**Graphene Devices for Beyond-CMOS Heterogeneous Integration**

Submitted in partial fulfillment of the requirements for

the degree of

Doctor of Philosophy

in

Electrical and Computer Engineering

Mohamed Darwish

B.S., Electronics and Communications Engineering, Ain Shams University

M.S., Electronics and Communications Engineering, Ain Shams University

Carnegie Mellon University

Pittsburgh, PA

September 2017

© Mohamed Darwish, 2017

All rights reserved

# Acknowledgements

# Abstract

Semiconductor manufacturing is the workhorse for a wide range of industries. It lies at the heart of consumer electronics, telecommunication equipment and medical devices. Most semiconductor electronics are made from Silicon, and are fabricated using CMOS technology. The versatility of semiconductor electronics stems from the ever-reducing cost of integrating more computing and memory functions on chip. The small cost for adding extra functions has been maintained in the past 50 years through transistor scaling. Transistor scaling focuses on shrinking the size of transistors integrated on chip. This reduction in transistor size, while keeping the overall cost of the chip fixed allowed us to reduce the cost per function with scaling, and is what is celebrated as Moore’s law. Scaling has been working gracefully up to the last decade, where the exponential rise in manufacturing cost and diminishing gains of scaling on device performance reduce its economic benefit. To revive the cost reduction trend, different techniques were proposed such as augmenting CMOS manufacturing with new materials (Beyond CMOS), 3D integration, and integrating more non-transistor elements on-chip (More than Moore).

In this work, we focus on the efficient implementation of several circuit functions using an allotropy of carbon known as graphene. Graphene, a single layer of carbon atoms arranged in a hexagonal lattice, has unique electronic properties that has been taken the solid-state electronics community by a storm since its first experimental conception in 2004. Despite its promising electronic properties, namely the very high charge-carrier mobility and reduced scattering by impurities, graphene circuits has been held back by a plethora of nonidealities and technological roadblocks that hamper its use in traditional transistor-based circuits. In this work, we attempt to leverage the unique physical properties of graphene to implement non von-Neumann neuromorphic computing architectures, low-loss diodes and evaluate the behavior of diffusive-transport graphene couplers. We focus on the the design, fabrication and characterization of graphene devices in the presence of the current performance-limiting technological nonidealities in heterogeneous graphene-CMOS systems. We present the design, fabrication and characterization of all-graphene resistive data converters devices and diodes, discussing their performance and application as building elements of all-graphene brain-inspired computing architectures. We evaluate the performance of graphene couplers operating in the diffusive transport regime, which serve as a method to analyze the cross-coupling between adjacent graphene interconnects. We also discuss the current technological limitations hampering the performance of graphene devices, and the roles of different processing non-idealities on the characteristics of graphene devices.

# Table of Contents

# List of Tables

# List of Figures

# Introduction

The Semiconductor Industry has evolved over the d

## Economics of Scaling and Moore’s Law

## Beyond-CMOS Materials and Devices

## Graphene as a Beyond-CMOS Material

## Overview of Document

# The Electronic Properties of Graphene

In this chapter, we discuss the electronic properties of graphene integral to this work. We start with a description of graphene’s band structure, elaborating on the its basic features. We then discuss the Klein tunneling in graphene, and how it differs from normal quantum mechanical tunneling and how it affects device performance. We then turn our discussion into experimental aspects that affect graphene’s performance, discussing how the presence of a substrate affects graphene. We then discuss the nature of charge transport in CVD graphene, followed by a discussion of how metal contacts induce graphene in their vicinity.

We then turn into the characterization methods of graphene, discussing the Raman analysis of graphene, which allows us to identify the number of layers and infer the information about the quality of the material. We then conclude by a description of the electrical model used to characterize and describe the devices presented in this work.

## Electronic Band Structure of Graphene

Graphene’s atoms lie in a hexagonal lattice with the atoms placed 1.42 Å[1], [2]. The hexagonal lattice corresponds to a two-atom unit cell in a triangular lattice as shown in Figure ‎2.1(a). The corresponding Brillouin zone is hexagonal, but the vertices of the hexagon are grouped into two distinct groups, K and K’. Each group contains the points that can be connected by integral multiples of the reciprocal lattice basis as shown in Figure ‎2.1(b). The high symmetry points of the Brillouin zone are the center (Γ), vertices (K and K’), and midpoint between the vertices (M)[2].

|  |  |
| --- | --- |
| (a) | (b) |

Figure . (a) Hexagonal placement of Carbon atoms in graphene can be analyzed as two interleaving triangular sublattices or a triangular lattice with a two-atom unit cell. The real spaces basis vectors are a1 and a2. (b) The Brillouin zone of graphene in reciprocal space is hexagonal. The alternate vertices of the hexagon belong to two different groups, each containing the vertices that are reachable from one another by integral multiples of the reciprocal lattice bases b1 and b2.

|  |  |
| --- | --- |
| (a) | (b) |

Figure . (a) The conduction and valence band touch at the vertices of the Brillouin zone. The points of contact have zero energy and are known as the Dirac points. The conduction and valence bands are symmetry around the zero energy plane. (b) Contour plot of the conduction band energy dispersion showing its symmetry. The zero energy points are shown in dark blue.

The electronic band structure of graphene shows the conduction and valence bands touch at the vertices of the Brillouin zone. These are the points of zero energy and are known as the Dirac points. Near the Dirac point, the energy is proportional to the magnitude of momentum, similar to massless Dirac fermions[2], and the dispersion relation looks conical up to energies in excess of 1 eV [3] as shown in Figure ‎2.3. All the plots assume a nearest-neighbor hopping energy of 2.7 eV[2].

The linear dispersion relation for energies less than 1 eV make the use of Dirac equation to describe charge carriers a very accurate approximation, as most devices operate in this range of energies[3]. In this range as well, the density of states is proportional to the energy of the charge carrier. This unique description of charge carriers in graphene as massless Dirac fermions leads to very interesting electronic properties[1]–[23].



Figure . Energy dispersion relation around the K point for energies between -1 and 1 eV. The energy is proportional to the magnitude of momentum away from the k point, making the dispersion relation look conical in 3D.

The symmetry of the Brillouin zone allows us to use the 3 high symmetry points to describe the energy describe the energy dispersion relation. The most commonly used cut is the ΓMKΓ cut path shown in Figure ‎2.4. The energy dispersion relation along the high symmetry cut is shown in Figure ‎2.5, clearly showing the Dirac point at the K point.

|  |  |
| --- | --- |
| (a) | (b) |

Figure . (a) Brillouin zone of graphene with the high-symmetry points labelled. The high-symmetry cut path ΓMKΓ is perimeter of the highlighted triangle. (b) a contour plot of the conduction band with the high symmetry cut path highlighted in red.



Figure . Energy dispersion relation across high-symmetry paths in graphene as obtained using the tight-binding approximation with a nearest-neighbor hopping energy of 2.7 eV. The blue curve represents the conduction band while the red curve represents the valence band.

## Klein Tunneling and the Absence of Backscattering

Klein tunneling is the absence of back scattering of massless relativistic particles incident normally on a potential barrier[24], [25]. As a virtue of graphene’s linear dispersion relation, charge carriers in graphene demonstrate Klein tunneling even in disordered graphene[2], [5], [6], [10], [11], [26]–[39].

Klein tunneling causes the lack of rectification in graphene PN junctions[33], [36], [37]. It allows graphene operate in the p or n-type doping regimes, or even cascade regions of different doping types without rectification[27], [33], [36]–[38], [40]–[43]. Klein tunneling is the basis of operation of the graphene neurons, synapses and diodes, which are presented in chapters 3 and 4, respectively.

## The Role of the Substrate

Simple theoretical analysis of graphene assumes no interaction between the graphene and the substrate. The use of a substrate breaks this assumption by adding different sources of scattering as substrate roughness, substrate phonons, and charges at substrate-graphene interface cause scattering, doping and creates hysteresis in the transfer characteristics, when changing the substrate gating voltage[6], [13], [14], [44]–[63]. The substrate roughness also induced substrate-induced structural distortion, which give rise to charge puddles[22], [56], [64].

The substrate role is important in exfoliated graphene, where the optical interference allows the quick identification of graphene optically[65]–[67]. Although graphene can be suspended between contacts[6], [12], [14], [38], [47], [55], the substrate acts as a mechanical support for the graphene ribbon, and substrate dielectric engineering allows the control of top and back gate strengths. Selection of the substrate affects the phonon-induced doping, and proper selection is crucial to prevent mobility degradation[68].

## Charge Transport in CVD Graphene

CVD grown graphene is subject to increased scattering from the grain boundaries, substrate toughness, charged-impurities from transfer chemistry and substrate interface charges[4], [6], [20], [23], [61], [62], [68]–[78]. Careful control of the transfer environment and removal of residuals of transfer residuals is crucial to reduce charged-impurity scattering[9], [48], [52], [57], [59], [68], [79]–[88].

The grain size of CVD grown graphene has risen consistently over the years, reaching few millimeters, allowing the direct measurement of grain-boundary resistance[89]–[94]. However, other effects as substrate roughness and phonon scattering affect the transport in CVD graphene[6], [13], [14], [44]–[63].

Due to the numerous sources of scattering in CVD graphene, charge carrier transport in CVD graphene is predominantly diffusive at room temperature. Charge carriers suffer successive scattering with a mean free path significantly shorter than the device length[95].

## Contact-Induced Doping

Metals in contact with the graphene induce doping and modifications in the band structure of graphene. Ab-initio calculations and experimental results verify that metals contacts dope the graphene, and some metals cause Fermi level pinning[8], [26], [31], [35], [38], [43], [57], [82], [96]–[135]. The shape of resistance vs gate voltage reveals the nature of interaction and its strength. A secondary peak in the resistance vs back gate voltage indicates Fermi level pinning, while asymmetry in the conductance around a single resistance peak reveals contact-induced doping[123], [135].

Contact-induced doping is a strong function of the metal-graphene interface and the metal depositions conditions. Klein tunneling allows graphene channels to have any doping type, but contact-induced doping causes asymmetry in the resistance vs back gate voltage, increasing the resistance at the doping side opposite to that caused by the contact[31], [33], [35], [38], [47], [63], [119], [124], [126], [129], [130], [135].

The effect of contact doping on channel transfer characteristics sets a lower bound on channel length. The finite density of states in graphene causes metal-induced doping up to 100 nm around the metal contact[130]. In the metal-doped region, the doping level depends strongly on the metal contact characteristics, weakening the control of the gate on the doping, similar to DIBL in deeply scaled CMOS. This is a physical effect due to the finite density of states of graphene should be kept in mind when designing deeply scaled graphene devices.

## Characterization of Graphene

Graphene characterization is essential to determine the quality of graphene, number of layers and for quantitative assessment of the quality of graphene devices. We describe non-destructive ways of determining the number of graphene layers through optical contrast of graphene over a dielectric and by Raman analysis. Raman analysis also allows us to determine the strain and doping in graphene. We also discuss the constant-mobility model, as a way to quantitatively assess the diffusive transport and contact characteristics of graphene devices.

### Optical Contrast

The contrast of graphene over a dielectric is enhanced by the proper choice of dielectric thickness and dielectric constant. In SiO2, 90 nm and 285 nm are optimal, giving a reflection contrast of up to 15% for single layer graphene; this is the reason why most commercial graphene and exfoliation is done over 90 or 285 nm of SiO2[65]–[67].

The transmission contrast of graphene is even more interesting; it is set by the Fine Structure constant to about 98% in single layer graphene, making it suitable for flexible and transparent electronics[67].

Both reflection and transmission contrasts are usable in identifying graphene, with the reflection coefficient being more suitable over opaque substrate.

### Raman Analysis

Raman analysis refers to the use of Raman scattering in studying the properties of materials. Raman scattering is an inelastic scattering process where photons scatter inelastically with phonons, causing a shift in the frequency between the incident and scattered photons[136]. This shift is known as Raman shift. Although Raman shift can be given as a difference in energy, frequency or wavenumber of the incident and scattered light, it is most commonly reported as a wavenumber shift in the unit of cm-1. The Raman shift in terms of wavenumber is given as: , were is the wavelength of the scattered photons, is the wavelength of the incident photon, is the Raman shift in in energy, and and are Planck’s constant and the speed of light, respectively[137].

Raman scattering can decrease the energy and frequency of the scattered photon due to absorption of some of the incident photon energy to excite a phonon state (Stokes shift), or less-commonly increase the energy of the scattered photon by lowering an already excited phonon state (Anti-Stokes shift). Most Raman processes reported are Stokes processes. Raman processes can be classified as resonant or non-resonant, depending on whether the interactions occur between stationary states of the material involved. Raman scattering occurs with a very low probability compared to elastic (Rayleigh) scattering, making their amplitude very small[137].

Due to the high frequency of photon electric field change, Raman scattering is usually electronically mediated, where the charge carriers in a material -typically the low mass electrons- get perturbed by the incident photons then interact inelastically with phonons before returning to their original state, emitting the scattered photon. The momentum conservation requires that the total momentum change in the Raman interaction be almost zero, due to the negligible momentum carried by photons, setting the fundamental Raman selection rule to photons with nearly zero momentum for first-order processes. Accordingly, only optical phonons cause appreciable energy changes through first-order Raman scattering[136], [137]. Multi-phonon and higher-order processes are complicated by the electron-phonon and electron-electron interactions[137]–[140].

Raman analysis is a very versatile tool when it comes to quantifying the properties. Raman analysis can be used to probe the number of graphene layers, strain and doping[101], [137]–[155]. The most prominent features of the Raman spectrum of graphene are the G, D an 2D peaks occurring at a Raman shift of ~1580, 1350 and 2700 cm-1, respectively[137]–[140]. An example of Raman spectrum of a low-defect graphene flake over hBN is shown in Figure ‎2.6.

The G peak is due to the in-plane transverse optical phonons (E2g symmetry modes), while the D peak is due to the defect-activated ring-breathing modes (A1g symmetry modes), in contrast to the 2D peak, which is a double-resonant process that does not require a defect for its activation. As such, G and 2D peaks are always present in any graphene or graphitic sample, while D peaks is mainly present in defective samples[137]–[140].



Figure . Raman spectrum of a graphene sample over hBN showing the distinct G and 2D peaks. The lack of a D peak around 1350 cm-1demonstrates the low defect density in the graphene flake.

The full-width at half-maximum (FWHM) of the 2D peak is good indicator of the number of layers of graphene. Single layer graphene shows a symmetry peak, with a FWHM < 50 cm-1[137], [138], [140], [151]. The Raman peaks shift with strain as well, allowing the peak position shift to be used as a strain measure[141], [142], [144], [148], [149], [154], [156], [157]. The ratio of the 2D and G peak values is a function of the substrate thickness and doping of the graphene sheet[137]–[140].

When analyzing measurement data, the use of the integrated area under a curve is less susceptible to noise, compared with the numerical value of the peak maximum, which is very prone to noise and fluctuations. As such, comparing the integrated area is more noise proof.

### Constant Mobility Model

Transport in CVD graphene at room temperature is diffusive. One of the simplest, yet most versatile models is the constant-mobility model[68]. Unlike more complicated models, it models the transport using bias and back-gate independent numbers, allowing straight-forward statistical analysis of measured data. The model gives a single number for the mobility (), Dirac point shift (), and charged-impurity concentration (). The contact resistance () is assumed to depend assumed to be back-gate voltage dependent, and can be extracted when dual-gated data is available. For single-gated devices, the contact resistance is assumed to be bias-independent.

The constant mobility model fits the results of measurement of device resistance to the equation:

|  |  |  |
| --- | --- | --- |
|  |  | (.) |

Where is the number of channel squares, and is the number of charge carriers in the channel capacitively-induced by voltages shifted from the Dirac point. The Dirac point shift can be included in by replacing it with . In the devices that have no top gate, we use the following Equation:

|  |  |  |
| --- | --- | --- |
|  |  | (.) |

Where is the capacitance per unit area of the back gate oxide. Figure ‎2.7 shows the results of fitting measured data to Equation (‎2.2), the value of the Dirac point voltage is about -12 V. All the important features of the data (peak position, broadening due to , and resistance dependence on ) are captured despite the imperfect fit. We will use Equation (‎2.2) throughout this work when analyzing data from devices without a back gate.

In Chapter 3, we analyze devices with multiple top gates over the channel and no top gates over the contacts. In such devices, we add the resistances of all sections in series to get the overall device resistance yielding:

|  |  |  |
| --- | --- | --- |
|  |  | (.) |

Where is the length of each section, and and is the areal capacitance and voltage applied on each top gate. We neglect the interface resistance between regions of different doping and assume all the gates share the same channel width .



Figure . Measured data and fit results computed using the constant mobility model for back-gated graphene FETs. The RMSE is 0.653 kΩ.

Throughout this work, we use the constant mobility model for characterizing the devices fabricated. As in all our devices, the gate covers only part of the channel, the data will be fit using a constant value for .

# All-Graphene Neuromorphic Computing Architectures

Neuromorphic or brain-inspired computing is a distributed computing paradigm proposed to overcome the bottlenecks of von Neumann machines[158]–[163]. In neuromorphic computing, the data storage and processing is distributed among a network of nodes called neurons, interconnected by a web of connections known as the synapses. There are different architectures for neuromorphic computing, which can be classified on how the synapses interconnect the neurons.

We focus on the implementation of an all-graphene neuromorphic computing system, based on a type of neural networks known as cellular neural networks. We start by a brief introduction to cellular neural networks (CNNs), followed by an overview of an all-resistance CNN architecture that can be efficiently implemented using graphene. We then discuss the implementation of the proposed neurons and synapses and present the results of measured prototype devices. We conclude by a discussion of the performance aspects and limitations of the proposed architecture.

## Cellular Neural Networks (CNNs)

There are different families of neural networks, such as Spike-timing dependent plasticity (STDP)[163]–[176], and Oscillator Neural Networks (ONNs)[159], [175], [177]–[181]. We focus on a different family of neural networks that is continuous-time and continuous-voltage analog cellular neural networks[182], [183].

In Cellular Neural Networks (CNNs), the neurons act as saturating summing nodes. The neurons sum and the input applied to them and the output of a neurons is the sum of the inputs saturated using a sigmoidal saturating function. The saturation of the neuron’s output bounds it and stabilizes the network. The inputs applied to a given neuron are the outputs of its neighbors, weighted by the synaptic weight connecting the neurons together, as well as an external, independent stimulus. In our work we focus on networks without an external stimulus. We can thus think of a “cell” as a central neuron with the synaptic weights used to connect it to its neighbors. This is shown schematically in Figure ‎3.1.

Unlike other neural network implementations where the synaptic weight is set by the spiking activity, a property known as synaptic plasticity [173], CNNs operate in a continuous-time, continuous-voltage fashion. The synaptic weights are determined offline and are used to set the function of the circuit; CNNs to have no synaptic plasticity.

CNNs are asynchronous by nature and all the nodes take part in computations simultaneously. The modularity of the architecture allows it to implement high-complexity, large systems. In addition, CNNs are highly robust against variabilities and non-idealities in its building units, making it very suitable for implementation using technologies with high variability and non-idealities in components[182]–[185].



Figure . Nearest-Neighbor CNN. The neurons (nodes) are shown as red squares, while the synapses are the blue arrows. The Cell is composed of a single neuron and the synapses connecting it to its neighbors.

Computations are performed by first presetting all the neurons to the initial values used in the computation (network input) and setting the synaptic weights to correspond to the function of interest. The network is then left to relax, reaching a steady state that corresponds to the computed output (network output). A more formal discussion of the theoretical foundations and mathematical model can be found in [182], [183].

## All-Resistance CNNs

CMOS implementations of CNNs has been marred by area and power requirements that limit the network size considerably[184], [186], [187]. One major bottleneck in analog CMOS implementations matching requirement between component values, which lead to large area overheads and the need to running circuitry[186], [187].

To overcome these requirements, another implementation using an all-resistive architecture that only requires local matching between components of the same cell has been proposed [40], [41]. It relies on using local potential divides to implement the synaptic weighting, and the neuron output is controlled using voltage-controlled resistors that set the voltage applied to a capacitor that stores the state (voltage) of the neuron. The all-resistive architecture is shown in Figure ‎3.2.

This architecture has several advantages: 1) the output saturation is inherent in the neurons, as the output DC voltage is set by the potential divider between and ; 2) the synaptic weights depend only on the potential divider of two digitally-controlled resistors, relaxing the global matching requirement in CMOS implementations; 3) digitally-controlled resistors used to set the synaptic weights, allowing us to store the synaptic weight settings in CMOS memory, which generally have a very high capacity and low footprint; 4) The synapses are composed of two resistors in series, with complementary code words applied on them making their sum fixed, thus there is no code-dependent loading on the neuron; and 5) the neuron output voltage is determined by the analog voltage-controlled resistors and , whose input is an analog voltage coming from neighboring neurons, without the need to do any analog-to-digital or digital-to-analog conversion. The synapse resistors can be made to have negligible loading on the neuron output, mitigating their effect on reducing the dynamic range of the circuit and reducing their contribution to the power consumption. () is labelled as the connected to the inhibitory (excitatory) outputs because its increase reduces (increases) the output voltage, which is labelled as inhibition (excitation) in neuromorphic terms.

In total we need 2 analog voltage-controlled resistors per neuron ( and ) and 2×2 digitally-controlled resistors per synapse. This means that we need one capacitor, 2 analog voltage-controlled resistors and 36 (2×2×9) digitally-controlled resistors per cell. This allows for a very compact implementation, since the matching is required locally only; the synapse potential divider elements need to be matched to have linear synaptic weighting with the applied code word.



Figure . All-resistance architecture CNN architecture. The potential dividers inside the red and blue boxes provide the synaptic weighting of the neuron output (capacitor voltage), using digitally controlled resistors. The pull-up- and pull-down resistors ( and ) are analog voltage-controlled resistors, with the resistance control inputs connected to the synoptically-weighted output of the neighboring neurons.

### Limitations of All-Resistance CNNs architecture

The proposed all-resistance architecture depends on the strength of the control terminal of the analog voltage-controlled resistors to set the output voltage of the neuron. Accordingly, the control terminal of the analog voltage-controlled resistors should be strong enough to pull the output voltage of the neuron to the required level. The strength of the control terminal is what provides the regeneration in the network; it should be strong enough to regenerate the analog voltage change at the input. Otherwise the signal levels will degrade and the network size would be very limited.

If we use graphene to implement those resistors, the maximal resistance change range is achieved by using strong gate (high capacitance) and having a low charged-impurity concentration. Quantitatively, neglecting the contract resistance, the resistance of a graphene sheet as per the constant mobility model is given as:

|  |  |  |
| --- | --- | --- |
|  |  | (.) |

The rate of change per units resistance is given as:

|  |  |  |
| --- | --- | --- |
|  |  | (.) |

Where we denote the gate voltage that would yield a charge density equal to the charged-impurity concentration . This function peaks at , and the maximum rate of change of resistance is given as:

|  |  |  |
| --- | --- | --- |
|  |  | (.) |

Where , and are the dielectric constant of the oxide, free space permittivity and thickness of gate oxide. The stronger the gate (higher and smaller ) and the lower the charged-impurity concentration the more resistance change can be achieved by the gate.

Another intuitive explanation is through the impact of charged-impurity concentration on the resistance change range. The higher the charged-impurity concentration the lower high-to-low resistance change range can be achieved by a given gate. If the high-to-low ratio is small enough the output voltage swing would be low and might not be enough to regenerate a small input change at the input.

This discussion elucidates the importance of low charged-impurity concentration and strong gates for successful implementation of graphene all-resistance CNNs.

## Graphene Neurons and Synapses

Graphene neurons can be implemented dual-gated graphene ribbons with equal length gates, while the synapses can be implemented by using binary-scaled gates. This can be explained by deriving the resistance of a graphene ribbon using the constant mobility model for a ribbon with multiple top gates and a single common back gate, similar to the one shown in Figure ‎3.3.



Figure . Cross-section of an n-gate graphene device. There are n top gates and one common back gate.

The resistance of such a structure, neglecting the contact resistance is given as:

|  |  |  |
| --- | --- | --- |
|  |  | (.) |

Where is the channel width, and are the sheet resistivity and length of the ungated sections, and the sheet resistivity and length of the i-th gates sections, and the voltage of the i-th top gate and back gate, is the interface tunneling resistance per unit width for the charge carrier to tunnel from the ungated section to the i-th gate and back out if the gating causes alternate doping types in the channel. can be neglected relative to the sheet resistance due to Klein-tunneling at the interfaces[36], [37]. The ability to have alternating dopant type across the channel without rectification is a direct result of Klein-tunneling in graphene, even for non-ballistic transport devices[30], [39], [188]. The use of a specific sheet resistance for each gate is due to the possibility of process gradients across the device, the presence of charge puddles and charge inhomogeneity, and because metal gates can alter the phonon scattering modes[22], [61].

In case of neurons, the gate lengths are equal (), making the contribution of the gates in the resistance change equal, while for synapses, they are binary scaled. For binary scaled gates, we have , making the device act as a string of binary-scaled resistors. Accordingly, if the binary scaled gates are connected to the corresponding bits of a digital control word, they will act as a digital to resistance converter. This can be quantified by setting for all gates, allowing us to rewrite Equation (‎3.4) as:

|  |  |  |
| --- | --- | --- |
|  |  | (.) |

If a digital signal having the values , corresponding to a digital bit , is applied on the top gates, the resistance change term () as a function of code word can be written as:

|  |  |  |
| --- | --- | --- |
|  |  | (.) |

Equation (‎3.6) shows how the binary scaling the gate lengths would make the device act as a digital-to-resistance converter, which is what we need for each resistor in the synapses.

Interestingly, this architecture allows for both positive and negative resistance change with the applied code word (voltage), depending on the relation between the back-gate voltage and the Dirac point. This is illustrated using Figure ‎3.4. Figure ‎3.4(a) shows the resistance of the ribbon when a at two top gate voltages V1 and V0 where V1 < V0. When V1 is applied on the top gate, the ribbon becomes more p-type relative to when V0 is applied, making the Dirac point shift right relative to its position when V0 is applied. The resistance change, defined as the resistance difference between the case when V1 is applied and when V0 is applied, is shown in Figure ‎3.4(b). When the back-gate voltage is such that the device is biased near the Dirac point when V1 is applied the resistance increases when V1 is applied relative to V0, and vice versa.

The dependence of the resistance change direction allows us also to use the graphene synapses as digital-to-resistance converters with positive or negative slope (resistance change with code word), depending on the applied back-gate voltage.

The ability to implement PN junctions without rectification, and creating both positive and negative resistance change after fabrication through an externally applied voltage are unique properties to graphene, which allows the device to be used for applications other than neuromorphic computing. This has various applications in other kinds of circuitries such as data converters[189].



Figure . (a) Resistance vs back gate voltage of a section of a graphene channel at two top gate voltages V1 and V0, and (b) resistance change, defined as R(V0)-R(V0), vs back gate voltage. Using a back-gate voltage around V1 would cause a positive resistance change, while using a back-gate voltage around V0 would cause a negative resistance change.

### Scaling limits of graphene all-resistance CNNs

The graphene ribbons used to implement the CNNs could be scaled in width until a band-gap is opened or line-edge roughness (LER) scattering becomes significant, increasing the resistance significantly. A band-gap of 10 meV opens in graphene when the width is about 30 nm[190], and LER-induced scattering degrades the mobility when the width is less than about 60 nm[77]. As such, we do not anticipate problems to occur down to device widths of 60 nm. If we are to keep are to design the neurons with an aspect ratio of 10, that would yield a neuron with a size of 60 nm × 600 nm or 0.036µm2. To our knowledge, the smallest CMOS implementation for a binary (2-level) synapse is 3.2 µm2, and it implemented integrate-and-fire networks using 45 nm CMOS[191]. Other non-CMOS implementations exits, but they focus on the implementation of STDP and ONN networks[163]–[165], [167], [170], [172], [192], [193][178]–[180], [194], [195]. This demonstrates the high scalability, and footprint of the proposed graphene all-resistance implementation of CNNs.

## Measurement Results

We fabricated prototype devices for a 3-bit synapse using the flow outlined in Appendix B. An optical microscope and false-color image of the fabricated device is shown in Figure ‎3.5. The devices were tested per the test setup shown in Appendix C. An important feature in Figure ‎3.5 is the use of common centroid in distributing the gates across the synapse. This reduces the impact of process gradient in modifying the strength of the gate and equalizes the strength of each gate[41]. This was verified experimentally where using direct binary scaled lengths showed unequal gate strengths in modifying the resistance of the synapse[40].

|  |  |
| --- | --- |
| (a) | (b) |

Figure . (a) Optical microscope, and (b) False color SEM image of 3-bit graphene synapse. The binary-scaled gate lengths are distributed in a common centroid fashion to reduce the effect of process gradient on gate strength. The spacing between the two contacts is 9 µm, and the width of the channel is 2.75 µm.

The longest gate was chosen to act as the Most Significant Bit (MSB) because it has the biggest impact on the resistance change compared to the other gates. The measured resistance of the device vs code word at two different back gate voltages corresponding to the maximum positive and negative resistance change is shown in Figure ‎3.6. The top gate voltage used in for binary 0/1 was 0/0.067 V respectively. The resistance changes by about 15.7 kΩ in the positive direction (19.5 – 35.4 kΩ) and 8.7 kΩ in the negative direction (19 – 27.7 kΩ). The inequality between the resistance change ranges and the negative resistance change at a lower back gate voltage despite applying a positive top gate voltage indicates the present of negative oxide charges affecting the field generated by the gate. This was confirmed by evaluating the amount of Dirac point voltage shift created by the top gate when a positive and negative voltages are applied, and it was found that the negative gate voltage cause more resistance change. The difference in strength was found to be due to a negative oxide charge of density of 8.74 × 1011 cm-2.

|  |  |
| --- | --- |
| (a) | (b) |

Figure . Resistance change vs applied digital code word at a back-gate voltage of (a) 22.67 V, and (b) 8V. These back gate voltage correspond to the maximum resistance change in each direction. The top gate voltage used was 0.067 V and 0V for binary 0 and binary 1, respectively. The dotted black line represents the best linear fit.

To evaluate the strength of each gate we extracted the resistance change per unit square of each gate. The extracted gate strengths per square, shown in Figure ‎3.7, indicate that all the gates show equal strength, with a behavior similar to that shown in Figure ‎3.4. This equality in strength is in line with the strong linear behavior and small non-linearity seen in the resistance vs code word measurements shown in Figure ‎3.6. Any inequality in gate strengths manifests itself as a non-linearity in the resistance vs code word characteristics[40], [41].



Figure . Extracted resistance change per unit square for each gate. The extracted values show that the gates have equal strength.

The measured devices were used to build Verilog-A models that were used for system-level simulations using Spectre. An architecture of 5×4 neurons was built and programmed to perform edge and line detection as well as pattern recognitions[40]. Different functions have different synaptic weights, which were determined by calculating them offline, using MATLAB. In each case, the network was precharged to the input value and the network was left to relax, converging to the output of the calculation. The system was able to perform edge and line detection correctly and recover(recognize) distorted patterns with up to 28% of distortion[40].

## Conclusion

We have demonstrated an area efficient of CNNs using graphene to implement an all-resistance architecture of synapses and neurons. The devices depend on the low charged-impurity concentration and high gate strength to be able to regenerate the voltage changes at the input. The graphene implementation of the all-resistance CNN architecture is highly scalable with very low area compared to CMOS implementations. We fabricated prototype devices for the synapses and used it to extract the information needed to build Verilog-A models for use in system-level simulations validating the viability of the architecture. The less the charged-impurity concentration and the stronger the device gates area, the bigger the network can be.

From a system perspective, more work need to be done on the size limits of the network and the limitations posed by device non-idealities. The maximum number of characters that can be stored in a network of a given size and its relation to the training scheme used to identify the required synaptic weights is still an open problem. Finally, there are no indicators on when a network has converged, but the network is assumed to have converged once the rate of change of all neuron outputs approaches zero[182], [183]. These problems need to be addressed and the impact of device non-idealities has to be evaluated for this architecture to be deemed mature enough for implementation.

# Graphene Low-Loss Diodes

Graphene’s unique electronic properties, make it attractive in designing devices that utilize the photon-like behavior of charge carriers. The photon-like behavior of charge carriers in graphene allows us to build zero turn-on voltage diodes, where the carrier trajectories can be controlled in an anisotropic manner. In this chapter, we propose an implementation of low-loss graphene diodes using asymmetric transmission of charge carriers across an oblique barrier. We present measurement results from prototype devices showing how our proposed diodes outperform other implementations and discuss the performance limitations of our proposed structure.

In comparing the performance of different low-loss diodes we use the current conduction asymmetry (), defined at the ratio of the forward to reverse current ratio at the same voltage magnitude:

|  |  |  |
| --- | --- | --- |
|  |  | (.) |

Where and represent the same voltage magnitude but applied in the forward and reverse directions, respectively[196]. The asymmetry of a perfect linear resistor is unity, and the asymmetry of an ideal diode approaches infinity as it indicates perfect blocking of reverse current.

We start by a brief overview of geometric diodes, outlining their principle of operation. We then propose a diode architecture based on the total internal reflection of charge carriers incident on an oblique barrier, discussing its principle of operation and performance limitations before presenting measurement results of fabricated prototype devices. We end the chapter by discussing ways to enhance the performance of the proposed device.

## Geometric Graphene Diodes

Geometric diodes use the asymmetric transmission of charge carrier across an aperture to achieve asymmetric conduction[196]–[201]. The principle of operation of such diodes can be explained using the toy model shown in Figure ‎4.1.

Charge carriers incident from the left to right get collimated by the geometry of the channel leading to the aperture, allowing most of them to pass through. On the other hand, charge carriers traveling from the left to right do not get collimated and only a small portion passes through the aperture. A formal analysis of geometric diodes can be found in[197].

The maximum asymmetry reported for geometric diodes was ~1.35 for exfoliated graphene and 1.15 for CVD graphene[196]. The reduction in asymmetry obtained in CVD graphene can be attributed to the increase in scattering, increasing the resistance of the forward path and allowing carriers to pass in the reverse direction[4], [5], [7], [9], [11], [60], [62], [69], [70], [72], [77], [88], [202]. This is a strong indicator of the impact of scattering on the performance any diode utilizing the photon-like behavior of charge carriers in graphene.



Figure . Toy model representing the asymmetric transmission across an aperture. Charge carriers (blue circles) get collimated into the aperture when traveling from left to right, causing large transmission. On the other hand, charge carriers travelling from right to left do not get collimated and transmit only through a small area.

## Oblique Incidence Diodes

Oblique incidence diodes depend on the concept of total-internal reflection of charge carriers from an oblique junction to cause asymmetric conduction. Klein tunneling in graphene causes normal incident carriers to always pass through, but obliquely incident carriers can be reflected. Asymmetric conductance across graphene oblique PN junction has been predicted theoretically and measured experimentally[28], [30], [32], [33], [39], [203]–[210].

The angular condition on reflection can be obtained by matching the boundary conditions at the barrier, which yields a Snell’s law-like relation[25], [30], [39], [188]s:

|  |  |  |
| --- | --- | --- |
|  |  | (.) |

The critical angle beyond which reflections of charge carrier occur can be obtained by setting to 90 degrees:

|  |  |  |
| --- | --- | --- |
|  |  | (.) |

Where is the energy of the charge carrier, and is the energy of the Dirac point in the incidence medium (subscript 1) and the transmission medium (subscript 2) respectively, and is the critical angle of incidence in the first medium. This is depicted in Figure ‎4.2 for a case when , causing transmission away from the normal.



Figure . Schematic presentation of charge carrier motion across an oblique carrier when , causing transmission away from the normal. Further increase in will cause to increase till it reaches 90 degrees. Further increase in causes total internal reflection of charge carriers.

The structure of an oblique incidence diode is shown in Figure ‎4.3. The oblique gate sets the angle of the potential barrier (θ) and the voltages on the top and back gate set the potential barrier height.

|  |
| --- |
| (a) |
| (b) |

Figure . (a) Cross-section, and (b) Top view of oblique incidence diode. The whole device lies on a back gate that is not shown for clarity. The top gate is oblique relative to the channel. The potential barrier can be set by controlling the voltage of the top and back gates.

The transmission coefficient across an abrupt junction vs the energy of charge carrier and angle of incidence when tunneling in different directions across regions with eV and eV is shown in Figure ‎4.4. Transmission asymmetry occurs except when .

|  |  |
| --- | --- |
| (a) | (b) |

Figure . Transmission coefficient when crossing abrupt barrier of between eV and eV. (a) from region 1 to region 2, and (b) from region 2 to region 1.

## Performance Limitations

### Lead Resistance

The transmission coefficient across the interface is used for calculating the interface resistance using the Landauer-Büttiker formalisms[211]–[213]. However, the asymmetry is controlled by the total resistance of the device in the forward and reverse direction. The lead resistance adds up to the resistance of the device in both directions and reduces the total resistance difference, and asymmetry of the device.

For example, Figure ‎4.4 shows that maximum asymmetry occurs when the Fermi level of the device causes one of the sides to be intrinsic. However, this causes the lead resistance of that section to be maximum, which can impact the overall asymmetry of the device. In summary, the lead resistance is an important factor that can limit the overall performance of the device.

### Thermal broadening of energy distribution

The transmission coefficient is a strong function of both the angle of incidence and energy of charge carriers. The thermal broadening of the Fermi-Dirac distribution at non-zero temperature means that a region of energies around the Fermi level take part in conduction[211]–[213]. As such, the total asymmetry depends on the energy thermal broadening. For example, if the device is biased to make one of the sides intrinsic, as the temperature increases, more charge carriers with non-zero transmission ratio take place in conduction, reducing the asymmetry of the interface resistance.

### Momentum Distribution Smearing by Scattering

Charge carriers in graphene are subject to many mechanisms of scattering[2], [4]–[7], [9], [10], [12], [15], [18], [20], [22]–[24], [27], [29], [30], [37], [43], [47], [48], [60]–[62], [64], [68]–[73], [75], [77], [78], [81], [87], [88], [97], [148], [156], [214]–[225]. Scattering smears the angular distribution of charge carriers[44], [226]–[229]. The angular distribution smearing should be included used in the Landauer-Büttiker to include the effect of such smearing. Accordingly, the Landauer-Büttiker equation for the interface conductance be written as:

|  |  |  |
| --- | --- | --- |
|  |  | (.) |

Where is the transmission probability, is the number of conducion modes, is the Fermi-Dirac distribution, and is the momentum smearing angular distribution function of the carriers.

This smearing function will cause more angles to take part in the conduction at the interface, which will noticeably increase the conduction of the interface in the reverse direction, reducing the asymmetry of the device.

Unfortunately, there are no analytical expressions for the momentum angular distribution, but it can be obtained using Monte Carlo simulation[44], [226]–[229]. Equation (‎4.4) highlights the profound effect scattering has on the performance devices relying on the photon-like properties of graphene charge carriers to achieve conduction asymmetry.

## Measurement Results

We fabricated 20 diode devices using CVD graphene on SiO2, following the same procedure given in Appendix B. The devices were fabricated with 4 oblique angles: 20°, 30°, 45° and 60°, with 5 devices fabricated at each angle. The measurements were performed using a setup similar to that shown in Appendix C. Figure ‎4.5 shows a false color SEM image of one of the fabricated devices with an oblique angle of 45°.



Figure . False color SEM image of one of the fabricated diode devices with a 45° oblique top gate.

The measurements shown in Figure ‎4.6 shows that the highest asymmetry corresponds to an oblique angle of 45°. The mean asymmetry at 45° is 1.65, and the smallest asymmetry was 1.47 at an angle of 30°; this asymmetry is 28% -43% higher than that obtained using CVD geometrical diodes, and 9%-22% higher than that obtained using exfoliated geometrical diodes. Interestingly, the mean asymmetry is a weak function of the gate angle.

The weak dependence of the asymmetry on the gate angle can be attributed to the smearing of the angular distribution of the carriers, distributing the incidence angles of the carriers over a broad range of angles. This angular broadening seems to be wide enough to smear the differences in angle over the range of the fabricated devices to the point of smearing out any trend.

In addition, the proposed diode outperforms the asymmetry of geometric diodes fabricated with either CVD or exfoliated graphene. This can be attributed to the effectiveness of oblique barriers in creating conduction asymmetry compared to apertures. The higher asymmetry however comes at the expense of adding an extra gate, which is not required in geometric diodes. The added gate control terminal is not necessarily a problem, and paves the road for oblique diodes to implement graphene PN junction multiplexers and logic[203]–[205], [207], [208], [230].

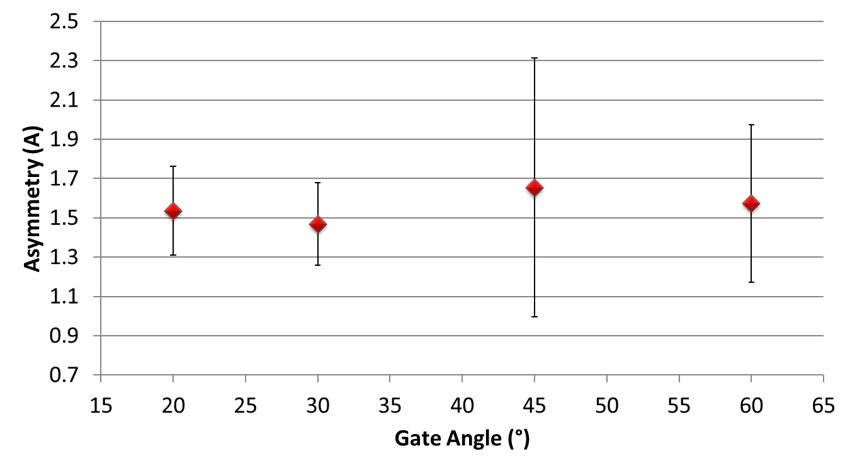


Figure . Extracted asymmetry for each oblique gate angle. The error bars represent the standard deviation of the data and the red diamonds represents the mean.

## Conclusion

Oblique incidence diodes outperform geometrical diodes in terms of the asymmetry. The added asymmetry comes at the expense of adding an extra terminal, making the oblique incidence diodes a 3-terminal device. The added terminal adds to the functionality of the device, allowing it to be used to implement graphene PN junction multiplexers and logic. The largest asymmetry was obtained for a gate angle of 45°.

Quantitatively, the asymmetry numbers are very low and need to be improved for the devices to be application worthy. The major culprit in limiting the asymmetry is the scattering, which smears the angular distribution of the carriers incident on the gate to the point of smearing the differences in angles from 20° to 60°. Higher performance might be achievable using better channel and dielectrics. For example, using exfoliated graphene encapsulated by hBN would reduce the scattering, achieve ballistic transport over a larger mean free path[223], [231]–[234], which could yield higher asymmetry.

# Diffusive-Transport Graphene Couplers

The similarity between the dispersion relation of photons and charge carriers in graphene appeals to designing optics-inspired electronic devices. An interesting optics-inspired analog is the electronic directional coupler. Optics directional couplers allow coupling light between two branches, where the coupling coefficient varies periodically with the coupling distance[235]. The electronic wave modes in graphene ribbons and the resistive coupling between graphene ribbons in close proximity has been analyzed for ballistic transport [10], [29], [236]. The problem of analyzing resistive coupling in diffusive graphene ribbons is different from that in ballistic graphene ribbons. The successive scattering events randomize the wavefunction phase information [211]–[213] prohibiting the direct application of the coupled-mode theory. A direct consequence of this phase randomization is losing the spatial periodicity of the coupling coefficient predicted in ballistic devices.

Modeling of resistive coupling is also crucial for deeply scaled interconnects, in which transport will inadvertently be diffusive due to line-edge roughness[70], [77], [237]. Graphene interconnects in close proximity has been studied previously to evaluate their cross-talk performance[238]–[244]. Prior work focused on analyzing the delay and energy metrics of a single graphene interconnect, accounting only for the capacitive coupling among interconnects.

In this chapter, we study the coupling between graphene ribbons operating in the diffusive transport regime. We start by developing an analytical model for the coupling resistance between two graphene ribbons separated by a dielectric, highlighting the impact of different fabrication non-idealities on the coupling. We then evaluate the spatial dependence of such coupling coefficient, showing its monotonic saturating behavior, and assess the impact of such coupling on the performance of deeply scaled interconnects.

## Modeling Diffusive-Transport Current Coupling

An electronic current coupler consists of two graphene ribbons in close proximity as shown in Figure ‎5.1. The phase incoherence associated with diffusive transport devices prohibits the direct application of coupled mode theory to evaluate the coupling between two coupled graphene ribbons. A more direct approach would be to model the coupling using the tunneling resistance between the two branches of the coupler; this emphasizes the diffusive nature of the transport and the lack of phase coherency in the associated wave functions.

The modeling of the tunneling resistance between the two graphene ribbons must take into account the difference in energy dispersion relations across the tunneling barrier. The energy dispersion relation of charge carriers changes from a linear dispersion relation in graphene, to a parabolic dispersion relation with an energy gap in the oxide regions, as shown in Figure ‎5.2. The lack of states in the dielectric energy gap translates to decaying wavefunctions from the graphene ribbons on either side of the dielectric. In other words, despite the linear energy dispersion relation of graphene, the lack of states in energy gap region of the parabolic dielectric gives rise to a decaying wavefunction, reminiscent of tunneling in parabolic systems.



Figure . Schematic representation of an electronic graphene coupler. The ribbons are spaced by a distance d apart, over a length of L, while each ribbon has a width of W. The graphene is shown in blue while the surrounding oxide is shown in pink. The oxide is only shown below the graphene ribbon for clarity.

The calculation of the tunneling resistance is based on the analysis of Graphene-Insulator-Graphene (GIG) junctions[245], [246]. The major difference between the prior work on GIG junctions and the current problem is that in this device tunneling occurs between the edges of the graphene ribbons rather than normal to them. The edge tunneling nature modifies the results presented in [245], [246] slightly, but follows its essence otherwise. We defer the estimation of the tunneling resistance to Section ‎5.3, but stress on the fact that it is a tunneling resistance whose value is very large compared to the ribbons’ resistance; its value is limited by how close the two ribbons can be spaced apart, and for all practical purposes, this tunneling resistance value is significantly larger than the resistance of the graphene ribbons. This observation will proof useful when analyzing the electrical model of the device.

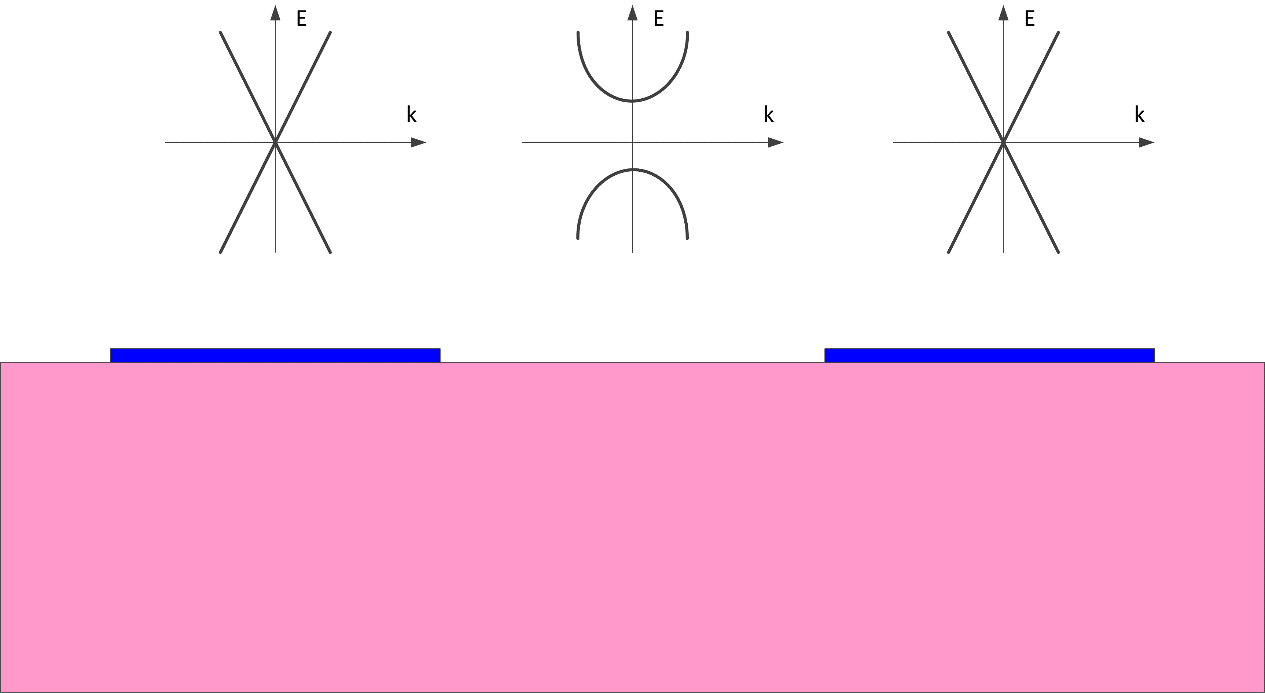


Figure . Cross Section of the graphene coupler with the bottom oxide only shown for clarity. The electronic energy dispersion relation for each region is shown above it; it is linear in each graphene region and parabolic with a band gap in the dielectric surrounding them. The energy gap in the parabolic region is significantly larger than the energy of charge carriers in each graphene ribbon.

The electrical model of the device is composed of three distributed resistors: a distributed resistor for each of the graphene ribbons with a distributed tunneling conductance connecting them together, as shown in Figure ‎5.3. We label one of the ribbons as the input ribbon, and the other as the output ribbon. For this analysis, we apply a current stimulus at the input ribbon and calculate the current at other end (output) of each ribbon.



Figure . Electrical model of graphene coupler. The graphene ribbons are modelled using two distributed resistors with a resistance per unit length of R1 and R2, and the tunneling resistance coupling them is modelled using a distributed conductance with conductance per unit length gc.

The coupling coefficient between the current in the two ribbons is defined as the ratio between the output ribbon and input ribbon branch currents as:

|  |  |  |
| --- | --- | --- |
|  |  | (.) |

A detailed analysis of the electrical model and a derivation of the coupling coefficient is provided in Appendix D. The current distribution and coupling are a strong function of the load at the output of each branch. This is expected due to the passive nature of the device that does not provide any buffering. Throughout this chapter, we assume that the ratio of the two loads matches the ratio of the ribbons’ resistance per unit length, that is . A more general analysis can be found in Appendix D.

Under the matching load condition, we can approximate the current distribution in each branch of the coupler is:

|  |  |  |
| --- | --- | --- |
|  |  | (.) |

Where .

## Dependence Current Coupling Coefficient on Coupling Distance

Under the condition of matching load ratios, , the current coupling coefficient is:

|  |  |  |
| --- | --- | --- |
|  |  | (.) |

In the limiting case when , Equation (‎5.3) reduces to:

|  |  |  |
| --- | --- | --- |
|  |  | (.) |

Equations (‎5.2), (‎5.3) and (‎5.4) provides a very intuitive way of explaining the behavior of the diffusive-transport coupler: given enough length, the coupler will divide the current by the ratio of the resistances of the two branches, just as if they shorted only at the input end. Unlike the ballistic-transport coupler or the optical directional coupler, the coupling coefficient does not show any periodicity on the coupling coefficient. The lack of coupling coefficient periodicity is due to the loss of the phase information due to successive scattering associated with diffusive transport. The diffusive-transport coupler rather acts as a current divider that divides the current with according to the ratio of the two branch resistances. However, rather than being an ideal current divider, the current division takes places over a special distance dictated by the characteristic length, which is a function of the ratio between the coupling and branches conductance.

An example of the spatial variation of the current coupling coefficient and the voltage across the branches of a balanced coupler if shown in Figure ‎5.6 as obtained using SPICE simulations.

|  |  |
| --- | --- |
| (a) | (b) |

Figure . (a) Spatial variation of the current and (b) Spatial variation of the voltage across a balanced coupler. The simulations were performed using SPICE over a discretized model comprised of 1000 sections.

## Estimation of the tunneling resistance

The tunneling resistance can be estimated using the Bardeen Hamiltonian approach as in other Graphene-Insulator-Graphene junctions [245], [246]. The tunneling current evaluated using the Bardeen Transfer Hamiltonian [245], [246] is given as:

|  |  |  |
| --- | --- | --- |
|  |  | (.) |

Where is the valley degeneracy factor (2 in graphene), is the spin degeneracy factor (), is the tunneling matrix element, and are all the states in the electrodes in between which tunneling occurs, and is the Fermi Dirac distribution in the input and output electrodes respectively, and is the energy of the state. The matrix element is given as:

|  |  |  |
| --- | --- | --- |
|  |  | (.) |

Where is the reduced Planck constant, is the free-electron mass, is the wavefunction of state , is the direction of tunneling current flow and the integral is performed on a plane midway between the two tunneling electrodes. The matrix element in the case of graphene electrodes is given in [245], [246] as:

|  |  |  |
| --- | --- | --- |
|  |  | (.) |

Where is the wavevector of the decaying exponential function inside the parabolic region barrier (also known as the extinction coefficient), is the two tunneling electrodes separation, is the wavefunction normalization constant, is a scaling factor in the order of unity that accounts for the wave vector misalignment between the overlapping tunneling wavefunctions, is the misalignment vector with an angle similar to the in accounting for rotational misalignment between the tunneling bands, and and is the wavevector between the input and output electrodes. For low field transport in graphene, the wavevector of the charge carriers taking part in tunneling is quite close to the Dirac points ( and ).

The lateral tunneling and the use of lithography to form the junction requires some modifications to tunneling matrix element that significantly simplify its analysis:

1) Lithography creates rough edges. An example of the edge roughness of metal lines obtained when creating metal lines using PMMA photoresist is shown in Figure ‎5.5. The edge roughness standard deviation is 3 nm and there is no correlation between the edges. As such, the tunneling distance and corresponding tunneling resistance is a random variable, and the differential equations used to describe the transport in the electrical model should replace a constant with a random variable, converting Equation (‎E.1) into a set of coupled stochastic differential equation. This level of detail is only required when the edge roughness represents a considerable portion of the ribbon width or ribbon spacing, but should be kept into consideration when developing statistical models for the coupler. When the roughness of the edge is small relative to the ribbons separation or the width of each ribbon, using the average separation distance to derive a single constant value for tunneling conductance is enough.

|  |  |
| --- | --- |
| (a) | (b) |

Figure . SEM images of 25 nm metal lines created using PMMA photoresist. The images shows an edge roughness standard deviation of 3 nm and the lack of correlation between edges in close proximity; (a) single metal wire, and (b) two metal wires with a separation of 25 nm.

2) In this device, tunneling is from a 1D edge to the opposite 1D edge across a 2D barrier, not from a 2D surface of a graphene sheet to another across a 3D barrier. The tunneling probability across a 2D barrier falls exponentially away from the normal connecting the two closest points across the barrier. As such, a 1D model is suitable to describe the tunneling probability between the edges. In the 1D model the tunneling probability is:

|  |  |  |
| --- | --- | --- |
|  |  | (.) |

Where is a point midway the 2D barrier assumed to be in the x direction.

3) The two branches of the coupler are fabricated from a single CVD graphene sheet over SiO2 with a single common back gate. Since the devices are in close proximity, their Fermi level should be closely aligned. However, the surface roughness of the substrate will cause the formation of charge puddles, which causes random shifts in the Fermi level, and the barrier height, across the graphene ribbons. Our first order estimate will neglect these charge puddles and assumes Fermi level alignment, but a more elaborate analysis should account for charge puddles and roughness.

4) The region between the two branches of the coupler is typically composed of an oxide. The effective mass of the oxide should be taken into account when calculating the tunneling matrix element, not the free-electron mass.

5) The tunneling resistance is a function of the Fermi level separation between the two ribbons, which is a function of the potential different between them. The voltage drop between the two ribbons is small, thus the variation of tunneling resistance due to the spatial variation of the Fermi level difference can be neglected. This is further justified by the fact that we are interested in the low-field transport phenomena occurring across lithographically defined barriers (roughly nm apart), a condition under which electrode charging occurs.

Taking the above points into consideration we can rewrite the tunneling current in the low-field transport regime at T=0 K as:

|  |  |  |
| --- | --- | --- |
|  |  | (.) |

Where is a constant in the order of unity that captures the effects of misalignment and normalization constants (it has the units of m-1), and V is the applied voltage. The extinction coefficient in this case is given as:

|  |  |  |
| --- | --- | --- |
|  |  | (.) |

Where is graphene’s work function, is the ribbon’s Fermi level, and is the electron affinity of SiO2. Equation (‎5.4) yields the tunneling current per unit length of the ribbons in the low field transport regime. The conductance per unit length of the ribbon is thus given as:

|  |  |  |
| --- | --- | --- |
|  |  | (.) |

Evaluating Equation (‎5.11) without the constant to estimate the order of and the extinction factor , we find it is a strong function of the separation distance and weak function of the ribbon’s Fermi level as shown in Figure ‎5.6.

|  |  |
| --- | --- |
| (a) | (b) |

Figure . (a) Logarithm of tunneling coupling conductance (), and (b) Extinction coefficient () while varying the Fermi level from 0 to 0.3 eV and the separation distance from 1 to 25 nm.

The numbers obtained show that the coupling is extremely weak. This is due to the use of a non-resonant tunneling device [245], [246]. In addition, with the separation distances achievable in the order of 10 nm, the extinction of the coupling wave function is very weak to yield appreciable coupling ( nS/m). This translates to a very large, impractical, coupling characteristic length, which means that the device operates essentially as two independent ribbons.

## Impact of Current Coupling on Deeply Scaled Interconnects

Resistive coupling between neighboring interconnects is a hazard to signal integrity. Deeply scaled graphene interconnects will not suffer because of current coupling before the spacing between them falls well below 5 nm laterally. However, care should be taken when discussing the limits on stacking graphene interconnects [238], [240]–[244], [247], [248]. Resonant-tunneling [245], [246], can play a significant role in the engineering of graphene interconnects.

Resonant tunneling between neighboring interconnects would increase the coupling conductance and cause resistive leakage of current among neighboring interconnects. and should be prevented if the interconnects do not belong to the same route.

Similarly, resonant tunneling can be used to reduce the overall interconnect resistance. It can be used to reduce the overall resistance of a stacked, multi-layer graphene interconnect, where the layers belong to the same route and carry the same signal. The large coupling distributed conductance will help shunt the resistance of each layer and reduce the overall resistance.

Proper design of stacked graphene interconnects, or different graphene-metallization layers in multi-layer graphene chips should take into account the resonant stacking constraints. If two ribbons are routed normal to each other on successive graphene-metallization layer (Manhattan Routing), the spacing between the layers should be carefully chosen to prevent resonant tunneling as long as they carry independent signals.

## Conclusion

A graphene-insulator-graphene electronic coupler operating in the diffusive-transport regime does not show a spatially periodic dependence on the coupler length. Diffusive-transport coupler show an asymptotic, monotonic dependence of the coupling coefficient on the coupling length due to the loss of wavefunction phase information by successive scattering events.

Diffusive-transport graphene couplers are not practically viable without creating a resonant tunneling structure. Otherwise, the tunneling-based coupling conductance between the graphene ribbons will be too small to yield any practically significant effect.

The analysis of the graphene coupler shows that graphene interconnects are highly scalable laterally, where resistive coupling will not be effective even when the inter-ribbon spacing falls below 5 nm. However, care should be taken when stacking graphene interconnects to prevent resonant tunneling between independent routes when using multi-layer graphene interconnects. Similarly, resonant tunneling can be used in multi-layer graphene interconnects to reduce the overall interconnect resistance.

# Current Technological Limitations on Graphene Devices and Circuits

The performance of graphene FETs and devices is hampered by numerous technological limitations. In this chapter, we discuss the technological limitations limiting the electrical performance of graphene FETs. We focus on electrical and thermal interface contact resistance, the evolution of stress in dual-gated graphene FETs, and the impact of photoresist chemistry, specifically metal-ion content, on the performance of graphene FETs.

## Contact Engineering

Graphene FET architectures typically utilize top contacts to graphene sheets as shown in Figure ‎6.1. Although not as good as edge contacts[234], they are simpler to fabricate. Contacts play an important role in the device’s performance. In this section, we discuss the current limitations in terms of electrical contact resistance and thermal interface resistance of metal-graphene contacts, and how they affect the performance of a graphene FET.

|  |  |
| --- | --- |
| (a) | (b) |

Figure . (a) False Color SEM image, and (b) Cross-Section of a typical graphene FET. The top gate oxide is used to passivate the device against environmental effects when used or measured in air.

### Electrical Resistance

Electrical contact resistance between metal and graphene has been studied extensively. Different contact models differ in the complexity and accuracy[26], [31], [35], [96], [101], [102], [110], [116], [120], [122], [129], [134], [135], [249], [250]. Experimental measurements of contact resistance to graphene are marred by many non-idealities, such as low adhesion of the contact metal to graphene, and a perceived negative contact resistance due to the metal contact doping the graphene region around it[251]. The use of circular TLM structure helps in mitigating adhesion issues due to the large surface contact area[133], [252]. The metal doping graphene around the contacts due to the finite density of states of graphene has been predicted theoretically[35], [38], [102], [108], [116], [119], [125], [126], [130], [135], measured experimentally[43], [87], [103], [110], [119], [122], [124], [128], and its impact on the interpretation of experimental data evaluated[251].

The proper choice of a metal contact to graphene is compromise between finding a metal that adheres well to graphene, but does not damage its electrical through excessive interaction[26], [31], [99], [101], [102], [108], [116], [119], [122], [123], [125], [127], [129], [130], [134], [135]. Palladium and nickel has been identified as the metals with lowest contact resistance with graphene[31], [96], [99], [102], [112], [119], [127], [129], [133], [135], [253], with Pd giving lower contact resistance results. The low adhesion of Pd to graphene necessitates the use of an adhesion layer. Titanium and chromium has been both proposed and used as adhesion layer candidates[101], [111].

To compare Ti/Pd vs Cr/Pd stacks, we fabricated 25 devices of each and fitted it to the constant mobility model[68], which provides a single value for the contact resistance. The fabrication procedure is similar to that outlined in Appendix B, without the top-gate metal deposition. The devices are similar to those shown in Figure ‎6.1, and where fabricated using CVD graphene on SiO2. We found that titanium consistently gives lower contact resistance, with less spread in statistical data; a box plot of the fitted data is shown in Figure ‎6.2.



Figure . Box plot of extracted contact resistance value for Cr/Pd and Ti/Pd metal stacks. The layer thickness was 1.5/40 nm for each stack. The data analysis was performed using measurements from 25 devices for each metal stack. The box plot whisker edges correspond to the extreme data points, the blue box limits correspond to the 75% and 25% percentile points and the red horizontal line is the median of the data. The results show the the Ti/Pd stack shows less spread and lower contact resistance compared to Cr/Pd.

The spread in Ti/Pd and Cr/Pd can be attributed to the formation of metal oxides at the adhesion layer/graphene interface[133]. The lower spread and median in Ti/Pd points towards less oxidation of Ti in our process relative to Cr. The spread of the data and the dependence of the stack contact resistance on the metal deposition conditions elucidates the importance of fully reporting the deposition conditions. When reporting contact resistance measurement data, deposition system-dependent characteristics (deposition rate, deposition pressure, purity of the metal source) can have a profound effect on the reported data, which can be contradictory if not information is given on the deposition conditions[8], [105], [110], [119], [123], [124], [127], [129], [133], [135].

### Thermal Interface Resistance

The thermal interface conductance affects how much Joule heating created by the electrical contact resistance flows through the channel, raising its temperature. Figure ‎6.3 shows a simplified contact thermal circuit depicting how Joule heating affects the temperature of the channel. Figure ‎6.3 is valid when a doped graphene FET is operating in low-field transport regime, causing a negligible amount of Joule heating inside the channel. This model is different from the high-field transport models, where the Fermi level spatial variation can cause the formation of a hot spot inside the channel at the position of the charge neutrality point[114], [254]–[256].



Figure . A simplified thermal circuit showing how heat flux generated by Joule heating distributes into the channel and the contact. The thermal current is the heat current generated due to Joule heating at the metal-graphene interface (), where is the electrical current flowing through the contact and is the electrical contact resistance. , , , and are the thermal resistances of the metal-graphene interface, metal contact, graphene channel, the graphene-substrate interface and substrate, respectively. All the grounds shown are thermal grounds, representing the temperature of the surrounding.

The amount of heat flux flowing through is given as:

|  |  |  |
| --- | --- | --- |
|  |  | (.) |

The temperature difference between the channel and the surrounding is that given as:

|  |  |  |
| --- | --- | --- |
|  |  | (.) |

The thermal resistance of the metal contact is typically low due to the high thermal conductivity of metals. We can thus approximate Equation (‎6.2) as:

|  |  |  |
| --- | --- | --- |
|  |  | (.) |

Equation (‎6.3) shows the thermal interface resistance plays a crucial role in determining how much heat flux flows into the channel, causing its temperature to increase. If the thermal interface resistance between the graphene and metal contact is low, most of the generated heat flux will flow out though the contact instead of through the channel. The metal-graphene thermal interface resistance is also important in the high-field transport regime, as it a low thermal interface resistance reduces the amount of temperature rise inside the channel[114].

Thermal interface resistance is a strong function of the interface and bonding between the metal and graphene. The interaction between the metal and the graphene plays a crucial role. Gold, palladium and nickel span the interaction spectrum from no interaction in the case of gold, weak interaction in the case of palladium, and strong interaction forming a carbide in the case of nickel. The corresponding thermal interface conductance varies from 20 MW/m2K for Au, 25 MW/m2K for Pd, and 45 MW/m2K for Ni.

Equation (‎6.3) highlights the importance of co-optimization of the electrical and thermal-interface contact resistances. The optimization of thermal contact resistance typically involves a proper choice of the metal contact material[117], [257], [258], and co-optimization should focus on the reduction of the electrical and thermal interface resistance product[117].

## Stress in dual-gated Graphene FETs during Processing

Stress in thin films is inversely proportional to the thickness of the film, and graphene being single layer, represents the lower limit of thickness. Graphene channels are typically sandwiched between two oxides; the fabrication of graphene FETs usually entails the deposition of a top-gate oxide, whether for use as a top-gate dielectric or for passivation[85], [259]. Even when fabricated over a single oxide layer, graphene being the lowest thickness element in the sack is the most prone to stress, and stress affects the electronic properties graphene[2], [137], [142].

To study the evolution of stress in graphene during fabrication, we used Raman analysis. We used the 2D peak position as a metric to evaluate the evolution of strain in the graphene channel after each process of fabrication outlined in Appendix B [137], [141], [142], [144], [145], [148], [149], [154], [157]. We evaluated the stress in the graphene channel before processing, after contacting, after ALD seed evaporation and after ALD of the top-gate oxide. We performed the experiment using 3 different photoresist and developer pairs and the results were consistent across all 3 pairs. The Raman analysis was performed on the photolithography defined graphene channel simislar to that shown in Figure ‎6.4. The Raman analysis was performed on data obtained from area scans over the exposed channels.



Figure . Optical microscope image of graphene FET device used to evaluate the stress evolution during fabrication. The graphene channels inside the red box was analyzed using Raman analysis to evaluate the stress during fabrication.

The 2D peak position was used as a metric for the strain in the channel. The 2D red shifts under tensile strain. The exact value of the strain was not calculated from the 2D peak shift due to the different values of the reported shift rates[148], [142], [149], [154], [144], [150].

The Raman spectrum from a single point and an aggregate of the 2D peak position after each fabrication step is shown in Figure ‎6.5. The single point spectrum shown in Figure ‎6.5(a) shows that the 2D peak red shift slightly after contacting indicating a slight tensile strain of the channel, followed by a distinctive shift in after the Al2O3 seed layer deposition followed by a return to its original position after the completion of the top gate oxide ALD.

|  |  |
| --- | --- |
| (a) | (b) |

Figure . (a) Raman spectrum around the 2D peak, (b) 2D peak data aggregate across the graphene channel after different fabrication steps. The arrows in (a) show clarify the evolution of the 2D peak position. The 2D peak position initially redshifts, indicating tensile strain, after contacting. The red shift is maximum after the Al2O3 seed layer deposition, indicating that this is the process inducing the maximum strain in the graphene channel. The 2D peak position almost moves back to position prior to the seed deposition after ALD. The 2D peak FWHM is not included in the analysis, as it is a weak function of the strain and due to its high sensitivity to impurities and substrate roughness.

The aggregate of 2D peak positions from the scan of the graphene channel shown in Figure ‎6.5(b) confirms the behavior shown in Figure ‎6.5(a). Initially, the 2D peak positions are scattered over a range of peak positions due to the roughness of the substrate[56]. Once the channel is contacted the spread of the points reduces significantly. This reduction of 2D peak position can be attributed to the contacts anchoring the graphene ribbon of the substrate with a given tension, because the contacts typically spread beyond the graphene, effectively pinning it to the substrate. After the depositions of the Al2O3 seed layer (1.5 nm of aluminum left to natively oxidize in air) the 2D peak positions redshift significantly. This can be attributed to the high tension in the graphene between the high-stress thin top oxide and the substrate. Graphene has a negative thermal expansion coefficient unlike Al2O3 and SiO2 [145], exposing it to a high tensile stress by the thin top oxide seed layer. After the completion of the ALD, which takes about 1 hour at 150 °C, the peak position almost move back to their original position, blue shifting by a small amount (~ 3 cm-1). This can be attributed to the thickness increase of the top gate oxide and the thermal processing causing a reduction in the overall stress in the top gate oxide, and the graphene channel subsequently. We cannot ascertain whether the long time anneal or the thickness increase is the cause of stress reduction, but experimenting with different growth times and temperatures could allow the evaluation of the effect of each parameter more accurately.

## Impact of Metal-Ion Containing Developers on the Performance of Graphene FETs

Photoresist developer chemistries can be divided into metal-ion free and metal-ion containing. The choice of a developer chemistry depends on its compatibility and if it has desirable properties with respect to the fabrication process.

We studied the impact of metal-ion free (MIF) developer and metal-ion containing (MIC) developers on the performance of graphene FET transistors. We used CD26 (TMAH-based developer) as the MIF developer, for use with Shipley S1805 photoresist with LOR3A lift-off layer, and used AZ Developer (Na-based developer) and AZ400K (K-based developer) as MIC developers, for use with AZ4110 photoresist. We fabricated 25 devices using each photoresist using commercial CVD graphene and evaluated the performance of the devices by fitting their characteristics to the constant mobility model[68]. Electrical measurements included a hysteresis sweep of the gate to extract the resulting hysteresis in Dirac point voltage occurring because of interface trapped charges between the graphene and SiO2 substrate[45], [46], [50]–[52], [54], [57], [58], [82], [85], [259]–[261]. The devices were fabricated using the same steps outlined in Appendix B, but without the top gate metal deposition. The devices were measured using the setup outlined in Appendix C. Each box corresponds to measurements from 25 devices. In all box plot figures in this section, the box plot whisker edges correspond to the extreme data points, the blue box limits correspond to the 75% and 25% percentile points and the red horizontal line is the median of the data.

Figure ‎6.6 shows the Dirac point voltage obtained using the different developers. The data shows that the MIF CD26 developer yields the lower Dirac point voltage. This is expected as the lack of metal ions in the MIF developer causes less doping compared to the MIC developers, which dope the graphene significantly due to the metal ion content.

The hysteresis of Dirac point voltage shown in Figure ‎6.7 shows that the MIC developers have the lowest magnitude of hysteresis. This can be attributed to the screening of graphene-dielectric interface charges by the metal-ion dopants, reducing their effect in causing hysteresis.

The extracted charged-ion concentration is shown in Figure ‎6.8. It confirms that the charged-impurity concentration of MIF developers is considerably less than that obtained using MIC developers. This results is expected, as the more metal ions are essentially charged impurities.

The extracted mobility is shown in Figure ‎6.9. We note that MIF developers yield significantly higher charge carrier mobility compared to MIC developers. This result can be attributed to the reduction in charged-impurity scattering when using MIF developers as they cause less charged-impurity concentration.

The extracted contact resistance normalized per device width is shown in Figure ‎6.10. Although the median values are close, the spread is significantly lower when using MIF developers. This indicates a more consistent metal-graphene interface as the contact resistance is a strong function of the state of the metal-graphene interface.

The results conclusively indicate the MIF developers are superior when processing graphene compared with MIC developers, unless the graphene channel is to be doped intentionally.



Figure . Dirac point voltage measured for the devices fabricated using different photoresist. Labels "Up" and "Down" correspond to sweeping the back-gate voltage up and down.

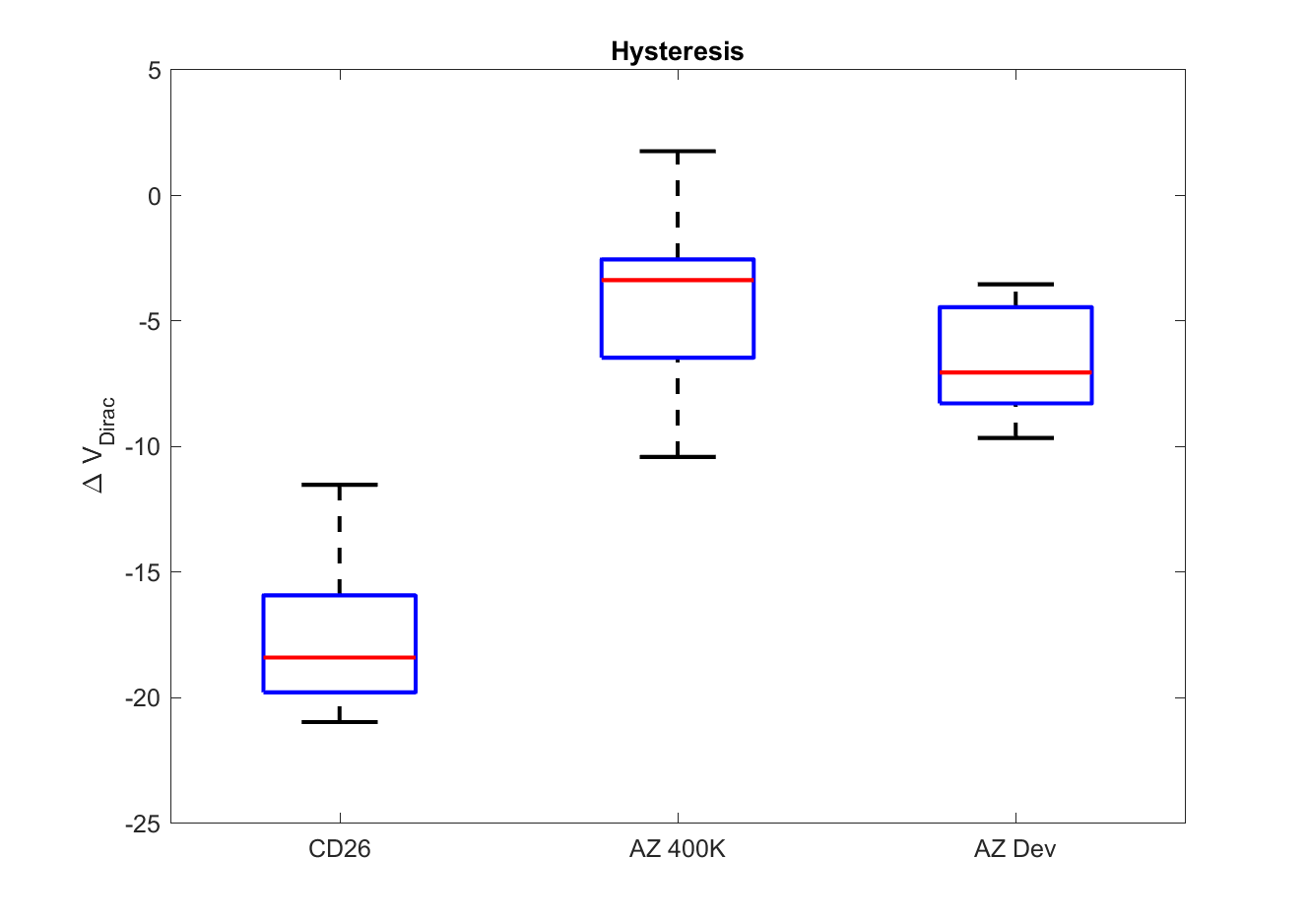


Figure . Hysteresis in Dirac point voltage, defined as the difference between the up and down voltage sweep Dirac point voltages. The MIC developers show small hysteresis, which can be attributed to the large density of metal-ion doping, screening the interface charges and reducing the hysteresis caused by them.



Figure . Extracted charged-impurity concentration using different developers. As expected, the MIF developers cause the least density of charged-impurities.



Figure . Extracted charge carrier mobility using different developers. MIF developers show higher mobility, which can be attributed to the less charged-impurity scattering as they yield lower charged-impurity concentration.



Figure . Extracted contact resistance. The median values are not far apart considerably, but MIF developers show the lowest spread, indicating the most consistent metal-graphene interface.

### Relation Between Charge Carrier Mobility and Charged-Impurity Concentration

Charge carrier mobility () and charged-impurity concentration () are inversely proportional[4]. It was shown theoretically that the charge carrier m [68]. The study of impact of MIF and MIC developers on graphene FET performance allowed us to experimentally evaluate the relation between charge carrier mobility and charged-impurity concentration.

The extracted charge carrier mobility and the inverse of charged-impurity concentration is shown in Figure ‎6.11. The data shows a clear inverse relationship between and . To evaluate the proportionality constant between the inverse relationship, the data was fit to using the equation . The best fit shows , which is significantly different from the theoretically predicted data. This result suggests that the charged-impurity scattering is stronger than what has been predicted theoretically.



Figure . Extracted charge carrier mobility vs inverse of charged-impurity concentration. the results show a clear inverse relation between the charge carrier mobility and charged-impurity concentration. The best fit equation is . The dashed line shows the extension of the fit to the origin.

## Photoresist Residuals on Graphene Channels

Photoresists are organic, which makes them adhere well to graphene. However, this high adhesion to graphene causes resist residuals to stay on the graphene surface, increasing the contact resistance and acting as a dielectric barrier between the graphene and top gate oxide. The use of oxygen plasma, or milder ozone, to remove the organic contaminants on the graphene and enhance the contact resistance has been reported in the literature[86], [113], [115], [131].

We studied the residual left over graphene channels after the channel definition step to evaluate how much residual is left by different residues. We used AFM analysis to evaluate the step height over the graphene channel right after removing the sacrificial photoresist layer used as an etch mask, as outlined in Appendix B. The step heights using the different photoresists and developers used in Section ‎6.3 are shown in Figure ‎6.12.

The CD26 developer with S1805 resist shows the smallest step height, 1.19 ± 0.16 nm, which is quite close to the idea AFM step height of graphene, indicating a very small residue[17], [260], [151].

|  |  |
| --- | --- |
| (a) | (b) |
| (c) | (d) |

Figure . (a) False color SEM image showing the graphene FET with the scan area shown as a yellow line. (b)-(d) AFM scan results for when using CD 26, AZ Developer and AZ400 developer, respectively.

On the contrary, the step height channels defined using AZ Developer and AZ400K with AZ4110 photoresist shows a step height in the excess of 10 nm. These results point towards high adhesion between the photoresist and graphene. The large step height also mean that the interface of graphene is covered with a relatively thick residue, which will certainly affect the transport and interface properties. This might be one of the reasons behind the large spread in extracted model values in Section ‎6.3 for the devices fabricated using AZ Developer and AZ400K.

# Conclusion and Future Work

## Alloyed Contacts for Optimized Thermal and Electrical Contact Resistance

## Oblique incidence diodes using exfoliated graphene encapsulated by hBN

## CVD Graphene over CVD hBN

## Mixing 2D Materials for Optimized Contacts

# References

1. Graphene Processing
   1. Preprocessing Anneal

CVD graphene transfer typically requires spinning a PMMA support layer over the graphene to hold it when etching away the growth substrate (typically copper)[83]. PMMA residue affect the transport properties of graphene and increase the contact resistance. To remove the PMMA residue, graphene post transfer is annealed in a forming gas at around 300 °C to remove the PMMA residuals[14], [79], [83], [115], [134], [262], [263].

Throughout this work, the preprocessing anneal was performed at a temperature of 270-300 °C, for at least 4 hours, 5% H2 in Ar by volume (25 sccm H2 in 475 sccm of Ar). The process was performed in a 2” quartz tube at atmospheric pressure.

* 1. Photoresist spin coat

The spin coating parameters depended on the photoresist used. The following parameters summarizes the photoresist spin coating parameters used throughout this work:

* S1805 Photoresist:
  1. Spin coat LOR3A lift-off layer at 4000 RPM for 45 seconds.
  2. Post spin bake at 190 °C for 5 minutes.
  3. Spin coat S1805 resist at 4000 RPM for 45 seconds.
  4. Post spin bake at 115 °C bake for 5 minutes.
* AZ4110 Photoresist:
  1. Spin coat AZ4110 photoresist at 5000 RPM for 90 seconds
  2. Post spin bake at 5 minute bake at 95 °C
  3. Photolithography

The lithography was performed using Karl Süss MA6 contact aligner in contact mode. The light source was an i-line, 320 nm source calibrated to an intensity of 5 mW/cm2. The photolithography masks used were 4”x4”x0.090” Quartz AR with a chrome mask layer, procured from Photronics Inc.

The photolithography parameters depended on the photoresist used. The following parameters summarize the photolithography exposure and development procedure used throughout this work. Note that prior to the exposure of the sample, a dummy sample was first exposed to ensure proper source calibration:

* S1805 Photoresist:
  + Exposure Time: 18 seconds
  + Development parameters: 1 minute in CD26 with gentle agitation followed by a 30 second rinse in gently running water.
* AZ4110 Photoresist with AZ400K developer:
  + Exposure Time: 45 seconds
  + Development parameters: 1 minute in 1:4 AZ400K:DI (by volume) with gentle agitation followed by a 30 second rinse in gently running water.
* AZ4110 Photoresist with AZ developer:
  + Exposure Time: 35 seconds
  + Development parameters: 1.7 minutes in 1:1 AZ400K:DI (by volume) with gentle agitation followed by a 30 second rinse in gently running water.

Lift-off was performed in Microposit 1165 remover kept in a closed beaker on a hot plate set to 70 °C.

* 1. Electron-Beam Lithography

PMMA A7 was used as an e-beam resist throughout this work. The following steps were used during e-beam lithography:

1. Spin coat PMMA A7 at 5000 RPM for 90 seconds.
2. Post spin bake at 180 °C or 190 °C for 90 seconds.
3. Exposure parameters:
   1. Working distance: 6.5-7.5 mm.
   2. Acceleration voltage: 30 kV.
   3. Spot Size: 3.
   4. Beam current: ~136 pA.
   5. Magnification: Depends on pattern size. Typically 1500x.
   6. Point-to-Point Spacing: 5 nm.
   7. Line-to-Line Spacing: 30 nm.
   8. Line dose: 1.6 nC/cm.
   9. Area dose: 450-600 µC/cm2.
4. Development parameters:
   1. Developer: 1:3 MIBK:IPA (by volume)
   2. Development time: 55 seconds with gentle agitation followed by 15 seconds in IPA with gentle agitation
   3. Electron-Beam Evaporation

Samples were loaded over a 3” carrier wafer using 90 °C heat release tape. The samples were left to pump out in vacuum for at least 3 hours, typically overnight reaching a base pressure in the low 10-7 Torr, in the Ultek E-Beam evaporator. The current then was increased slightly till the pressure started going down (this only occurs for the first layer after the pump down when the evaporated metal reacts with the water vapor and radicals sticking to the evaporation chamber walls), while ensuring the e-beam is properly focused and centered on the source crucible. The shutter was opened once the chamber pressure started going up.

The rate for all deposition over graphene was kept between 0.1 – 0.3 Å/sec to enhance the adhesion between the metal and the graphene. This was kept for the first 3-5 nm then the rate was increased to the values typically used with that metal in the evaporation chamber.

The typical evaporation currents were as follows:

* Cr: 20 mA
* Ti: 30-40 mA
* Pd: 40-50 mA
* Al: 40 mA
* Au: 50 mA

After the evaporation run was complete the chamber was let to cool down for 15 minutes before venting.

* 1. Top Gate Oxide/Passivation Deposition

The top gate oxide used throughout this work was Al2O3 due to its high reported mobility[68]. To facilitate the ALD of oxide on graphene we used a seed layer of evaporated aluminum left to oxidize in air[68]. The top gate oxide deposition parameters throughout this work was as follows:

1. Evaporate 1.5 nm of aluminum then leave it in air for at least 30 minutes.
2. Load the sample into ALD machine with the growth chamber temperature set and stable at 150 °C.
3. Grow 8.5 nm of Al2O3 using a TMAH based precursor with H2O as an oxygen source (thermal growth NOT plasma assisted growth). The growth rate was about 1 Å/cycle. The 8.5 nm deposition took around an hour.

The sample was removed from the chamber once the run was done.

* 1. Graphene etching

Graphene etching was done using O2 plasma in a Reactive-Ion Etching machine (Plasma-Therm 790 RIE). We used this machine because initial experimentation with the IPC Barrel Etcher showed excessive lateral etching underneath the etch mask sacrificial layer. On the other hand, the RIE machine yielded cleaner etches with less lateral etching. The etch parameters in the Plasma-Therm 790 RIE are as follows:

* Recipe name: mdgphn1m
* Chuck used: Aluminum
* Etch time: 1 minute
* Power: 20 W
* Gases: 14:6 O2:Ar
* Pressure: 1 mT
* DC Voltage: ~ 180V (automatically controlled)
  1. Backside oxide etch

Commercially available graphene and thermally-grown SiO2 over silicon substrates come with an oxide on the on both the top-side and back-side of the substrate, that is, the oxide is on both sides of the silicon carrier wafer. The top-side oxide is grown over the finished silicon face and acts as the back-gate oxide. The top-side oxide is the back-gate oxide that carrier the transferred graphene. The back-side oxide is a byproduct of the thermal growth process. The back-side oxide prevents electrical contact to the silicon substrate and should be removed prior to any measurements. This can be confirmed by optically inspecting the back side of the carrier wafer: if the backside is any color other than grey then it is covered with an oxide. Pictures depicting the back side of the sample (graphene facing down) are shown in Figure ‎A.1, showing the back side before and after the removal of the back-side oxide.

|  |  |
| --- | --- |
|  |  |

Figure .(a) Before, and (b) after the removal of the back-side oxide from the silicon sample. The back-gate oxide is deposited on the top-side of the sample, while the back-side oxide growth is a byproduct of the thermal oxide growth process.

Before doing any processing on the back side, the top-side oxide and graphene should be covered by spin coating PMMA or photoresist over it to prevent scratching the graphene during the etch of the back-side oxide. The back-side oxide can be removed using RIE or by floating it over BHF. We prefer the use of RIE etching because it is more safe and less prone to accidental damage to the graphene features if the sample falls into the BHF. When using an RIE oxide etch recipe, the etch time and recipe should be carefully set up to prevent, and remove, any polymer formation on the silicon surface as it would increase the contact resistance to the back gate (the silicon).

1. Fabrication of a Dual-Gated Graphene FET

We go through the details of fabricating a dual-gated graphene FET. This procedure is identical for neurons, synapses and diodes, except for the e-beam mask used to shape the top gate. The processing parameters follow the values outlines in Appendix A. The steps are numbered according to their order. Note that the devices presented in this work were all fabricated using CVD graphene over 285 nm of thermal SiO2, procured from Graphene Supermarket. The top gate oxide was 10 nm of Al2O3, and the top gate metal was 25 nm thick, e-beam evaporated aluminum. The contact stack was Ti/Pd with thicknesses 1.5/40 nm.

Preprocessing Anneal

The preprocessing anneal is crucial to remove any PMMA residual from the graphene transfer step and is typically performed for at least 4 hours before any fabrication.

Contact Photolithography

1. Spin coating the photoresist and contact layer lithography as shown in Figure ‎B.2

|  |  |
| --- | --- |
| (a) | (b) |

Figure . Sample cross-section after (a) Spin coating, and (b) photolithography and development of contact layer.

1. Contact metal stack evaporation and lift-off as shown in Figure ‎B.3.

|  |  |
| --- | --- |
| (a) | (b) |

Figure . Sample cross-section after (a) contact stack evaporation, and (b) contacts lift-off.

Figure ‎B.4 shows an optical microscope image of the sample after steps (1) and (2).

|  |  |
| --- | --- |
| (a) | (b) |

Figure . Optical microscope image after (a) contact photolithography, and (b) contacts lift-off.

Channel Definition

1. Spin coating the photoresist and the channel etch mask as shown Figure ‎B.5.

|  |  |
| --- | --- |
| (a) | (b) |

Figure . Sample cross-section after (a) Spin coating, and (b) photolithography and development of channel etch mask.

1. O2 plasma etching of the exposed graphene using the RIE and removal of the etch mask as shown in Figure ‎B.6.

|  |  |
| --- | --- |
| (a) | (b) |

Figure . Sample cross-section (a) during channel definition, and (b) after etch mask removal. The red arrows signifiy the O2 plasma etching the exposed graphene.

Figure ‎B.7 shows an optical microscope image of the sample after steps (1) and (2). Figure ‎B.7(a) shows the tears in the graphene channel prior to the etch. After the etch no graphene is visible outside the etch mask as seen in Figure ‎B.7(b).

|  |  |
| --- | --- |
| (a) | (b) |

Figure . Optical microscope image after (a) contact photolithography, and (b) contacts lift-off.

Top Gate Oxide/Passivation Deposition

ALD seed layer is evaporated then the ALD is performed over the seed as shown in Figure ‎B.8. The top gate oxide also acts as a passivation to protect the channel from the environment. In our experiments, 10 nm of Al2O3 was enough to protect the graphene from environmental dopants.

|  |  |
| --- | --- |
| (a) | (b) |

Figure . Sample cross-section after (a) seed layer evaporation, and (b) top oxide ALD.

Top Gate Lithography

This step is only required if a top gate is needed. The step is used for the fabrication of the neurons, synapses or diode oblique gates. The only difference is the e-beam mask used to pattern the top gate.

1. Spin coating the e-beam resist and the e-beam lithography of top gate as shown in Figure ‎B.9.

|  |  |
| --- | --- |
| (a) | (b) |

Figure . Sample cross-section after (a) Spin coating, and (b) e-beam lithography and development of top gate deposition mask.

1. Evaporation of the top gate metal stack and lift-off as shown in Figure ‎B.10.

|  |  |
| --- | --- |
| (a) | (b) |

Figure . Sample cross-section after (a) top gate metal stack evaporation, and (b) top gate lift-off.

Figure ‎B.11 shows an optical microscope image and a false color SEM image of a graphene synapse after step (2).

|  |  |
| --- | --- |
| (a) | (b) |

Figure . (a) Optical microscope image of graphene synapse, and (b) False color SEM image of the same device.

Back-side oxide etch

1. Spin coating the top side of the sample (devices side) with photoresist or PMMA and bake it.
2. Use the RIE to etch the back side oxide, while putting protected device-side down, or put the sample gently over BHF, while keeping the protected device-side up, making sure to put it down softly so as it floats over the BHF. The BHF will etch only the back side oxide. Sometimes, the BHF vapor seeps under the protection layer damaging some of the oxide and devices near the edges and corners of the sample.

Contact window opening

This step is optional and only required if the top gate oxide it too thick or too hard for the probes to punch through it, or if the devices are to be wire bonded.

1. Spin coating the photoresist and the photolithography of contact openings as shown in Figure ‎B.12.

|  |  |
| --- | --- |
| (a) | (b) |

Figure . Sample cross-section after (a) Spin coating, and (b) photolithography and development of oxide etch mask.

1. Ion milling or RIE etching of the exposed oxide and removal of the etch mask as shown in Figure ‎B.13. It is recommended to use O2 plasma etching to remove any the top most crust of the photoresist prior to dissolving the etch mask as the crust is typically hardened by ion milling and removing it enhances dissolving of the etch mask.

|  |  |
| --- | --- |
| (a) | (b) |

Figure . Sample cross-section (a) during oxide etching, and (b) after etch mask removal. The blue arrows signifiy the accelerated ions used for ion millng or the reactive species if RIE is used for etching the exposed oxide.

Measurement Setup for Dual-Gated Graphene FETs

Derivation of the Current Coupling Coefficient in a Diffusive-Transport Graphene Coupler



Figure . Electrical model of graphene coupler. The graphene ribbons are modelled using two distributed resistors with a resistance per unit length of R1 and R2, and the tunneling resistance coupling them is modelled using a distributed conductance with conductance per unit length gc.

The electrical model of the device is composed of three distributed resistors: a distributed resistor for each of the graphene ribbons with a distributed tunneling conductance connecting them together, as shown in Figure ‎5.3. We label one of the ribbons as the input ribbon, and the other as the output ribbon. For this analysis, we apply a current stimulus at the input ribbon and calculate the current at other end (output) of each ribbon.

We start by solving the voltage differential equation for the distributed system then relating it to the current.

Applying KCL at any node the node (i), we get:

This can be rewritten as:

Taking the limit as , we get:

|  |  |  |
| --- | --- | --- |
|  |  | (.) |

Substituting from the two equations together to decouple the equations we get:

Accordingly, the differential equation for each branch after substitution:

The resulting decoupled equations are:

|  |  |  |
| --- | --- | --- |
|  |  | (.) |

|  |  |  |
| --- | --- | --- |
|  |  | (.) |

The two differential equations are the same. The general solution is written as:

|  |  |  |
| --- | --- | --- |
|  |  | (.) |

The above equation can be rewritten by setting to be:

|  |  |  |
| --- | --- | --- |
|  |  | (.) |

Current conservation dictates the boundary conditions on the current as:

|  |  |  |
| --- | --- | --- |
|  |  | (.) |

Ohm’s law relates the current and voltage at any given position as:

|  |  |  |
| --- | --- | --- |
|  |  | (.) |

In addition, Ohm’s law as relates the voltage and current at the output end of each ribbon:

|  |  |  |
| --- | --- | --- |
|  |  | (.) |

The coupling coefficient between the current in the two ribbons is defined as the ratio between the output ribbon and input ribbon branch currents as:

|  |  |  |
| --- | --- | --- |
|  |  | (.) |

Equations (‎E.1) and (‎E.4) can be solved for the relation between the constant to give:

|  |  |  |
| --- | --- | --- |
|  |  | (.) |

Before proceeding to solve the equation, we note that the voltage and current equation present a system of two coupled second order linear equations, reducible to two fourth order decoupled ordinary differential equations. Ohm’s law relates the voltage gradient to the current and hence, the current continuity equation poses a condition on the first derivative of the voltage, while Equation (‎E.8) serves as a Robin boundary condition relative the voltage to its derivative at the boundary. Accordingly, the current system cannot be completely solved analytically; we will not be able to obtain the values of the four constants needed to fully determine a unique solution, but it can be solved numerically. In this discussion, we provide an incomplete solution that does not determine all the unknown constants, but reveals the functional form of the solution.

By letting , we can write the current in each branch using the left side of Equation (‎E.7) as:

|  |  |  |
| --- | --- | --- |
|  |  | (.) |

From the current conservation equation (‎E.6) we obtain the requirement on the constant , allowing us to rewrite Equation (‎E.11) as:

|  |  |  |
| --- | --- | --- |
|  |  | (.) |

The voltage across each ribbon is this given as:

|  |  |  |
| --- | --- | --- |
|  |  | (.) |

Although it is quite tempting to null the increasing exponential constant , its presence is important in maintaining the consistency of the equations. This is can be seen through applying the boundary condition given by Equation (‎E.8):

|  |  |  |
| --- | --- | --- |
|  |  | (.) |

If the constant, is set to zero the while , Equation (‎E.14) reduces to:

|  |  |  |
| --- | --- | --- |
|  |  | (.) |

Equation (‎E.15) can only be satisfied if D is also nulled. This result is erroneous as it means that the current will not change regardless of the values of if the output of the device is shorted. In line with Equation (‎E.14), we can extract the value of as:

|  |  |
| --- | --- |
|  | (.) |

Equation (‎E.16) shows that the coefficient of the exponential increasing term decays exponentially with the length of the device and will not cause an unphysical increase in the voltage or current across the device length.

To sum up, we can write the functional form of the voltage and current across the coupler as:

|  |  |  |
| --- | --- | --- |
|  |  | (.) |

|  |  |  |
| --- | --- | --- |
|  |  | (.) |

The solution reveals the functional form of the current to distribute between the two lines in an asymptotic fashion. The current asymptote is roughly given by the current division ratio had the two branches been connected only at the input end. The asymptotic behavior roughly follows an exponentially decaying function with a characteristic length.

To demonstrate the functional behavior, we study the case when the coupler branches have matched impedance with their loads, i.e. when. In such a case, Equation (‎E.16) reduces to:

|  |  |  |
| --- | --- | --- |
|  |  | (.) |

For all practical purposes, the value of and thus the contribution of the exponentially increasing term in Equation (‎E.17) and Equation (‎E.18) can be neglected to give:

|  |  |  |
| --- | --- | --- |
|  |  | (.) |

|  |  |  |
| --- | --- | --- |
|  |  | (.) |

A useful approximation that simplifies the analysis considerably is to assume . This assumption is valid especially when. This assumption is especially valid in our analysis, as is a tunneling conductance that is considerably small relative to the conductance of the either branches of the coupler. Under this assumption, and thus we can rewrite Equation (‎E.21) as:

|  |  |  |
| --- | --- | --- |
|  |  | (.) |

Equation (‎E.22) demonstrates the behavior of the coupler under the matching approximation, while neglecting . The current in each branch asymptotically approaches its value had the two ribbons been connected only at the input side, with an asymptotic behavior following an exponential function with a characteristic length of . In this case the current coupling coefficient is given as:

|  |  |  |
| --- | --- | --- |
|  |  | (.) |

In the limit when , the current coupling coefficient at the output end of the coupler is given as:

|  |  |  |
| --- | --- | --- |
|  |  | (.) |