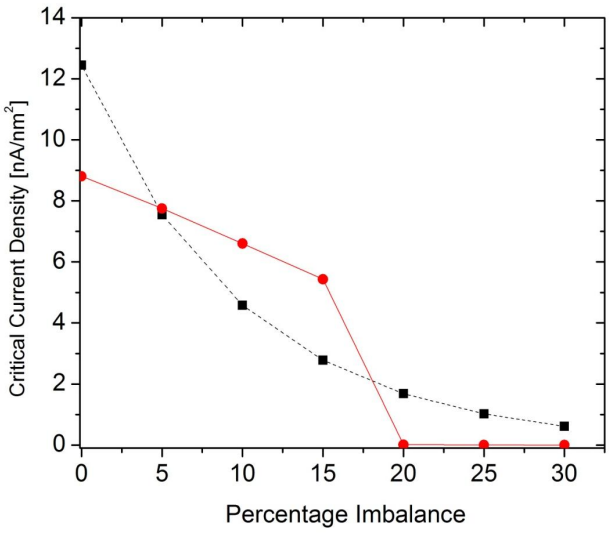

Figure 3.8: Critical current as a function of charge imbalance (as defined in Figure)from self-consistent exchange (red line with circles) compared to exponential dependence assumed in the BiSFET model (black line with squares).

ing the same gates to create and simultaneously control the condensate via voltages on the scale of $k_BT/q \approx 26$ mV becomes problematic. Therefore, the proposed BiSFET 2 design of Fig. 3.6(b) provides the large carrier densities necessary in the region of condensate formation via substantial but fixed voltages to well-insulated gates in a low-k gate stack environment. As a result, the voltages to these gates do not contribute to the switching energy or steady-state energy. Moreover, there is little capacitive coupling to these gates as well to further reduce switching energies. Instead, current crowding between two parallel condensates is employed. A gate is provided above one of the graphene leads to one of two condensate regions, where the gate effectively varies the contact resistance to that region. By increasing the effective contact

resistance to one of the two condensate regions, the current density into the other condensate is increased and the overall critical current is reduced: Once the condensate without gated contact resistance reaches its critical current and the current though it drops, the current through the condensate region without the gate contact resistance will also be pushed to its critical value in a chain reaction. Note that the "ON/OFF" ratio for this gating can be quite small. As for BiSFET 1, the gates are there only establish which device reaches its critical current first in circuits.

For a BiSFET 2 compact model, I simply add the model of gated contact resistance shown in Figure 3.19(a) to one lead for the compact model of Figure 3.6(a) for BiSFET 1, where the contact resistances of the other leads are assumed fixed. The capacitive coupling to the contacts used to create the condensate are substantially reduced. And, as noted, these latter gates are contacted only by fixed voltages.

3.4 BiSFET Logic

In this section, I illustrate how Boolean logic gates could be implemented using BiSFETs. We consider both BiSFET 1 and BiSFET 2 variants. Most of these circuits, except in Section 3.4.6, employ the BiSFET 1 variant simply because the original work was performed before the BiSFET 2 variant was developed. However, with both variants exhibiting a conductance peak intrinsically centered about zero interlayer voltage and a gateable critical current followed by NDR, one may expect that the essential circuit architectures and

scale of power consumption would be similar. And, indeed, in Section 3.4.6 I quickly revisit inverter and NAND Gates with the more recently proposed BiSFET 2, using the same basic circuit architecture and finding very similar results. For circuit simulations, I rely on SPICE-level simulations implemented with the previously described device models in Spectre [47].

Because of the negative differential resistance (NDR) characteristics, i.e., current decay—vs. saturation in MOSFETs—with increase of interlayer voltage, BiSFET or any device with similar NDR characteristics, cannot be used as a drop-in replacement for MOSFETs in CMOS based logic circuits. For such NDR devices, an entirely different way of implementing logic is required. Consider, for example, two BiSFETs, B1 and B2 in a CMOS-inverter-like layout with a fixed power supply voltage V_s of 25 mV, as shown in Figure 3.9(a). The solid curves in Figure 3.9(b) are the $I-V$ behavior of B1 for three different input voltages $V_{in} = 0$ mV, 12.5 mV and 25 mV, as obtained from Eqs. (3.2) and (3.3). The dotted curves, which represent the load line for B1, are the $I-V$ behavior of B2 at the same input and supply. The voltage axis for the $I-V$ characteristics of BiSFETs B1 and B2 is indicated by the arrows in Figure 3.9(b). The three points of intersection for each input voltage indicate stable points of operation, with associated output voltages of $V_{out} \approx 0$ mV, 12.5 mV, or 25mV. However, as can be seen, these output voltages show little dependence on the input/gate voltage. That is, the input signal, V_{in}, to the gates will not switch the output signal under these conditions. As shall be seen, this insensitivity to the input voltage can be useful for memory. However,

this insensitivity clearly illustrates the incompatibility of BiSFETs as simple drop-in replacement for MOSFETs in conventional CMOS logic elements.

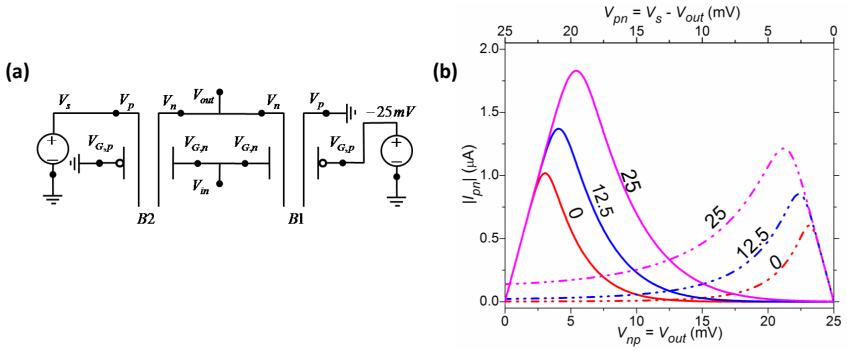

Figure 3.9: (a) BiSFET-based inverter and (b) I-V characteristics of B1 (solid) and B2 (dash) for fixed a fixed supply voltage V_s of 25 mV, and three different input voltages V_{in} (mV) shown along the curves. The magnitude of current, $|I_{pn}|$, across the devices B1 and B2 is plotted on Y-axis. The voltage across the terminals for B1 and B2 are shown on bottom X-axis and top X-axis, respectively. The intersection points of the curves for B1 and B2 indicate possible stable operating points(Reprinted with permission from [12], copyright (2010) by the IEEE).

3.4.1 Inverter

Despite the above discussion, a BiSFET-based inverter can be realized within the CMOS-like complementary layout of Figure 3.9(a). However, as suggested by the preceding discussion, the supply voltage V_s cannot be held fixed. Rather, it must be ramped up and down as a function of time t. Consider first, for simplicity, low frequency operation with a clocked power supply voltage $V_s(t)$ varying between 0 mV and 25 mV, as illustrated in Figure 3.10. Figure 3.10(a) shows quasi-static $I - V$ characteristics for BiSFETs B1 and

B2 of Figure 3.9(a) for values of the now clocked supply voltage $V_s(t)$ increasing from 0 mV to 25 mV, for a fixed ("high" or "one") 25 mV gate input voltage V_{in}. The voltage axis for the I-V characteristics of BiSFETs B1 and B2 is shown by the solid and dotted arrows in Figure 3.10(a). The numbered (non-uniformly) time-ordered intersection points of the $I-V$ characteristics of B2 with the lower edge of the figure indicate the supply voltage; the intersection points of the $I-V$ characteristics of the two BiSFETs with each other indicate the corresponding output voltage. Figure 3.10(b) shows the same for a fixed ("low" or "zero") 0 mV V_{in}. Fig. 3.10(c) shows the clocked supply voltage $V_s(t)$, an input signal $V_{in}(t)$ shaped to illustrate that only the input signal during the upwards clock ramp matters, and the corresponding output voltage $V_{out}(t)$, as a function of time. In the case of a high gate input voltage (logic 1), Figure 3.10(a), and $V_s(t) = 0$ mV, the interlayer voltage drop across both BiSFETs is zero, along with the output voltage. As $V_s(t)$ begins to increases with both BiSFET in their low-resistance states, $V_s(t)$ is split approximately equally across the two BiSFETs and $V_{out}(t)$ begins to increases, as the intersection point of the two BiSFET $I-V$ characteristics begins to move up and to the right. However, at $V_s(t)$ $\approx$8 mV, the current in the BiSFET with the (more) unbalanced bilayer charge densities, B2, reaches its maximum allowed value. Beyond this value of $V_s(t)$, B2 enters into its NDR regime and the current through both serially connected BiSFETs must decrease, and the intersection point of the two $I-V$ curves now begins to move back down and to the left. By the time $V_s(t)$ reaches 25 mV, B2 is in a high resistive state

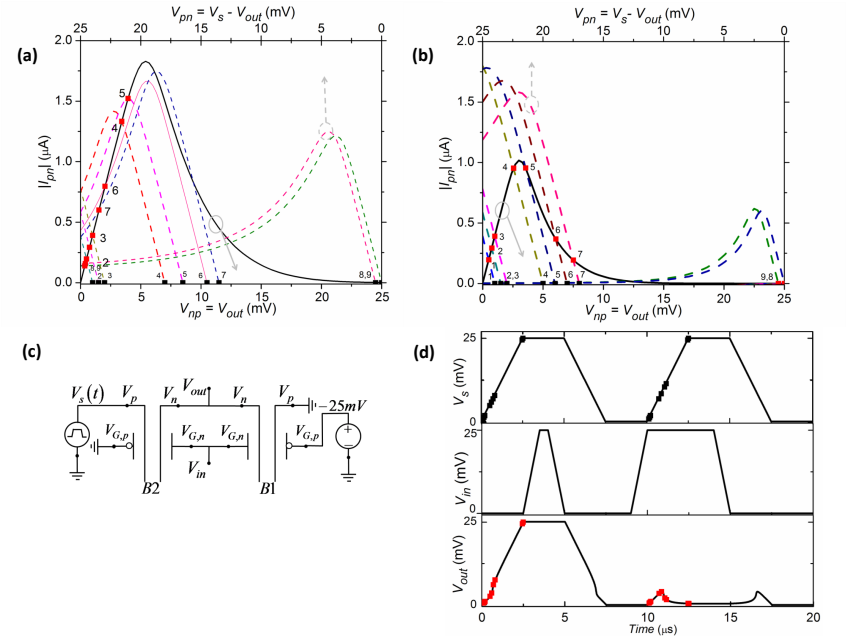

Figure 3.10: Quasi-static I-V characteristics of BiSFETs B1 (solid line) and B2 (dashed lined) of the inverter shown in (c) with adiabatically varying clocked supply voltage $V_s(t)$ and fixed V_{in} of (a) 25 mV, (b) 0 mV. (d) Results of SPICE-based simulation of the BiSFET inverter showing the, here, low-frequency clocked supply voltage V_s, deliberately aperiodic input signalto illustrate that only input signal during the upwards clock ramp mattersand the corresponding output voltage V_{out}. The small squares in (d) correspond to the similarly marked time-ordered numbered intersection points in (a) and (b). Reprinted with permission from [12], copyright (2010) by the IEEE.

well into it's NDR regime while B1 remains in a low resistance state, $V_s(t)$ is dropped predominately across B2 accordingly, and $V_{out}(t)$ approaches 0 mV (logic 0).

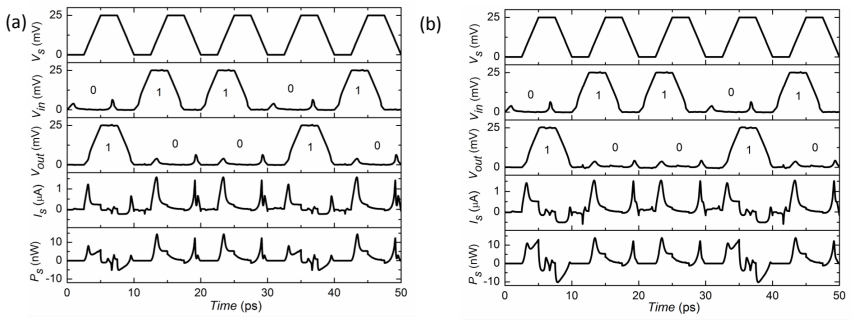

Figure 3.11: SPICE simulation results for inverter with (a) one inverter load and (b) four inverter load. I_s is the current supplied and $P_s = V_s I_s$ is the instantaneous power supplied by V_s (Reprinted with permission from [12], copyright (2010) by the IEEE).

The operation of inverter with $V_{in} = 0$ (logic 0), Fig 3.10(b), can understood in the similar way. This time, however, B1 has the unbalanced charge densities and reaches its maximum current value and enters into its NDR regime with increasing $V_s(t)$. Accordingly, by the time $V_s(t)$ reaches 25 mV, $V_s(t)$ is dropped predominately across B1 and $V_{out}(t)$ approaches 25 mV too (logic 1).

In either case, as illustrated in Figure 3.10(d) the input is required to set the output signal, but not to maintain it. And in terms of timing, the output signal follows the clock signal and not the input signal. In other words, each gate shows a self-latching behavior such that once the output of a gate is determined, the prior gate can be released for processing subsequent

information. In a sequence of BiSFET-based gates, there is no need to hold the input to the first gate until the output of the final gate is determined as in CMOS. The cost is having to clock each gate.

Note that Fig 3.10(c) merely represents one possible inverter configuration. For example, which gate of which BiSFET has a fixed signal and which is switched is of secondary importance. Of primary importance is only that the charge distribution in one BiSFET is initially (more) balanced and the other (more) unbalanced as the clock signal is ramped on.

At higher frequencies, displacement currents required to quickly charge the various capacitances must also be considered and the current flows and output voltages will no longer precisely follow the intersection points of the quasi-static curves of Figs 3.10(a) and (b). And at sufficiently high combinations of frequencies and capacitive loads, these displacement currents will (as for conventional CMOS) cause the logic to fail, here by, e.g., pushing B2 into its NDR region during the ramp-up of $V_s(t)$via load currents running through it but not B1, even when it has the larger of the two gate-controlled values of I_{max}.

Still, I have been able to run this BiSFET inverter gate at a 100 GHz in SPICE simulations with a seven inverter fan-out load representing some combination of subsequent gates and interconnect capacitances. Results for e.g., four inverter load is shown in Fig. 3.11(b) (and will be further discussed below). In these simulations, the clocked power supply $V_s(t)$ signal was taken as trapezoidal in time and delayed relative to the input signal $V_{in}(t)$—supplied

by a preceding inverter with the same load—by the 2.5 ps rise time of the clock, allowing the latter to be set before the inverter is clocked. In contrast, the timing of the output signal $V_{out}(t)$ simply follows $V_s(t)$. I have confirmed that further fan out is possible by reducing the clock rate and/or by optimizing the fixed gate voltages. It should also be possible by optimizing the shape of $V_s(t)$ and/or the relative peak/charge-balanced values I_{max} for the two BiSFETs.

To calculate the energy consumed by the BiSFETs themselves per switching in such an inverter, ignoring parasitics, we integrated the instantaneous power supplied by $V_s(t)$, $P_s(t) = V_s(t)I_s(t)$ over one clock cycle in the SPICE simulations, where $I_s(t)$ is the current flowing out of the clocked supply to one BiSFET inverter with its output $V_{out}(t)$ serving as the logical input signal $V_{in}(t)$ to one following inverter, as shown in Figure 3.11(a). This way, both resistive energy losses across the BiSFETs, which occur during switching and under quasi-static conditions, and capacitive charging energies for the BiSFET gates which occur only during switching were considered. The energy used per inverter was then averaged over the two logical states and divided by two to obtain the average switching energy per BiSFET to be approximately 7 zepto-Joules (zJ = 10^{-3} aJ) per clock period, for these 100 GHz SPICE simulations. For comparison, according to the ITRS [16] current CMOS logic consumes 100 aJ per switching, and the "end of the roadmap" CMOS in 2020 will consume about 5 aJ. However, I have ignored the power dissipation required to deliver the clocked supply voltage. And, because whether or not the logical input state changes between clock cycles has little if any effect on power

consumption, the gate activity factor is effectively unity so long as there is a clocked $V_s(t)$ provided.

The switching energy quoted above, however, does contain some small negative instantaneous power $P_s(t)$contributions to the energy integral as can be seen in Figure 3.11(a). While of little consequence in that case, as number of subsequent gates and/or interconnect capacitances increase, along with an expected increase in total energy consumed, the importance of the negative power contributions increases as well. For example, if we increase the load to four inverters as per Figure 3.11(b), the calculated average power per BiSFET to the original inverter increases to either approximately 8 zJ or 12 zJ, depending whether or not those negative instantaneous power contributions are included in the energy integral. Comparing the time-dependent power $P_s(t)$ consumed during the clock cycle for high and low output states, significantly more energy is consumed during the $V_s(t)$ ramp-up for a high output, as required to charge the gates of the subsequent inverters. However, particularly during the $V_s(t)$ ramp-down as the subsequent gates are discharged, current flows back into the supply and some of that energy could be returned. To what degree, will depend on how the clocked power supply $V_s(t)$ is generated and how it is shared among gates—where gates in a low output state are still net consumers of power during the ramp-down of $V_s(t)$—which are issues not addressed here. However, the difference is not critical in this work where the purpose of such power calculations is only to provide very rough estimates, and in either case the total energy consumed is quite small.

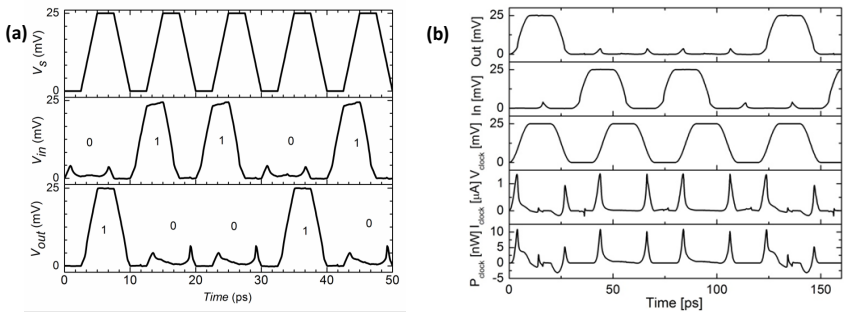

Figure 3.12: Supply voltage, input voltage signal, and inverted output signal obtained using SPICE for an inverter (a) with a 25 mV, 100 GHz clock using the BiSFET $I-V$ model of Eq. (3.4) (Reprinted with permission from [12], copyright (2010) by the IEEE), (b) with a 25 mV, 25 GHz clock with four inverter load with the Hartree-Fock based decay model shown in Figure 3.8

Of course, power consumption in CMOS MOSFETs depends critically on the static source-to-drain leakage current for an OFF-state MOSFET. The analogous current here is perhaps the interlayer tunneling current I_{pn} within the NDR region at the peak value of $V_s(t)$. Recall that, for example, a successful inverter operation requires one BiSFET in balanced state and other in imbalanced state. In other words one should have higher peak current than other and this is essentially achieved by unbalancing the other device. A weaker decay of critical current with imbalance would in principle imply a need for lower clock frequency. Otherwise, due to the capacitive currents inverter logic will fail. In order to understand the above mentioned possible issues I have included the decay behavior shown in Figure 3.8 into the spice model and re-checked some of earlier BiSFET simulations. I find that the inverter works with a 25 GHz clock as shown in Figure 3.12(b). The energy per operation is about 36 zJ which is nearly 4 times the energy per operation

reported earlier for BiSFET model using exponential decay for critical current. I have also considered the power consumed–and verified the logical operation of inverters, see Figure 3.12(a)–with another expected model of decay in the NDR region of the BiSFET characteristic but with the exponential decay of the critical current with percentage charge imbalance, replacing

$$I_{pn} = G_o V_{pn} \left[1 + \left(\frac{V_{pn}/V_{\max}}{V_{\max}/\left|V_{pn}\right|} \right)^4 \right]^{-1/4} \tag{3.4}$$

which decays in the NDR region as only $I \approx \pm G_o V_{\max}[V_{\max}/|V_{pn}|]$. The energy consumed per clock cycle per BiSFET was about 24 zepto-J due to increased resistive losses. While not insignificant, the large on/off ratio change is far less important here than it would be for CMOS because little time is spent under quasi-steady-state conditions; the input is set, the clocked supply signal is ramped on setting the output in the process, the output is held only long enough to set the output of the next gate, then the supply signal is ramped back off.

Also, a 25 mV signal level used through most of this work was chosen as a rounded off approximation to $k_B T$ at room temperature. With the assumed device characteristics and parameters of Section 3.2, the basic logic operation works for still lower voltages as illustrated in Figure 3.13(a) which shows the inverter characteristics with a peak clock voltage of only 15 mV, although with the clock pulsed at only 50 GHz. However, the energy consumed per BiSFET per clock cycle with a one BiSFET inverter actually increased to 38 zJ due to increased steady-state leakage power over a longer period of time.

Of course, at lower voltage, gate and interconnect charging energies would decrease. Accordingly, optimal supply voltage level would undoubtedly depend on application, as for CMOS, and, of course, actual BiSFET characteristics. Thus, I emphasize that the energy numbers provided here are intended only to ballpark the potential of BiSFET based logic.

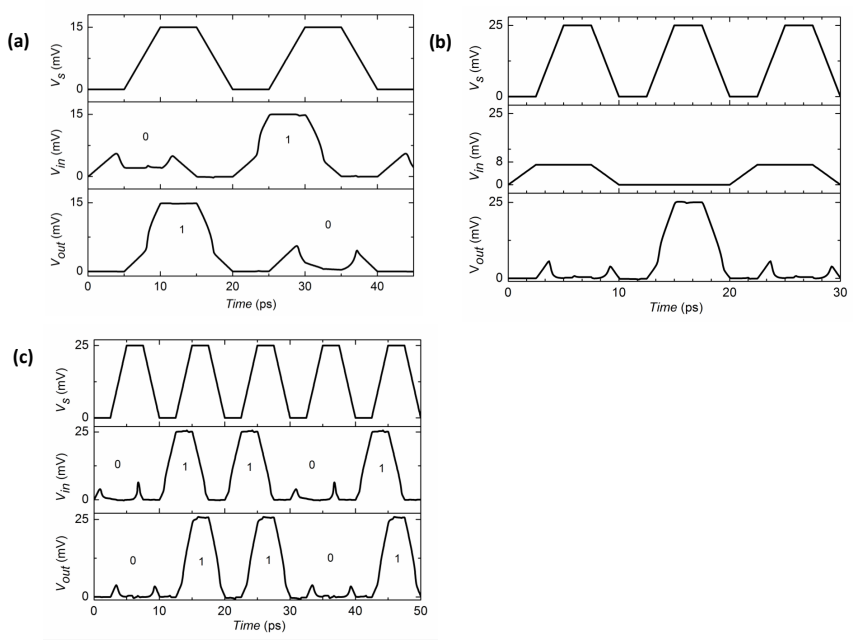

Figure 3.13: Supply voltage, input voltage signal, and inverted output signal obtained using SPICE for an inverter (a) with a 15 mV, 50 GHz clock, (b) with a 25 mV, 100 GHz clock but degraded input voltage, with full output signal restoration and (c) Signal follower with four followers as load. Reprinted with permission from [12], copyright (2010) by the IEEE.

The inverter (and BiSFET gates in general) also show gain and signal restoration. Indeed, there is a transition input voltage determined by the fixed gate voltages, below which all values of input voltage $V_{in}(t)$ will be processed as

logic-0, and above which all voltages above are processed as logic-1, without appreciably affecting the output signal $V_{out}(t)$ voltage swing. For example, with the p-gate fixed voltage of B1 adjusted to -13.5 mV, the full 25 mV output signal can be driven by an only 8 mV input signal, as shown in Fig 3.13(b) for a four-inverter load. This simulation was intended to mimic a case where the input signal from a previous stage is degraded, possibly due to voltage drops across long interconnects.

The inverter of Figure 3.10(c) can readily be converted to a signal follower by, e.g., setting the p-gate voltage of B1 to 0 mV, and B2 to -25 mV (or even 0 mV as it turns out), as shown via the result of Figure 3.13(c). Such a gate would be useful in BiSFET circuits for adding delays along a signal line to synchronize signals, as well as for increasing the fan-in or fan-out for a gate.

3.4.2 Inverter-based OR, AND, NOR and NAND gates

A conceptually simple way to create a NOR gate is to use the output of one inverter with Input A to power/clock the second with Input B, as illustrated in Figure3.14(a). When Input A is low, the first/upper inverter's output signal clocks the second inverter in phase with the original clock signal. If Input B is also low, the second/lower inverters output then follows its clock signal and, thus, the original clock signal, providing a high output from the gate. In contrast, if Input B is high, the output of the second/lower inverter and, thus, the gate will remain low independent of the output of the first/upper inverter determined by Input A. If Input A is high, there will be no clock

signal to the second/lower inverter and gate the output will again remain low independent of Input B. I have verified the logical outputs for all four inputs for such a NOR gate via SPICE simulation with the same clock signal and four inverter load considered above, as show in Figure 3.14(b). However the W/L for the first inverter had to be doubled to 4 to provide sufficient drive current to power the second inverter and subsequent loads, increasing the capacitive load for the preceding gate supplying Input-A to the NOR gate. To again consider just the power requirement of the NOR gate ignoring parasitics, the energy consumed was calculated, via the integration of $P_s(t) = V_s(t)I_s(t)$ in the SPICE simulations while supplying an_Input A and an Input B to the following NOR gate. The average calculated energy consumed per clock cycle per NOR gate (not BiSFET) was 40 zJ or 48 zJ, depending on whether or not the negative instantaneous power contributions were considered, respectively. Inverting the outputs of the above NOR gate, of course, creates an OR gate.

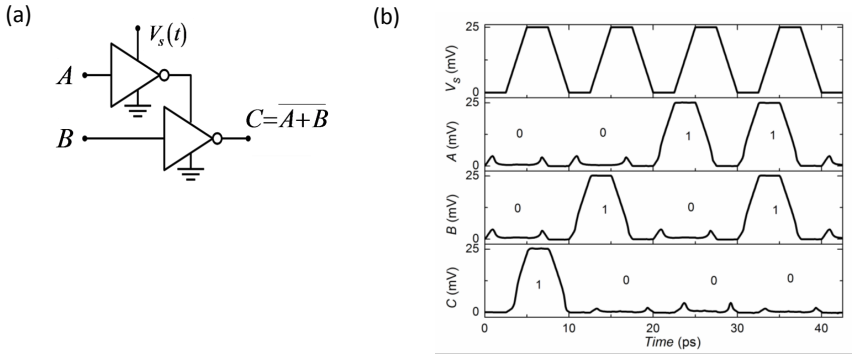

Figure 3.14: (a) BiSFET Inverter based NOR Gate and (b) Clock signal (100 GHz), input voltage signals A and B, and output signal C (Reprinted with permission from [12], copyright (2010) by the IEEE).

Inverting the inputs creates an AND gate, as well as eliminating the problem for reduced fan-in for Input A. Inverting both inputs and outputs produces a NAND.

3.4.3 Programmable NAND/OR gate

It may also possible to take a more direct and flexible route to creating such logic gates. Figure 3.15(a) shows a potentially programmable BiSFET-based NAND/OR gate. In this realization, Input A and Input B are applied to the n-gates of BiSFETs B1 and B2 which have a W/L ratio of 2, so there is no increase in capacitive load for the preceding gates. Meanwhile the Boolean functionality of this gate, NAND or OR, depends on the voltage, $V_{NAND/OR}$ applied to the p-gates of both B1 and B2, potentially allowing one to program the functionality of this gate these "other" gates are indeed contactable. There are no input signals to the larger BiSFET B3, however, but it's (here) n-layer gate does become part of the of the "output" signal load.

The operating principle of this circuit is similar to that of the inverter discussed in Section 3.4.1, in that there will be two $I-V$ curves whose intersection point is the output voltage. However, with no input signals, the $I-V$ curve for B3 will be fixed when the ramp-up of $V_s(t)$ begins. The other $I-V$curve represents the parallel combination of B1 and B2, and its shape changes with Input A, Input B and $V_{NAND/OR}$. A $V_{NAND/OR}$ of -25 mV produces a NAND gate functionality; a $V_{NAND/OR}$ of 0 mV produces an OR gate functionality. In this implementation, the W/L ratio of B3 was chosen to be

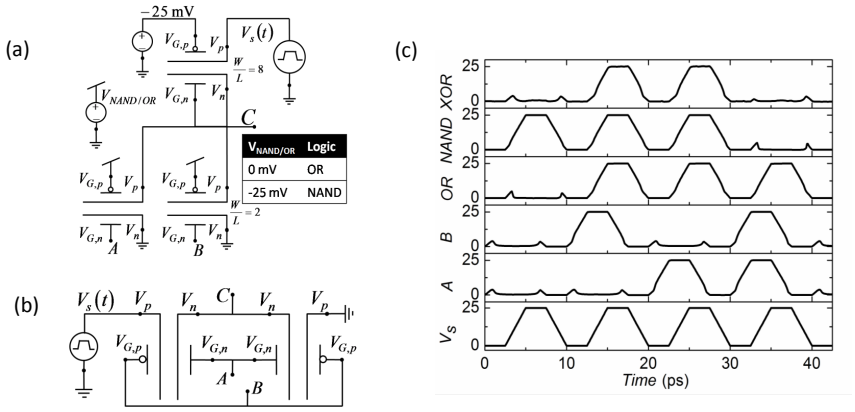

Figure 3.15: (a) BiSFET Inverter based programmable NAND/OR gate, (b) XOR gate and (c) SPICE simulations of the OR, NAND and XOR gates

8. If the W/L ratio was smaller than 6, the maximum current that B3 could supply was insufficient to drive both B1 and B2 and the load when the input conditions are such that both B1 and B2 have largest possible I_{max}. This condition leads to an always low output, irrespective of input conditions. On the other hand, if the W/L of B3 was very large, the output was always high. SPICE simulations for this gate and the XOR gate is shown Figure 3.15(c) again with a 100 GHz, 25 mV V_{peak} clock signal with an inter-gate delay of 2.5 ps, and a fan-out of 4 inverters. To once again just calculate the power requirement of the gate ignoring parasitics, the energy consumed was calculated, via the integration of $P_s(t) = V_s(t)I_s(t)$ in the SPICE simulations for one NAND/OR supplying an Input A and an Input B to subsequent NAND/OR gates. The average calculated energy consumed per clock cycle per gate (not BiSFET) was 16 zJ or 34 zJ depending on whether or not the negative instantaneous power contributions were considered, respectively, for the NAND

gate, and was 38 zJ or 50 zJ, respectively, for the OR gate.

3.4.4 Functional Block: A 4-bit Ripple Carry Adder

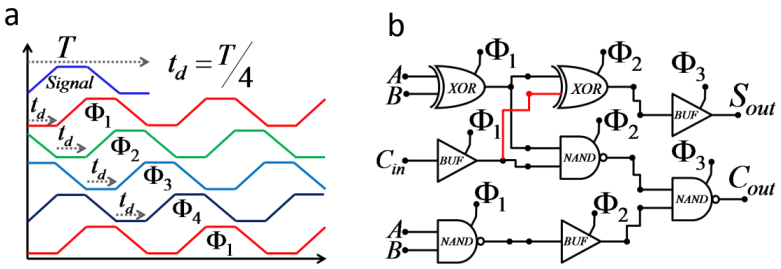

Figure 3.16: (a)Four phase clocking scheme for BiSFET based logic circuits and (b) Circuit schematic of a 1-bit full adder (Reprinted with permission from IOP Publishing: Journal of Physics. D [1], copyright (2011)).

Finally putting it all together, I present the largest BiSFET based circuit I have simulated, a 4-bit ripple carry adder. As shown in Figure 3.16(b), a 1-bit full adder is a three-stage circuit. In contrast, all of the above-mentioned circuits are single stage circuits requiring only one clock signal. However, as previously noted, the input signal to a gate must be set before the gate is clocked. Here I use a quarter clock period delays between input and clock signals for BiSFET based logic circuits. This delay implies a requirement for a four stage clocking scheme for gate connected in series, as shown in Figure 3.16(a), with a one-quarter cycle clock delay between stages. The result of addition is obtained after a delay of $3/4T$ where T is the clock period. And using this full adder I have implemented a four bit ripple carry adder which is shown in Figure 3.17(a), with SPICE-simulation verified functionality, as shown in Figure 3.17(b) for addition of each bit of two random 4 bit numbers.

At each of the four stages, the output is obtained after the delay of $3T/4 = 7.5$ ps, so that the total computation delay is $3T$. However, unlike for CMOS based adders, the inputs need not been held constant for the duration of the calculation. Because, each constituent BiSFET based logic gate also acts as a latch, new inputs can be considered each clock cycle in principle. This basic ability to consider new inputs at each clock cycle holds in BiSFET based logic circuits no matter how deeply logic stages might be stacked for some logic function.

3.4.5 Robustness: Noise and Jitter Studies

As already noted, these BiSFET gates intrinsically show signal restoration, so that variation in the input signal up to half of the supply signal can be tolerated without affecting the output. However, what about variations in the clock signal? All of the above mentioned SPICE simulation results were obtained using an ideal clock, i.e., no noise and jitter in the signal. As a preliminary step towards studying the effect of clock noise on the functionality of the gates, I have used a very high frequency, frequency modulated (FM) sine wave as noise. Figure 3.18 show the SPICE simulated response of various gates with a noisy clock, a noisy input signal, and with a delay scheme shown in Figure 3.18(a). For a nominal operation the clock edge should be delayed with respect to input signals by 2.5 ps or quarter clock period for a 100 GHz clock. For the jitter studies I have looked at the window of operation by determining the extreme limits of the clock edge drift in time in either direction from the

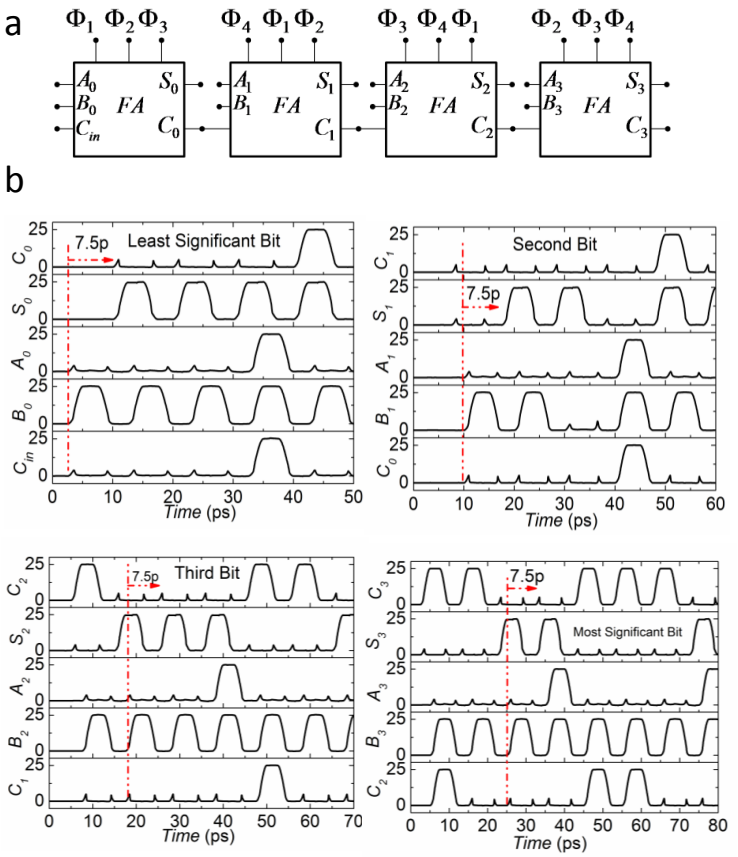

Figure 3.17: (a) Circuit schematic of a four bit ripple carry adder and (b) SPICE simulation based verification of the full adder functionality (Reprinted with permission from IOP Publishing: Journal of Physics. D [1], copyright (2011)).

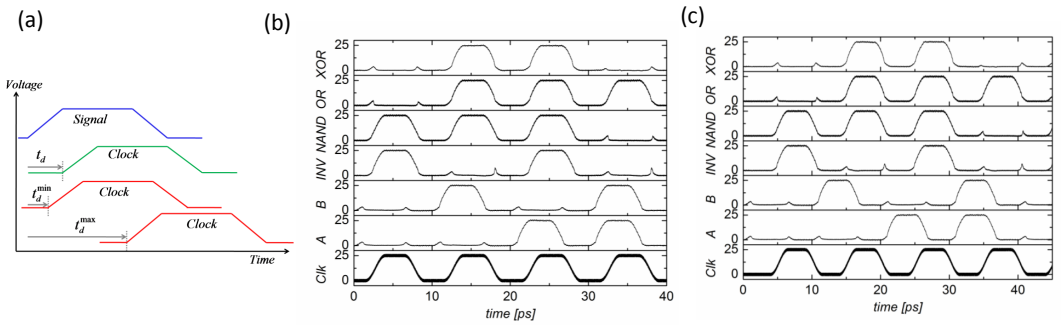

Figure 3.18: (a) Illustration of the delay window for clocking scheme, SPICE simulation results of the basic BiSFET based logic gates: Inverter, NAND, OR, and XOR with a 100 GHz noise clock (noise = 1.5 mV, 1THz FM sine wave) for two different delays: (b) nominal 1.5 ps, (c) nominal + 1.5 ps. Note that the input signals A and B are also noisy as they are outputs of a inverter with the same noisy clock.

nominal clock edge position with respect to the input signal. Figure 3.18(b) and (c) show the verification of functionality of the logic gates when the clock edge comes as early as 1ps after the input or as late as 4.0 ps after the input giving a 3ps window for the clock edge arrival. I have also considered the effect of relaxing the clocking scheme by increasing the clock period from 10 ps (corresponding to 100 GHz clock signal) to 15 ps (67 GHz) while keeping the ramp times at 2.5 ps. In this case the widow for jitter rises from an already relative large 3 ps up to a still larger 4.5 ps with the relaxed clocking scheme. However, the OR gate fails in functionality. The results of these studies begin to address the potential sensitivity of the basic BiSFET logic gates' functionality with respect to noise and jitter in the clock signal. However, the effect of noise and clock jitter on the BiSFET based circuits functionality requires an even more detailed SPICE model including parasitics and realistic clock

circuits. This may be taken up in future work.

3.4.6 Inverter and NAND gates revisited with BiSFET 2

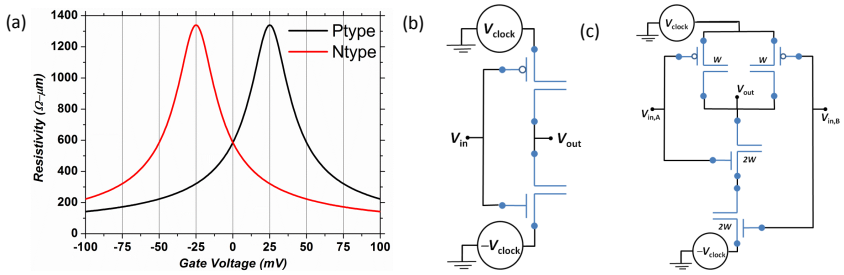

Figure 3.19: (a) "Contact" FET compact model used for SPICE simulation, (b) schematic of inverter circuit, and (c) schematic of NAND circuit. In (a) for the 25 mV supply voltage of interest, only the gate voltage region between ±25 mV affects the operation. I have also considered much smaller "ON/OFF" ratios without affecting simulated switching

As noted above, both BiSFET variants exhibit a conductance peak intrinsically centered about zero interlayer voltage and a gateable critical current followed by NDR, and therefore one might expect circuit architectures and scale of power consumption to be shared. As previously noted, Figure 3.19 (a) shows the model we have used for the gate-modulated resistance of graphene. Figure 3.19(b) and Figure 3.19(c) show the schematic of Inverter and the NAND circuit based on BiSFET 2. Figure 3.20 shows the SPICE simulated response of an inverter and a NAND gate with a 25 mV, 10 GHz clock, verifying their functionality. Energy per operation per BiSFET for the inverter with fan-out of four inverters is about 10 zJ, much as for BiSFET 1. Energy per operation per NAND gate as a whole with no load is about 140 zJ, more than for BiSFET 1 simulations, but still quite small.

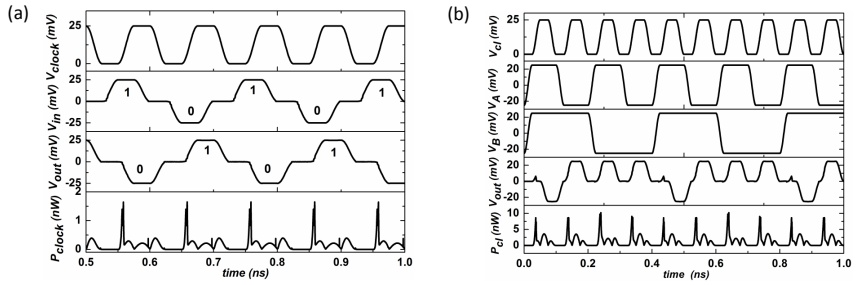

Figure 3.20: SPICE simulated response of (a) inverter and (b) NAND with 10 GHz, 25 mV clock for BiSFET 2. The energy per operation per BiSFET for an inverter with fanout of 4 is about 10 zJ, and the energy per operation for the NAND gate as a whole with no load is 140 zJ

3.5 BiSFET Memory

As discussed at the beginning of Section 3.4, when a BiSFET is used in a CMOS-like inverter configuration with a fixed supply voltage, the output is locked at one of the three possible operating points, at least for limited output loads. The output can, however, be pulled high or low even with the fixed supply voltage by effectively "grounding" it as though adding to much fan-out to the desired value. With the "ground" removed, the output signal will again be locked.

This behavior can be exploited to implement a 2-BiSFET static memory cell, as shown in Figure 3.21(a) along with a peripheral circuit to the right used in the SPICE simulations to test the functionality of the memory element. Note that all of the BiSFET gates are grounded—that is no contact to the "gates" is actually required—while the only "input" to the BiSFETs is the output/memory ("mem") signal. When ON, the transmission gate, Tx,

couples the data signal to the bit line. The access gate, Ac, in turn couples the bit signal to the memory element. Writing consists of turning on Tx to charge the bit-line capacitance, and turning Ac ON to write the data onto the particular memory element from a low impendence source. That is to effectively grounding the Mem output node of that element high or low to exceed the I_{max} of B2 or B1, respectively, to place it in its NDR region. Reading consists of turning on Ac to write signal stored in the memory the bit line, and turning on Tx to read the signal from the bit line via, e.g, a high input impedance sense amplifier, so as not to exceed the I_{max} of the currently ON/low resistance BiSFET.

Memory operation can be seen in the transient simulation result shown in Figure 3.21(b), which is demarcated into six rows (R1-R6) and four columns (C1-C4) for ease of understanding. For these simulations a pre-charge circuit has been added. In the Write cycle for a "one", the transmission gate is turned on with the Tx pulse (see R1 and C1). When the data signal arrives (R2 and C1), the bit line goes high (R5 and C1) and then the access gate is turned on (R3 and C1), causing the voltage at the memory node Mem to go high (R4 and C1). When the access is turned off the BiSFET Mem node is still high i.e., a "one" is stored in the memory. The data stored in the memory can be read as follows: First the pre-charge gate Pc is turned on to charge the bit line to a known voltage level. Then, when the access gate is turned on (R3 and C2) the voltage on bit line goes high (R5 and C2) implying that a "one" is stored in the memory. Similarly, columns C3 and C4 demonstrate

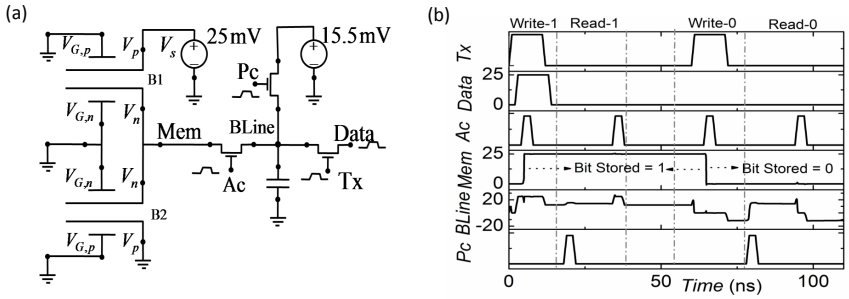

Figure 3.21: (a) BiSFET-based 1-bit static memory cell along with the peripheral circuits used for testing (b) SPICE simulated read and write cycles for the memory cell. Tx, Ac, and Pc are the gate voltages for the MOSFETs as shown in the circuit schematic. The peak value of gate pulses to MOSFETs is 550 mV. Mem, Data and Bline are the voltages in mV at the nodes Mem, Data and Bline respectively, as shown in circuit schematic. The supply voltage is fixed.(Reprinted with permission from [12], copyright (2010) by the IEEE).

the write and read cycles for a "zero". Note during the read cycle for the data bit "one" or "zero" stored in the memory, the bit line goes high or low, respectively. This variation of the bit line voltage can be sensed using a sense amplifier to read out the stored bits from the memory. For proof of concept, we have used conventional MOSFETs for the peripheral circuits in order to perform SPICE simulations, and BiSFETs only to implement the memory cell itself. The gate pulses used for the MOSFETs are 550 mV high. Pulse width for the access gates is 4 ns, which by no means represents the upper limit on the basic two-BISFET memory element as evidenced by the prior simulations of logic gates. The results of this simulation should only be read as qualitative verification of functionality. Quantitative aspects such as times for read write cycles, sensitivity and stability, etc. have not yet been tested.

However, I note that the power consumed by BiSFET memory elements

will be much more sensitive than BiSFET-based logic gates to the rate of current decay in the NDR tail of the BiSFET $I - V$ characteristic. But, as the voltage scale is defined by $V_{\max}$, which can be less than $k_B T$ in principle, low voltage and low static power storage may be possible.

3.6 Conclusion

I have reviewed the essential physics underlying BiSFET operation. I have described the evolving BiSFET design, and SPICE models thereof as required for circuit simulation. I have then provided BiSFET-compatible circuit designs for various Boolean logic gates as well as a simple memory element. These designs and their dependence on a clock power supply, and, of course the scale of the power supply, are quite different than CMOS circuit designs. In particular, I have exhibited potential device switching energies on the 10 zJ scale, two to three orders of magnitude below even end-of-the-Roadmap CMOS switching energies. I also have pointed out substantial challenges, both theoretical and technological, to realization of BiSFETs. I believe, however, that the potential benefit of the former justifies the risk associated with addressing the latter.

Chapter 4

Interlayer Tunnel Field Effect Transistor

4.1 Introduction

Electronic devices have been explored in the past based on resonant single-electron CB (conduction band) to CB tunneling between parallel quasi-two dimensional (2D) quantum wells within III-V heterostructures and their accompanying negative differential resistance (NDR) [19]. In resonance, both energy and in-plane crystal momentum can be conserved during interlayer tunneling and interlayer conductance can be high; out of resonance, both cannot be conserved simultaneously and interlayer conductance is ideally zero. At least that is the case to the extent that both energy and crystal momentum are good quantum numbers; when they are not, or at least not entirely so, such as due to various scattering processes, the peak is broadened and the conductance decays more-or-less smoothly away from the resonance peak. The result is the afore-mentioned NDR.

Such devices are attractive for high speed electronics, and digital logic circuits also have been demonstrated using a combination of conventional and such NDR FETs [64]. As I have discussed in Chapter 3, the BiSFET is also based on an NDR current-voltage characteristic. Based on SPICE simula-

tions, I have shown that BiSFET can be a plausible alternative for CMOS for ultra low voltage and low power operation. The BiSFET device concept is based on many-body enhanced tunneling that can occur in the double layer graphene system under appropriate conditions. The tunneling current in the BiSFET is due to strong coupling of conduction band states (electrons) in one layer to valence band states (holes) in another layer via Coulomb interaction. However more conventional single-particle resonant tunneling could also occur in sufficiently closely spaced graphene layers even in the absence of such Coulomb-mediated many body interactions via resonant single particle tunneling, either via CB to CB tunneling or equally well via valence band (VB) to VB tunneling, given the band-structure symmetry in energy about the Dirac points. The latter possibility, for example, suggests the possibility of "complementary" circuit architectures like CMOS, in at least, that way.

Recently, NDR has been experimentally observed in current flow between graphene layers [65]. In this Chapter, I re-evaluate a device concept [19] which we refer to as Interlayer Tunnel Field Effect Transistor, (ITFET) (and previously referred to as a Double Electron Layer Tunneling Transistor [19]) considering also the graphene material system, and ultra-short, sub-100 nm channel lengths. The atomically near-perfect 2D nature of the component graphene layers, low scattering rates and, as noted above, the conduction/valence band symmetry offer advantages over III-Vs.

As I have shown in Chapter 3, a device with strong NDR characteristics can not be a drop-in replacement for MOSFETs, instead requiring rather novel

approaches for logic circuit design. The logic design approach for BiSFETs using multi-phase clocked power supply presented in Chapter 3 in principle can be used for any device with similar NDR characteristics. Because ITFETs have an NDR characteristic, there is a possibility of shared circuit architectures with BiSFETs. Thus, logic design approaches for BiSFETs using multi-phase clocked power supply presented in Chapter 3 may also be appropriated for ITFETs. On the other hand, there remain qualitative differences in the $I-V$ characteristics of these devices due to the fundamentally different physics underlying the NDR in the BiSFET and the ITFET systems, so required detailed logic gate design could vary substantially. Furthermore, due to finite channel regions of overlap/coupling between the two layers of graphene or other well materials, I will illustrate that there are short-channel resonance-broadening effects associated with Heisenberg position-momentum uncertainty translating into possibly less-sharp resonances and higher voltage operation, which may be particularly problematic for graphene. Thus, in this chapter I explore, compare and contrast the essential physics of ITFETs based on III-V double quantum well and dielectrically separated bilayer graphene systems using analytic and numerical non-equilibrium Green's function (NEGF) based calculations.

4.2 Basic Device Physics

A schematic of the ITFET is shown in Figure 4.1(a) with independent Contacts C1 and C2 to Layers 1 and 2, respectively, and a simple equivalent circuit model is shown in Figure 4.1(b). Gates G1 and G2 are used to control

the potential difference between the layers. Carriers that enter the Layer 1 via Contact C1, may tunnel to Layer 2 and then exit via Contact C2, or may reflect back out of contact C1 (with or without having tunneling to Layer 2 along the way). As an example, consider an ITFET with graphene layers. The band structure of the Layer 1 and Layer 2 at Potential φ_1 and Potential φ_2, respectively, are shown in Figure 4.1(c). When the interlayer potential difference is zero, the layers are in resonance. The current is due to all the states that satisfy energy and momentum conservation and, in the limit of zero temperature for clarity of explanation only, lie between the Fermi levels of the layers as shown by the shaded region between the Fermi surfaces in Figure 4.1(d). At higher temperatures, the energy range of current flow is smeared beyond the Fermi levels a bit, but this smearing does not directly contribute to broadening of the resonance. Only to the extent scattering increases with temperature, energy and momentum thus become less precise quantum numbers, will the resonance be broadened by temperature increases.

Qualitative comparison of the current voltage $(I - V)$ characteristics of an ITFET and the BiSFET is shown in Figure 4.2(a) and (b) respectively. The peak current in an ITFET at resonance occurs when the interlayer potential difference is zero and the peak current is due to resonant tunneling of all the states that can tunnel from one layer to another at that interlayer bias, V, as shown in Figure 4.2(c). When out of resonance with $\Delta\varphi \neq 0$ as shown in Figure 4.2(d), current is reduced as it flows only to the extent that energy and/or crystal momentum are broadened. Unlike the BiSFET, the

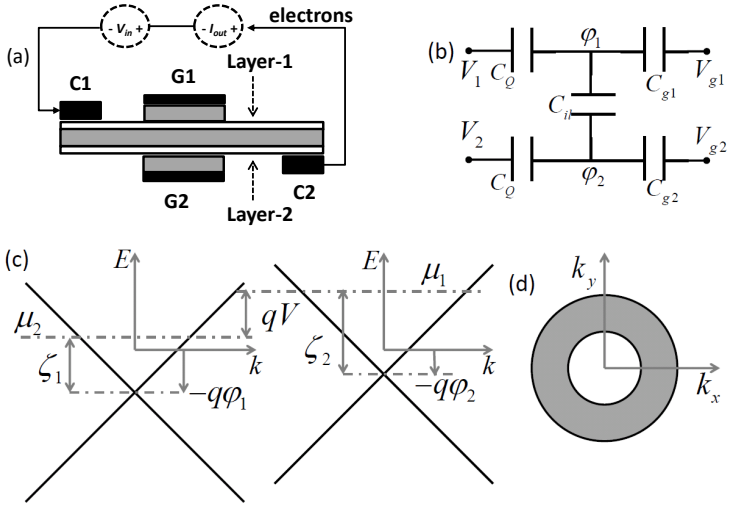

Figure 4.1: (a) Illustrative drawing of the interlayer tunnel FET, (b) Simple equivalent capacitance circuit of the ITFET, (c) Band structure of the graphene layers 1 and 2 at resonance showing the Fermi levels of each layer and (d) Fermi surface of Layer 1 and Layer 2 at resonance. The current is due to all states in the gray annular area at 0K. The edges of the annular region are smeared at higher temperatures, but this is not directly relevant to the resonance condition. (figure b,c and d reprinted with permission from [13], copyright (2012) by the IEEE).

peak conductance condition, equals resonant condition here, is not centered around zero interlayer bias for all gate voltages. Rather, resonance occurs at an appropriate interlayer bias and gate voltage conditions that result in zero interlayer potential difference subject to the self-consistent electrostatics of coupling of the layers to each other, the gates, and the source and drain. This qualitative difference in I-V characteristics means that, even assuming NDR-based-switching and a similar clocking scheme as for BiSFETs, required detailed logic gate design could vary substantially.

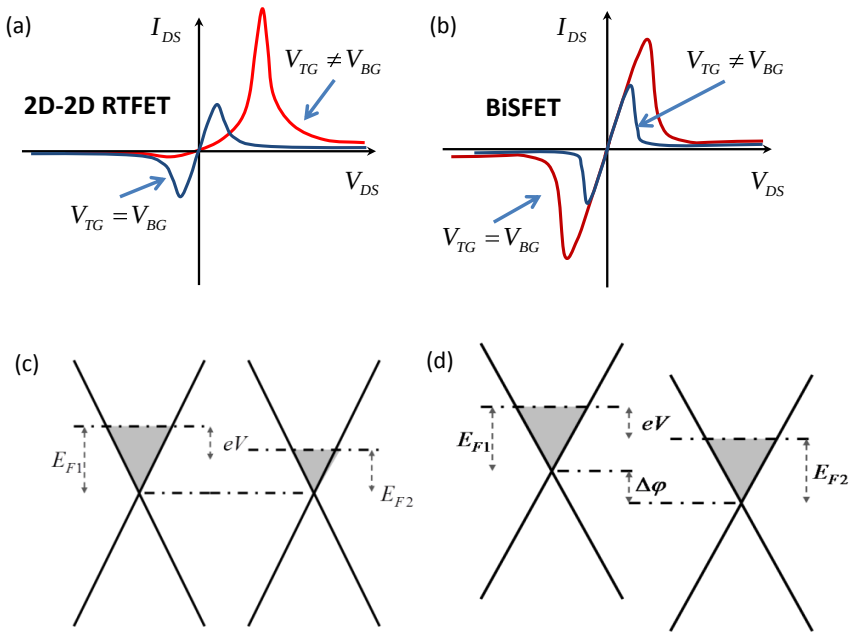

Figure 4.2: Qualitative comparison of expected I-V characteristics. (a) 2D-to-2D resonant tunneling, (b) BiSFET, (c) Band alignment of graphene layers in resonance and (d) Band alignment of graphene layer out of resonance for the same interlayer voltage difference (the same Fermi level difference).

4.3 Capacitance-Voltage Model

Consider the double layer graphene device whose schematic is shown in Figure 4.1(a) with applied gate voltages V_{g1} and V_{g2} and the interlayer bias $V = V_2 - V_1$. Let the gate voltages corresponding to charge neutrality in graphene layers be $V_{g1} = V_{FB1}$ and $V_{g2} = V_{FB2}$, respectively, for Layers 1 and 2. Using the equivalent circuit model show in Figure 4.1(b), we can write the charge voltage equations for the device as,

$$\begin{pmatrix} Q_1(\zeta_1) \\ Q_2(\zeta_2) \end{pmatrix} = -\begin{pmatrix} C_{g1} & 0 \\ 0 & C_{g2} \end{pmatrix} \begin{pmatrix} V_{g1} - V_{FB1} - \frac{\zeta_1}{e} - V_1 \\ V_{g2} - V_{FB2} - \frac{\zeta_2}{e} - V_2 \end{pmatrix} + \begin{pmatrix} C_{il} & -C_{il} \\ -C_{il} & C_{il} \end{pmatrix} \begin{pmatrix} \frac{\zeta_1}{e} + V_1 \\ \frac{\zeta_2}{e} + V_2 \end{pmatrix} \tag{4.1}$$

where C_{g1}, C_{g2}, and C_{il} are the capacitance of the Gate 1 to Layer 1, Gate 2 to Layer 2 and the interlayer tunnel dielectric capacitance, respectively. The layer potentials are given by $\varphi_1 = \zeta_1/q + V_1$ and $\varphi_2 = \zeta_2/q + V_2$, where the conduction band referenced Fermi levels $\zeta_{1/2}$ are defined as shown in Figure 4.1c. The charge density in graphene layers is $Q(\zeta) = qn_i\left(F_1(-\zeta/k_BT) - F_1(\zeta/k_BT)\right)$ where n_i is the intrinsic carrier density in graphene, q is the electron charge and F_1 is the Fermi integral of order 1.

4.4 Current Voltage Model

To first provide a simple model (which I will go beyond in later sections) of the current voltage characteristics of the graphene based ITFET, I used a perturbative tunneling Hamiltonian approach [66,67]. The Hamiltonian of the

coupled system, assuming A-B coupling, is given by

$$H = \sum_{ks} \varepsilon_{1sk} a^{\dagger}_{1sk} a_{1sk} + \sum_{ks} \varepsilon_{2sk} a^{\dagger}_{2sk} a_{2sk} + \frac{1}{2} \sum_{k} t_k \left(a^{\dagger}_{1ck} - b^{\dagger}_{1vk} \right) \left(a_{2ck} + b_{2vk} \right) + h.c \tag{4.2}$$

where the first two terms represent the Hamiltonian of Layers 1 and 2. The summation over momentum k is around the Dirac points in the respective layers. The summation over $s = c, v$ is for conduction and valence bands, respectively. The conduction band energy is given by $\varepsilon_{ick} = \hbar v_F k - q\varphi_i$ and the valence band energy is given by $\varepsilon_{ivk} = -\hbar v_F k - q\varphi_i$ with $i = 1, 2$, where again, $\varphi_{1/2}$ is the potential in Layers 1 and 2 respectively. The third term in Eq. (4.2) is the coupling between the layers obtained by assuming, for specificity, AB coupling of strength t between the graphene layers. The interlayer coupling is given by $t_k = te^{(-i\theta_{1k} - i\theta_{2k})/2}$ where θ_{1k} and θ_{2k} is the direction of the Bloch momentum in layers 1 and 2 with respect to corresponding Dirac points. Assuming perfectly aligned layers, $t_k = te^{-i\theta_k}$.

The tunneling current between the graphene layers for an applied bias $\mu_1 - \mu_2 = qV$ is given by,

$$I = -\frac{q}{\hbar} \int_{-\infty}^{\infty} T\left(E\right) \left[f\left(E - \mu_1\right) - f\left(E - \mu_2\right) \right] \frac{dE}{2\pi} \tag{4.3}$$

where q is the charge of electron, $\mu_{1/2}$ is the Fermi level in layers one and two, $f(E)$ is the Fermi distribution function. $T(E)$ is the transmission between the layers given by,

$$T\left(E\right) = \sum_{k;ss'} \left|t_k\right|^2 A_{1s}\left(k, E\right) A_{2s'}\left(k, E\right) \tag{4.4}$$

where A_1 and A_2 are the spectral density functions for Layers 1 and 2. The spectral density of states is assumed to be of the form,

$$A_s(k, E) = \frac{2\Gamma}{(E - \varepsilon_{sk})^2 + \Gamma^2} \tag{4.5}$$

where Γ represents the energy broadening of the states and, for scattering only, represents the inverse mean free lifetime of the carriers. See Appendix E for more details on the derivation of the above set of equations. For example, in the limit of an infinite lifetime Γ , the $A(k, E)$ becomes delta function $\delta(E\epsilon_{sk})$, and therefore $T(E)$ becomes a delta function $\delta(\epsilon_{1sk}\epsilon_{2sk})$, and a perfect resonance condition is achieved, again independent of temperature. Note that Eq. (4.3) to Eq. (4.5) are not specific to graphene but are applicable to tunneling between weakly coupled systems assuming particle conservation.

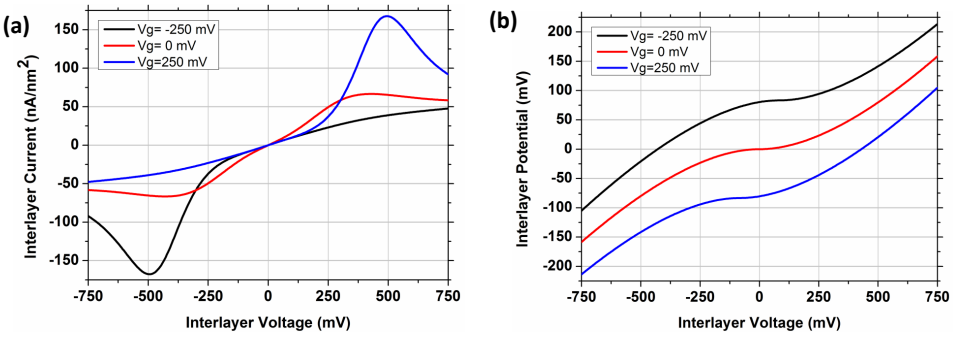

Figure 4.3: (a) Interlayer tunneling current as a function of interlayer voltage illustrating the negative differential resistance and (b) Interlayer potential as function of interlayer voltage for three different gate voltages split equally with opposite polarity between the gates. The data is obtained using t=25 meV and Γ = 10 meV and effective oxide thickness (EOT) of 0.8 nm for the gate and interlayer dielectrics. The graphene layers are assumed to be undoped. (Reprinted with permission from [13], copyright (2012) by the IEEE).

Figure 4.3(a) shows the interlayer tunneling current as a function of interlayer bias obtained by numerically integrating Eq. (4.3) for three different gate voltages. The dependence of interlayer potential on interlayer bias obtained by solving Eq. (4.1) is shown in Figure 4.3(b). The graphene layers are in resonance and the interlayer current peaks in Figure 4.3(a) when the interlayer potential difference is zero. The magnitude of the current also is—roughly allowing for the thermal smearing of the Fermi surfaces at 300K—proportional to the number of states in the annular area shown in Figure 4.1(d) between the Fermi surfaces defined by the voltage difference between layers.

The above discussion illustrates that large area graphene based device with tunneling limited transport illustrate the expected NDR current-voltage characteristic. In the above calculations, the value of interlayer coupling and the broadening of density of states is chosen arbitrarily. Figure 4.4 illustrate the effect of the gate voltages, interlayer hopping strength, T_{hop}, and the density of states broadening Γ on the current voltage characteristics. I have assumed the graphene layers are electron doped a "flat-band" gate voltage, that required to get the layers to charge neutral due to residual doping, of $V_{FB} = 500$ mV, and an effective oxide thickness (EOT) of 0.8 nm for the gate and interlayer dielectrics. Due to the capacitive coupling, the interlayer potential difference is roughly proportional to the gate voltage difference, $\Delta V_g = V_{g1} - V_{g2}$. In this model, the peak current scales as the square of the interlayer coupling, where as the broadening of the NDR $I - V$ is governed by the Γ. (For reference, in contrast, the peak/critical current BiSFET scales

linearly with the interlayer coupling as discussed in Chapter 3.) Figures 4.4(a) and (c) show the current response to interlayer bias with same gate bias to gates $G1$ and $G2$. The position of the peak current is only altered slightly with changing gate bias since $\Delta V_g = 0$, and the interlayer bias is weakly coupled to interlayer potential via the quantum capacitance of the graphene layers. Figures 4.4(b) and (d) show the calculated current response when the applied gate bias is split equally between the gates with opposite polarities. The position of the peak current is now a strong function of the gate bias and the magnitude of the peak current at resonance is proportional to the interlayer bias at which the resonance occurs. Furthermore, a sharp resonance, associated with a small Γ produces a large peak-to-valley current ratio, as shown in Figure 4.4(a) and (b) whereas a larger Γ leads to a weaker NDR, i.e., low peak to valley ratio, as shown in Figures 4.4(c) and (d)

From a device application point of view, it is important to recognize the mechanisms that prominently contribute to the broadening Γ. A typical tunneling current voltage characteristic, for example, for a parabolic band system including higher order perturbation terms might be of the representative form

$$I_{il} \propto A_1 t_1^2 \frac{V_{il}}{\Gamma_{net,1}^2 + \varphi_{il}^2} + \ldots + A_n t_n^2 \frac{V_{il}}{\Gamma_{net,n}^2 + \left(\varphi_{il} - \Delta E_n/q\right)^2} + \ldots \qquad (4.6)$$

where the first term in the above equation is current due to lowest order perturbation due to coupling between the layers. The current calculated in Eq. (4.3) is equivalent to the first term in Eq. (4.6). The higher order terms in Eq. (4.6)

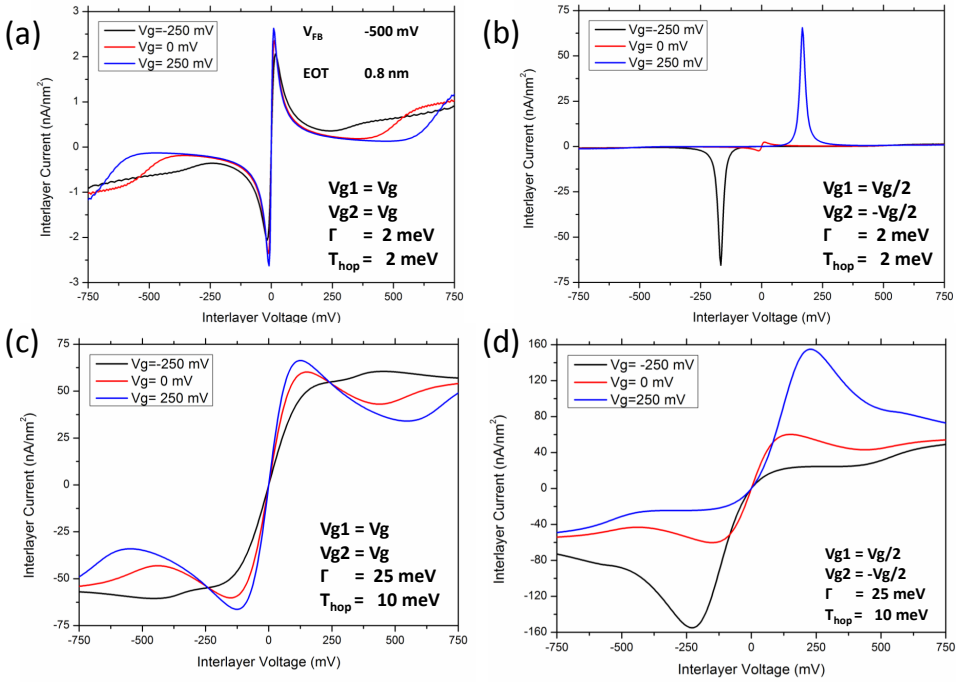

Figure 4.4: Interlayer tunneling current as a function of interlayer voltage illustrating the effects of gate voltages, interlayer coupling and density of states broadening. The data is obtained using an effective oxide thickness (EOT) of 0.8 nm for the gate and interlayer dielectrics and doped graphene layers such that $V_{FB} = -500mV$.

are contributions due to excited state subband resonances or phonon assisted resonances. The net broadening due to various sources of scattering is of the form $\Gamma^2_{net} = \Gamma^2_{space} + \Gamma^2_{inhomogenous} + \Gamma^2_{homogenous}$ where $\Gamma_{inhomogenous}$ denotes the broadening due to inhomogenous scattering sources such as impurities, surface roughness and defects. Broadening due to homogenous scattering sources such as intra-well phonon scattering is given by $\Gamma_{homogenous}$. One can in principle add to the Hamiltonian in Eq. (4.2) terms accounting for various scattering sources and calculate the current in perturbation limit perhaps including

higher order terms as well. But the calculations are involved and, in any case, for current calculations for stronger interlayer coupling one must go beyond the perturbation limit. In the rest of this Chapter, I present non-equilibrium Green's function (NEGF) based calculations to further understand the essential physics of the ITFET.

4.5 NEGF Simulation and Finite Length Effects

The NEGF calculations are performed in ballistic limit and do not account for broadening due to scattering sources such as phonons, impurities and surface roughness. However, the calculations allow to go beyond the perturbation limit. I will also consider short channel effects which introduce a new and perhaps dominant form of broadening, label it although it may not entirely fit the simple broadening model of Eq. (4.5), which may set practical, if material dependent, scaling limits for such devices.

Model device structures of graphene and III-V based ITFETs used for the NEGF based simulations are illustrated in Figures 4.5(a) and (b) respectively. For the graphene device I use an atomistic p_z orbital based tight binding Hamiltonian with coupling between the layers only in the region between $-L/2$ and $L/2$. Although the graphene layers are separated by a dielectric, they are modeled as Bernal stacked for specificity but only weakly coupled. For the III-V based device, I use an effective mass Hamiltonian and coupled mode space approach to solve for the Green's function [68]. See Appendix F for more details on the effective mass Hamiltonian construction used for

the NEGF simulations. The Hamiltonian for the bilayer graphene system is implemented using the procedure described in Appendix D.

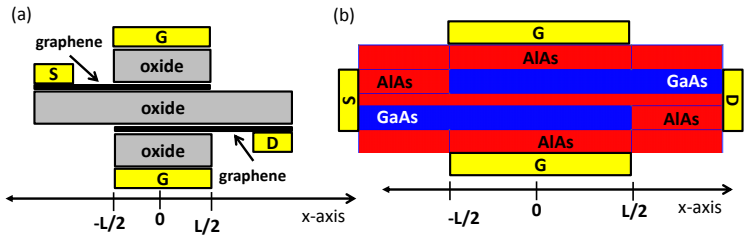

Figure 4.5: Device schematic of (a) Graphene ITFET and (b) IIIV based ITFET used for NEGF simulations

Figure 4.6(a) shows the zero field transmission for the graphene device with overlap length $L = 30$ (blue solid line), 60 (red solid line) and 120 (black solid line) nm and an interlayer coupling of 20 meV. Figures 4.6(b)-(c)show the band structure of the graphene in the left lead, overlap region and the right lead. The graphene strip is about 6 nm wide and has armchair configuration. The width is chosen such that the graphene in the leads is metallic. (I note that the basic operation in no way depends on the use of nano-ribbons. However, the use of metallic nano-ribbons with their constant velocity carriers in the ground-state subband also simplifies the interpretation of some results. In addition, for these fully 2D simulations within the layers, practical simulation considerations limit the graphene width.) As a result, there is only one available subband for any injection energies between approximately ± 0.35 eV relative to the Dirac point. Figure 4.6(a) shows that the transmission has a sinc function like behavior at low injection energies and a cosine function like behavior at higher energy. Also, the period of the cosine function is propor-

tional to the length of the overlap region as illustrated by the doubling of the dips in transmission on doubling the overlap length. For example, in Figure 4.6(a) between 0.1 eV and 0.18 eV on the energy axis, there are two dips in the red curve between two dips of the blue and two dips in the black curve between two dips of the red curve.

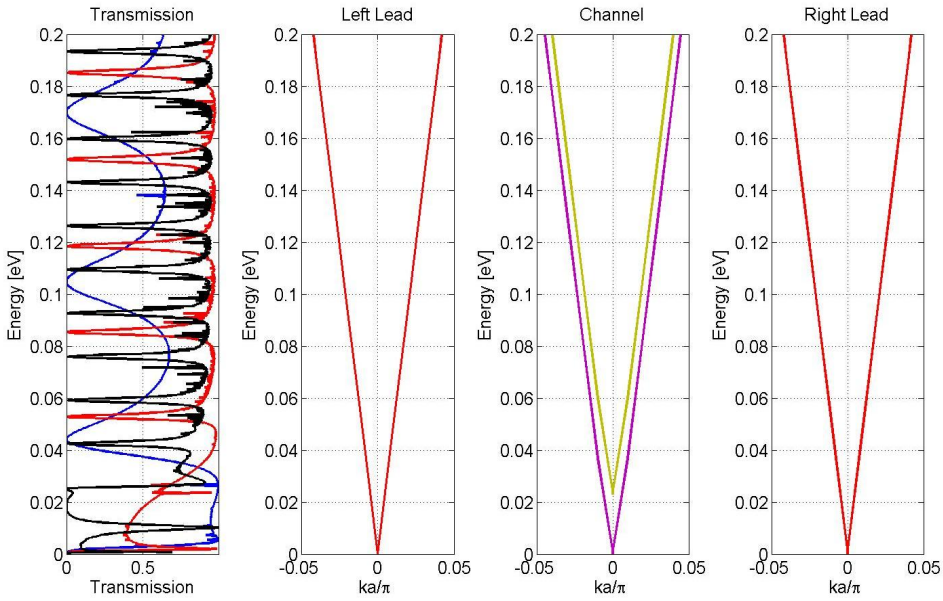

Figure 4.6: (a) Zero field/resonant transmission in a Graphene ITFET for three different gate lengths, and band structure above the Dirac point of graphene in (b) left lead, (c) channel and (d) right lead under resonant conditions, where interlayer coupling results in a Bernal-like graphene band structure but at lower energies.

Figure 4.7 shows the zero field transmission for the same device but with interlayer coupling for 25 meV (blue curve) and 100 meV (red curve), and for three different channel length of 30, 60 and 120 nm. It can be observed from Figure 4.7(a) that the transmission increases (roughly as the square of the

coupling strength) with increasing coupling strength initially, and ultimately saturates to a maximum value of one. Also, for a given coupling strength, for example the red curve with 25 meV of coupling, the transmission at a given energy between layer increases with increasing overlap length.

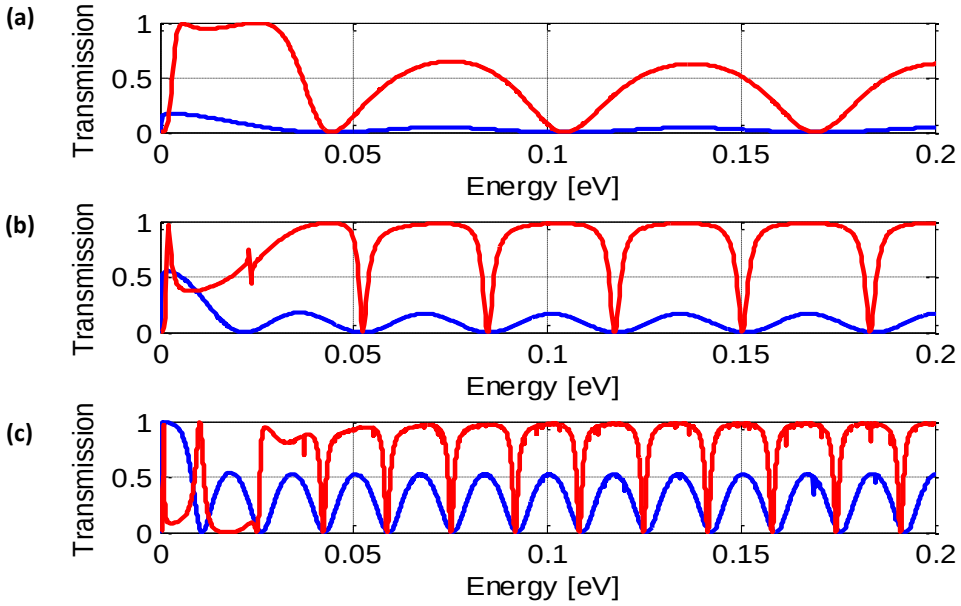

Figure 4.7: Zerofield/resonant transmission vs injection energy for two different inter layer coupling: blue(100 meV) and red (25 meV) . (a) 30 nm, (b) 60 nm and (c) 120 nm

Figure 4.8(a) shows the band energy diagram at 100 meV interlayer potential difference of the III-V ITFET across the middle of the device The interlayer potential difference is assumed to drop linearly in the AlAs interlayer. Unless otherwise mentioned the thickness of GaAs quantum wells and the AlAs gate barriers is 4 nm. The thickness of the AlAs interlayer is used as a variable to control the strength of coupling between layers. The barrier

height between GaAs and AlAs for the results in Figure 4.8 is 594 meV and 214 meV for rest of the III-V FET simulation results in this Chapter. The effective mass for electrons in GaAs was $0.067m_e$ and in AlAs was $0.1085m_e$, where m_e is the electron free-space rest mass. The onset of the transmission is determined by the position of the lowest sub-band in the quantum well. Energy levels corresponding to first two sub-bands are shown by the blue (E1)

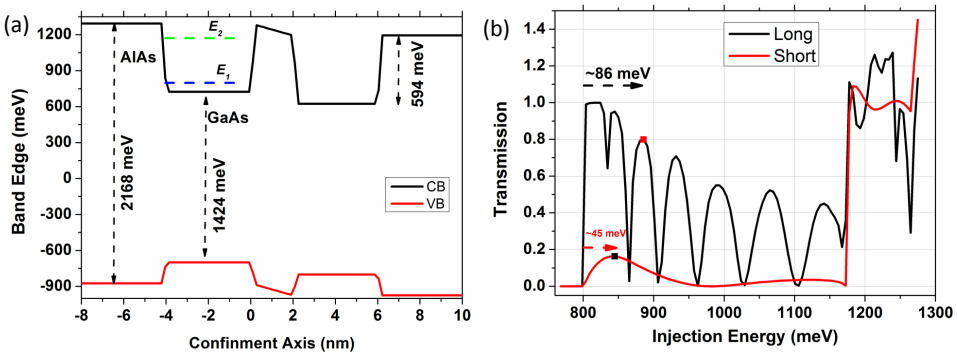

Figure 4.8: (a) Band energy diagram along the confinement direction in the middle of the device at 100 mV interlayer potential difference. The first two energy levels in the well are shown by the dashed horizontal blue and green lines respectively. The Γ-valley conduction band barrier height and the Γ-valley band gaps for AlAs, GaAs are shown by the vertical dashed arrows. (b) Zero field transmission vs injection energy at normal incidence ($k_z = 0$) for channel lengths of 30 nm (black curve) and 5 nm (red cuve)

and green dashed lines (E2).

The zero interlayer field/resonant transmission at normal incidence (i.e., crystal momentum directed straight into the channel) for 30 nm (black curve) and 5nm (red curve) channel lengths are shown in Figure 4.8(b). Energy levels are referenced to the mid-gap of GaAs. The transmission is dependent on the length of the overlap region, and exhibits injection energy dependent

oscillations, both qualitatively similar to that of the graphene results shown in Figure 4.6. The sharp increase of the transmission probability, actually the transmission probability for injection totaled over all subbands in which injection can occur, after ≈1200 meV is due to onset of contributions from the second sub-band. Modes with nonzero out of plane momentum also show a similar behavior, but the onset of transmission and the later sharp increase are shifted to higher total energy of the mode as shown in Figure 4.9.

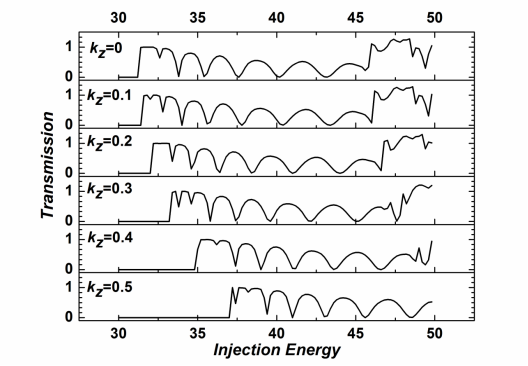

Figure 4.9: Zero field transmission vs injection energy (in units of kT = 25.6 meV) at different out of plane momentum values scaled to unit less dimensions. Energy levels referenced to mid-gap of GaAs and the GaAs-AlGaAs conduction band barrier height $\tilde{2}$14 meV.

Figure 4.10 shows the zero field transmission at normal incidence for the III-V device structure shown in Figure 4.5(b) for overlap lengths of 30, 60 and 90 nm. The coupling for the III-V quantum wells is controlled by changing the interlayer thickness. The energy levels in the transmission plots are now referenced to the nominal conduction band energy of the GaAs. In Figures 4.10(a)-(c) the blue curve illustrating weaker coupling is for the 4 nm

AlAs layer and the red curve illustrating stronger coupling is for the 3 nm AlAs interlayer. It can be observed from Figure 4.10 (a) that the transmission increases with increasing coupling strength, ultimately saturating to a maximum value of one per sub-band. Similarly, for a given coupling strength, the transmission at a given energy increases with increasing overlap length. And in either case, saturation occurs first at lower energies where the carrier resides in the channel longer, which contrasts to the behavior for the graphene metallic nanoribbion where the carrier velocity and, thus, the channel reside time is constant.

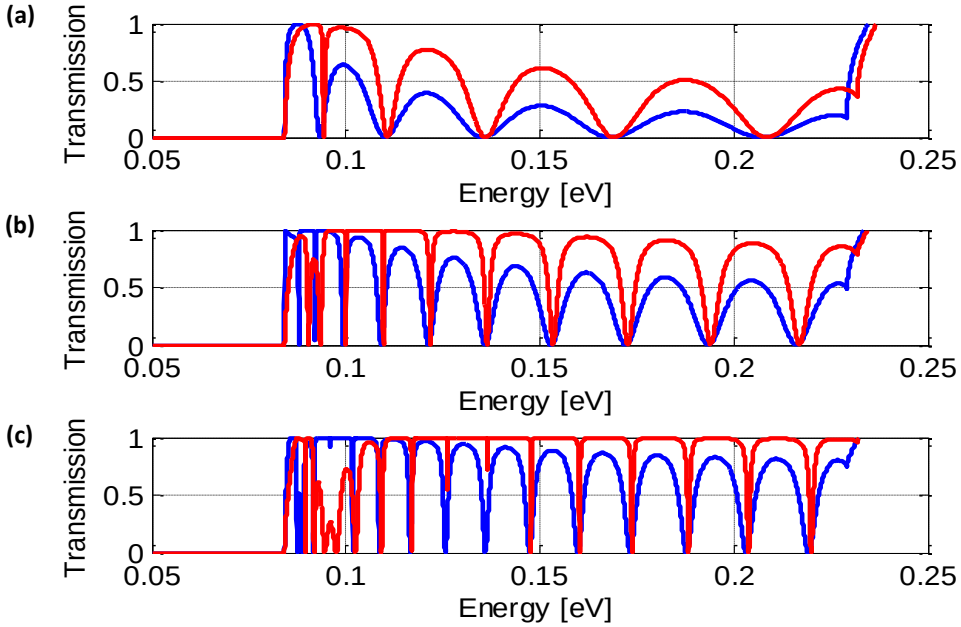

Figure 4.10: Transmission probability vs injection energy for a III-V ITFET with two different interlayer thickness (blue 4 nm) and red (3 nm) and three gate lengths (a) L = 30 nm, (b) L = 60 nm and (c) L = 90 nm

The length dependent dips in the transmission of the graphene and III-

V devices is due to the standing wave patterns formed in the electron layers. To understand the observed behavior, consider the case when the source and drain channels are infinitely long with coupling only in the region of length under the gate, as shows in Figure 4.5, and coupling remains weak/perturbative. The wave function in source layer, ψ_s, and drain layer, ψ_d, will be of the form

$$\psi_s(x) = ae^{k_s x} \text{ and } \psi_d(x) = be^{k_d x} \tag{4.7}$$

In this weak coupling limit, the matrix element t_k in Eq. (1.4) will be proportional to the inner product of the source and drain wave functions. Consequently the transmission at resonance will be of the form

$$T \propto \left| ab^* \frac{1}{2} L \frac{\sin(k_S - k_D)L/2}{(k_S - k_D)L/2} \right|^2 \tag{4.8a}$$

$$\propto \frac{|ab^*|^2}{4} L^2 \text{ at resonance, } k_S = k_D \tag{4.8b}$$

However, with the closed boundaries at $\pm L/2$, the wave function in source layer and drain layer, will be of the form given below

$$\psi_s(x) = a\left(e^{k_s(x-L/2)} - e^{-k_s(x-L/2)}\right) \tag{4.9a}$$

$$\psi_d(x) = b\left(e^{k_d(x+L/2)} - e^{-k_d(x+L/2)}\right) \tag{4.9b}$$

and in the weak coupling limit, the transmission between layers at resonance

will be of the form

$$T \propto |ab*|^2 \left| L\frac{\sin(k_S - k_D)L/2}{(k_S - k_D)L/2}\cos(k_S + k_D)L/2 \right.$$
$$\left. - L\frac{\sin(k_S + k_D)L/2}{(k_S + k_D)L/2}\cos(k_S - k_D)L/2 \right|^2 ;\ k_S = |k_S|\,,\ k_S = |k_D| \quad (4.10a)$$
$$\propto |ab*|^2 L^2 \left|\cos(kL) - \frac{\sin kL}{kL}\right|^2 \text{ at resonance, } k = \ k_S = k_D \quad (4.10b)$$

Such a simple analysis is not expected to fully explain the NEGF results. Nevertheless, the transmission in the above equation shows a similar qualitative behavior as the transmission obtained via NEGF simulations. At lower energies the transmission has a sinc function like behavior and at higher energies it is periodic with a overlap length dependent period.

However, while there are oscillations in the resonant transmission probability with injected carrier energy, and overall increases in tunneling probability with increasing channel length as to be expected, there is a much more problematic effect of short channel length that must be considered. In the perturbation limit, the transmission essentially depends on the matrix element for the interaction and the convolution of the spectral density of states. Ideally, the spectral density of states approaches a delta function i.e., $A(k, E) = \delta(E - \epsilon_k)$ and the transmission is non zero only if both energy and momentum are conserved. As discussed previously, scattering can be a source of broadening in the spectral density function. However, Eqs. (4.8) and (4.10) show that finite channel lengths L also produce broadening. For example in Eq. (4.8) in the limit of large L, the sinc function tends to a delta function requiring $k_S = k_D$,

but then broadens inversely with the L as L decreases. In equation (4.10), the behavior is more complicated but the basic result is the same. This broadening can be understood as essentially Heisenberg position-momentum uncertainty with the position uncertainty corresponding to the channel length.

To illustrate the resonance broadening effects, I performed non-self-consistent NEGF simulations to study the effect of the interlayer potential difference on the transmission at a given energy. For the simulation results of this section, the applied interlayer potential difference between the layers is split equally with opposite polarities. The applied potential is assumed to be constant along the transport direction in each layer. For the III-V FET the potential is assumed to vary linearly across the tunnel barrier between the wells. For example, Figure 4.8(a) shows the conduction band edge across the middle of the device for a 100 meV interlayer potential difference. Figure 4.11(a) shows the source to drain transmission at injection energy of 200 meV with respect to the zero interlayer potential Dirac point in the lead, for a graphene device with three different channel lengths of 30, 60 and 120 nm. The transmission at resonance is broadened with broadening decreasing with increasing overlap length as expected. Note that the transmission in Figure 4.11(a) is scaled by the maximum value of the corresponding channel length to better resolve the relative broadening vs channel length. The interlayer coupling strength was 10 meV. Figure 4.11(b) shows the transmission vs interlayer potential for a III-V device with 2 nm AlAs interlayer. The injection energy for the 30 nm channel length device (black curve) and the 5 nm channel length

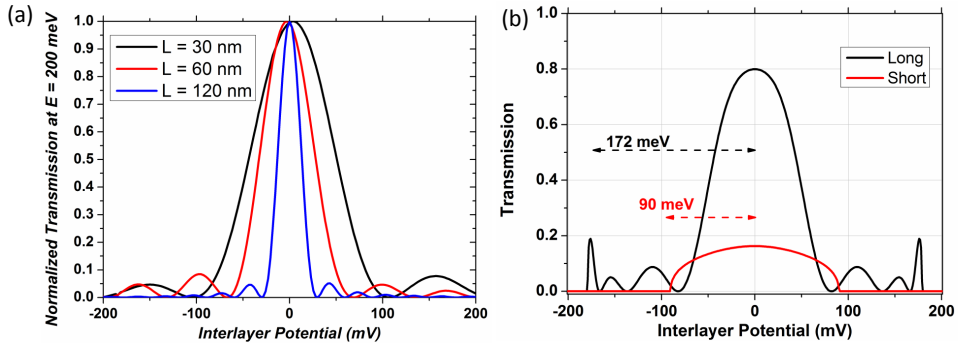

Figure 4.11: (a) Transmission at injection energy of 200 meV vs interlayer potential for a graphene ITFET for three different channel lengths of 30 nm, 60 nm and 120 nm. The width of graphene is about 6 nm (b) Transmission vs. interlayer potential difference for a long (black) and red(short) channel III-V device at a fixed injection energies relative to mid barrier, corresponding to resonant current peaks indicated by square symbols in Figure 4.8(b).

device (red curve) is shown by the red and black squares respectively in the transmission plot shown in Figure 4.8(b). It can be observed that the transmission is qualitatively similar to that of the graphene device but with a cutoff at high energy. Due to the symmetry in applied potential, the transmission is cut off when interlayer potential difference is more than two times $E_{in} - E_1$, where the latter were about 86 meV and 45 meV for the long and short devices in these illustrative simulations. Finally, note that a non-self-consistent two terminal current response at a given interlayer potential difference can be obtained by integrating the transmission as function of energy.

To further illustrate the source of the broadening in transmission, Figure 4.12 exhibits the transmission vs. interlayer potential for the above mentioned graphene FET, but now with channel length of about 100 nm and

varying interlayer coupling from 5 meV to 60 meV. The transmission is nor-

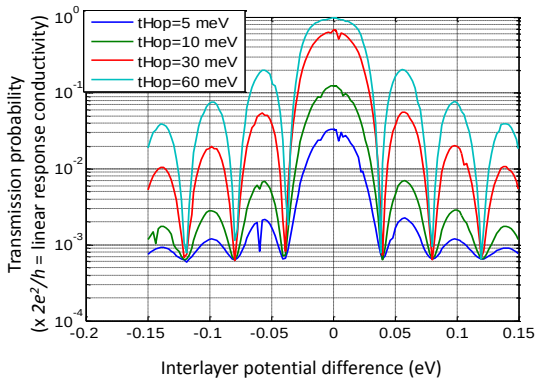

Figure 4.12: Transmission at injection energy of 200 meV vs interlayer potential for a 6 nm with graphene ITFET with an overlap length of ≈100 nm for different interlayer coupling from 5 mev to 60 meV.

malized with the square of hopping strength corresponding to each curve. At t_{hop} = 5 and 10 meV, peak transmission probability scales as t_{hop}^2 and tends to saturate at higher coupling strength. The minimum probability floor in the transmission is due to a small imaginary energy used to shift the poles of real axis for the retarded NEGF calculations of the transmission probability. The dips in the transmission occur at energies which almost satisfy the $L_c(E - E_o)2v_g = n$, where n is any non zero integer, v_g is the group velocity of the carriers and L_c is the channel length.

Similarly, Figure 4.13 shows the non-self-consistent NEGF based ballistic transmission at a single injection energy, as a function of interlayer potential difference, for AB coupling between 5.4 nm wide metallic graphene nanoribbons with a 104.3 nm long overlap/channel region. Figure 4.13(a) illustrates the expected $|sinc|^2$ function model fit (solid line) to the transmission

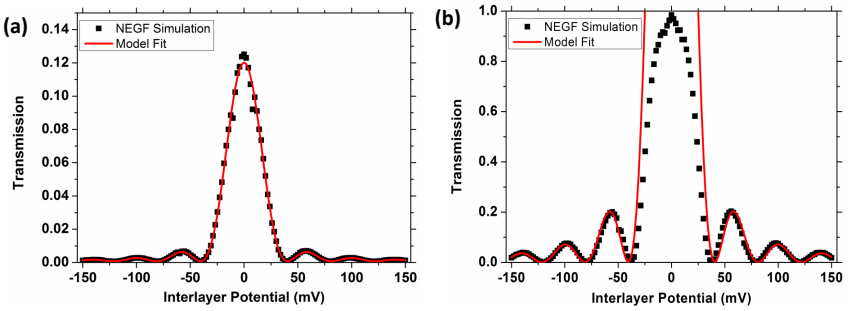

Figure 4.13: NEGF based transmission as function of interlayer voltage at an injection energy of 200 meV for an AB bilayer graphene with width 5.4 nm and length 104.3 nm with (a) a weak inter layer coupling thop =10 meV and (b) a strong interlayer coupling thop =60 meV. In (a) the red solid line is the model fit to the NEGF transmission data (black square). In (b) the red solid line is the red curve in (a) scaled by 36 times. (Reprinted with permission from [13], copyright (2012) by the IEEE).

data for 10 meV interlayer coupling. Figure 4.13(b) illustrates the injection-limited transport as the transmission probability saturates toward unity due to a higher, 60 meV, interlayer coupling. The reference solid lines in Figure 4.13(b) are a fit to the expected form of Heisenberg uncertainty associated with the finite channel lengths, but are not adjusted for injection-limited transport. In the results of Figures 4.12 and 4.13, it can be seen that the zeros of the transmission probability do not vary with the coupling strength and that overall, outside of the effects of saturation toward unity, the rate of decay of the transmission probability peaks with increasing interlayer potential are not dependent on the interlayer coupling strength. This does not contradict the concept of finite lifetime broadening per se; rather this result simply shows that in the tails of the distribution, particularly at the zeros, the tunneling time is again quite long, infinite at the zeros, and, thus, leads to no broaden-

ing there. In short, there is little or no penalty to pay in terms of additional broadening for allowing near unity transmission on resonance.

Figure 4.14 illustrates the relative effects of short-channel broadening, showing the transmission vs interlayer potential for a graphene and III-V FETs, exhibiting relatively much poorer behavior in the graphene devices, which we can trace to the large group velocity that is otherwise generally considered a benefit in graphene. Unfortunately for these material systems, the

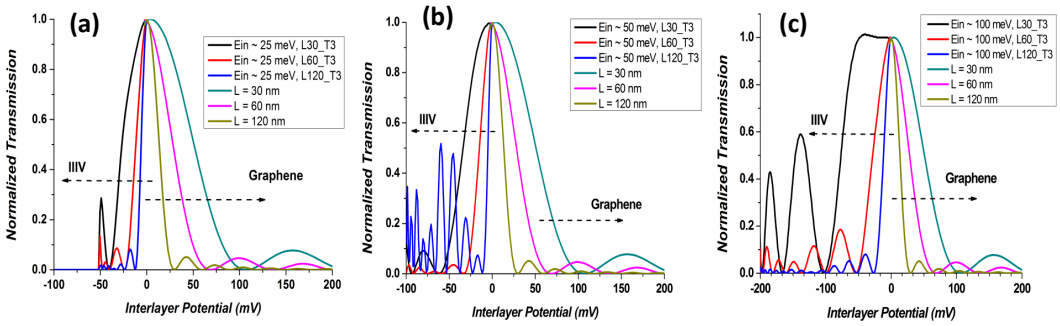

Figure 4.14: Normalized transmission as function of interlayer voltage for a III-V FET with electrons injected at (a) 25 meV (b) 50 meV and (c) 100 above the first subband, compared to results form graphene. For graphene FET in all cases the mode is injected at 200meV from the Dirac point of the lead, but this is not relevant to the results given the fied group velocity, as discussed in the text. The overlap length for the devices is 30, 60 and 100nm. Interlayer thickness for III-V FET was 3 nm and the coupling strength of the graphene FET was 10 meV

position-momentum uncertainty is translated to position-energy uncertainty via the carrier group velocity, dE/dk, in the transport direction, and, thus, is exacerbated for fast carriers such as in graphene or at higher energies in III-Vs. Thus, in Figure 4.14, as the injection energy increase in the GaAs-AlAs, the carrier velocity increases, increasing the broadening. For graphene, the

carrier velocity is large and independent of energy, and the broadening is still more substantial. (The detail resonances structure is further complicated by standing wave patterns created within the channel region due to reflections from the ends of each layer.) Therefore, for short channel devices, as I will illustrate in the next section,—vs. some large experimental devices—there will be a trade-off between device size and voltage of operation as allowed by the resonance width.

4.6 Inverter with III-V ITFET

As a proof of concept and challenges, I implement an inverter with III-V based ITFET, using the compact model discussed earlier in the chapter using only a short-channel contribution to broadening Γ, much as discussed in Section 4.4. The broadening in the current model is based on a fit to the non-self-consistent current vs interlayer potential obtained via NEGF simulation approach described in the previous section. A correction for saturation/lead limited injection has also been added. Specifically, the equations used to fit the current data are

$$I = I_{inj}\left(1 - \exp\left(-\frac{L_g}{L_o}\right)\right) \tag{4.11}$$

where $I_{inj} = \frac{V_{il}}{R}$ is the lead limited injection current, L_g is the gate length, and $L_o = I_{inj}/I_{tun}$. The nominal interlayer tunneling current I_{tun} for a given interlayer bias V_{il} and interlayer potential φ_{il} is given by

$$I_{tun} = A\frac{\Gamma V_{il}}{\left(V_{il} - \phi_{il}\right)^2 + \Gamma^2} \tag{4.12}$$

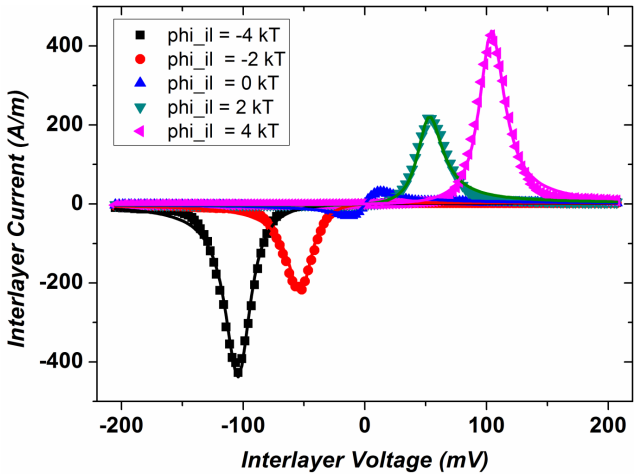

Figure 4.15: Non-self consistent Interlayer current vs Interlayer voltage for 60 nm Length and 3nm barrier thickness III-V Tunnel FET for 5 different inter layer potential offsets (in units of kT = 25.6 mV) between wells at zero interlayer Bias. The offsets are introduced to illustrate the gate control of the I-V characteristics in a non-self consistent simulation. The peak current occurs when the combined effect of interlayer bias and gate voltage results in zero interlayer potential difference.

where the pre-factor A is the scale factor for the tunneling current and Γ is the effective short-channel broadening. In the limit of large coupling between the layers, the current is mainly limited by contact resistance i.e., when $I_{tun} >> I_{inj}$, L_o is small and the exponential term tends to zero. Figure 4.15 shows the model fit (solid lines) to the NEGF based interlayer current (symbols) for a 60 nm long III-V ITFET with 3 nm thick tunnel barrier. To illustrate the effect of gate voltage, each I-V curve is calculated at a fixed interlayer potential offset φ_{il}, as shown in the legend. To fit the NEGF current data with the above model, Γ, A and R are used as fitting parameters and obtained to be, $A = 52.87$ A/m, $\Gamma = 12.44$ mV, and $R = 141.50\,\Omega$.

The I-V model given by Eqs. (4.11) and (4.12) along with the equivalent C-V model shown in Figure 4.1(b) was implemented as verilog-A compact model for use in a circuit simulator (Spectre from Cadence). As a proof of concept, I have implemented two designs for a III-V ITFET based inverter. The first one has configuration similar to the BiSFET-1 based inverter using only one polarity clock but the bottom FET has W/L 3 times that of the top transistor. The second design is similar to the BiSFET-2 based inverter and requires dual polarity clocking as discussed in detail in Chapter 3. Both the inverter designs are verified for functionality as illustrated by the Spice simulated responses shown in Figure 4.16.

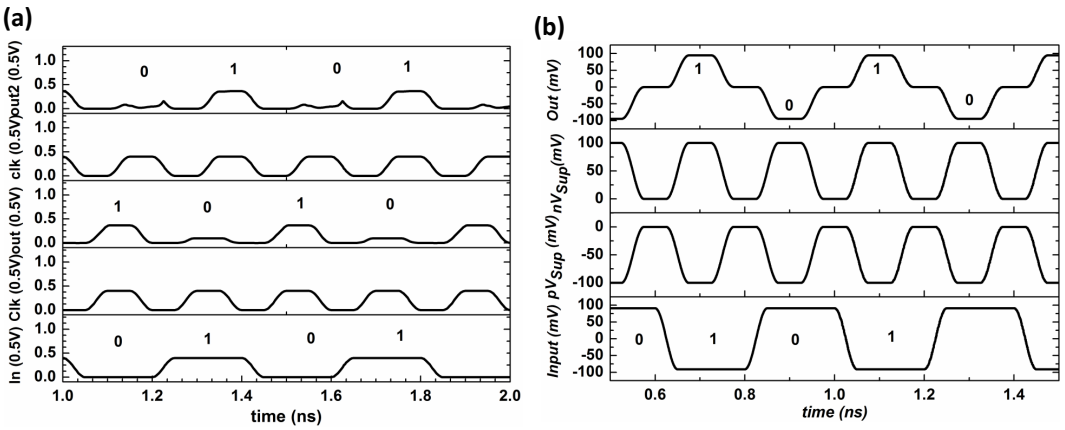

Figure 4.16: (a) SPICE-level Verilog-A simulation of III-V Tunnel FET with a one polarity, 200 mV, 5GHz clock, and one inverter load. W/L of the bottom transistor is 3 times W/L of top transistor. (b) SPICE-level Verilog-A simulation of III-V Tunnel FET Inverter with a two-polarity 100 mV, 5GHz dual polarity clock one inverter load. W/L of the bottom transistor is 1 times W/L of top transistor. L=60 nm, W = 30nm. Half width of the lorentzian is 15 meV.

Finally, Figure 4.17 shows the effect of broadening on the power con-

sumption and inverter functionality for the two polarity clocked device. As the width of the Lorentzian broadening (Γ) in I_{tun} is increased, the power consumed by the inverter increases, as shown by Figure 4.17 (a). Figure 4.17 (b) shows the output signal for the inverter with one inverter load for varying values of the broadening Γ . Observe that as the gamma increases, the output voltage in Figure 4.17 (b) does not go all the way to the either supply voltage. For $\Gamma = 15$ meV, the corresponding energy is on the scale of 1 fJ, which is over an order of magnitude larger than current CMOS switching energies. Note that for the simulation results in Figure 4.17 , the length of the FET is not scaled commensurate to the width of the Lorentzian, so with reduced gate capacitances to charge in the shorter devices, the power differences would be much less.

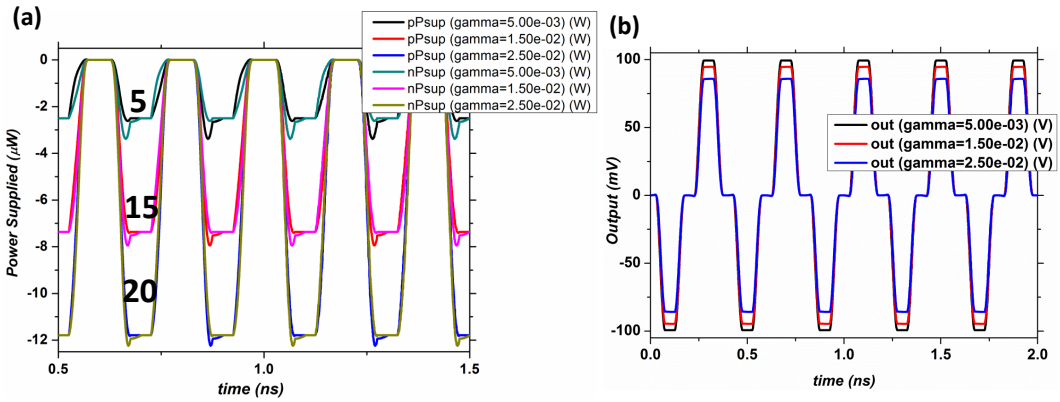

Figure 4.17: (a) SPICE simulation of III-V tunnel FET inverter with one inverter load showing instantaneous power supplied by the positive clock (pPsup) and negative clock (nPsup) for three different Γ : 5 meV, 15 meV and 25 meV, and (b) SPICE simulation of IIIV tunnel FET inverter with one inverter load showing the output signal for varying width of Lorentnzian broadening width.

4.7 Conclusions

The ITFET device concept was explored in the past for applications in digital logic circuits. However, the experimental devices were large. In this Chapter, I have explored the essential physics of the sub-100 nm channel length ITFETs. I have shown that there are substantial short-channel induced resonance broadening effects, but that for even for strong interlayer coupling leading to saturation of resonant peak there is no additional finite-lifetime broadening of the tails. Moreover, I have shown that high carrier velocities considered good in MOSFETs, are actually counterproductive for resonant broadening-based ITFETs. To demonstrate all of this, I have developed and performed NEGF simulations in III-V and graphene systems; used analytic models appropriate in the perturbative limit to confirm the short-channel effects seen in the NEGF simulations, and have developed an essential physics based compact model for a IIIV-based ITFET for use in SPICE-level simulations. The latter have shown the detrimental effects of short-channel broadening on device performance.

In terms of two-dimensional material systems, materials with much slower carriers such as the transition metal dichalcogenides (TMD) would be substantially better, with short channel broadening perhaps a factor of 3 less than that in the simulated AlAs-GaAs systems, for a given channel length, based on typical effective mass ratios. However, scattering induced resonance broadening in these latter systems could be even more pronounced. There is some possibility that with better materials, switching energies could ap-

proach CMOS like. However, BiSFET-like performance projections are not possible, nor were they expected to be. Neither system is intrinsically subject to thermal smearing. However, BiSFETs are also not subject to short-channel resonance broadening as tunneling is not resonant, and superfluid systems can be fairly robust to disorder and scattering which lead to broadening of resonant tunneling.

Chapter 5

Time Dependent Quantum Transport in Graphene

5.1 Introduction

Although the lack of a band gap in Graphene continues to challenge CMOS-like logic applications, high carrier velocities, the ultimate thin body and potential process compatibility with silicon technologies still make graphene a promising candidate for radio frequency (RF) applications [26–28] and perhaps for novel "beyond CMOS" applications. An example of beyond CMOS applications is the BiSFET device concept discussed in Chapter 3. Not only are BiSFET logic circuits expected to be clocked at perhaps 10 GHz or greater but BiSFETs may be capable of producing multi-THz oscillations beyond their critical voltages, in a manner analogous to the AC Josephson effect. For such devices, one must go beyond quasi-static to truly time-dependent analysis to fully understand their intrinsic frequency limitations and dynamic response. Towards this end, I present a numerical method for modeling time-dependent quantum transport in graphene via a time-dependent non-equilibrium Green's function (NEGF) method. Here I describe the essential elements of the method—Hamiltonian, time evolution scheme and open boundary conditions—and illustrate them via simple MATLAB-based simulations for clarity. With these elements

given, future, e.g., electrostatically self-consistent simulation of graphene MOSFETs with a thermal distribution of source-injected carriers will add only to the computational burden, not the technical one.

5.2 Solutions of Time-Depndent Schrödinger Equation

Within the tight-binding formalism, the time-dependent Schrodinger equation is of the form,

$$i\hbar\frac{d}{dt}\psi = \mathbf{H}\psi \tag{5.1}$$

where ψ and $\mathbf{H}$ are the wave-function column matrix and Hamiltonian square matrix, respectively. Here we consider on-site and nearest-neighbor π-bonding only on the two-dimensional (2D) hexagonal lattice of graphene (Figure 5.1), with nearest-nearest neighbor matrix elements $H_{i,j} = \tau_o = 3.03$ eV, on-site matrix elements $H_{i,i}$ defined by an electrostatic potential, and all other matrix elements set to zero. See appendix D for more detailed discussion of constructing monolayer and bi-layer graphene Hamiltonian for easy implementation in code.

To numerically solve Eq. (5.1), we first consider discretization of time via the well-proven semi-implicit Crank-Nicolson (CN) method [69]. This approach provides accurate, unitary and time reversible evolution and serves as a reference here. Assuming a time step of Δt for time integration, the CN is

$$i\hbar\frac{\psi\left(t+\Delta t\right)-\psi\left(t\right)}{\Delta t} = \mathbf{H}\left(t+\frac{\Delta t}{2}\right)\left(\frac{\psi\left(t+\Delta t\right)+\psi\left(t\right)}{2}\right) \tag{5.2}$$

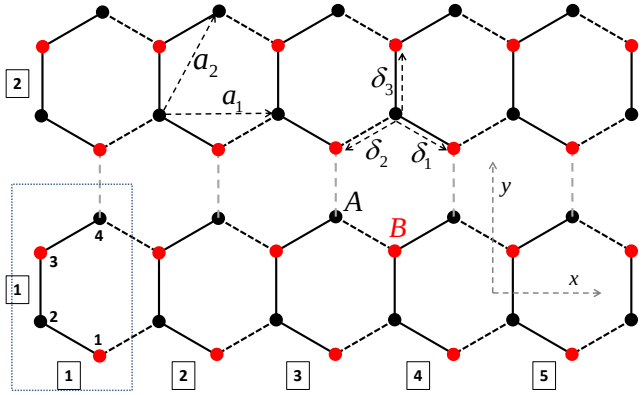

Figure 5.1: Illustration of the graphene crystal lattice, represented by the points where the colors distinguish the sublattice, and the nearest tight-binding coupling of the tight-binding Hamiltonian, represented by the lines (solid or dashed, black or gray). For one implementation of an alternating direction implicit scheme, ADI1 (see Section III), the associated four-atom unit cell is shown in the rectangle. In this case, coupling between atoms within the same unit cell are indicated by the black solid lines, coupling between unit cells in the x direction by black dashed lines and coupling between unit cells in y direction by dashed gray lines. (Reprinted with permission from [14], copyright (2012) by the IEEE).

Denoting $\psi^n = \psi(n\Delta t)$ and $\mathbf{H}^{n+1/2} = \mathbf{H}(n\Delta t + \Delta t/2)$, Eq. (5.2) can be rewritten for each time step as

$$\left(\mathbf{I} + i\Delta t\mathbf{H}^{n+1/2}\right)\psi^{n+1} = \left(\mathbf{I} - i\Delta t\mathbf{H}^{n+1/2}\right)\psi^{n} \tag{5.3}$$

5.2.1 Gaussian Wave Packet on Graphene

As an illustration of time evolution, consider a Gaussian wave-packet centered around one of the Dirac points $\mathbf{k}_D$ in wave-vector space and initially well-localized in real-space, as per Fig. 5.2(a),

$$\psi(\mathbf{r}, t=0) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-|\mathbf{r} - \mathbf{r}_o|^2/2\sigma^2 + i\mathbf{k}_D\right) \begin{pmatrix} 1 \\ i \end{pmatrix} \tag{5.4}$$

The two-element column matrix is the so-called 1 x 2 pseudo-spin matrix on

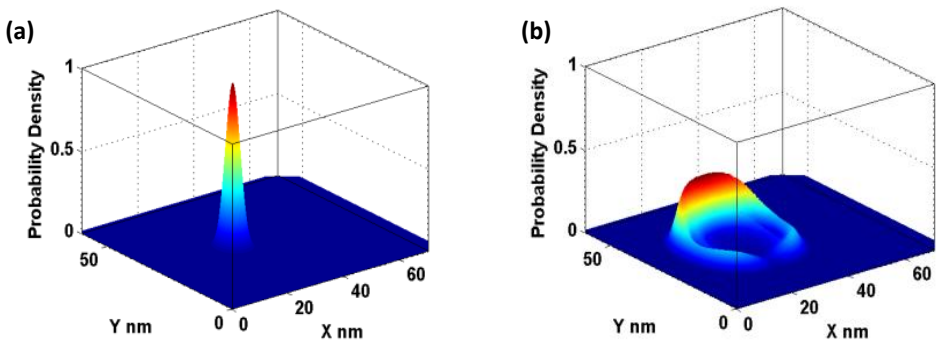

Figure 5.2: Snapshots of time evolution of initially Gaussian wave-packet at (a) t = 0 and (b) t = 14 fs.(Reprinted with permission from [14], copyright (2012) by the IEEE).

the far right-hand side of Eq. (5.4) describes the relation between the complex amplitude coefficients of the two sub-lattices of graphene (and is not to be confused with the which layer psuedospin considered for the BiSFET). While fixed

initially here, in general it can be a function of **r** as well. This wave-packet contains equal contributions from the conduction and valence band, and extends well into both. Due to the predominately linear dispersion/constant carrier speed in graphene within the relevant energy range, the wave-packet spreads from the center in a asymmetric ring-like shape with little change in the ring thickness with time, as shown in Fig. 5.2(b) at t = 14 fs. The principle direction of motion is associated with the given initial pseudo-spin phase. This behavior, confirmed by analytic results for this simple system [70], contrasts markedly to the broadening-Gaussian evolution characteristic of particles in a single band of well-defined effective mass.

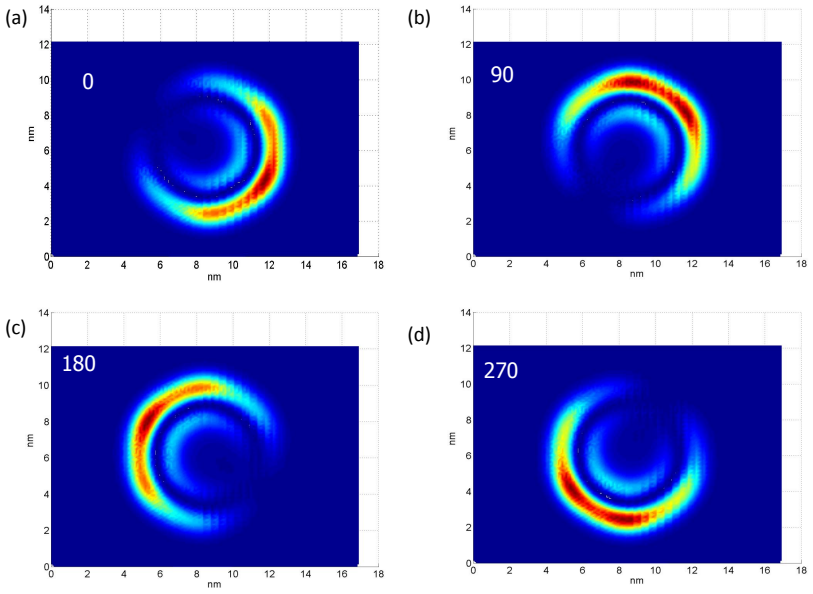

Figure 5.3: Figure shows the snapshot of time evolution of the wave packet for four different initial pseudospin phases.

A Gaussian wave packet centered around $\mathbf{k}_o$ on a rectangular lattice has

a dispersive motion with the peak of the wave packet moving in the direction of $\mathbf{k}_o$. However, in case of graphene, due to linear dispersion at low energies the magnitude of velocity of the components constituting the Gaussian wave packet is independent of the Bloch momentum. As a result the wave packet spreads in the shown ring like pattern. Figure 5.3 shows the snapshot of time evolution of the wave packet for four different initial pseudospin phases, θ. It can be observed that the wave packet does spread almost equally in all directions except the direction anti parallel to the pseudospin phase. Consequently, it appears to be moving in the direction of pseudo-spin. The following analysis explains the observed behavior: We can decompose the initial Gaussian wave packet as,

$$\psi\left(r, t=0\right) = \frac{1}{\sqrt{\pi\sigma^2}} e^{-\frac{r^2}{2\sigma^2}} e^{ik_o \cdot r} \begin{pmatrix} 1 \\ e^{i\theta} \end{pmatrix} \tag{5.5a}$$

$$= \frac{1}{\sqrt{\pi\sigma_k^2}} \int \frac{dk}{2\pi} e^{-\left[\frac{(k-k_o)^2}{2\sigma_k^2}\right]} \psi_k\left(r\right) \tag{5.5b}$$

where $\sigma\sigma_k = 1$. It can be verified that the above equation is satisfied if

$$\psi_k\left(r\right) = \frac{1}{\sqrt{2}} e^{ik\cdot r} \begin{pmatrix} 1 \\ e^{i\theta} \end{pmatrix} \tag{5.6a}$$

$$= \frac{1}{\sqrt{2}} e^{ik\cdot r} \sum_{s=\pm 1} C_{sk}\left(\theta\right) \begin{pmatrix} 1 \\ s e^{i\varphi_k} \end{pmatrix} \tag{5.6b}$$

where

$$C_{sk}\left(\theta\right) = \frac{1}{2}\left(1 + s e^{i(\theta - \varphi_k)}\right) \tag{5.7}$$

The above decomposition clearly illustrates that the Gaussian wave packet on graphene lattice has both electron and hole states which are weighted

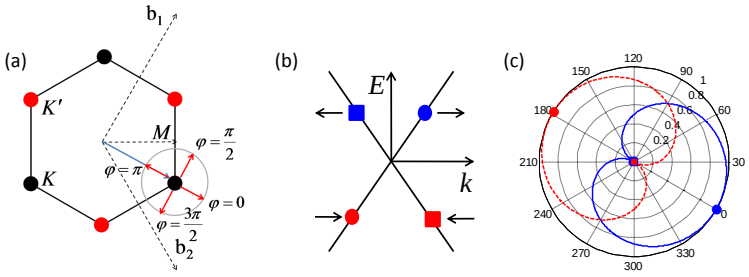

Figure 5.4: (a) The first Brillouin zone showing the directions corresponding to pseudospin angle φ, (b) slice of Graphene's low energy band structure along k_x-axis and (c) Polar plot of the function C_s for s=1(blue solid) and s=-1(red dotted) as function of φ for $\theta = 0$.

by C_{sk} in addition to the Gaussian distribution function in momentum space centered around $\mathbf{k}_o$. As we can see from Figure 5.4, the weight function strongly suppress the conduction band states which are orthogonal to the pseudo spin θ and valence band states which are orthogonal pseudo spin $\theta + \pi$. Consequently, the shape of the time evolved wave packet in Figure 5.3 reflects the shape of the weight function shown in Figure 5.4(c). From this example of Gaussian wave packet on graphene, we see that modeling such transport requires a multi-band model most reliably available through an atomistic tight binding Hamiltonian. As a result, the computational cost per time step increases due to the large system of equations that needs to be solved. Furthermore, the very high velocity of carriers in graphene and the atomistic lattice put an upper bound on the time step.

For an explicit time-evolution scheme, the fixed carrier speed of 1 nm/fs and nearest-neighbor inter-atomic spacing on the order of 0.1 nm would suggest the use of a time step less than or equal to about 0.1 fs simply to track a wave

front. In practice, I have also found this estimate to be appropriate for our half-implicit scheme(s). As shown in Figure 5.5, the use of a 1.0 fs time step results in slowed and qualitatively inaccurate motion of the wave-packet by comparison to the result for a 0.1 fs time step. Note that in the latter case (Figure 5.5(a)) the leading edge of the wave-packet travels about 3 nm in 3 fs as expected. With the basic CN method, the computational cost per

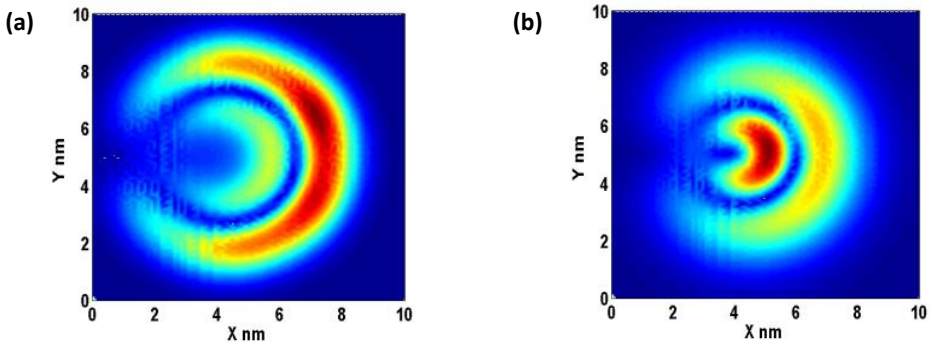

Figure 5.5: Snapshot of a time evolved initially Gaussian wave-packet on Graphene at 3 fs for a time step of (a) 0.1 fs and (b) 1 fs. (Reprinted with permission from [14], copyright (2012) by the IEEE).

time step increases super-linearly with size of the considered simulation region due to the far-from-diagonal nonzero elements of the Hamiltonian required in Eq. (5.2). To reduce the computational effort, I introduce two split operator alternating-direction implicit (ADI) methods in the following section.

5.2.2 Alternate Direct Implicit method for Graphene

Consider a graphene nanoribbon with $N_{L/C/R}$ layers in left, central and right region and each layer with N_y rectangular unit cells as defined in Fig. 5.1.

The size of the Hamiltonian is $N = N_U \times N_y \times N_x$ where $N_x = N_L + N_C + N_R$ and number of atoms per unit cell $N_U = 4$. Even though the Hamiltonian is sparse, it has a large band width. For example, for an arm-chair nanoribbon of length 51.12 nm ($N_x = 120$) and width 7.37 nm ($N_y = 30$) the resulting Hamiltonian has bandwidth of 91 after sparse reverse Cuthill-McKee (RCM) ordering. In order to reduce the computational time per time step I use split operator alternate direction implicit (ADI) method. I introduce two ADI methods for graphene based on the splitting of the Hamiltonian. In the first approach I split the Hamiltonian operator H into two operators H_x and H_y such that $H_{x/y}$ represents the hopping only in the x and y directions respectively. Using equations (D.2), (D.3), and (D.4) and setting $\alpha_o = 0.5\alpha_1$, $\beta_y = 0$, $\beta_x = \beta_2$ we obtain the operator H_x. Similarly the operator H_y can be obtained by setting $\alpha_o = 0.5\alpha_1$, $\beta_y = \beta_1$, $\beta_x = 0$. With such splitting we can write $H = H_x + H_y$. The process is equivalent to mapping the graphene to a two dimensional rectangular lattice with 4 atom basis as indicated in Figure 5.1. The time step is divided into two sub steps as given below such that

$$\left(I + i\frac{\Delta t}{2}H_x^{n+1/2}\right)\psi^{n+1/2} = \left(I - i\frac{\Delta t}{2}H_y^{n+1/2}\right)\psi^n \tag{5.8a}$$

$$\left(I + i\frac{\Delta t}{2}H_y^{n+1/2}\right)\psi^{n+1} = \left(I - i\frac{\Delta t}{2}H_x^{n+1/2}\right)\psi^{n+1/2} \tag{5.8b}$$

in the first sub step I consider explicit hopping in y direction and implicit hopping in x direction and vice versa in the second sub step. In this way the initial problem of solving a sparse linear system with large bandwidth at every time step is reduced to solving two linear systems of smaller band

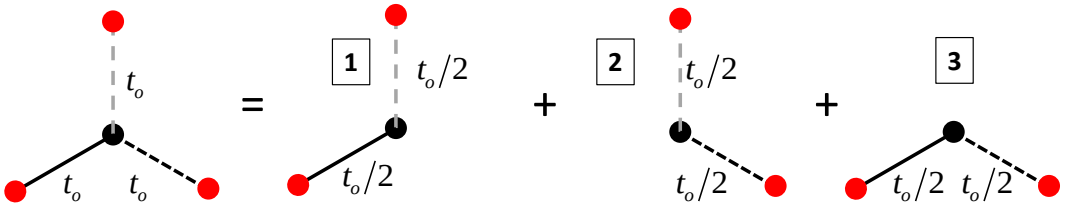

Figure 5.6: Illustration of bond breaking to split the Hamiltonian into three parts

width. For the particular example of a nanoribbon mentioned above, the split operators $H_{x/y}$ after RCM ordering are tridiagonal and pentadiagonal, respectively, which makes the time step relatively less expensive with respect to the non-ADI method. Another approach to split the Hamiltonian operator uses the inherent symmetry of lattice. Due to the sp^2 bonding of the carbon atoms in graphene, each atom is connected to three other atoms as shown in the left hand side of Fig. 5.6. Consequently, in the nearest neighbor tight binding approach each row of the Hamiltonian has at most 3 non zero values. We can construct the Hamiltonian H_i by ignoring all bonds in the direction δ_i for each atom where $i = 1, 2, 3$. The procedure is illustrated in Figure 5.6 where the bonds that are considered in the Hamiltonian for each direction are shown. Since each bond appears in two Hamiltonians, I replace the hopping strength with half the hopping strength of the graphene. With such splitting procedure we can write $H = H_1 + H_2 + H_3$. Like the two sub step procedure mentioned in Eq. (5.8), I now split the time step into three sub steps as given

below

$$\left(I + i\frac{dt}{3}\eta H_1\right)\psi^{n+1/3} = \left(I - i\frac{dt}{3}\left(H - \eta H_1\right)\right)\psi^{n} \tag{5.9a}$$

$$\left(I + i\frac{dt}{3}\eta H_2\right)\psi^{n+2/3} = \left(I - i\frac{dt}{3}\left(H - \eta H_2\right)\right)\psi^{n+1/3} \tag{5.9b}$$

$$\left(I + i\frac{dt}{3}\eta H_3\right)\psi^{n+1} = \left(I - i\frac{dt}{3}\left(H - \eta H_3\right)\right)\psi^{n+2/3} \tag{5.9c}$$

where sub step is a valid approximation of the original Schrödinger equation. The difference between the three sub steps is only in the partitioning of the hopping into explicit and implicit terms. For example, in the sub step one I perform implicit time evolution for bonds contributing to H_1 and explicit in the remaining bonds. The operator H_1 contributes to 1/3 of the hopping of the total Hamiltonian which will imbalance the amount of implicit and explicit hopping. Therefore, in order to balance the amount of implicit hopping and explicit hopping at each atom I use a factor $\eta = 3/2$.

Figure 5.7(a)-(c) illustrate the accuracy of these ADI methods, showing essentially identical snapshots obtained at 3 fs from an initially Gaussian wave-packet using the reference non-ADI, ADI1, and ADI2 methods, respectively. Figure 5.7(d) shows the computational effort for the three different methods as a function of the number of atoms in the simulation region for an approximately 20 nm wide graphene ribbon of varying lengths. Both ADI methods exhibit only linear growth in simulation time with simulation region size, in contrast to the super-linear growth with the non-ADI method. (That the effort required does not diverge earlier is a testament to the UMFPACK matrix solver used in MATLAB). While the ADI1 method requires slightly

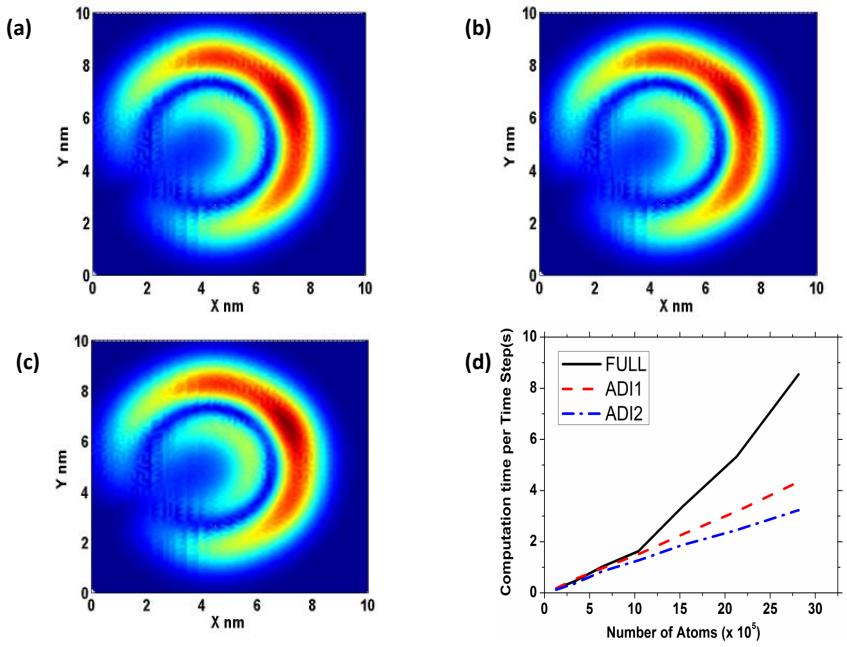

Figure 5.7: Snapshot of an initially Gaussian wave-packet with a pseudospin angle of 60o at 3fs, obtained using (a) non-ADI, (b) ADI1, and (c) ADI2 methods, and (d) computational time per time step for the nonADI (black), ADI1 (dashed red), and ADI2 (dash-dot blue) methods as a function of simulation region size. (Reprinted with permission from [14], copyright (2012) by the IEEE).

less effort per time step, a more detailed analysis finds the ADI2 method to be slightly more accurate for a given time step. The differences are small enough that convenience may be the best determinant of which to use.

5.3 Absorbing and Injection Boundary Conditions

For the sake of the current discussion, consider a central (C) simulation region through which I wish to consider dynamic probability/charge current flow, coupled to open left (L) and right (R) "leads," where the wave-function in each region is ψ_C, ψ_L, and ψ_R, respectively. Consider injection first, from the left lead for specificity. The total wave-function in the left lead can then be written as $\psi_L = \psi_L^{in} + \psi_L^r$ where ψ_L^{in} is the assumed-given incident wave-function whose time evolution in the lead is known, and ψ_L^r is any reflected wave. The time-dependent Schrödinger's equation is transformed into the inhomogeneous/NEGF form,

$$i\hbar\frac{d}{dt}\begin{bmatrix}\psi_L^r\\ \psi_C\\ \psi_R\end{bmatrix} = \begin{bmatrix}\mathbf{H}_L & \mathbf{H}_{LC} & 0\\ \mathbf{H}_{CL} & \mathbf{H}_C & \mathbf{H}_{CR}\\ 0 & \mathbf{H}_{RC} & \mathbf{H}_R\end{bmatrix}\begin{bmatrix}\psi_L^r\\ \psi_C\\ \psi_R\end{bmatrix} + \psi_S \tag{5.10}$$

Here, H_C, H_L, and H_R are the Hamiltonians of the isolated central, left and right regions, respectively. The off-diagonal elements represent the coupling between the adjacent regions. The probability source term ψ_S on right-hand-side is given by

$$\psi_S = \begin{pmatrix}\left(\mathbf{H}_L - i\hbar\frac{d}{dt}\right)\psi_L^{in}\\ \mathbf{H}_{LC}\psi_L^{in}\\ 0\end{pmatrix} \tag{5.11}$$

Consider next absorption by the leads. In principle, one could derive

boundary self-energies based on quantum transmitting boundary conditions (QTBCs) as for time-independent NEGF. However, non-stationary QTBCs are required for the time-dependent system. In principle, creation of these requires keeping track of, and integrating over at each time step, the past history of the wave-function at the boundary, or at least a significant period thereof for a reasonable approximation. It may be possible to provide non-stationary but at least local-in-time approximate QTBCs via extrapolation of the time-dependent wave-function within the central region to (just) across the boundary, based on assumptions about its approximate form (such as in [71] perhaps aided by a transverse mode expansion). However, multi-band transport and quasi-non-dispersive transport, such that abrupt variations in the wave-function induced far from the lead boundary may remain at the boundary, make this latter approach to QTBCs challenging.

For these reasons, for simplicity, and for flexibility, I employ stationary but spatially non-local position-dependent complex absorbing potentials (CAP)—self-energies—within leads of finite length, such as used in electromagnetics [72]. To be effective, however, there are some basic requirements that the combination of complex potential and lead length must meet. To avoid reflection from the end of the finite leads, the average complex potential within the lead must be sufficiently large to completely absorb any injected—and very fast in graphene—wave-function before it can reach the end of the lead and reflect all the way back to the simulation region. However, the rate of change of the complex potential with position, particularly near the boundary between

the central region and the lead, cannot be so fast as to cause back reflection in and of itself. See appendix G for an illustrative more detailed discussion of the application of the above discussed methods for a one dimensional problem.

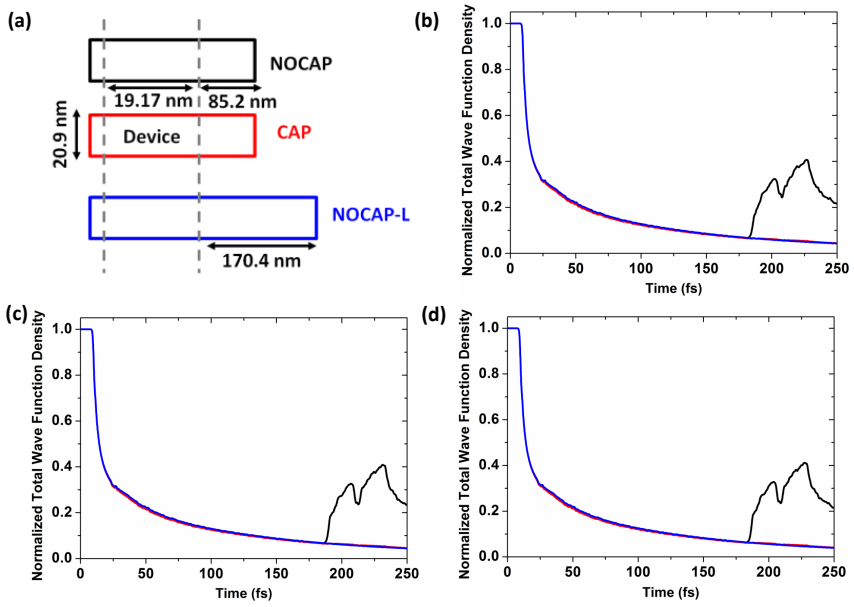

Figure 5.8: (a) Geometries used for simulations with complex absorbing potential (CAP), no complex absorbing potentials (NOCAP) and long (effectively infinite over the time period considered) devices with no complex absorbing potentials (NOCAP-L),as described in text. Total probability function density in the central region as a function of time for simulation with (b) Non ADI, (c) ADI 1 and, (d) ADI 2 methods, for NOCAP (solid black) and CAP (solid red) and NOCAP-L (solid blue) simulations.The NOCAP simulations exhibit back reflection into the central simulation region of interest. The CAP simulations results show no such reflections are essentially identical to the NOCAL-L simulations. (Reprinted with permission from [14], copyright (2012) by the IEEE).

To first illustrate the absorption boundary conditions only, the time evolution of an initially predominately right-directed Gaussian wave-packet out of an approximately 21 nm wide by 19 nm long section of graphene is

simulated. I consider three scenarios for the right lead, as illustrated in Figure 5.8(a): an 85 nm long lead region with no complex absorbing potential (NOCAP); the same lead region with an added complex absorbing potential (CAP) linearly ramped from zero at the boundary to the central region up to a purely imaginary 15 meV at its right-side hard-wall boundary; and an $\approx$ 170 nm long lead region with no complex absorbing potential (NOCAP-L). (With the predominately right-directed wave, the left lead is essentially of no consequence). Figures 5.8(b)-(d) show the probability density within the simulation region as a function of time for the non-ADI, ADI1 and ADI2, methods. The reflection from the lead end back into the simulation region in the NOCAP case is clear. In the NOCAP-L case, the simulation region is effectively a perfect infinite lead within the considered 250 fs simulation period as there is no time for the wave-function of maximum velocity 1 nm/fs to traverse back and forth across the lead region, which makes the associated result the reference ideal result. The agreement between the CAP and NOCAP results, therefore, demonstrates the accuracy of the absorbing potential approach. However, although accurate here, the combination of absorbing potential and channel length has not yet been fully optimized to minimize the latter and, thus, the computational burden. Finally, I note that the required length of the CAP leads is independent of the size of the central region, so the relative computational overhead associated with the CAP region will decrease with larger central regions that will be required for device simulation.

To illustrate injection along with the absorption, consider transient

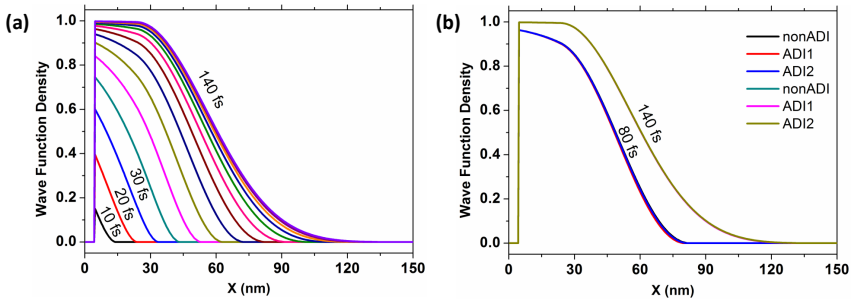

Figure 5.9: (a) Snapshots of probability density as function of position within the central region and a long right absorbing lead over 140 fs with the non-ADI implementation, with a left-injected nominally propagating eigenmode of the ribbon, although ramped up exponentially toward steady-state with a 20 fs time constant, serving as a source term in the left lead. (b) A subset of the results of (a) but using non-ADI, ADI1 and ADI2 methods with essentially indistinguishable results. (Reprinted with permission from [14], copyright (2012) by the IEEE).

through steady-state transport through an ≈21 nm wide metallic armchair graphene ribbon, with an 23 nm long central region. ψ_L^{in} nominally corresponds to a mono-energetic 95 meV electron relative to the Dirac point within the metallic subband. However, the amplitude of the incident wave-function is ramped exponentially toward saturation with a 20 fs time constant, so it is not truly mono-energetic, at least prior to saturation. Also, the source term ψ_s was defined as nonzero only within the "slice"–corresponding to the thickness of the gray rectangle along the x direction in Figure 5.1—of the left lead immediately adjacent to the central region, which is all that is necessary. As shown in Figure 5.9, near-steady-state conditions are achieved within the central region well within the 140 fs simulation period. Moreover, there is no apparent reflection back from the right lead at any time, as would be readily observable via a

standing wave pattern. (To show more clearly the decay of the probability with position in the lead, the length of the right lead was extended beyond what was necessary; again, the probability need only be absorbed before traversing both ways across the length of the lead.) The results of Figure 5.9(b) also evidence the continued effectiveness of the proposed ADI methods.

Continuing the previous example, now we look at the non-self consistent current response due to a sudden 200 mV bias on the drain of a two terminal device. The potential in the graphene is assumed to be flat and zero. Each of the injected transverse mode is normalized such that it carries current of $4q/h$. Figure 5.10(a) on the left shows the zero field transmission of the 21 nm wide metallic armchair graphene nanoribbon along with the band structure of the left lead, channel and right lead. The quasi-static current obtained by integrating the transmission is shown by the flat dashed-dot black line in Figure 5.10(b) on the right. The current calculated using time dependent formalism using the nonADI and the two ADI methods is shown by the solid lines. Clearly, the current level in the steady state agrees quite well with the current obtained by quasi static method.

Finally, I wish to point out that, although discussed and illustrated above in terms of a central simulation region and left and right leads, Hamiltonian is not subdivided by region, and source terms and, separately, absorbing regions could be placed anywhere within a plane of graphene as needed to simulate a particular system of interest.

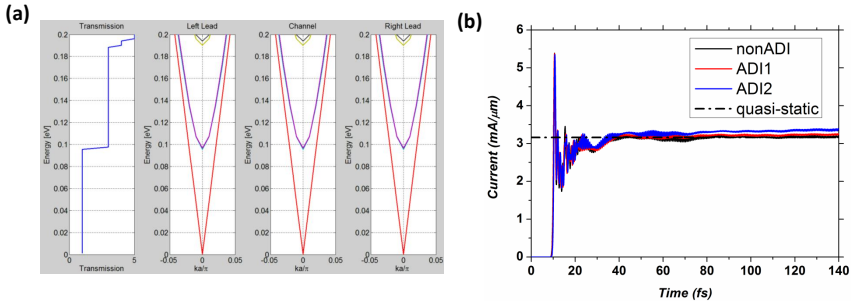

Figure 5.10: (a) Transmission as a function of injection energy and the band structure of the source lead, channel and right lead for a 20.9 nm wide armchair graphene nanoribbon (b) Non-self consistent current response in the middle of the channel for 200 mV step bias

5.4 Conclusion

I have presented a framework for an efficient time-dependent NEGF method for simulation of dynamic through steady-state quasi-ballistic intra- and inter-band quantum transport in graphene using an atomistic-tight-binding Hamiltonian, novel alternating direction semi-implicit numerical time-evolution schemes, and injecting and absorbing boundary conditions. The essential elements of this framework were demonstrated via illustrative simulations. Extension to multi-graphene-layer systems should also be possible.

Chapter 6

Conclusion

6.1 Research Motivation

There are many semiconductors with nominally superior electronic properties compared to silicon. However, silicon became the material of choice for MOSFETs due to its robust native oxide. With Moore's observation as a guiding principle, the semiconductor industry has come a long way in scaling the silicon MOSFETs to smaller dimensions every generation with engineering ingenuity and technological innovation, so that complementary metal-oxide semiconductor (CMOS) has become the ubiquitous technology for logic, as well as being increasingly used for analog and mixed signal applications. As per the 2012 International Technology Roadmap for Semiconductors (ITRS), the MOSFET is expected to be scaled to near 6 nm gate length by 2025. Alternate channel materials such as III-V, Germanium and/or device structures such as gate all around FETs, nanowire FETs may be used to boost the performance of the FETs. Engineering MOSFETs to their physical limitations will be more challenging than ever, and will be more so if we are to use III-V or germanium as channel material. More profoundly important however, are those physical limitations, which are based not on fabrication technology but on such things as tunneling and thermionic emission.

Novel two-dimensional materials such as Graphene and transition metal dichalcogenides (TMDs) are being explored as to further improve device performance. Promising process compatibility with existing CMOS technologies, fast carriers with high mobilities, and symmetric conduction and valence bands, have led to graphene being considered as a possible alternative to silicon. However, large area graphene FETs cannot be turned strongly off due to graphene's essentially zero energy band gap, and reliably creating band gaps via, e.g., use of nanoribbons is challenging at best. However, as a channel material, graphene may remain a good candidate for radio frequency (RF) applications where speed may be more important than ON/OFF ratio. Perhaps much more important in the long term—only time and research will tell—are the possibilities for novel device concepts that take advantage of the novel materials such as graphene, so call beyond CMOS devices.

6.2 Research Summary

In my research, I explored three graphene devices options for applications in RF and digital logic circuits. A recurring theme in the research work has been bottom up multi-scale modeling from basic physics to elementary circuit blocks. A schematic summarizing the typical workflow for my research work, using the BiSFET as an example, is shown in Figure 6.1. I have used analytical and NEGF simulations and measured data of real devices where available to understand essential device physics. I have then used such first principles work, mine and/or that of others, as a foundation for compact mod-

els capturing the essential physics. I have then used these compact models in Spice-level circuit simulations to explore potential benefits, limitations and requirements, of using the devices. I believe such an approach is necessary and can provide invaluable early insight into how the basic physics of yet to be demonstrated device concepts may translate into advantages or disadvantages at the application level. Device alternatives for beyond CMOS logic may not be drop in replacements for the MOSFET in CMOS logic circuits. There may not be MOSFET-equivalent device figures of merit (FOM). However, potential circuit performance must be compared, as for elementary logic blocks realized in CMOS or via some more exotic technology such as the BiSFET. While it might seem premature to perform circuit work for devices that do not yet, and may never, exist, the results of exploratory circuit level work can provide impetus for continuing, redirecting, or discontinuing device level research. The BiSFET and ITFET discussed in Chapter 3 and Chapter 4, respectively, illustrate this principle. ITFET-like devices have been demonstrated in the past, but only with large device dimensions. Understanding how scaling will affect device and circuit performance, therefore, becomes critical. The BiSFET is a novel device based on novel physics in a novel materials system, and neither a BiSFET nor just a superfluid condensate in dielectrically separated graphene layers has yet to be realized, making this device a clearly high-risk proposition. While the physics involved is highly interesting, it is only circuit level simulations, exhibiting order of magnitude reductions in switching power as compared to end of the roadmap CMOS, that identifies the potential

engineering reward of continuing along this path.

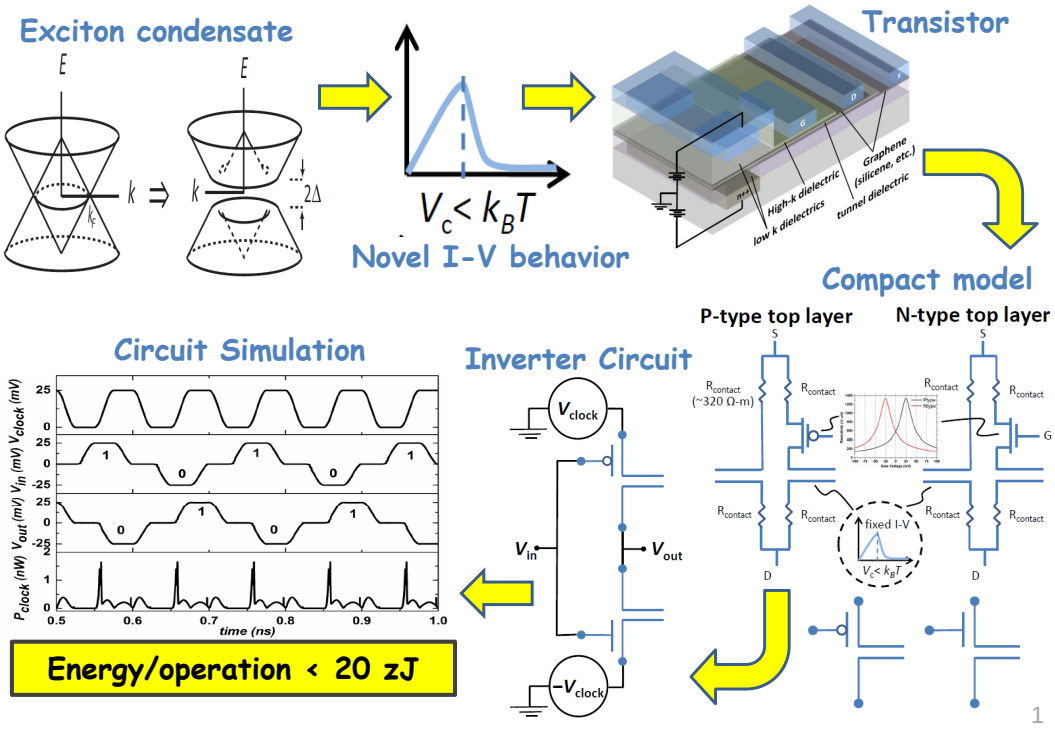

Figure 6.1: Graphical illustration of workflow from basic physics to circuits used for evaluating BiSFET

In Chapter 2, I presented a hardware-correlated compact model for large area graphene FETs based on experimental results as well theory.

In Chapter 3, I reviewed the essential physics on which the BiSFET is based as motivation for the compact models I have developed. I show that the BiSFET, if realized, can be used for implementing ultra-low power digital logic circuits with few 10s of zepto-Jule (10^{-21} J) switching energies consuming orders of magnitude less power than end of road map CMOS for the same logic functions. However, due to the NDR current voltage characteristics, I

showed that the BiSFETs are not a drop in replacement for the MOSFET. I discussed the implementation of basic logic circuits, from simple inverters to ripple carrier adders, and a memory element using BiSFETs. Once again due to the unique characteristics for the BiSFET, I showed that the logic circuits require multi-phase low-voltage clocked power supplies.

In Chapter 4, I have discussed the essential physics of the finite channel length effects that are important for short channel resonant-tunneling-based ITFETs which also exhibit NDR, with the resonance width in energy varying inversely with channel length. I also showed that fast carrier velocities are counterproductive here, leading to resonance broadening that increases linearly with the carrier velocity. In other words, short channels and fast carriers as generally considered in most devices, are intrinsically counterproductive here. From this point of view, I have suggested–but not modeled here–that TMDs with their much slower carriers would provide substantially better short channel effects, although TMD based devices might be limited by substantial scattering/finite lifetime induced resonance broadening. On the positive side, I showed that even strong coupling between layers allowing currents to approach lead-limited in resonance do not lead to additional resonance broadening. I implemented a compact model capturing the short channel induced resonance and lead injection limited saturation. I have shown an example of Inverter design for III-V ITFET device with dual polarity clocking, following the circuit architecture for the BiSFETs. I showed that for short channel devices there will be a trade-off between device size and voltage of operation as allowed by

the resonance width. Finally, I showed that, unfortunately, at least for the considered graphene and III-V material systems, switching power would likely exceed that of CMOS for scaled devices.

Quasi-static simulations are not sufficient to understand the intrinsic performance limitations for the graphene FETs for RF applications or to fully understand the device switching characteristics of the BiSFET which to date has been modeled with quasi-static device characteristics as a function of dynamic terminal voltages. Therefore, in Chapter 5, I presented a framework for efficient simulation of time-dependent quantum transport in graphene-based devices, considering efficient norm-conserving time evolution via alternating direction semi-implicit schemes appropriate and compatible with the employed hexagonal-lattice atomistic tight-binding model of graphene. I provided open boundary conditions addressing both carrier injection and absorption that are consistent with time dependent simulation, and again compatible with the hexagonal-lattice atomistic tight-binding model of graphene.

6.3 Future Research Directions

The BiSFET, which exhibit theoretically sub-k_BT onset NDR based on many-body superfluid excitonic condensation, could have extremely low power operation and is a promising alternative for conventional CMOS. Of course, there are many technological and theoretical challenges to its realization. But if realizable, I have shown that ultralow-power logic circuits, and memory elements are possible, even while the device design and thus device and circuit

models have had to evolve.

However, there is much room for exploration of other applications, and the approach to all potential applications. While I have shown a way in which the devices can be employed for these purposes, there may be other still better ways, such as based on gated resonant tunneling diode logic but at much lower voltages, or based on Josephson junction-like logic but at room temperatures with gating. Moreover, one might consider digital functional blocks such as programmable logic gates, parity generators, and analog applications such as frequency multipliers, analog to digital converters. With the compact BiSFET device model of this work, such possibilities can be evaluated via SPICE simulation. Moreover, the compact model itself and perhaps even the device design will need to be developed further as we learn more about the physics of the exicton condensates in graphene bilayers, or consider other 2D material systems.

A critical aspect of the BiSFET model that needs further refinement is the behavior of the device as it switches into and oscillates within the NDR region. However, understanding the device current voltage characteristics in the NDR region requires a study of the condensate dynamics. To the best of my knowledge, a time-dependent simulation of many body electron-electron interactions in a device structure of size of BiSFET has never been done. With the efficient formalism for time dependent transport in graphene presented in Chapter 5 as a starting point, one may be able to extend it to a time dependent Hartree-Fock (HF) based simulation of the condensate dynamics. (The long

range carrier-carrier interactions which result in a dense Hamiltonian perhaps could be approximated by a local model to reduce the matrix band width of the HF Hamiltonian). For the device simulations, one may be able to interface the code developed as a part of my research with the open source time-dependent density functional theory code (octopus).

I have also explored and captured the essential physics of the sub 100 nm ITFETs in to a compact model. This work can be extended with further refinement of the model with fully self-consistent current-voltage simulations of the graphene and III-V based ITFETs. In my experience, achieving self consistency in the NEGF simulation of the ITFET has proved to be very difficult due to strong density oscillations caused by resonances and standing wave patterns in the device. The problem is further complicated due to possible multiple solutions to the coupled transport and Poisson system of equations. For example, Figure 6.2 shows one of the converged self-consistent scaled electron density profile for a 30 nm long GaAs-AlAs with 2 nm AlAs interlayer. However, due to uniform dopant density (along the transport direction) in the source drain leads, one expects a uniform electron density profile in the leads, whereas the NEGF solution shows an oscillating solution. Charge density calculation is computationally expensive due to the adaptive integration scheme required for resolving the resonances which in turn are sensitive to the potential profile. I have implemented a parallel adaptive integration scheme and solution stabilizing method known as Anderson mixing to improve the convergence, but this work is not complete. One can take it forward from here

to iron out the convergence issues for the self-consistent NEGF simulations if

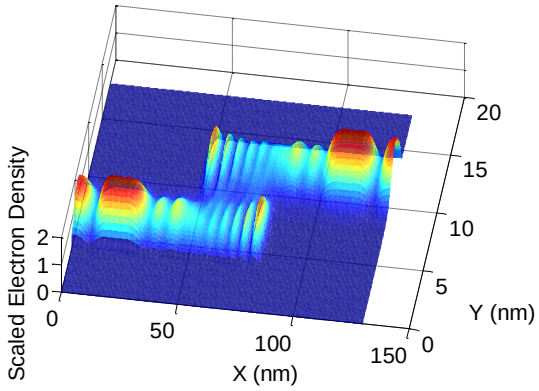

Figure 6.2: Snapshot of a converged density of a 30 nm long GaAs-AlAs based ITFET with 2 nm AlAs interlayer.

ITFETs, issues which also may be important for BiSFET simulation.

However, based on preliminary results for III-V and graphene systems, it is not clear if the latter work is justified in terms of resonant-tunneling-based ITFETs alone. However, the materials database for the code can be extended to include transition metal dichaclcogenides (TMD), to study the TMD-based ITFETs. (Note that the III-V Hamiltonian is based on effective-mass parabolic band structure of the valley; extending it to multiple valleys in the effective mass approximation would be conceptually simple, although detailed knowledge of the simulation code will be required.) Also, while we have considered resonant tunneling-based ITFETs only, band-to-band tunneling, particularly in gapped 2D material systems such as TMDs should be treatable with much the same simulations tools, where converting from electron-like to hole-like band structures on alternate layers should not be problematic within

the effective mass based calculations.

Finally, the code for the work presented in Chapter 5 can be used essentially as is to perform self-consistent time dependent NEGF simulations for graphene FETs. A more challenging problem is to generalize the alternate direct implicit method (ADI) presented for the graphene case to other crystal structures. Traditionally, ADI methods were implemented by splitting the separable operators into orthogonal grid directions or by splitting the operators based on the physical variables in a coupled system of equations. The ADI method I developed for the graphene exploits the symmetry of the graphene crystal lattice to write the nearest neighbor atomistic Hamiltonian into three parts each of which is used to time evolve the wave function in along a particular direction. One can in principle extend this method to other crystalline materials such as Silicon or GaAs and develop split operator methods to perform time dependent NEGF transport simulations there as well. During each sub-step of the ADI method, the operators must be carefully partitioned into explicit and implicit operators to ensure norm conserving time evolution. However, speedup of the computation is not a given and it should be carefully examined. A straightforward implementation that I can think of is to extend the method used in ADI1 by mapping the unit cell of the crystal to a node on two dimensional or three dimensional orthogonal grid.

Appendices

Appendix A

Graphene carrier density

The number density of electrons(n) in graphene is given by,

$$n = \int_{E_D}^{\infty} g(E) f(E) dE \tag{A.1}$$

$$= \frac{N_{cv}(T)}{(k_B T)^2} \int_{E_D}^{\infty} \frac{E - E_D}{1 + e^{\frac{E-\mu}{k_b T}}} dE \tag{A.2}$$

$$= N_{cv}(T)\, \Gamma(2)\, F_1\left(\frac{\mu - E_D}{k_B T}\right) \tag{A.3}$$

where F_1 is the Fermi integral (of index 1). I have assumed that the linear dispersion is valid unto large energies ($\rightarrow \infty$). This assumption is reasonable as the Fermi distribution falls to zero with in few $k_B T$ from μ which is typically less than 300 meV and the linear dispersion is a valid approximation upto 1 eV. Fermi Integral of index j is given by,

$$F_j(z) = \frac{1}{\Gamma(j+1)} \int_0^{\infty} \frac{x}{1 + e^{x-z}} dx \tag{A.4}$$

Similarly, the hole density is given by,

$$p = N_{cv}(T)\, \Gamma(2)\, F_1\left(\frac{E_D - \mu}{k_B T}\right) \tag{A.5}$$

The carrier density in the graphene in the limit of zero temperature, $T \to 0$, is given by

$$n = \frac{g_s g_v}{2\pi} \frac{\zeta^2}{4\left(\hbar v_F\right)^2} \left(1 + \mathcal{S}\left(\zeta\right)\right) \tag{A.6}$$

$$p = \frac{g_s g_v}{2\pi} \frac{\zeta^2}{4\left(\hbar v_F\right)^2} \left(1 - \mathcal{S}\left(\zeta\right)\right) \tag{A.7}$$

where $\zeta = \mu - E_D$ is the Dirac point referenced Fermi level. Substituting the value of the spin and valley degeneracy, the above equations can be simplified as

$$n = \frac{\zeta^2}{2\pi\left(\hbar v_F\right)^2} \left(1 + \mathcal{S}\left(\zeta\right)\right) \tag{A.8}$$

$$p = \frac{\zeta^2}{2\pi\left(\hbar v_F\right)^2} \left(1 - \mathcal{S}\left(\zeta\right)\right) \tag{A.9}$$

where $\mathcal{S}(z)$ is the signum function given by

$$\mathcal{S}(x) = \begin{cases} -1 & x < 0 \\ 0 & x = 0 \\ 1 & x > 0 \end{cases} \tag{A.10}$$

In the zero temperature approximation, the total charge density in graphene as a function of ζ is given by

$$Q_g\left(\zeta\right) = -e \frac{\zeta\left|\zeta\right|}{\pi\left(\hbar v_F\right)^2} \tag{A.11}$$

where $|\cdot|$ is the absolute value function. At finite temperature, the charge density is given by,

$$Q_g\left(\zeta\right) = e n_i \left(F_1\left(-\frac{\zeta}{k_B T}\right) - F_1\left(\frac{\zeta}{k_B T}\right)\right) \tag{A.12}$$

Appendix B

Poisson equation for Graphene MIS

Figure B.1 shows the schematic drawing of vertical cross section of a metal-insulator-semiconductor (MIS) structure where the semiconductor is a monolayer of graphene. Consider a gate dielectric with dielectric constant ϵ_{di} and thickness t_{di}. Assuming no charge density in the gate dielectric, the electric filed in the dielectric is constant. Denoting the surface charge density on the metal as Q_m, the electric filed E_{di} (in x direction) in the dielectric is given by

$$E_{di} = -\frac{\varphi_{dg} - \varphi_m}{t_{di}} = \frac{Q_m}{\epsilon_{di}} \tag{B.1}$$

where φ_m and φ_{dg} denote the potential at the metal and dielectric-graphene interface respectively. Assuming the interface charge density at dielectric and graphene interface as Q_i^{gd}, the boundary condition at the dielectric and

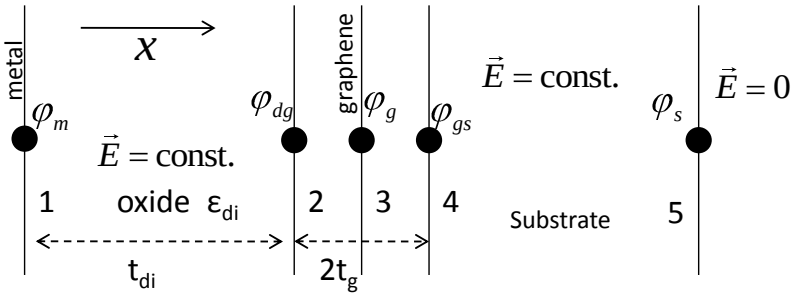

Figure B.1: Vertical cross-section of a graphene MIS structure

graphene interface is given by

$$\epsilon_g E_1 - \epsilon_{di} E_{di} = Q_i^{gd} \tag{B.2}$$

where E_1 is the electric field on the graphene side of dielectric-graphene interface and is given by

$$E_1 = -\frac{\varphi_g - \varphi_{dg}}{a} \tag{B.3}$$

assuming no interface charge at the substrate and air interface, electric field in the substrate is zero (assumption). Since graphene is a 2D material, boundary condition for graphene interface is given by

$$\epsilon_g E_2 - \epsilon_g E_1 = Q_g \tag{B.4}$$

Also, assuming an interface charge density Q_i^{gs} at graphene and substrate interface we have

$$0 - \epsilon_g E_2 = Q_i^{gs} \tag{B.5}$$

where

$$E_2 = -\frac{\varphi_{gs} - \varphi_g}{a} \tag{B.6}$$

Denoting χ_g as the electron affinity of graphene, the conduction band energy level E_c in presence of potential φ_g is given by

$$E_c = -\chi_g - e\varphi_g \tag{B.7}$$

Denoting the conduction band referenced Fermi energy as ζ, the Fermi level μ_g in graphene is given by

$$\mu_g = \zeta + E_c = \zeta - \chi_g - e\varphi_g \tag{B.8}$$

Similarly, the Fermi level in the metal is given by,

$$\mu_m = -W_m - e\varphi_m \tag{B.9}$$

Under an applied gate bias V_G with respect to graphene, we have

$$\mu_m - \mu_g = -eV_G \tag{B.10}$$

The above equation can be rewritten as,

$$-eV_G = -W_m - \zeta + \chi_g + e\left(\varphi_g - \varphi_m\right) \tag{B.11}$$

The unknowns in the above equation are the conduction band referenced Fermi energy of graphene ζ, and the potential difference between graphene and metal. The later can be obtained as follows: Let us define $C_{di} = \frac{\epsilon_{di}}{t_{di}}$ and $C_g = \frac{\epsilon_g}{a}$, to obtain,

$$Q_m = \epsilon_{di} E_{di} \tag{B.12}$$

$$Q_i^{gd} = \epsilon_g E_1 - \epsilon_{di} E_{di} \tag{B.13}$$

$$Q_g = \epsilon_g E_2 - \epsilon_g E_1 \tag{B.14}$$

$$Q_i^{gs} = -\epsilon_g E_2 \tag{B.15}$$

As a sanity check we can see that the above equations satisfy the charge neutrality condition, $Q_m + Q_i^{gd} + Q_g + Q_i^{gs} = 0$. Using Eq. (B.13) we obtain an expression for φ_{dg},

$$\frac{\varepsilon_g}{a}\left(\varphi_g - \varphi_{dg}\right) - \frac{\varepsilon_{di}}{t_{ox}}\left(\varphi_{dg} - \varphi_m\right) = -Q_i^{gd} \tag{B.16}$$

$$\varphi_{dg} = \frac{C_g\varphi_g + C_{di}\varphi_m + Q_i^{gd}}{C_g + C_{di}} \tag{B.17}$$

Using the expression for φ_{dg} in Eq. (B.14) and(B.15) gives,

$$\varphi_g - \varphi_m = \frac{C_g + C_{di}}{C_g C_{di}}\left(Q_g + Q_i^{gs}\right) + \frac{Q_i^{gd}}{C_{di}} \tag{B.18}$$

Since the distance a between graphene layer and the interface is very small, $C_g >> C_{ox}$. Therefore, in the limit of large C_g, potential difference between graphene and metal is given by,

$$\varphi_g - \varphi_m = \frac{Q_g + Q_i^{gs} + Q_i^{gd}}{C_{di}} \tag{B.19}$$

Note that in the above equation the charge in graphene Q_g is a function of ζ. The interface charges can be written as $Q_i^{gs} + Q_i^{gd} = Q_F + Q_{it}\left(\zeta\right)$, where Q_F is the fixed charge density, and Q_{it} is the ζ dependent charge density due to interface traps. Finally, from Eq. (B.11) and the last equation above, we have

$$eV_G = W_m - \chi_g + \zeta - e\frac{Q_g\left(\zeta\right) + Q_F + Q_{it}\left(\zeta\right)}{C_{di}} \tag{B.20}$$

Defining the gate voltage at which the graphene is charge neutral($\zeta = 0$ and Q_g=0) as the flat band voltage V_{FB}, we then have

$$V_{FB} = \frac{W_m - \chi_g}{e} - \frac{Q_F + Q_{it}\left(\zeta = 0\right)}{C_{di}} \tag{B.21}$$

Subtracting Eq. (B.21) from Eq. (B.20) gives the charge voltage relationship,

$$V_G - V_{FB} = \frac{\zeta}{e} - \frac{Q_g\left(\zeta\right)}{C_{di}} - \frac{Q_{it}\left(\zeta\right) - Q_{it}\left(\zeta = 0\right)}{C_{di}} \tag{B.22}$$

where $Q_g\left(\zeta\right)$ is given by Eq. (2.2). Assuming a constant interface trap density D_{it} gives an interface trap charge density $Q_{it} = -eD_{it}\zeta$ and interface trap

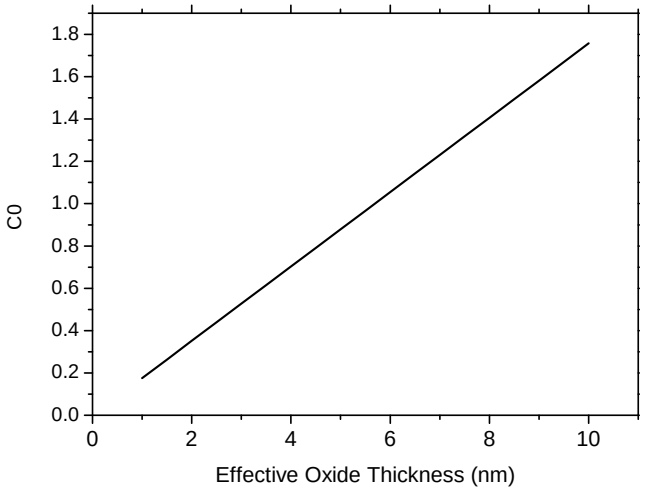

Figure B.2: Variation of parameter C0 with effective oxide thickness.

capacitance $C_{it} = e^2 D_{it}$. Equation (2.7) can be rewritten in dimensionless form as,

$$\frac{V_G - V_{FB}}{V_{th}} = \frac{\zeta}{k_B T}\left(1 + \frac{C_{it}}{C_{di}}\right) - e^2 \frac{n_i}{k_B T C_{di}}\left(F_1\left(-\frac{\zeta}{k_B T}\right) - F_1\left(\frac{\zeta}{k_B T}\right)\right) \tag{B.23}$$

where the thermal voltage is $V_{th} = \frac{k_B T}{e}$. Denoting the effective gate voltage in units of thermal voltage as $x = \frac{V_G - V_{FB}}{V_{th}}$, the conduction band referenced Fermi energy as $z = \frac{\zeta}{k_B T}$, and $C_0 = \frac{e^2 n_i}{k_B T C_{di}}$ and $C_1 = \frac{C_{it}}{C_{di}}$, gives,

$$x = z(1 + C_1) - C_0\left(F_1(-z) - F_1(z)\right) \tag{B.24}$$

Figure B.2 and Figure B.3 shows the variation of the parameters C_0 and C_1 respectively with effective oxide thickness of the oxide. For different values of parameters with ranges of values as shown in Fig (B.3) and Fig (B.2) and for a 2 volt variation around flat band voltage, x varies between -80 to 80 at room

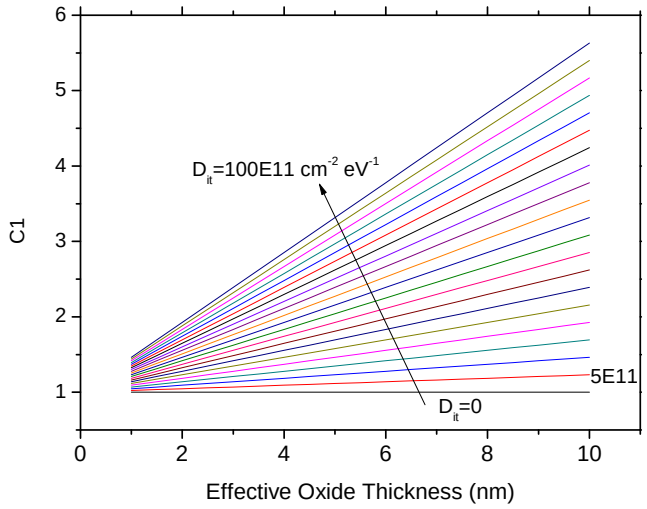

Figure B.3: Variation of parameter C1 with effective oxide thickness.

temperature, Eq. (B.24) must be solved numerically to obtain z. Note that the normalized conduction band referenced Fermi energy, z, is also referred to as surface potential. Note that the implicit Eq. (B.24) has no closed form solution for finite temperature. Whereas, assuming zero temperature, using Eq. (A.11), Eq. (B.23) can be written as,

$$V_G - V_{FB} = \frac{\zeta}{e}\left(1 + \frac{C_{it}}{C_{di}}\right) + e\frac{\zeta\,|\zeta|}{\pi C_{ox}\left(\hbar v_F\right)^2} \tag{B.25}$$

and the in dimensionless form as,

$$x = z\,(1 + C_1) + C_0\frac{z\,|z|}{4} \tag{B.26}$$

For scaling purposes in the above equation, T is assumed to be $300K$. Equation (B.26) can be solved analytically and the Fermi energy as function of gate

voltage is given by,

$$z = 2\mathcal{S}(x)\left(\frac{\sqrt{(1+C_1)^2 + C_0|x|} - (1+C_1)}{C_0}\right) \tag{B.27}$$

B.1 Terminal Charges

Charge on the gate at a distance x from source is given by,

$$Q_g(x) = C_{di}\left(V_g - V_{FB} - V(x) - \zeta(x)\right) \tag{B.28}$$

where

$$\zeta(x) = \frac{V_{th}a\left(V_g - V_{FB} - V(x)/V_{th}\right)}{b + \sqrt{|V_g - V_{FB} - V(x)|/V_{th}}} \tag{B.29}$$

$$\zeta(y) = \frac{V_{th}a\left(y/V_{th}\right)}{b + \sqrt{|y|/V_{th}}} \tag{B.30}$$

The gate charge is given by

$$Q_G = WC_{di}\int_0^L \left(V_g - V_{FB} - V(x) - \zeta(x)\right)dx \tag{B.31}$$

$$= -WC_{di}\int_{V_{gs,eff}}^{V_{gd,eff}} \left(y - \zeta(y)\right)\left(\frac{dx}{dV(x)}\right)dy \tag{B.32}$$

$$= -W\frac{L}{V_{ds}}C_{di}\int_{V_{gs,eff}}^{V_{gd,eff}} \left(y - \zeta(y)\right)dy \tag{B.33}$$

$$= -W\frac{L}{V_{ds}}C_{di}\int_{V_{gs,eff}}^{V_{gd,eff}} \left(y - \zeta(y)\right)dy \tag{B.34}$$

$$= -W\frac{L}{V_{ds}}C_{di}V_{th}^2\frac{(x_d^2 - x_s^2)}{2} + W\frac{L}{V_{ds}}C_{di}V_{th}^2\int_{x_s}^{x_d} z dx \tag{B.35}$$

In the limit of $V_{ds} \to 0$, charge on the gate is given by

$$Q_G = WLC_{di}\left(V_{gs} - V_{FB} - \zeta(x_s)\right) \tag{B.36}$$

Appendix C

Current Model Equations

The carrier densities $n = n_i F_{-z}$ and $p = p_i F_z$ can be obtained as follows: From Eq. (B.24) we have

$$F_1(-z) - F_1(z) = \frac{x - z(1 + C_1)}{C_0} \tag{C.1}$$

and the Fermi integral of order 1 also satisfies,

$$F_1(-z) + F_1(z) = \frac{z^2}{2} + 2F_1(0) \tag{C.2}$$

From Eqs. (C.1) and (C.2) we can obtain,

$$F_z = \frac{z^2}{4} + F_1(0) - \frac{x - z(1 + C_1)}{2C_0} \tag{C.3}$$

$$F_{-z} = \frac{z^2}{4} + F_1(0) + \frac{x - z(1 + C_1)}{2C_0} \tag{C.4}$$

where z is the solution of the implicit equation, Eq. (B.24) and has the form,

$$z = a\frac{x}{b + \sqrt{|x|}} \tag{C.5}$$

where a and b are fitting parameters which deepened on C_0 and C_1. The approximate solution is obtained based on asymptotic behavior of the solution of the implicit equation, Eq. (B.24). For large x, $z \propto \sqrt{x}$ and for small x, $z \propto x$. In other words, the charge density varies linearly with gate voltage V_G

when the gate voltage is much larger than Dirac voltage V_D, and has nonlinear dependence around Dirac point.

To obtain the current equation as a function of gate voltage and drain voltage we require following integrals,

$$\int \frac{x}{\sqrt{x}+b}dx = 2b^2\sqrt{x} - bx + \frac{2x^{3/2}}{3} - 2b^3 \log[b+\sqrt{x}] \tag{C.6}$$

and,

$$\int \left(\frac{x}{\sqrt{x}+b}\right)^2 dx = \frac{2b^5}{b+\sqrt{x}} - 8b^3\sqrt{x} + 3b^2x - \frac{4}{3}bx^{3/2} + \frac{x^2}{2} + 10b^4 \log[b+\sqrt{x}] \tag{C.7}$$

Appendix D

Graphene Hamiltonian

Graphene's crystal structure can be visualized as a triangular lattice with a two atom basis or as two interspersed triangular lattices, say A and B, with a one atom basis as shown in Figure D.1. Graphene's lattice is also referred to as honeycomb lattice. The carbon atoms on one lattice are connected to three atoms from other lattice via sp^2 hybridized sigma bonds shown by vectors $\delta_i, i = 1, 2, 3$ which are 120° apart as shown in Figure D.1. The strong

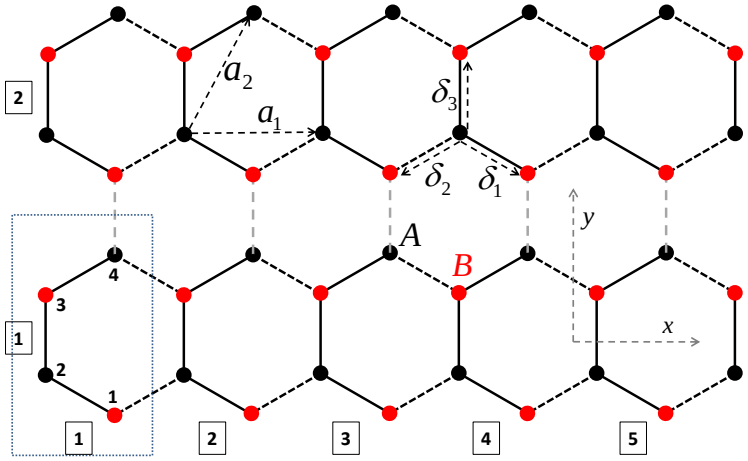

Figure D.1: Illustration of unit cells for tight binding Hamiltonian in graphene nanoribbon with N_x=5 unit cells in x direction and N_y=2 unit cells in y direction. The four atom unit cell is shown in dotted rectangle. Black dashed bonds show the connectivity between unit cells in x direction and dashed gray lines show the connectivity between individual unit cells in y direction.

sigma bonds formed by the carbon atoms result in deep lying valence bands of graphene's electronic structure and are responsible for the excellent mechanical properties of this material. The electrons in the unhybridized $2p_z$ orbitals, which are responsible for the optical and electronic properties of graphene, delocalize on the surface of the plane to form π(bonding) and π^* (anti-bonding) bands.

D.1 Graphene Tight Binding Hamiltonian

D.1.1 Real Space Representation

Atomistic p_z orbital based tight binding Hamiltonian for the graphene can be written as

$$H = \sum_{ll'} t_{R_l,R_{l'}} a^\dagger_{R_l} b_{R_{l'}} + t_{R_{l'},R_l} b^\dagger_{R_{l'}} a_{R_l} \tag{D.1}$$

where $a^\dagger$ and $b^\dagger$ are the fermion creation operators for lattice sites A and B respectively. The sum is over all the lattice sites of the A and B sublattice, $t_{R_l,R_{l'}}$ is the coupling strength between atoms at lattice position vector R_l on sublattice A and $R_{l'}$ on sublattice B. Assuming a nearest neighbor coupling, the tight binding Hamiltonian of Eq. (D.1) for the graphene nanoribbon shown in Figure D.1 can be written in a one dimensional nearest neighbor tight binding form

$$H_{ij} = \beta^\dagger \delta_{i-1j} + \alpha\delta_{ij} + \beta\delta_{i+1j} \quad i,j= 1 \text{ to } N_x \tag{D.2}$$

where α and β are the on site and inter-site hopping parameters and δ_{ij} is the Kronecker delta. However, α and β are matrices, unlike the 1D system with

one atom basis described by (G.6), where they are numbers. The matrix α describes the connectivity of the atoms along each slice in y direction, which again can be written in a 1D tight binding form

$$\alpha_{ij} = \beta_y^\dagger \delta_{i-1j} + \alpha_o \delta_{ij} + \beta_y \delta_{i+1j} \quad i,j= 1 \text{ to } N_y \tag{D.3}$$

and the matrix β which describes the inter-slice connectivity is given by

$$\beta_{ij} = \beta_x \delta_{ij} \quad i,j= 1 \text{ to } N_y \tag{D.4}$$

where the cell connectivity matrices $\alpha_o = \alpha_1$, $\beta_y = \beta_1$, $\beta_x = \beta_2$. The atom connectivity matrix α_1 for the unit cell and inter cell connectivity matrices β_1 and β_2 are defined as

$$\alpha_1 = \begin{pmatrix} 0 & t_o & 0 & 0 \\ t_o & 0 & t_o & 0 \\ 0 & t_o & 0 & t_o \\ 0 & 0 & t_o & 0 \end{pmatrix} \beta_1 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ t_o & 0 & 0 & 0 \end{pmatrix} \beta_2 = \begin{pmatrix} 0 & t_o & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & t_o & 0 \end{pmatrix} \tag{D.5}$$

D.1.2 Momentum Space Representation

The real space Fermion operators of Eq. (D.1) can be transformed to momentum space using the Fourier transforms given below

$$a_R = \frac{1}{\sqrt{N}} \sum_k a_k e^{ik\cdot R} \tag{D.6}$$

$$b_R = \frac{1}{\sqrt{N}} \sum_k b_k e^{ik\cdot R} \tag{D.7}$$

where N is the number of real space lattice sites. The Hamiltonian in (D.1) can be Fourier transformed to

$$H = -t \sum_k \begin{pmatrix} a_k^\dagger & b_k^\dagger \end{pmatrix} \begin{pmatrix} 0 & f(k) \\ f^*(k) & 0 \end{pmatrix} \begin{pmatrix} a_k \\ b_k \end{pmatrix} \tag{D.8}$$

by considering only nearest neighbor hopping with hopping strength t. The following Hamiltonian is obtained by linearizing the above equation bear the Dirac points K and K' in first Brillouin zone

$$H = \sum_{k,s=\pm 1} \begin{pmatrix} a_{sk}^{\dagger} & b_{sk}^{\dagger} \end{pmatrix} \begin{pmatrix} 0 & \varepsilon_k e^{-is\theta_k} \\ \varepsilon_k e^{is\theta_k} & 0 \end{pmatrix} \begin{pmatrix} a_{sk} \\ b_{sk} \end{pmatrix} \tag{D.9}$$

where $s = +1$ corresponds to low energy Hamiltonian near K point and $s = -1$ for K' point. We can diagonalize the 2×2 matrix in the above equation with the following unitary transformation

$$\begin{pmatrix} a_{sk} \\ b_{sk} \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} e^{-is\phi_k} & e^{-is\phi_k} \\ e^{is\phi_k} & -e^{is\phi_k} \end{pmatrix} \begin{pmatrix} a_{sck} \\ b_{svk} \end{pmatrix} \tag{D.10}$$

where c/v stand for conduction and valence bands, $2\phi_k = \theta_k = tan^{-1}(k_x/k_y)$ and $\varepsilon_k = \hbar v_F |k|$. The low energy Hamiltonian for the graphene is given below

$$H = \sum_{k,s=\pm 1} \begin{pmatrix} a_{sck}^{\dagger} & a_{svk}^{\dagger} \end{pmatrix} \begin{pmatrix} \varepsilon_{ck} & 0 \\ 0 & \varepsilon_{vk} \end{pmatrix} \begin{pmatrix} a_{sck} \\ a_{svk} \end{pmatrix} \tag{D.11}$$

where $\varepsilon_{ck} = \hbar v_F |k|$ and $\varepsilon_{vk} = -\hbar v_F |k|$. The number density of carriers in graphene is given by

$$N = \sum_{l} a_{R_l}^{\dagger} a_{R_l} + b_{R_l}^{\dagger} b_{R_l} = \sum_{k,s=\pm 1} a_{sk}^{\dagger} a_{sk} + b_{sk}^{\dagger} b_{sk} = \sum_{k,s=\pm 1} a_{sck}^{\dagger} a_{sck} + a_{svk}^{\dagger} a_{svk} \tag{D.12}$$

Appendix E

Tunneling Hamiltonian and Linear Response I-V Model

E.1 A-B Coupled Bilayer Graphene

For the bilayer graphene we can write the Hamiltonian as the sum of layer Hamiltonians H_1, H_2 and the interlayer coupling H_T given below

$$H = H_1 + H_2 + H_T \tag{E.1}$$

where

$$H_1 = \sum_{ll'} t_{1;R_l,R_{l'}} a^\dagger_{1R_l} b_{1R_{l'}} + t_{1;R_{l'},R_l} b^\dagger_{1R_{l'}} a_{1R_l} \tag{E.2}$$

$$H_2 = \sum_{ll'} t_{2;R_l,R_{l'}} a^\dagger_{2R_l} b_{2R_{l'}} + t_{2;R_{l'},R_l} b^\dagger_{2R_{l'}} a_{2R_l} \tag{E.3}$$

$$H_T = \sum_{ll'} t_{12;R_l,R_{l'}} a^\dagger_{1R_l} b_{2R_{l'}} + t_{21;R_{l'},R_l} b^\dagger_{2R_{l'}} a_{1R_l} \tag{E.4}$$

The subscripts of the operators and the summation have the same definition as in the previous section and the numbers in the subscripts correspond to layer number. The interlayer coupling Hamiltonian has only terms representing coupling between atoms at A sublattice on one layer to B sublattice in another layer with coupling strength t_{12}. The terms in Hamiltonian in (E.1) can be

Fourier transformed to

$$H_1 = -t_1 \sum_k \begin{pmatrix} a_{1k}^\dagger & b_{1k}^\dagger \end{pmatrix} \begin{pmatrix} 0 & f(k) \\ f^*(k) & 0 \end{pmatrix} \begin{pmatrix} a_{1k} \\ b_{1k} \end{pmatrix} \tag{E.5}$$

$$H_2 = -t_2 \sum_k \begin{pmatrix} a_{2k}^\dagger & b_{2k}^\dagger \end{pmatrix} \begin{pmatrix} 0 & f(k) \\ f^*(k) & 0 \end{pmatrix} \begin{pmatrix} a_{2k} \\ b_{2k} \end{pmatrix} \tag{E.6}$$

$$H_T = t_{12} \sum_k b_{1k}^\dagger a_{2k} + a_{2k}^\dagger b_{1k} \tag{E.7}$$

We can diagonalize H_1 and H_2 with the unitary transformation given by equation D.10 and the tunneling Hamiltonian is accordingly transformed using

$$b_k^\dagger = \frac{e^{-i\phi_k}\left(a_{ck}^\dagger - b_{vk}^\dagger\right)}{\sqrt{2}} \tag{E.8}$$

$$a_k = \frac{e^{-i\phi_k}\left(a_{ck} + b_{vk}\right)}{\sqrt{2}} \tag{E.9}$$

into

$$H_T = \frac{1}{2}\sum_k t_k \left(a_{1ck}^\dagger - b_{1vk}^\dagger\right)\left(a_{2ck} + b_{2vk}\right) + h.c \tag{E.10}$$

where $t_k = t_{12}e^{-i\theta_k}$. Finally, the low energy Hamiltonian for bernal (A-B) stacked bilayer graphene is given below

$$H = \sum_{ks} \varepsilon_{1sk} a_{1sk}^\dagger a_{1sk} + \sum_{ks} \varepsilon_{2sk} a_{2sk}^\dagger a_{2sk} + \frac{1}{2}\sum_k t_k \left(a_{1ck}^\dagger - b_{1vk}^\dagger\right)\left(a_{2ck} + b_{2vk}\right) + h.c \tag{E.11}$$

where the first two terms are the diagonalized Hamiltonians for layers 1 and 2 and the third term is the coupling between the layers.

E.2 Interlayer Tunneling : Linear Response

Consider a case of a graphene monolayers separated by an interlayer dielectric. We calculate the current with in the linear response regime where

we assume the tunneling term to be a weak perturbation. As an aside, note that the Fermion operators satisfy the following anti-commutation relation

$$\left\{a^{\dagger}_{isk}, a_{js'k'}\right\} = \delta_{ij}\delta_{ss'}\delta_{kk'} \tag{E.12}$$

where the layer index $i = 1, 2$, the band index $s = c, v$ and k is the Brillouin zone vector with respect to Dirac point K. The current through the device is defined by the rate of change of total number of electrons in the layer 1 or 2, $I_e = -e\langle I\rangle$ where

$$I = \dot{N}_1 = \frac{dN_1}{dt} = i\left[H, N_1\right] \tag{E.13}$$

and $[\cdot,\cdot]$ is the commutator. The total number for electrons in layer 1 is given by

$$N_1 = N_c + N_v = \sum_k a^{\dagger}_{1ck}a_{1ck} + \sum_k a^{\dagger}_{1vk}a_{1vk} \tag{E.14}$$

where $N_{c/v}$ is the total number of particles in conduction and valence band. Because the number operator commutes with single particle band Hamiltonians for layer 1 and 2, the current is given by $I = \dot{N}_c + \dot{N}_v$ and

$$\dot{N}_c = i\left[H_T, N_c\right] = \frac{1}{2}\sum_k\sum_{k'}\left[\left(t_k c^{\dagger}_{1k}c_{2k} + t^{*}_k c^{\dagger}_{2k}c_{1k}\right), a^{\dagger}_{1ck'}a_{1ck'}\right] \tag{E.15}$$

$$\dot{N}_v = i\left[H_T, N_v\right] = \frac{1}{2}\sum_k\sum_{k'}\left[\left(t_k c^{\dagger}_{1k}c_{2k} + t^{*}_k c^{\dagger}_{2k}c_{1k}\right), a^{\dagger}_{1vk'}a_{1vk'}\right] \tag{E.16}$$

$$\tag{E.17}$$

where $c_{1k} = a_{1ck} - a_{1vk}$ and $c_{2k} = a_{2ck} + a_{2vk}$. Using the anti-commutation relation given by equation E.12 we have the following

$$\dot{N}_{1c} = i\left[H_T, N_{1c}\right] = -i\frac{1}{2}\sum_k t_k a^\dagger_{1ck} c_{2k} - t^*_k c^\dagger_{2k} a_{1ck} \tag{E.18}$$

$$\dot{N}_{1v} = i\left[H_T, N_{1v}\right] = i\frac{1}{2}\sum_k t_k a^\dagger_{1vk} c_{2k} - t^*_k c^\dagger_{2k} a_{1vk} \tag{E.19}$$

and similarly for particles in layer 2 we have

$$\dot{N}_{2c} = i\left[H_T, N_{2c}\right] = i\frac{1}{2}\sum_k t_k c^\dagger_{1k} a_{2ck} - t^*_k a^\dagger_{2vk} c_{1k} \tag{E.20}$$

$$\dot{N}_{2v} = i\left[H_T, N_{2v}\right] = i\frac{1}{2}\sum_k t_k c^\dagger_{1k} a_{2vk} - t^*_k a^\dagger_{2vk} c_{1k} \tag{E.21}$$

clearly we have $\dot{N} = \dot{N}_1 + \dot{N}_2 = 0$. We can write the $\dot{N}_1$ in compact form as follows

$$\dot{N}_1 = -i\sum_k t_k c^\dagger_{1k} c_{2k} - t^*_k c^\dagger_{2k} c_{1k} \equiv -i\left(L - L^\dagger\right) \tag{E.22}$$

Using the Kubo formalism the particle current is given by

$$\langle I\rangle(t) = \int_{-\infty}^{\infty} dt' C^R_{I_p H_T}(t, t') \tag{E.23}$$

$$C^R_{I_p H_T}(t, t') = -i\theta(t - t')\left\langle\left[I_p(t), H_T(t')\right]\right\rangle_0 \tag{E.24}$$

where the time development is governed by unperturbed Hamiltonian $H = H_1 + H_2$. The retarted correlation function C^R_{I,H_T} can be simplified as

$$C^R_{I_p H_T}(t, t') = -\theta(t - t')\left\langle\left[L(t) - L^\dagger(t), L(t') + L^\dagger(t')\right]\right\rangle_0 \tag{E.25}$$

$$= -\theta(t - t')\left[\left\langle\left[L(t), L(t')\right]\right\rangle_0 - \left\langle\left[L^\dagger(t), L(t')\right]\right\rangle_0 + c.c.\right] \tag{E.26}$$

The first term in the above equation does not conserve particles. Therefore we are left with

$$I_p(t) = 2\,\mathrm{Re}\int_{-\infty}^{\infty} dt'\theta(t-t')\left\langle\left[L^\dagger(t), L(t')\right]\right\rangle_0 \tag{E.27}$$

$$= 2\,\mathrm{Re}\int_{-\infty}^{\infty} dt'\theta(t-t')\sum_{k,k'} t_k^* t_{k'}\left\langle\left[c_{2k}^\dagger(t)\, c_{1k}(t), c_{1k'}^\dagger(t')\, c_{2k'}(t')\right]\right\rangle_0 \tag{E.28}$$

$$= 2\,\mathrm{Re}\int_{-\infty}^{\infty} dt'\theta(t-t')\sum_{k,k'} t_k^* t_{k'}\begin{pmatrix} \left\langle c_{1k}(t)\, c_{1k'}^\dagger(t')\right\rangle_0\left\langle c_{2k}^\dagger(t)\, c_{2k'}(t')\right\rangle_0 \\ -\left\langle c_{1k'}^\dagger(t')\, c_{1k}(t)\right\rangle_0\left\langle c_{2k'}(t')\, c_{2k}^\dagger(t)\right\rangle_0 \end{pmatrix} \tag{E.29}$$

Changing the variable $t' = t' + t$, we can rewrite the above integral as

$$I_p = 2\,\mathrm{Re}\int_{-\infty}^{0} dt' e^{i(-\mu_{12})t'}\sum_{k;ss'}|t_k|^2\left(G_{1s}^{>}(k,-t')\,G_{2s'}^{<}(k,t') - G_{1s}^{<}(k,-t'')\,G_{2s'}^{>}(k,t')\right) \tag{E.30}$$

which can be Fourier transformed to

$$I_p = \int_{-\infty}^{\infty}\frac{d\omega}{2\pi}\sum_{k;ss'}|t_k|^2\left(G_{1s}^{>}(k;\omega)\,G_{2s'}^{<}(k;\omega+eV) - G_{1s}^{<}(k;\omega)\,G_{2s'}^{>}(k;\omega+eV)\right) \tag{E.31}$$

writing the lesser and grater Green's functions in terms of spectral densities we obtain the following expression for the tunnel current

$$I_p = \int_{-\infty}^{\infty}\frac{d\omega}{2\pi}\sum_{k;ss'}|t_k|^2 A_{1s}(k,\omega)\,A_{2s'}(k,\omega)\left[n_F(\omega-\mu_1) - n_F(\omega-\mu_2)\right] \tag{E.32}$$

Appendix F

Effective Mass Hamiltonian for NEGF Formalism

Consider the time independent Schrodinger equation given below

$$(E - H)\,\psi\,(x, y, z) = 0 \tag{F.1}$$

where the effectivemass Hamiltonian operator H is given by

$$H = -\frac{\hbar^2}{2}\frac{\partial}{\partial x}\frac{1}{m_x\,(x,y)}\frac{\partial}{\partial x} - \frac{\hbar^2}{2}\frac{\partial}{\partial y}\frac{1}{m_y\,(x,y)}\frac{\partial}{\partial y} - \frac{\hbar^2}{2}\frac{\partial}{\partial z}\frac{1}{m_z\,(x,y)}\frac{\partial}{\partial z} + U\,(x,y) \tag{F.2}$$

where $1/m_{x,y,z}$ are the diagonal terms of the effective mass tensor, $U = E_c(x,y) - q\phi(x,y)$ is the net potential energy. The conducton band edge of the material is given by E_c and the electrostaic potential is ϕ. For this chapter, the variables are assumed to be dimensionless using the scaling given below

$$(x, y, z) \rightarrow L\,(x, y, z) \tag{F.3}$$

$$(m_x, m_y, m_z) \rightarrow m_o\,(m_x, m_y, m_z) \tag{F.4}$$

$$(E, U, H) \rightarrow k_B T\,(E, U, H) \tag{F.5}$$

where L is a length scale, m_o is the mass of free electron, and k_BT is the thermal energy. Define $\lambda = \frac{\hbar^2}{2m_oL^2k_BT}$. The wave function ψ can be written as

$$\psi(x,y,z) = \int \frac{dk_z}{2\pi} e^{-ik_z z} \sum_n \psi_n(x)\,\xi_n(y,x) \tag{F.6}$$

where the modes ζ are the eigen modes of the following equation

$$\left(-\frac{\hbar^2}{2}\frac{\partial}{\partial y}\frac{1}{m_y(x,y)}\frac{\partial}{\partial y} + U(x,y)\right)\xi_n(y,x) = E_n(x)\,\xi_n(y,x) \tag{F.7}$$

and form a complete set satisfying the following equations

$$\langle \xi_m \mid \xi_n \rangle = \int \xi_m^*(y,x)\,\xi_n(y,x)\,dy = \delta_{m,n} \tag{F.8}$$

$$\sum_n \xi_n(y,x)\,\xi_n(y',x) = \delta(y-y') \tag{F.9}$$

Using equation (F.6) in equation (F.1) we can rewrite the Schrodinger equation as

$$-\lambda\frac{\partial}{\partial x}\frac{1}{m_x(x,y)}\frac{\partial}{\partial x}\Phi + \sum_n E_n\psi_n\xi_n + \lambda k_z^2\frac{1}{m_z(x,y)}\Phi = E\Phi \tag{F.10}$$

where

$$\Phi(x,y) = \sum_n \psi_n(x)\,\xi_n(y,x) \tag{F.11}$$

Equation (F.10) can be rewritten as an effective one dimensional coupled mode space Schrodinger equation given below

$$\sum_n H_{mn}^0\partial_x^2\psi_n + \sum_n H_{mn}^1\partial_x\psi_n + \sum_n H_{mn}^2\psi_n + \sum_n H_{mn}^3\psi_n = E\psi_m \tag{F.12}$$

where

$$H^0_{mn} = -\lambda \left\langle \xi_m \left| \nu_x \right| \xi_n \right\rangle \tag{F.13}$$

$$H^1_{mn} = -2\lambda \left\langle \xi_m \left| \nu_x \right| \partial_x \, \xi_n \right\rangle - \lambda \left\langle \xi_m \left| \partial_x \nu_x \right| \xi_n \right\rangle \tag{F.14}$$

$$H^2_{mn} = -\lambda \left\langle \xi_m \left| \nu_x \right| \partial_x^2 \xi_n \right\rangle - \lambda \left\langle \xi_m \left| \partial_x \nu_x \right| \partial_x \xi_n \right\rangle \tag{F.15}$$

$$H^3_{mn} = \lambda k_z^2 \left\langle \xi_m \left| \nu_z \right| \xi_n \right\rangle + E_m \delta_{mn} \tag{F.16}$$

The matrix element term such as $\langle \xi_m \left| A \right| \xi_n \rangle$ is defined as

$$\left\langle \xi_m \left| A \right| \xi_n \right\rangle = \int dy \xi_m^* \left(x, y \right) A \left(x, y \right) \xi_n \left(x, y \right) \tag{F.17}$$

Discretizing the channel along x-axis with uniform grid size $\Delta x = L/N$ we can write (F.12) in discrete from using central difference scheme for the derivative operators. The grid points are indexed as $x_i = i\Delta x$ where $i = 0$ to N such that $x_{0/N}$ is the source and drain boundary respectively. The

F.1 Lead Spectrum

Inside the leads the inverse effective masses ν_x and ν_z and potential U do not vary along transport direction. Therefore H^1, H^2 is zero in the leads and (F.12) is reduced to

$$H^0 \partial_x^2 \psi + H^3 \psi = E\psi \tag{F.18}$$

where ψ is column vector of size N_M which is the number of the modes. We seek solutions to (F.18) of the form $\psi = \chi e^{-i\lambda x}$. Therefore, we have to solve the following eigenvalue equation

$$-\mu^2 H^0 \chi = \left(EI - H^3 \right) \chi \tag{F.19}$$

where $\mu^2 = 2(1 - cos(\lambda\Delta x))/\Delta x^2$ and in the continuum approximation $\mu = \lambda$. Let us denote $A = (EI - H^3)$ and $B = -H^0$. The eigenvectors are normalized as

$$\langle \chi_m \left| B \right| \chi_n \rangle = \delta_{mn} \tag{F.20}$$

We can now write the right traveling wave ψ as a linear combination with weights C_m

$$\psi_n(x) = \sum_m C_m \chi_{nm} e^{i\lambda_m x} = \sum_m \chi_{nm} e^{i\lambda_m x} \langle \chi_m \left| B \right| \psi(0) \rangle \tag{F.21}$$

where χ_{nm} is the n^{th} component of m^{th} eigenvector of (F.19) with eigenvalue μ_m^2. If the effective mass ν_x and ν_z is independent of y, then H^0 and H^3 are diagonal matrices. Then the eigenvalue equation (F.19) has the solution

$$\mu_m^2 = \frac{E - E_m - \lambda\nu_z k_z^2}{\lambda\nu_x} \text{ and } \chi_{nm} = \delta_{nm} \tag{F.22}$$

F.2 Mode Space Greens Function

We can write the descretized one dimensional mode space Schrodinger equation as

$$\left(E\psi^i - H^{i,i-1}\psi^{i-1} - H^{i,i}\psi^{i-1} - H^{i,i+1}\psi^{i+1}\right) = 0 \tag{F.23}$$

where $H^{i,j}$ is block matrix of size N_M, $\psi^i = \psi(x_i)$ and index i runs from $-\infty$ to ∞. On restricting the Schrodinger equation to device region i.e., $i = 1$ to N_x equation (F.23) takes the form

$$(EI - H - \Sigma)\,\phi = S \tag{F.24}$$

where $\phi^i = \psi^i$ is a column vector of of size N_x, S is the source vector due to injection from source and drain boundaries. The self energy term Σ is obtained by imposing absorbing boundary conditions at $i = 1$ and $i = N$. We require that any outward going wave at the boundaries be propagated with out back reflections into the leads. For example, at the source boundary an outward going wave in the lead is given by

$$\psi_n^j = \sum_{n'} \sum_{m'} \chi_{nm'} e^{-i\lambda_{m'} x_j} \chi^\dagger_{m'n'} \psi^0_{n'} \tag{F.25}$$

The self energy term is obtained by requiring

$$H^{0,-1}\psi^{-1} = \sum_j \Sigma^{0,j} \psi^j \tag{F.26}$$

using equation (F.25) we have

$$H^{0,-1}\psi^{-1} = -\sum_k \frac{\lambda}{\Delta x^2} \langle \xi_m | \nu_x | \xi_k \rangle \sum_{m'} \chi_{km'} e^{-i\lambda_{m'} x_j} \left\langle \chi_{m'} | B | \psi^0 \right\rangle \tag{F.27}$$

$$= -\sum_n \frac{\lambda}{\Delta x^2} \sum_{k',m',n'} \langle \xi_m | \nu_x | \xi_{k'} \rangle \chi_{k'm'} e^{i\lambda_{m'}\Delta x} \chi^\dagger_{m'n'} B_{n'n} \psi^0_n \tag{F.28}$$

From the above set of equations we can observe that the self energy term for source is given by

$$\Sigma^{0,j}_{mn} = -\delta_{0,j} \frac{\lambda}{\Delta x^2} \sum_{k',m',n'} \langle \xi_m | \nu_x | \xi_{k'} \rangle \chi_{k'm'} e^{i\lambda_{m'}\Delta x} \chi^\dagger_{m'n'} B_{n'n} \tag{F.29}$$

similarly, the self energy term for drain lead is given by

$$\Sigma^{N,j}_{mn} = -\delta_{N,j} \frac{\lambda}{\Delta x^2} \sum_{k',m',n'} \langle \xi_m | \nu_x | \xi_{k'} \rangle \chi_{k'm'} e^{i\lambda_{m'}\Delta x} \chi^\dagger_{m'n'} B_{n'n} \tag{F.30}$$

F.3 Real Space Charge Density

We can calculate the carrier density at a location (x, y) in the channel using the injected wave function at an energy E and transverse momentum k_z given below

$$\psi_{E,k_z}(x, y, z) = \Phi_{E,k_z}(x, y) \frac{e^{ik_3 z}}{\sqrt{L_z}} \tag{F.31}$$

where Φ is expressed in mode basis as follows

$$\Phi_{E,k_z}(x, y) = \sum_n \psi_n(x, E, k_z) \xi_n(y) \tag{F.32}$$

The electron charge density is given by

$$\begin{aligned}
\rho(x, y) &= 2L_z \iint \frac{dk_z}{2\pi} dE \left| \frac{\Phi_{E,k_z}(x, y) e^{ik_3 z}}{\sqrt{L_z}} \right|^2 f(E - \mu) \\
&= 2 \iint \frac{dk_z}{2\pi} dE \sum_m \psi_m^*(x, E, k_z) \xi_m^*(y) \sum_n \psi_{n,E,k_z}(x, E, k_z) \xi_n(y) f(E - \mu) \\
&= 2 \iint \frac{dk_z}{2\pi} \frac{dE}{2\pi} \sum_{m,n} \xi_m^*(y) G_{mn}^n(x, E, k_z) \xi_n(y)
\end{aligned} \tag{F.33}$$

The factor of 2 in the above expression is for spin and G^n is the mode space electron Green's function.

Appendix G

One Dimensional Schrodinger Equation: Numerical Simulation

G.1 Time Dependent Schrödinger Equation: Numerical Simulation

The time dependent Schrödinger equation (G.1) governs the time evolution of a quantum state $|\psi(t)\rangle$ of a quantum mechanical system described by Hamiltonian operator H.

$$i\hbar\frac{\partial|\psi(t)\rangle}{\partial t} = H|\psi(t)\rangle \tag{G.1}$$

Assuming a time independent Hamiltonian, the solution of Eq. (G.1) can be formally written as

$$|\psi(t)\rangle = e^{-i\frac{H}{\hbar}(t-t_o)}|\psi(t_o)\rangle \tag{G.2}$$

where $|\psi(t_o)\rangle$ is the state of a system at time t_o in the past. If the initial state of system is an eigenstate with energy E such that $H|\psi(t_o)\rangle = E|\psi(t_o)\rangle$, then Eq. (G.1) has a closed form solution, $|\psi(t)\rangle = e^{-i\omega(t-t_o)}|\psi(t_o)\rangle$ where $\omega = E/\hbar$. Such closed form solutions are possible if the initial state is an eigenstate or a linear combination of eigenstates and the Hamiltonian is time independent. However if the initial state is not an eigenstate or the Hamiltonian is time dependent, Eq. (G.1) can only be solved numerically. For example, consider a

problem where we are interested in the response, source to drain current, of a nano-scale field effect transistor due to sudden switching of the voltage of the gate terminal. Clearly, in such a scenario the initial steady state wave function is not an eigenstate of the system. Also, the Hamiltonian is time dependent because the potential in the channel is time dependent as a response to time varying terminal voltages. The potential energy, $U(x,t) = -qV(x,t)$, in the channel at any time t can be obtained by solving the Poisson equation given below, with appropriate boundary conditions,

$$-\nabla^2 V = \rho(x,t) \tag{G.3}$$

where $\rho(x,t) = -q\left|\psi(x,t)\right|^2$ is the charge density, q is the charge of electron and the wave function $\psi(x,t) = \langle x|\ \psi(t)\rangle$ is coordinate representation of the state of the system. For practical reasons, Eq. (G.1) is usually solved in coordinate representation

$$i\hbar\frac{\partial\psi(x,t)}{\partial t} = H\psi(x,t) \tag{G.4}$$

where the Hamiltonian, $H = -\frac{\hbar^2}{2m}\nabla^2 + U(x,t)$, is sum of the kinetic energy and potential energy operators in coordinate representation, and m is the effective mass of electron. To numerically solve Eq. (G.4), the spatial operator is approximated with central difference scheme. As an illustrative example, consider a one dimensional lattice which is divided in into left, central and right zones as shown in Figure G.1. The points on the lattice are equidistant from each other with lattice spacing, $a = \Delta x$. Let the wave function at lattice point $x = j\Delta x, j = (\ldots, -1, 0, 1, \ldots)$ and at time t be denoted by

$\psi_j(t) = \psi(j\Delta x, t)$. Then with central difference approximation Eq. (G.4) can be rewritten as a set of ordinary differential equations,

$$i\hbar \frac{d\psi_j(t)}{dt} = -\frac{\hbar^2}{2m}\left(\frac{\psi_{j-1}(t) - 2\psi_j(t) + \psi_{j+1}(t)}{\Delta x^2}\right) + U_j(t)\psi_j(t) \tag{G.5}$$

where $j = (\ldots, -1, 0, 1, \ldots)$. Unless the spatial domain is bounded with closed boundary conditions, one must solve an infinitely large number of coupled ordinary differential equations which is a computationally prohibitive task. In other words, it can be seen from Eq. (G.5) that time evolution of the wave function at a given lattice point depends on the wave function value at its neighboring points as well. One can restrict Eq. (G.5) to a finite spatial domain, only if the wave function is specified at the boundaries via a Dirichlet or a Neumann boundary condition. For example, in the example transistor problem posed earlier one is usually interested only in time evolution of the wave function in the channel region between source and drain. In this scenario, an electron injected from source travels thorough the channel and finally exits thorough the drain. Systems in which particles can enter and exit are called open systems. To accurately model the physical behavior of such open systems, time evolution of the electron wave function must be followed in the exterior regions as well i.e., in the semi-infinite drain and source domains in the transistor example. However, to make the problem computationally feasible, the domain must be restricted to the channel region with boundary conditions such that time evolution in the channel is similar to the solution

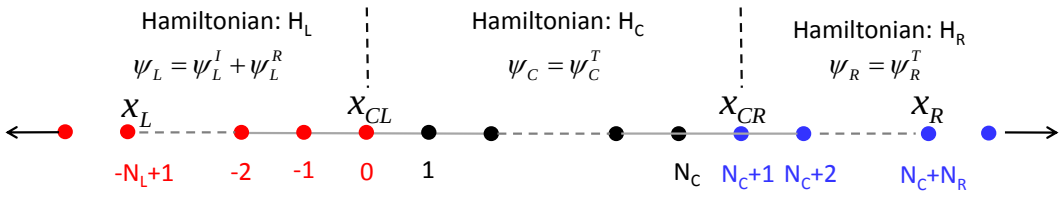

Figure G.1: A one dimensional lattice partitioned into left, central and right physical domains and showing the indexed lattice sites.

obtained if one were to solve the infinite domain problem. Clearly, one cannot use Dirichlet boundary conditions as it results in complete reflection of the wave function from the source and drain in the channel which is unphysical. The imposed boundary conditions should minimize or completely get rid of spurious reflections from the source and drain contacts into the channel region. Such boundary conditions are called open or transparent boundary conditions. The coupled set of equations (G.5) can be written in the matrix form

$$i\hbar\frac{d}{dt}\begin{bmatrix}\vdots\\ \psi_{j-1}\\ \psi_j\\ \psi_{j+1}\\ \vdots\end{bmatrix} = \begin{bmatrix}\ddots & \ddots & & & \\ \beta^\dagger & \alpha_{j-1} & \beta & & \\ & \beta^\dagger & \alpha_j & \beta & \\ & & \beta^\dagger & \alpha_{j+1} & \beta \\ & & & \ddots & \ddots\end{bmatrix}\begin{bmatrix}\vdots\\ \psi_{j-1}\\ \psi_j\\ \psi_{j+1}\\ \vdots\end{bmatrix} \tag{G.6}$$

where $\alpha_j = U_j - 2t_o$,$\beta = t_o$ and $t_o = -\frac{\hbar^2}{2m\Delta x^2}$. Equation (G.6) can be partitioned according to the physical domains shown in Figure G.1 to rewrite in the form,

$$i\hbar\frac{d}{dt}\begin{bmatrix}\psi_L\\ \psi_C\\ \psi_R\end{bmatrix} = \begin{bmatrix}H_L & H_{LC} & 0\\ H_{CL} & H_C & H_{CR}\\ 0 & H_{RC} & H_R\end{bmatrix}\begin{bmatrix}\psi_L\\ \psi_C\\ \psi_R\end{bmatrix} \tag{G.7}$$

where $\psi_{L/C/R}$ and $H_{L/C/R}$ are the time dependent total wave function and the Hamiltonian in left, center and right regions respectively. The coupling

between the left and center regions is described by the term $H_{LC} = H_{CL}^{\dagger}$. The coupling between the right and center regions is described by the term $H_{RC} = H_{CR}^{\dagger}$. Note the in the above partition, the Hamiltonian in the central region is of finite size N_C, whereas left and right Hamiltonians are of infinite size.

Having obtained the Hamiltonian after spatial discretization we must discretize time in Eq. (G.7) to get a numerically solvable from. Because the time evolution operator must be unitary for a Hermitian Hamiltonian, the discretized equation must preserve this property. Assuming the Hamiltonian varies negligibly slowly for a small time Δt, one can approximate the time evolution operator

$$U\left(t + \Delta t, t\right) = e^{-iH(t)\Delta t} \tag{G.8}$$

where,e^A the exponential of the operator A ,is formally defined by Taylor series expansion

$$e^A = \sum_{n=0}^{\infty} \frac{A^n}{n!} \tag{G.9}$$

In literature (for example, see [73] and references herein) there are various procedures for approximating the exponential of an operator. For example, by truncating to certain number of terms of the trivial Taylor series expansion or Chebychey polynomial based expansion, Krylov subspace projection based methods such as Short Iterative Lancoz method or Split operator method coupled with FFT. Then there are methods based on numerical approximation to the differential operator such as the Crank-Nicolson(CN) method. the CN method is a simple to implement procedure which gives a unitary and time

reversible evolution. Assuming a time step of Δt, let us denote the wave function at n^{th} time step as $\psi^n = \psi(n\Delta t)$. To obtain the CN propagator we use a implicit midpoint rule for integration,

$$i\hbar\frac{\psi(t+\Delta t)-\psi(t)}{\Delta t} = H\left(t+\frac{\Delta t}{2}\right)\left(\frac{\psi(t+\Delta t)+\psi(t)}{2}\right) \tag{G.10}$$

The above equation can be rearranged to get the CN propagator,

$$\left(I + i\Delta t H^{n+1/2}\right)\psi^{n+1} = \left(I - i\Delta t H^{n+1/2}\right)\psi^n \tag{G.11}$$

where $H^{n+1/2} = H(n\Delta t + \Delta t/2)$. As can be seen from Eq. (G.11), one must solve a linear equation of the form $Ax = b$ at every time step, which can be done using direct or iterative solvers.

G.2 Absorbing Boundary Conditions

As we have seen in the previous section, for open systems, one must solve an infinitely large system to accurately model the physics. The general structure of a Schröedinger equation for a open system with two terminals is given by Eq. (G.6) which can be rearranged as

$$\begin{aligned} i\hbar\frac{d}{dt}\psi_C = H_C\psi_C + \int_0^t dt'\Sigma^R(t,t')\,\psi_C(t') \\ + i\sum_{\alpha=L,R} H_{C\alpha}g_\alpha^R(t,0)\,\psi_\alpha(0) \end{aligned} \tag{G.12}$$

where $g_\alpha^R, \alpha = L, R$ is the retarded greens function of the contacts and $\Sigma^R = \Sigma_{\alpha=L,R} H_{C\alpha} g_\alpha^R H_{\alpha C}$. Equation (G.12) gives an exact solution of the

open system. However, it is computationally expensive to solve due to the presence of second and third terms that are nonlocal in time and space. There are procedures one can use to approximate the convolution terms that appear when using otherwise exact boundary conditions such as above. One can also introduce artificial boundary conditions which can approximately satisfy the requirement of no unphysical reflections at the artificial boundaries created by domain truncation to region of interest. Examples of such approaches include procedures such as extrapolation of the wave function into the exterior domain [71], phase matched layer (PML) boundary conditions, use of complex imaginary potentials (self-energies). For the present work we use complex absorbing potential as absorbing boundary conditions.

G.2.1 Complex Absorbing Potential

The complex absorbing potential (CAP) method [72] is a simple method which involves introduction of an artificial negative imaginary potential profile into the computational domain. The principle behind this method is that the probability is not conserved for a non Hermitian Hamiltonian. For example, for a free particle, a constant negative (positive) imaginary potential results in decay (growth) of probability density. One advantage of using the CAP approach to as artificial boundary conditions is that the method does not have restrictions based on the shape of the computational domain. However, one must extend the computation domain beyond the region of interest which will increase the computational burden, although perhaps less so than the

nominally exact methods. The potential profile in the extended region must be smooth enough to minimize the reflections. Since the extended region is closed, we use Dirichlet boundary conditions at the external ends which can result in reflections. The length of the region should be chosen such that the wave function decays significantly during a round trip in the extended region before it reenters the region of interest.

As an illustration, we consider the 1D problem shown in Figure G.1 and truncate the domain to N_L points in the left region and N_R points in the right region. Time evolution of the wave packet on this lattice is now described by Eq. (G.6). However the size of Hamiltonian $H_{L/R}$ is now $N_{L/R}$. Effectively, now one has to solve a closed system of size $N = N_L + N_C + N_L$ instead of an open system with size N_C. Also, we add the artificial CAP

$$V = \begin{cases} -iV_L \frac{f(x_{CL}-x)}{f(x_{CL}-x_L)} & x_L < x \leq x_{CL} \\ 0 & x_{CL} < x < x_{CR} \\ -iV_R \frac{f(x-x_{CR})}{f(x_R-x_{CR})} & x_{CR} \geq x < x_R \end{cases} \tag{G.13}$$

to the Hamiltonian, where $V_{L/R}$ is the strength of the absorbing potential in left and right regions respectively. The coordinates of the boundaries, x_L, x_{CL}, x_{CR}, and x_R are as shown in Figure G.1. The profile function $f(x)$ can be chosen to be smoothly varying function such as x^n or $(1 + e^{-\alpha x})^{-1}$, referred to as Saxon-Woods potential.

G.3 Injecting Boundary Condition

If a wave packet is incident from the left region then the total wave function,$\psi_L = \psi_L^I + \psi_L^R$, in the left region is given by sum of the incident wave ψ_L^I and reflected wave ψ_L^R. The total wave function in the right region is the transmitted wave ψ_R^T. Since the incident wave function in the left region is known for all times, Eq. (G.7) must be solved only for the unknown reflected and transmitted waves. Correcting for the known wave function, Eq. (G.7) can be rewritten as,

$$i\hbar\frac{d}{dt}\begin{bmatrix}\psi_L^R\\ \psi_C\\ \psi_R\end{bmatrix} = \begin{bmatrix}H_L & H_{LC} & 0\\ H_{CL} & H_C & H_{CR}\\ 0 & H_{RC} & H_R\end{bmatrix}\begin{bmatrix}\psi_L^R\\ \psi_C\\ \psi_R\end{bmatrix} + \psi_S \tag{G.14}$$

The source term, ψ_S, on right hand side of the above equation is given by,

$$\psi_S = \begin{bmatrix}H_L\psi_L^I - i\hbar\frac{d\psi_L^I}{dt}\\ H_{CL}\psi_L^I\\ 0\end{bmatrix} \tag{G.15}$$

Now, e.g., let the injected wave be a m^{th} transverse eigenmode with longitudinal momentum k given by

$$\psi_L^I(x,t) = \varphi_{mk}e^{ik\cdot x - i\omega_{mk}t} \tag{G.16}$$

where the transverse mode φ_{mk} satisfies the eigenvalue equation

$$\left(\beta^\dagger e^{-ika} + \alpha + \beta e^{ika}\right)\varphi_{mk} = E_{mk}\varphi_{mk} \tag{G.17}$$

and $\omega_{mk} = E_{mk}/\hbar$. Then the only non zero terms in the source term are

$$\psi_S(j=0,t) = -\beta e^{ika}\psi_L^I(x_{CL},t) \tag{G.18a}$$

$$\psi_S(j=1,t) = \beta^\dagger\psi_L^I(x_{CL},t) \tag{G.18b}$$

where x_{CL} is the coordinate of the $j = 0$ node in Figure G.1 and a is the lattice constant. To avoid significant oscillations in the wave function density due to sudden introduction of the source, the source term in Eq. (G.14) is introduced adiabatically by pre-multiplying the amplitude in Eq. (G.18) with $g(t) = 1 - e^{-t/t_o}$, where t_o is a chosen time constant to minimize oscillations.

Using a imaginary quadratic potential profile given by Eq. (G.13) we solve for time evolution of an initial Gaussian wave packet on a finite one dimensional lattice shown in Figure G.1. Figure G.2(a) shows the snap shots of the wave packet at three different times illustrating the smooth exit of the wave packet from the (central) region of interest. In the absence of complex absorbing potential in the left and right regions, we can see the reflected Gaussian wave entering the central region in the snapshot at 30.5 fs shown in Figure G.2(b). The total wave function density in the central region is same for the simulations with and without complex absorbing potential as shown in Figure G.2(c). However, as it can be seen from Figure G.2(c) that the total wave function density increase after some time as the reflected Gaussian wave travels back into the central region. Also, note that due to CN procedure used for time discretization the total wave function density is conserved during the initial time until the wave function starts leaving the central region when it starts decreasing. Figure G.2(d) shows the time evolution of the wave function density in the simulation region due to an adiabatically introduced source term.

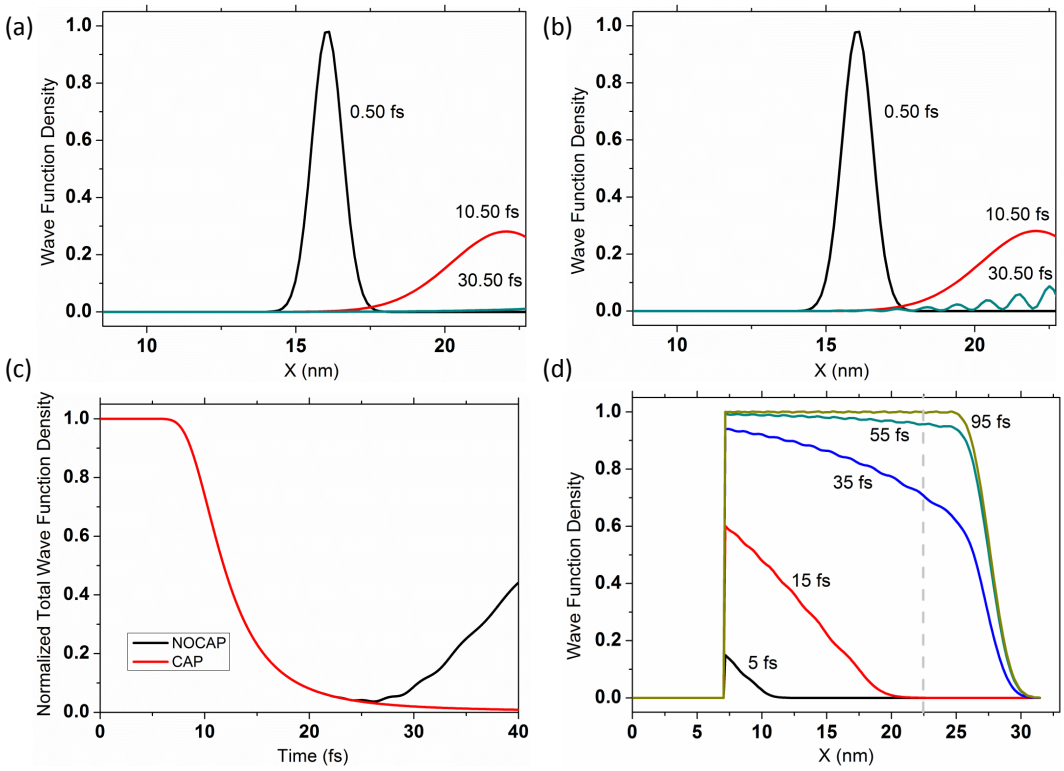

Figure G.2: Time evolved wave function density of an initially Gaussian wave packet in the central region of the 1D lattice at 0.5, 10.5 and 30.5 fs with (a) complex absorbing potential and (b) no complex absorbing potential in the in the left and right extended regions. (c) Normalized total wave function density in the central region vs time for CAP (red solid) and noCAP (black solid). (d) Snapshots of wave function density in the simulations region with source term injecting a plane wave at 5, 15, 35, 55 and 95 fs. Note that the amplitude of the source term adiabatically increase to 1.

Bibliography

[1] D. Reddy, L. F. Register, G. D. Carpenter, and S. K. Banerjee, "Graphene field-effect transistors," *Journal of Physics D: Applied Physics*, vol. 44, p. 313001, Aug. 2011.

[2] E. McCann, D. S. L. Abergel, and V. I. Falko, "Electrons in bilayer graphene," *Solid State Communications*, vol. 143, no. 1-2, p. 110115, 2007, 0038-1098 doi: DOI: 10.1016/j.ssc.2007.03.054.

[3] Y. Zhang, T.-T. Tang, C. Girit, Z. Hao, M. C. Martin, A. Zettl, M. F. Crommie, Y. R. Shen, and F. Wang, "Direct observation of a widely tunable bandgap in bilayer graphene," *Nature*, vol. 459, no. 7248, p. 820823, 2009, 10.1038/nature08105.

[4] A. H. C. Neto, F. Guinea, N. M. R. Peres, K. S. Novoselov, and A. K. Geim, "The electronic properties of graphene," *Reviews of Modern Physics*, vol. 81, no. 1, p. 10954, 2009.

[5] M. Y. Han, B. zyilmaz, Y. Zhang, and P. Kim, "Energy band-gap engineering of graphene nanoribbons," *Physical Review Letters*, vol. 98, no. 20, p. 206805, 2007, copyright (C) 2010 The American Physical Society Please report any problems to prola@aps.org PRL.

[6] D. Basu, M. J. Gilbert, L. F. Register, S. K. Banerjee, and A. H. MacDonald, "Effect of edge roughness on electronic transport in graphene nanoribbon channel metal-oxide-semiconductor field-effect transistors," *Appl. Phys. Lett.*, vol. 92, p. 042114, 2008, 10.1063/1.2839330.

[7] D. Basu, L. Register, D. Reddy, A. MacDonald, and S. Banerjee, "Tight-binding study of electron-hole pair condensation in graphene bilayers: Gate control and system-parameter dependence," *Physical Review B*, vol. 82, no. 7, Aug 2010. [Online]. Available: http://link.aps.org/doi/10.1103/PhysRevB.82.075409

[8] X. Mou, L. F. Register, and S. K. Banerjee, "Quantum transport simulation of bilayer pseudospin field-effect transistor (bisfet) on tightbinding hartree-fock model," in *Simulation of Semiconductor Processes and Devices, International Conference on*, Glasgow, Scotland, 2013, accepted.

[9] D. Basu, L. Register, A. MacDonald, and S. Banerjee, "Effect of interlayer bare tunneling on electron-hole coherence in graphene bilayers," *Physical Review B*, vol. 84, no. 3, Jul 2011. [Online]. Available: http://link.aps.org/doi/10.1103/PhysRevB.84.035449

[10] I. Sodemann, D. Pesin, and A. MacDonald, "Interaction-enhanced coherence between two-dimensional dirac layers," *Physical Review B*, vol. 85, no. 19, May 2012. [Online]. Available: http://link.aps.org/doi/10.1103/PhysRevB.85.195136

[11] X. Mou, L. F. Register, and S. K. Banerjee, "unpublished."

[12] D. Reddy, L. F. Register, E. Tutuc, and S. K. Banerjee, "Bilayer pseudospin Field-Effect transistor: Applications to boolean logic," *IEEE Transactions on Electron Devices*, vol. 57, no. 4, pp. 755–764, 2010.

[13] D. Reddy, L. F. Register, and S. K. Banerjee, *Bilayer graphene vertical tunneling field effect transistor*. IEEE, Jun 2012, p. 7374. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6256932

[14] D. Reddy, P. Jadaun, A. Valsaraj, L. F. Register, and S. K. Banerjee, "Time dependent quantum transport in graphene," in *Simulation of Semiconductor Processes and Devices, International Conference on*, Denver, USA, 2012, pp. 51–54.

[15] R. Dennard, F. Gaensslen, V. Rideout, E. Bassous, and A. LeBlanc, "Design of ion-implanted MOSFET's with very small physical dimensions," *IEEE Journal of Solid-State Circuits*, vol. 9, pp. 256–268, Oct. 1974.

[16] (2012) International technology roadmap for semiconductors. [Online]. Available: http://www.itrs.net

[17] G. Bourianoff, M. Brillouet, R. K. Cavin, T. Hiramoto, J. A. Hutchby, A. M. Ionescu, and K. Uchida, "Nanoelectronics research for beyond CMOS information processing," *Proceedings of the IEEE*, vol. 98, no. 12, pp. 1986–1992, 2010.

[18] S. K. Banerjee, L. F. Register, E. Tutuc, D. Reddy, and A. H. MacDonald, "Bilayer PseudoSpin Field-Effect transistor (BiSFET): a proposed new logic device," *Electron Device Letters, IEEE*, vol. 30, no. 2, pp. 158–160, 2009.

[19] J. Simmons, M. Blount, J. Moon, W. Baca, J. Reno, and M. Hafich, "Unipolar complementary bistable memories using gate-controlled negative differential resistance in a 2d-2d quantum tunneling transistor." IEEE, Dec 1997, p. 755758. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=650492

[20] K. S. Novoselov, A. K. Geim, S. V. Morozov, D. Jiang, Y. Zhang, S. V. Dubonos, I. V. Grigorieva, and A. A. Firsov, "Electric field effect in atomically thin carbon films," *Science*, vol. 306, no. 5696, p. 666669, 2004.

[21] K. S. Novoselov, A. K. Geim, S. V. Morozov, D. Jiang, M. I. Katsnelson, I. V. Grigorieva, S. V. Dubonos, and A. A. Firsov, "Two-dimensional gas of massless dirac fermions in graphene," *Nature*, vol. 438, no. 7065, p. 197200, 2005, 10.1038/nature04233.

[22] Y. Zhang, Y.-W. Tan, H. L. Stormer, and P. Kim, "Experimental observation of the quantum hall effect and berrys phase in graphene," *Nature*, vol. 438, no. 7065, p. 201204, 2005, 10.1038/nature04235.

[23] M. I. Katsnelson, K. S. Novoselov, and A. K. Geim, "Chiral tunnelling and the klein paradox in graphene," *Nat Phys*, vol. 2, no. 9, pp. 620–625, 2006.

[24] H. Min, R. Bistritzer, J.-J. Su, and A. H. MacDonald, "Room-temperature superfluidity in graphene bilayers," *Physical Review B*, vol. 78, no. 12, p. 121401, 2008, copyright (C) 2010 The American Physical Society Please report any problems to prola@aps.org PRB.

[25] S. Y. Zhou, G. H. Gweon, J. Graf, A. V. Fedorov, C. D. Spataru, R. D. Diehl, Y. Kopelevich, D. H. Lee, S. G. Louie, and A. Lanzara, "First direct observation of dirac fermions in graphite," *Nat Phys*, vol. 2, no. 9, pp. 595–599, 2006.

[26] I. Meric, N. Baklitskaya, P. Kim, K. L. Shepard, and Ieee, "RF performance of top-gated, zero-bandgap graphene field-effect transistors," in *IEEE International Electron Devices Meeting 2008, Technical Digest*, ser. International Electron Devices Meeting, San Francisco, CA, 2008, pp. 513–516.

[27] Y. M. Lin, K. A. Jenkins, A. Valdes-Garcia, J. P. Small, D. B. Farmer, and P. Avouris, "Operation of graphene transistors at gigahertz frequencies," *Nano Letters*, vol. 9, no. 1, pp. 422–426, 2009.

[28] Y. M. Lin, C. Dimitrakopoulos, K. A. Jenkins, D. B. Farmer, H. Y. Chiu, A. Grill, and P. Avouris, "100-GHz transistors from Wafer-Scale epitaxial graphene," *Science*, vol. 327, no. 5966, p. 662, 2010.

[29] Y.-W. Son, M. L. Cohen, and S. G. Louie, "Half-metallic graphene nanoribbons," *Nature*, vol. 444, no. 7117, p. 347349, 2006, 10.1038/nature05180.

[30] N. Tombros, C. Jozsa, M. Popinciuc, H. T. Jonkman, and B. J. van Wees, "Electronic spin transport and spin precession in single graphene layers at room temperature," *Nature*, vol. 448, no. 7153, p. 571574, 2007, 10.1038/nature06037.

[31] J. Moser, A. Barreiro, and A. Bachtold, "Current-induced cleaning of graphene," *Appl. Phys. Lett.*, vol. 91, p. 163513, 2007, 10.1063/1.2789673.

[32] R. Murali, Y. Yang, K. Brenner, T. Beck, and J. D. Meindl, "Breakdown current density of graphene nanoribbons," *Applied Physics Letters*, vol. 94, no. 24, p. 2431143, 2009.

[33] N. M. R. Peres, "Graphene, new physics in two dimensions," *Europhysics News*, vol. 40, no. 3, p. 1720, 2009.

[34] P. R. Wallace, "The band theory of graphite," *Physical Review*, vol. 71, no. 9, p. 622, 1947, copyright (C) 2011 The American Physical Society Please report any problems to prola@aps.org PR.

[35] K. Nakada, M. Fujita, G. Dresselhaus, and M. S. Dresselhaus, "Edge state in graphene ribbons: Nanometer size effect and edge shape dependence," *Physical Review B*, vol. 54, no. 24, p. 17954, 1996, copyright (C) 2011 The American Physical Society Please report any problems to prola@aps.org PRB.

[36] K. Wakabayashi, M. Fujita, H. Ajiki, and M. Sigrist, "Electronic and magnetic properties of nanographite ribbons," *Physical Review B*, vol. 59, no. 12, p. 8271, 1999, copyright (C) 2011 The American Physical Society Please report any problems to prola@aps.org PRB.

[37] Y. Yoon and J. Guo, "Effect of edge roughness in graphene nanoribbon transistors," *Applied Physics Letters*, vol. 91, no. 7, 2007, yoon, Youngki Guo, Jing.

[38] H. Wang, D. Nezich, J. Kong, and T. Palacios, "Graphene frequency multipliers," *IEEE Electron Device Letters*, vol. 30, no. 5, pp. 547–549, 2009.

[39] L. Liao, Y. Lin, M. Bao, R. Cheng, J. Bai, Y. Liu, Y. Qu, K. L. Wang, Y. Huang, and X. Duan, "High-speed graphene transistors with a self-aligned nanowire gate," *Nature*, vol. 467, pp. 305–308, Sep. 2010.

[40] Y. M. Lin, A. Valdes-Garcia, S. Han, D. B. Farmer, I. Meric, Y. Sun, Y. Wu, C. Dimitrakopoulos, A. Grill, P. Avouris, and K. A. Jenkins, "Wafer-Scale graphene integrated circuit," *Science*, vol. 332, no. 6035, pp. 1294–1297, Jun. 2011.

[41] Y. Lin, D. B. Farmer, K. A. Jenkins, Y. Wu, J. L. Tedesco, R. L. Myers-Ward, J. C. R. Eddy, D. K. Gaskill, C. Dimitrakopoulos, and P. Avouris, "Enhanced performance in epitaxial graphene with optimized channel morphology," *IEEE Electron Device Letters*, 2011(accepted).

[42] H. Pao and C. Sah, "Effects of diffusion current on characteristics of metal-oxide (insulator)-semiconductor transistors," *Solid-State Electronics*, vol. 9, no. 10, pp. 927–937, Oct. 1966.

[43] I. Meric, "Current saturation in zero-bandgap, top-gated graphene field-effect transistors," *Nature Nanotech.*, vol. 3, pp. 654–659, 2008, 10.1038/nnano.2008.268.

[44] S. J. Han, A. Valdes-Garcia, A. A. Bol, A. D. Franklin, D. B. Farmer, E. Kratschmer, K. A. Jenkins, and W. Haensch, "Graphene technology with inverted-t gate and rf passives on 200mm platform," *IEDM*, 2011(to appear).

[45] S. Han, Z. Chen, A. A. Bol, and Y. Sun, "Channel-Length-Dependent transport behaviors of graphene Field-Effect transistors," *IEEE Electron Device Letters*, vol. 32, no. 6, pp. 812–814, Jun. 2011.

[46] A. Konar, T. Fang, and D. Jena, "Effect of high-κ gate dielectrics on charge transport in graphene-based field effect transistors," *Physical Review B*, vol. 82, Sep. 2010.

[47] "Virtuoso spectre circuit simulator user guide," cadence Design Systems, Inc., San Jose, CA. Product Version 6.1.3.

[48] C. H. Zhang and Y. N. Joglekar, "Excitonic condensation of massless fermions in graphene bilayers," *Physical Review B*, vol. 77, no. 23, p.

233405, 2008, copyright (C) 2010 The American Physical Society Please report any problems to prola@aps.org PRB.

[49] Y. E. Lozovik and V. I. Yudson, "Feasibility of superfluidity of paired spatially separated electrons and holes; a new superconductivity mechanism," *Soviet Journal of Experimental and Theoretical Physics Letters*, vol. 22, p. 274275, Dec 1975.

[50] E. E. Mendez, L. Esaki, and L. L. Chang, "Quantum hall effect in a two-dimensional electron-hole gas," *Physical Review Letters*, vol. 55, no. 20, p. 2216, 1985, copyright (C) 2010 The American Physical Society Please report any problems to prola@aps.org PRL.

[51] U. Sivan, P. M. Solomon, and H. Shtrikman, "Coupled electron-hole transport," *Physical Review Letters*, vol. 68, no. 8, p. 1196, 1992, copyright (C) 2010 The American Physical Society Please report any problems to prola@aps.org PRL.

[52] M. Pohlt, M. Lynass, J. G. S. Lok, W. Dietsche, K. v. Klitzing, K. Eberl, and R. Muhle, "Closely spaced and separately contacted two-dimensional electron and hole gases by in situ focused-ion implantation," *Applied Physics Letters*, vol. 80, no. 12, p. 21052107, 2002.

[53] J. A. Seamons, D. R. Tibbetts, J. L. Reno, and M. P. Lilly, "Undoped electron-hole bilayers in a gaas/algaas double quantum well," *Applied Physics Letters*, vol. 90, no. 5, p. 0521033, 2007.

[54] P. Avouris, Z. Chen, and V. Perebeinos, "Carbon-based electronics," *Nature Nanotechnology*, vol. 2, no. 10, p. 605615, Sep 2007.

[55] I. B. Spielman, J. P. Eisenstein, L. N. Pfeiffer, and K. W. West, "Resonantly enhanced tunneling in a double layer quantum hall ferromagnet," *Physical Review Letters*, vol. 84, no. 25, p. 5808, 2000, copyright (C) 2010 The American Physical Society Please report any problems to prola@aps.org PRL.

[56] D. Nandi, A. D. K. Finck, J. P. Eisenstein, L. N. Pfeiffer, and K. W. West, "Exciton condensation and perfect coulomb drag," *Nature*, vol. 488, no. 7412, p. 481484, Aug 2012.

[57] J.-J. Su and A. H. MacDonald, "How to make a bilayer exciton condensate flow," *Nat Phys*, vol. 4, no. 10, p. 799802, 2008, 1745-2473 10.1038/nphys1055 10.1038/nphys1055.

[58] L. Tiemann and e. al, "Critical tunneling currents in the regime of bilayer excitons," *New Journal of Physics*, vol. 10, no. 4, p. 045018, 2008.

[59] D. Basu, "Quantum transport and bulk calculations for graphene-based devices," Ph.D. dissertation, The Univeristy of Texas at Austin, 2010.

[60] Y. E. Lozovik and A. A. Sokolik, "Electron-hole pair condensation in a graphene bilayer," *JETP Letters*, vol. 87, no. 1, p. 5559, Apr 2011.

[61] G. Liu, J. Velasco, W. Bao, and C. N. Lau, "Fabrication of graphene p-n-p junctions with contactless top gates," *Applied Physics Letters*, vol. 92, no. 20, p. 203103, 2008.

[62] J. Park and C. Hu, "Air-spacer mosfet with self-aligned contact for future dense memories," *IEEE Electron Device Letters*, vol. 30, no. 12, p. 13681370, Dec 2009.

[63] K. Wu, A. Sachid, F.-L. Yang, and C. Hu, "Toward 44% switching energy reduction for finfets with vacuum gate spacer," in *Simulation of Semiconductor Processes and Devices, International Conference on*, Denver, Colrado, 2012, pp. 253–256.

[64] P. Mazumder, S. Kulkarni, M. Bhattacharya, S. Jian Ping, and G. I. Haddad, "Digital circuit applications of resonant tunneling devices," *Proceedings of the Ieee*, vol. 86, no. 4, p. 664686, 1998.

[65] L. Britnell, R. V. Gorbachev, A. K. Geim, L. A. Ponomarenko, A. Mishchenko, M. T. Greenaway, T. M. Fromhold, K. S. Novoselov, and L. Eaves, "Resonant tunnelling and negative differential conductance in graphene transistors," *Nature Communications*, vol. 4, p. 1794, Apr 2013.

[66] L. Zheng and A. MacDonald, "Tunneling conductance between parallel two-dimensional electron systems," *Physical Review B*, vol. 47, no. 16, p. 1061910624, Apr 1993.

[67] N. Turner, J. T. Nicholls, E. H. Linfield, K. M. Brown, G. A. C. Jones, and D. A. Ritchie, "Tunneling between parallel two-dimensional electron gases," *Physical Review B*, vol. 54, p. 10614, 1996, 15.

[68] J. Wang, E. Polizzi, and M. Lundstrom, "A three-dimensional quantum simulation of silicon nanowire transistors with the effective-mass approximation," *Journal of Applied Physics*, vol. 96, no. 4, p. 2192, 2004.

[69] J. Crank and P. Nicolson, "A practical method for numerical evaluation of solutions of partial differential equations of the heat-conduction type," *Advances in Computational Mathematics*, vol. 6, no. 1, p. 207226, Dec 1996.

[70] G. Maksimova, V. Demikhovskii, and E. Frolova, "Wave packet dynamics in a monolayer graphene," *Physical Review B*, vol. 78, no. 23, Dec 2008. [Online]. Available: http://link.aps.org/doi/10.1103/PhysRevB.78.235321

[71] L. F. Register, U. Ravaioli, and K. Hess, "Numerical simulation of mesoscopic systems with open boundaries using the multidimensional time-dependent schr[o-umlaut]dinger equation," *Journal of Applied Physics*, vol. 69, no. 10, pp. 7153–7158, 1991.

[72] C. Leforestier and R. E. Wyatt, "Optical potential for laser induced dissociation," *The Journal of Chemical Physics*, vol. 78, no. 5, p. 2334, 1983.

[73] A. Castro, M. A. L. Marques, and A. Rubio, "Propagators for the time-dependent KohnSham equations," *The Journal of Chemical Physics*, vol. 121, no. 8, p. 3425, 2004.

Vita

Dharmendar Reddy Palle received his B.Tech degree in Electrical Engineering from the Indian Institute of Technology, Kanpur and his M.S.E and Ph.D in Solid State Electronics from The University of Texas in Austin in 2006,2008 and 2013 respectively. He is currently an Engineer in the Advanced Logic Lab at Samsung Semiconductor R & D Center in Austin since July 2013. He is one of the recipient of the Ben Streetman prize for outstanding research in Electronic and Photonic Materials and Devices at the University of Texas at Austin in 2010. He also received IBM Ph.D scholarship in 2010 and NRI inventor recognition award in 2009. His current research interests include alternative switching methods and/or state variables for beyond CMOS, quantum transport modeling and device design for post 5 nm technology nodes.

Permanent address: E-mail: dharmareddy84@gmail.com

This dissertation was typeset with LaTeX[†] by the author.

[†]LaTeX is a document preparation system developed by Leslie Lamport as a special version of Donald Knuth's TeX Program.