

TUNNEL FIELD EFFECT TRANSISTORS: FROM THEORY TO APPLICATIONS

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Mingda Li

May 2018

© 2018 Mingda Li

ALL RIGHTS RESERVED

TUNNEL FIELD EFFECT TRANSISTORS: FROM THEORY TO
APPLICATIONS

Mingda Li, Ph.D.

Cornell University 2018

The performance of computing systems has been increasingly choked by power consumption and memory access time within and between system components. Meanwhile, the explosion of artificial intelligence requires massive data-heavy computation. Therefore, it is crucial to develop energy efficient computing from devices to architectures. This work is developed along three streams: a steep device with low operation voltage, a novel device enabling complex logic operation, and an efficient modeling algorithm to quickly incorporate emerging devices into circuit designs. On the first front, tunnel field effect transistors (TFETs), which switch by modulating quantum tunneling, promise sub-60 mV/dec subthreshold swing and operate at low power consumption. Based on the unique properties of atomically thin 2D layered materials, two-dimensional heterojunction interlayer tunneling field effect transistor (Thin-TFET) was proposed as a ultra-scaled steep transistor. On the second front, we converted the “undesirable” ambipolar behavior in TFETs into XNOR logic operation, and proposed a one-transistor XNOR design: TransiXNOR. On the third front, we structured artificial neural networks with awareness of device physics, and developed an accurate, efficient, and generic device compact modeling algorithm: physics-inspired neural network (Pi-NN).

Mingda Li

CONTACT INFORMATION	2250 N. Triphammer Rd., Apt. H2F Ithaca, NY 14850	574-339-1802 ml888@cornell.edu
RESEARCH FOCUS	Developing algorithms for physical and empirical modeling of electronic devices	
EDUCATION	Cornell University , Ithaca, NY Ph.D., Electrical and Computer Engineering, <i>Expected: May 2018</i> <ul style="list-style-type: none"> • Thesis Topic: <i>Tunneling Field Effect Transistor: From Theory to Applications</i> • Committee: Prof. Huili Grace Xing, Prof. Claire Cardie and Prof. Debdeep Jena University of Notre Dame , Notre Dame, IN M.S., Electrical Engineering, Jan 2015	
	Fudan University , Shanghai, P. R. China B.S., Microelectronics, July 2012	
PUBLICATIONS	<ol style="list-style-type: none"> 1. Qin Zhang, Mingda Li, Edward B. Lochocki, Suresh Vishwanath, Xinyu Liu, Rusen Yan, Huai-Hsun Lien, Małgorzata Dobrowolska, Jacek Furdyna, Kyle M. Shen, Guangjun Cheng, Angela R. Hight Walker, David J. Gundlach, Huili G. Xing, and N. V. Nguyen, “Band offset and electron affinity of MBE-grown SnSe₂”, <i>Applied Physics Letter</i>, 112, 042108, 2018 2. Mingda Li, Ozan İrsoy, Claire Cardie, and Huili Grace Xing, “Physics-Inspired Neural Networks for Efficient Device Compact Modeling”, <i>IEEE Journal on Exploratory Solid-State Computational Devices and Circuits</i>, vol. 2, pp. 44-49, 2016. 3. Mingda Li, Rusen Yan, Debdeep Jena, and Huili Grace Xing, “Two-dimensional Heterojunction Interlayer Tunnel FET (Thin-TFET): From Theory to Applications”, <i>IEEE International Electron Devices Meeting (IEDM)</i>, pp. 19.2.1-19.2.4, 2016. 4. Mingda Li, Shudong Xiao, Rusen Yan, Suresh Vishwanath, Susan Fullerton-Shirey, Debdeep Jena, and Huili Grace Xing, “Fermi Level Tunability of a Novel 2D Crystal: Tin Diselenide (SnSe₂)”, <i>Device Research Conference (DRC)</i>, pp. 1-2, 2016. 5. Nhan Nguyen, Mingda Li, Suresh Vishwanath, Rusen Yan, Shudong Xiao, Huili Xing, Guangjun Cheng, Angela Hight Walker, Qin Zhang, “Internal Photoemission Spectroscopy of 2-D Materials”, <i>APS March Meeting</i>, abstract R46.001, 2016 6. Mingda Li, David Esseni, Gregory Snider, Debdeep Jena, and Huili Grace Xing, “Two-dimensional Heterojunction Interlayer Tunneling Field Effect Transistors (Thin-TFETs)”, <i>IEEE Journal of the Electron Devices Society</i>, vol. 3, no. 3, pp. 200-207, 2015. 7. Rusen Yan, Sara Fathipour, Yimo Han, Bo Song, Shudong Xiao, Mingda Li, Nan Ma, Vladimir Protasenko, David A. Muller, Debdeep Jena, and Huili Grace Xing, “Esaki Diodes in van der Waals Heterojunctions with Broken-Gap Energy Band Alignment”, <i>Nano Letters</i> 15(9), 5791-5798, 2015 8. Mingda Li, David Esseni, Debdeep Jena, and Huili Grace Xing, “Lateral Transport in Two-dimensional Heterojunction Interlayer Tunneling Field Effect Transistor (Thin-TFET)”, <i>Device Research Conference</i>, pp. 17-18, 2014. 	

9. Shudong Xiao, **Mingda Li**, Alan Seabaugh, Debdeep Jena and Huili Grace Xing, “Vertical heterojunction of MoS₂ and WSe₂”, *Device Research Conference*, pp. 169-170, 2014
10. **Mingda Li**, David Esseni, Gregory Snider, Debdeep Jena, and Huili Grace Xing, “Single Particle Transport in Two-dimensional Heterojunction Interlayer Tunneling Field Effect Transistor”, *Journal of Applied Physics* vol.115, pp. 074508, 2014.
11. Debdeep Jena, **Mingda Li**, Nan Ma, Wan Sik Hwang, David Esseni, Alan Seabaugh, Huili Grace Xing, “Electron transport in 2D crystal semiconductors and their device applications”, *2014 Silicon Nanoelectronics Workshop (SNW)*, Honolulu, HI, pp. 1-2, 2014
12. Esseni, David, Marco G. Pala, Alberto Revelant, Pierpaolo Palestri, Luca Selmi, **Mingda Li**, Gregory Snider, Debdeep Jena, and Huili Grace Xing. ”Challenges and Opportunities in the Design of Tunnel FETs: Materials, Device Architectures, and Defects.” *ECS Transactions* 64, no. 6: 581-595, 2014
13. Huili Grace Xing, Guangle Zhou, **Mingda Li**, Yiqing Lu, Rui Li, Mark Wistey, Patrick Fay, Debdeep Jena, Alan Seabaugh, “Tunnel FETs with tunneling normal to the gate”, *Third Berkeley Symposium on Energy Efficient Electronic Systems (E3S)*, pp. 1-1, 2013
14. Guo, Jiaojiao, **Mingda Li**, Qingqing Sun, Wen Yang, Peng Zhou, Shijin Ding, and David Wei Zhang. “A Water-free Low Temperature Process for Atomic Layer Deposition of Al₂O₃ Films.” *Chemical Vapor Deposition* 19, no. 46: 156-160. 2013
15. Yongjun Li, **Mingda Li**, Jianshuang Liu, Qingqing Sun, Peng Zhou, Pengfei Wang, Shijin Ding, David Wei Zhang, “Atomic scale investigation of the abnormal transport properties in bilayer graphene nanoribbon”, *Applied Physics Letters* 100 (1), 013110, 2012

AWARDS	Best poster award at The 2nd International Symposium on Devices and Application of Two-dimensional Materials	2016
PATENT APPLICATION	Two-dimensional Heterojunction Interlayer Tunneling Field Effect Transistors (Thin-TFET) U.S. Patent Application No. 14/629,222	2015
WORK EXPERIENCE	Research Scientist, Facebook • Responsibility: Developing algorithms and models for personalized ranking systems.	2018

This dissertation is dedicated to my parents and Jiayan.

ACKNOWLEDGEMENTS

Looking back on the day I first met Dr. Huili Grace Xing and Dr. Debdeep Jena in Fudan University, Shanghai in 2012. It is their passion for scientific research, relentlessness in the pursuit of new knowledge and rigorous attitude toward science and engineering that I remembered and have been trying to follow for the last five and half years.

I would like to first express my thanks to Dr. Huili Grace Xing for her invaluable guidance in research and life. Whenever I find myself doubting how far I can reach during my Ph.D., I was always lucky enough to have Dr. Xing's encouragement to keep me moving on. I appreciate her constructive advices on my research, which greatly improved the quality and impact of my research. I am also thankful for the countless hours she has spent on revising my manuscripts, abstracts and dissertation, mentoring on my writing and presentation skills. It is only through these productive hours could I have become a better researcher.

Dr. Debdeep Jena's device physics classes are the most enjoyable and fruitful physics courses I have ever token. The skills and knowledge I learned from him, through both the classes and discussions, have been a important part of my research. His expertise in semiconductor device physics has greatly widen the scope of my research.

I am also fortunate to have Dr. Claire Cardie on my committee to make my studies more comprehensive and rigorous. Her introduction to natural language processing class sparks my interests in machine learning. Her inputs help me complete my first machine learning related paper, which in turn opens me to a career in machine learning after graduation.

I would also like to thank Dr. David Esseni for generous collaboration on

my first paper, Dr. Gregory Snider for teaching me the discipline of clean room works, Dr. Alan Seabaugh and Dr. Susan Fullerton for insightful suggestions, Dr. Mert Sabuncu for mentoring me in the lung cancer detection competition and introducing me several opportunities in medical research, and Dr. Lorenzo Alvisi for lovely coffee time talk about Italian cars and motorcycles.

I would also like to thank my fellow group members, who have generously offered much appreciated help on my research. These include: Mingda Zhu, Bo Song, Shudong Xiao, Xiang Li, Nan Ma, Guangle Zhou, Rusen Yan, Vladimir Protasenko, Zongyang Hu, Jashan Singhal, Hyunjea Lee, Malavika Attaluri, Ozan Irsoy, Wenshen Li, Suresh Vishwanath, Kazuki Nomoto, Kevin Lee, Jia Guo, Yuanzheng Yue, Guowang Li, Pei Zhao, Amit Verma, Meng Qi, Satyaki Ganguly, Moudud SM Islam, Wenjun Li, Wansik Hwang, Jimy Joe Encomendero Risco, Alexander Chaney, Brian Schutter, Sam Bader, Henryk Turski, Zexuan Zhang, Nicholas Tanen, Reet Chaudhuri, Shyam Bharadwaj, Joseph Casamento, John Wright.

I am thankful to the Center for Low Energy Systems Technology (LEAST) sponsored by the Semiconductor Research Corporation (SRC) and the Defense Advanced Research Projects Agency (DARPA), and the EFRI 2-DARE program funded by National Science Foundation. Sincere thanks go to them for creating these great research programs, which not only allowed me to carry out the research, but also to have interacted with experts in different fields.

I would also like to thank all the staff members in CNF at Cornell and NDNF at Notre Dame. Without their professional work that keeps the facilities in good operating conditions, I would not be able to complete the experiments presented in this work.

Last but not least, I would like to expressed my most sincere gratitude to my

parents and Jiayan for their unconditional support and encouragement. I thank them for making this work possible and a whole lot more.

TABLE OF CONTENTS

Dedication	6
Acknowledgements	7
Table of Contents	10
List of Tables	13
List of Figures	14
1 Introduction	1
1.1 Scalability of Power Consumption	1
1.2 Steep Slope and Tunneling	2
1.3 Previous Works on TFETs	5
1.4 Devices and Machine Learning	8
1.5 Brief Outline of This Work	10
Bibliography	11
2 Physical Modeling of Two-dimensional Heterojunction Interlayer Tunneling FETs (Thin-TFETs)	19
2.1 Introduction	19
2.2 Modeling of the Tunneling Transistor	22
2.2.1 Device Concept and Electrostatics	22
2.2.2 Transport Model	25
2.2.3 Effects of Energy Broadening	30
2.2.4 Rotational Misalignment and Tunneling Between In-equivalent Extrema	33
2.2.5 An Analytical Approximation for the Tunneling Current .	37
2.3 Numerical Results for the Tunneling Current	40
2.3.1 Parabolic Band Approximation	40
2.3.2 Effects of Correlation Lengths, Interlayer Thicknesses and Energy Broadening	41
2.4 N-type and P-type Thin-TFETs	46
2.4.1 Effects of Non-uniform van der Waals Gap Thickness and Access Resistance	51
2.4.2 Capacitance Evaluation	53
2.4.3 Benchmarking	54
2.5 Discussion and Conclusions	58
2.5.1 Experimental Insights	58
2.5.2 Conclusions	59
Bibliography	62

3 Comparative Study of Intrinsic Capacitances of Thin-TFETs and p-in-TFET	70
3.1 Enhanced Miller Effect of TFETs	70
3.2 Effects of TFET Geometries: “lateral” TFETs vs. “vertical” TFETs .	70
3.3 Numerical Simulations of C-V Curves	74
3.3.1 Simulation Methods	74
3.3.2 Simulation Results with Different Undercut/Underlap Lengths	76
3.4 Complimentary TFET Inverters	80
3.5 Conclusion	84
Bibliography	85
4 XNOR-enabled Transistor (TransiXNOR) for Binarized Neural Network Accelerator	86
4.1 Introduction	86
4.2 Dual-gated XNOR-enable Transistor: TransiXNOR	90
4.2.1 Device Working Principle	90
4.2.2 Simulation Approach	92
4.2.3 Results and discussion	93
4.3 TransiXNOR Crossbar Architecture for Binary matrix-vector Multiplication	96
4.4 Conclusion	99
Bibliography	101
5 Artificial Neural Networks (ANNs) for Device Compact Modeling	105
5.1 Introduction	105
5.2 Previous Works	108
5.2.1 Low Current Regime Challenge	110
5.3 The Idea of Pi-NN: Structured Physical System	113
5.4 Adjoint Sensitivity Network	116
5.5 Weighted L1 Loss Function	122
5.6 Experiments	124
5.6.1 Modeling of GaN HEMT	124
5.6.2 Modeling of Thin-TFET	132
5.7 Conclusion	133
Bibliography	135
6 Future Works	138
6.1 Non-ideal effects in Thin-TFETs	138
6.2 Experimental Demonstration of TransiXNOR	140
6.3 Adjoint Network as Regularization in Pi-NN	141

LIST OF TABLES

2.1	The band gaps, electron affinities and effective masses used for MoS ₂ and WTe ₂	41
2.2	Benchmarking Parameters	56
3.1	The material and device parameters of pin-TFETs and Thin-TFETs in the simulation.	76

LIST OF FIGURES

1.1	(a) Strong scaling of power consumption: Using the same number of transistor, compute a fixed-size of operations per second with N times less power consumption; (b) Weak scaling of power consumption: Using N times more transistors, compute a N times bigger size of operations per sec with the same power consumption.	3
1.2	(a.0) The schematic structure of an n-channel MOSFET; the band diagrams of source and channel when the MOSFET is (a.1) OFF and (a.2) ON. The orange shapes represent free electron distributions in the source conduction band. (b.0) The schematic structure of a n-type TFET; the band-diagrams of source and channel: (b.1) the TFET has no tunnel window, however the leakage current is due to the band tail states; (b.2) the TFET has tunnel window, however due to the long tunneling distance, the tunnel current is still small; (b.3) the TFET is ON. The orange shades represent the band tail states.	5
1.3	The interactions between physic, device, system and machine learning.	8
2.1	(a) Schematic device structure for the Thin-TFET, where V_{TG} , V_{BG} and V_{DS} are the top gate, bottom gate and drain to source voltages; (b) sketch of the band diagram, where $\Phi_{M,T}$, $\Phi_{M,B}$ are the work-functions and $E_{F,MT}$, $E_{F,MB}$ the Fermi levels of the metal gates, while $\chi_{2D,T}$, $\chi_{2D,B}$ are the electron affinities, E_{FT} , E_{FB} the Fermi levels, E_{CT} , E_{CB} the conduction band edges and E_{VT} , E_{VB} the valence band edges respectively in the top and bottom 2D layer. V_{TOX} , V_{IOX} and V_{BOX} are the potential drops respectively across the top oxide, interlayer and bottom oxide.	23
2.2	Sketch of the band alignments in a Thin-TFET between the top and bottom 2D layer in: (a) OFF state and (b) ON state.	24
2.3	Sketch of a possible rotational misalignment between the top and bottom 2D layer, x-y is the reference coordinate for the bottom 2D layer and x'-y' is the reference coordinate for the top 2D layer. θ is the rotational misalignment angle. We assume the top layer and the bottom layer have the same lattice constant a_0	34
2.4	(a) Band structure for hexagonal monolayer MoS ₂ and (b) hexagonal monolayer WTe ₂ as obtained using DFT method described in the paper of C. Gong et.al. ¹⁸ The dashed lines represent the analytical approximation obtained with a parabolic effective mass model.	41

2.5	Numerical results of (a) band alignment versus the top gate voltage V_{TG} and (b) tunnel current density versus the top gate voltage V_{TG} for different values of the correlation length L_c . The parameters used in (b) are: matrix element is $M_{B0} = 0.01 \text{ eV}$; decay constant of wave-function in the interlayer is $\kappa = 3.8 \text{ nm}^{-1}$; energy broadening is $\sigma = 10 \text{ meV}$ and interlayer thickness is $T_{IL} = 0.6 \text{ nm}$ (e.g. 2 atomic layers of BN). $V_{BG} = 0$ and $V_{DS} = 0.3 \text{ V}$ in both (a) and (b).	43
2.6	Numerical calculations for: (a) current density versus V_{TG} with several interlayer thicknesses; (b) current density versus V_{TG} with different values of energy broadening σ . The insert shows that SS increases with σ , and a SS value of 60 mV/dec corresponds to a energy broadening as high as 40 meV. The matrix element is $M_{B0} = 0.01 \text{ eV}$; the decay constant of wave-function in the interlayer is $\kappa = 3.8 \text{ nm}^{-1}$. In (a) the energy broadening is $\sigma = 10 \text{ meV}$. In (b) the interlayer thickness is $T_{IL} = 0.6 \text{ nm}$ (e.g. 2 atomic layers of BN). $V_{BG} = 0$ and $V_{TG} = 0.3 \text{ V}$ in both (a) and (b).	44
2.7	An example to realize both <i>n</i> -type and <i>p</i> -type Thin-TFETs using one pair of 2D semiconductors (2H-WSe_2 and 1T-SnSe_2) with near broken gap band alignment. For the <i>n</i> -type Thin-TFET, SnSe_2 is the top (i.e. drain) 2D layer and WSe_2 is the bottom (i.e. source) 2D layer, along with the top and back gate labeled as <i>n</i> -type in blue. While for the <i>p</i> -type Thin-TFET, WSe_2 is the top (i.e. drain) 2D layer and SnSe_2 is the bottom (i.e. source) 2D layer, along with the top and back gate labeled as <i>p</i> -type in red; Band gaps, electron affinities, effective masses are shown for WSe_2 and SnSe_2 . The <i>n</i> -type and <i>p</i> -type metal work functions are tuned to give symmetric threshold voltages for the <i>n</i> -type and <i>p</i> -type Thin-TFETs.	47
2.8	For the <i>n</i> -type and <i>p</i> -type Thin-TFETs shown in Fig. 2.7: (a) the band alignment versus V_{TG} ; (b) Current density versus V_{TG} , the average SS is calculated from $10^{-3} \mu\text{A}/\mu\text{m}$ to $10 \mu\text{A}/\mu\text{m}$; (c) the current density versus V_{DS} at various V_{TG} ; (d) the transconductance versus V_{TG} ; (e) the carrier concentration in the top and bottom 2D layers versus V_{TG} at various V_{DS} ; (f) the quantum capacitances of the top and bottom 2D layers versus V_{TG} at various V_{DS} ;	49
2.9	Effect of van der Waals gap thickness variation on a <i>p</i> -type Thin-TFET: (a) tunnel current density versus V_{TG} for different van der Waals gap thicknesses T_{vdW} ; (b) differential SS versus current density assuming an evenly distributed van der Waals gap thickness T_{vdW} in the specified range.	51

2.10	Effect of total access resistance on a <i>p</i> -type Thin-TFET: (a) I_D versus V_{TG} and (b) I_D versus V_{DS} with various total access resistance R_C values.	53
2.11	Capacitance network model of the Thin-TFET	53
2.12	For the <i>p</i> -type Thin-TFET, (a) C_{GD} and C_{GS} versus V_{DS} at $V_{TG}=-0.2, -0.3, -0.4$ V; (b) C_{GD} and C_{GS} versus V_{TG} at $V_{DS}=-0.2, -0.3, -0.4$ V.	55
2.13	The intrinsic switching energy and delay for HP CMOS, LP CMOS, HetJTFET, HomJTFET and Thin-TFETs with $V_{DD}=0.2, 0.3, 0.4$ V and $R_C=52, 320 \Omega\mu\text{m}$	57
3.1	Schematic structure of (a) n-type pin-TFET and (a) n-type Thin-TFET.	71
3.2	Comparison between pin-TFET and Thin-TFET: (Column 1) Device schematics with the tunneling area highlighted. Thin-TFET has a larger tunneling area, which can potentially render a higher tunnel current; (Column 2) 1D intrinsic capacitance networks, where C_G is the gate capacitance of the pin-TFET, C_Q is the quantum capacitance of the channel material in the pin-TFET, $C_{T(B)G}$ is the top (bottom) gate capacitance of the Thin-TFET, $C_{Q,T(B)}$ is the quantum capacitance of the top (bottom) material, and $C_{Interlayer}$ is the interlayer capacitance between the top and bottom materials in the Thin-TFET; (Column 3 & 4) analytical expressions for C_{GD} and gate efficiency.	74
3.3	(a) C_{GD} versus V_G at different V_{DS} for both pin-TFETs (black lines) and Thin-TFETs with different underlap and undercut lengths respectively; (b) C_{GS} versus V_G at different V_{DS} or both pin-TFETs (black lines) and Thin-TFETs with different underlap and undercut lengths respectively; (c) C_{GG} versus V_G at different V_{DS} or both pin-TFETs (black lines) and Thin-TFETs with different underlap and undercut lengths respectively; (d) C_{GD} versus underlap/undercut length at different V_{DS} and $V_G = 0.4$ V.	78
3.4	(a) Gate efficiencies versus the drain voltages V_{DS} for both pin-TFETs and Thin-TFETs, the solid lines are the gate efficiency at the threshold while the dash lines are the average gate efficiency when swiping V_G from 0 to 0.4 V; (b) the threshold voltages versus the drain voltages V_{DS} for both pin-TFETs and Thin-TFETs, the increasing threshold voltages at smaller V_{DS} lead to the non-linear onset in the output characteristics.	80

3.5	(a) The threshold voltages versus the drain voltages for Thin-TFETs with different undercut lengths. The red solid line is the threshold voltages computed at the center of the channel (shown in (b)), the dash lines are the threshold voltages computed at the drain-side edge (shown in (b)). The differences between the dash lines and the red solid line indicate the non-uniformity of the threshold voltages along the channel of Thin-TFETs.	81
3.6	The schematic layout of the complementary TFET (CTFET) inverter. $C_{GD,n}$ and $C_{GD,p}$ are the gate-to-drain capacitance of the <i>n</i> -type and <i>p</i> -type TFETs. C_L is the load capacitance.	81
3.7	(a)-(d) the input and output voltages of pin-TFETs and Thin-TFETs based CFET inverter versus time with different ON current density and load capacitance; (e)-(h) the current density of pin-TFETs and Thin-TFETs in the inverters versus time with different On current density and load capacitance.	83
4.1	The RRAM crossbar architecture. X is the input voltages signals, W is the weight matrix whose elements are the RRAM conductivities, and Y is the output current signals. The relationship of X, W, Y is shown in Eq.4.1	88
4.2	(a) The schematic structure of TransiXNOR; (b) the band diagrams at different gate bias conditions of TransiXNOR when V_{DS} equals V_{DD} : (<i>left</i>) the channel/drain tunnel junction is ON when both V_{TG} and V_{BG} are 0; (<i>right</i>) the source/channel tunnel junction is ON when both V_{TG} and V_{BG} are V_{DD} ; (<i>middle</i>) both the channel/drain tunnel junction and source/channel tunnel junction are OFF at the bias conditions such as both V_{TG} and V_{BG} are $V_{DD}/2$, or one gate is V_{DD} and the other is 0; (c) The schematic mapping of transiXNOR ON/OFF states at different V_{TG} and V_{BG} when V_{DS} is V_{DD} , which resembles XNOR logic.	91
4.3	(a.1) The I-V curve of the drain current I_{DS} versus V_{TG} when $V_{BG} = 0$ V and $V_{DS} = 0.2$ V; (a.2) The band diagram and (a.3) the current spectrum when $V_{DS} = 0.2$ V and both $V_{TG} = V_{BG} = 0$ V. (b.1) THe I-V curve of the drain current I_{DS} versus V_{TG} when $V_{BG} = 0.1$ V and $V_{DS} = 0.2$ V; (b.2) The band diagram and (b.3) the current spectrum when $V_{DS} = 0.2$ V and both $V_{TG} = V_{BG} = 0.1$ V. (c.1) THe I-V curve of the drain current I_{DS} versus V_{TG} when $V_{BG} = 0.2$ V and $V_{DS} = 0.2$ V; (c.2) The band diagram and (c.3) the current spectrum when $V_{DS} = 0.2$ V and both $V_{TG} = V_{BG} = 0.2$ V.	94

4.4	(a.1) The family characteristic of TransiXNOR with various V_{TG} at $V_{BG} = 0.2$ V; (a.2) The band diagram and (a.3) the current spectrum when $V_{DS} = 0.1$ V and both $V_{TG} = V_{BG} = 0.2$ V. (b.1) The family characteristic of TransiXNOR with various V_{TG} at $V_{BG} = 0$ V; (b.2) The band diagram and (b.3) the current spectrum when $V_{DS} = 0.1$ V and both $V_{TG} = V_{BG} = 0$ V.	96
4.5	A grid of 2D mappings of I_{DS} along both V_{TG} and V_{BG} axes at different V_{DS} . The coloring represents the current density in logarithm. When V_{DS} is larger than 0.1 V (half V_{DD}), the transiXNOR resembles XNOR logic; and when V_{DS} is smaller than 0.1 V, the transiXNOR resemble AND logic.	97
4.6	The XNOR cell built with TransiXNOR and RRAM. The bit line and work line are used to write to the RRAM. The RRAM is in series with a regular resistor. After writing each element $W_{k,n}$ of the 2D binary weight matrix W to the each RRAM, the word line is set floating, and the bit line is set to ground. During the computing, the source line is set to V_{DD} , and each element X_k of the input vector X is set through each input line in parallel. The current entering the output line represent the XNOR result of $W_{k,n}$ and X_k	98
4.7	The XNOR array to compute $Y=W \times X$ in the constant time.	98
5.1	The Multiplayer Perception (MLP) neural network model	111
5.2	A training procedure for Artificial Neural Network (ANN) device compact modeling.	112
5.3	The compact model of the n-type Thin-TFET derived based on the MLP neural network widely used in previous works, ³⁻⁵ (a) the training errors and test errors for a variety of hyperparameters; (b) the MLP neural network with 7 $tanh$ neurons in the first and second hidden layers. From (c) to (f), the I-V curves generated by the MLP neural network shown in (b) are plotted along with the training data and the test data: (c) I_{DS} versus V_{DS} at different V_{TG} ; (d) I_{DS} versus V_{TG} at different V_{DS} in linear scale; (e) I_{DS} versus V_{DS} at different V_{TG} around $V_{DS} = 0$ V, the embedded plot shows unphysical I_{DS} - V_{DS} relationships around V_{DS} equals 0; (f) I_{DS} versus V_{TG} at different V_{DS} in semi-log scale, unphysical oscillation of I_{DS} around zero appears in the sub-threshold region and when $V_{DS} = 0$ V.	114
5.4	The architecture of Pi-NN. The shaded area indicates a Pi-NN block, which is the building block of Pi-NN network.	115
5.5	The Physics-Inspired Neural Network (Pi-NN) model.	117
5.6	118

5.7	(a) The adjoint network of a fully connected (FC) layer with sigmoid activation functions, where $\beta=\nabla_\gamma y$, γ is the output of FC layer, and y is the outputs of the neural network; (b) The adjoint network of a Pi-NN block, where $\beta=\nabla_\gamma y$ and $\alpha=\nabla_\delta y$, γ is the output of FC layer in <i>sig</i> subnet, δ is the output of FC in <i>tanh</i> subnet, and y is the outputs of the neural network.	121
5.8	Construction of the weighted L1 loss with max scale loss limit.	123
5.9	The I-V characteristics of a GaN HEMT: (a) I_{DS} versus V_{DS} at different V_{GS} in the linear scale and (b) in the log scale; (c) I_{DS} versus V_{GS} at different V_{DS} in the linear scale and (d) in the log scale.	125
5.10	(a) The weighted L1 metric versus max loss scale. Each red circle represents the training weighted L1 metric, and each blue circle represents the evaluation weighted L1 metric. The blue and red line are the average value of each runs. (b) The L1 metric versus max loss scale. Each red circle represents the training L1 metric, and each blue circle represents the evaluation L1 metric. The blue and red line are the average value of each runs.	126
5.11	Each blue circuit represents one run, and the dash line is the average value of multiple runs. (a) The train weighted L1 losses (with max loss scale limit) versus different base learning rates; (b) The evaluation weighted L1 losses (with max loss scale limit) versus different base learning rates; (c) The train L1 metric versus different base learning rates; (b) The evaluation L1 metric versus different base learning rates; (a) The train weighted L1 metric (without max loss scale limit) versus different base learning rates; (b) The evaluation weighted L1 metric (without max loss scale limit) versus different base learning rates.	127
5.12	(a) The training/evaluation loss (weighted L1 loss with max loss scale limit) versus epochs; (b) the training/evaluation weighted L1 metric (without max loss scale limit) versus epochs.	128
5.13	The I-V curves generated by the Pi-NN model are plotted along with the training data (blue circles) and the evaluation data (red circles): (a) I_{DS} versus V_{DS} at different V_{GS} in the linear scale and (b) in the log scale; (c) I_{DS} versus V_{GS} at different V_{DS} in the linear scale and (d) in the log scale.	130
5.14	The I-V curves generated by the Pi-NN model are plotted along with the training data (blue circles) and the evaluation data (red circles) for I_{DS} versus V_{DS} at different V_{GS} in the linear scale. V_{DS} for the model are extended to 80% beyond the training V_{DS} range and V_{TG} extended to +/- 40% beyond the training V_{GS} range.	131

CHAPTER 1

INTRODUCTION

1.1 Scalability of Power Consumption

What is a device? Different researchers may have their own answers. I would argue devices are the implementation of physics. A device utilizes a set of physical effects to control a set of physical state variables with another set of physical state variables. The first generation computers used vacuum tubes for circuitry and magnetic drums for memory. Vacuum tubes utilized thermionic emission of electrons to control current with voltage (as in tetrode tubes¹) or with direction of electrons (as in diodes²). Memory drums utilized ferromagnetism to control magnetic orientation with electric fields. Since then, more than 70 years went by, enormous amount of devices have been proposed, implemented, and become ubiquitous. In this work, we introduce an electronic device: tunnel field effect transistors (TFETs), which utilizes quantum tunneling and thermal equilibrium statistics to control current with voltage.

TFETs are proposed to battle the high power consumption challenge in very-large-scale integration systems (VLSI). Power consumption in VLSI consists of both static and dynamic power consumption. Dynamic power is proportional to CfV^2 , where C is the load capacitance, f is the clock frequency, and V is the supply voltage. Static power consumption is due to the off-state leakage. It mainly consists of two parts: subthreshold leakage and the gate leakage. High-k gate dielectrics³ has been developed to mitigate the gate leakage problem. Thus the subthreshold leakage is the most important static power problem. Therefore, solving the power consumption problem comes down to lower the operation

voltage while keeping the subthreshold leakage very low and ON/OFF ratio high. Inevitably, the operation voltage is limited by the subthreshold slope, namely the sharpness of the ON/OFF switching. In practice, there are two types of power consumption scaling: *strong scaling* and *weak scaling* (shown in Fig.1.1). When two transistors have the same leakage current and ON current, steeper subthreshold slope means smaller operation voltage is required. Smaller operation voltage means less power consumption. Same ON current means same performance (operations per second). Therefore, we can compute a fixed-size of operations per second more energy efficiently using the transistor with steeper subthreshold slope. This is called *strong scaling of power consumption*. However, sometimes the transistor with steeper subthreshold slope has smaller ON current value. If we operate at lower ON current value, steeper subthreshold slope can lead to smaller operation voltage. Although we won't be able to achieve the same performance due to lower ON current, the reduced energy consumption per transistor can allow us to compute a bigger size of operations per second with more transistors within the same energy budget. Current, TFETs have only achieved weak scaling.

1.2 Steep Slope and Tunneling

Current VLSIs use metal-oxide-semiconductor field-effect transistors (MOSFETs) as the fundamental building blocks. Consider the basics of MOSFET operation, figure 1.2(a.0) shows a schematic structure of an n-channel MOSFET. When the device is OFF (shown in Fig.1.2(a.1)), a high energy barrier exists between the source and drain. The electron distribution is a product of the Fermi-Dirac distribution and the electron density of states (DOS). Only the electrons

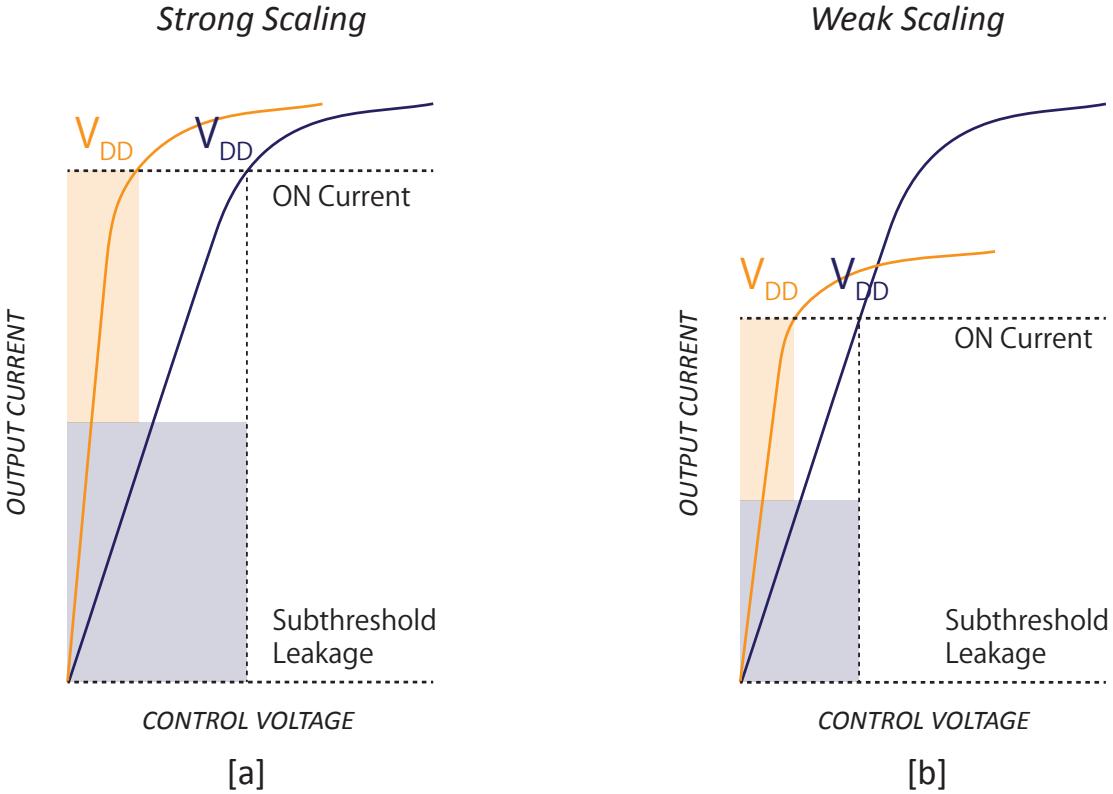


Figure 1.1: (a) Strong scaling of power consumption: Using the same number of transistor, compute a fixed-size of operations per second with N times less power consumption; (b) Weak scaling of power consumption: Using N times more transistors, compute a N times bigger size of operations per sec with the same power consumption.

above the energy barrier in the source can flow to channel. This leakage (diffusion) current flows from source to drain. When positive gate voltage is applied (shown in Fig.1.2(a.2)), the barrier for majority carrier diffusion from source to channel is reduced. The lower the barrier, the more electrons can flow from source to channel, so is the diffusion current. The device is therefore ON. Since DOS is a slowly increasing function in the conduction band, the increasing rate of the electron density above the barrier is dominated by the Fermi-Dirac distribution. At room temperature (i.e. 300 K), the cumulative probability of Fermi-Dirac distribution increases at the rate of 60 mV/dec, which gives the famous 60 mV/dec subthreshold swing limit in MOSFETs. On the other hand, the sub-

threshold slope of TFETs is no longer limited by the Fermi-Dirac distribution by using tunneling instead of thermal emission as the transport mechanism. Figure 1.2(b.0) shows a schematic structure of an n-type TFET. When the device is OFF (shown in Fig.1.2(b.1)), the source valance band source is below the channel conduction band, therefore, ideally there is no empty state the electron in the source can tunnel into. However, in the real materials, the band edge has a finite broadening, which we refer to as the band tail. Therefore, due to the electrons in the band tail of the source valance band tunneling into the empty states in the band tail of the channel conduction band, the resulting current in the OFF state is the leakage current of the TFETs. The subthreshold slope of TFET is fundamentally limited by the sharpness of the band tails. When positive voltage is applied (shown in Fig.1.2(b.2-3)), the source valence band edge moves below the channel conduction band edge, therefore electrons in the source valence band can tunnel into the empty states in the channel conduction band. We refer the energy difference between the source valance band edge and the drain conduction edge *the tunneling window*. In MOSFET, the transport probability of electrons above the energy barrier is approximate one. Unfortunately, for TFET, there is always a finite energy barrier in the transport direction. Therefore, the transport probability is much smaller than one and exponentially dependent on the height and the thickness of the tunneling barrier. Therefore, the tunneling barrier modulation also affects the subthreshold slope of TFETs. From Fig.1.2(b.2) to Fig.1.2(b.3), the widening tunneling window and the thinning tunneling barrier contributes to increasing tunneling current. Ideally, we prefer the ON/OFF of a TFET purely controlled by the tunneling window. In Chapter 2, we will introduce a novel TFET where the tunneling current is purely controlled by the tunneling window.

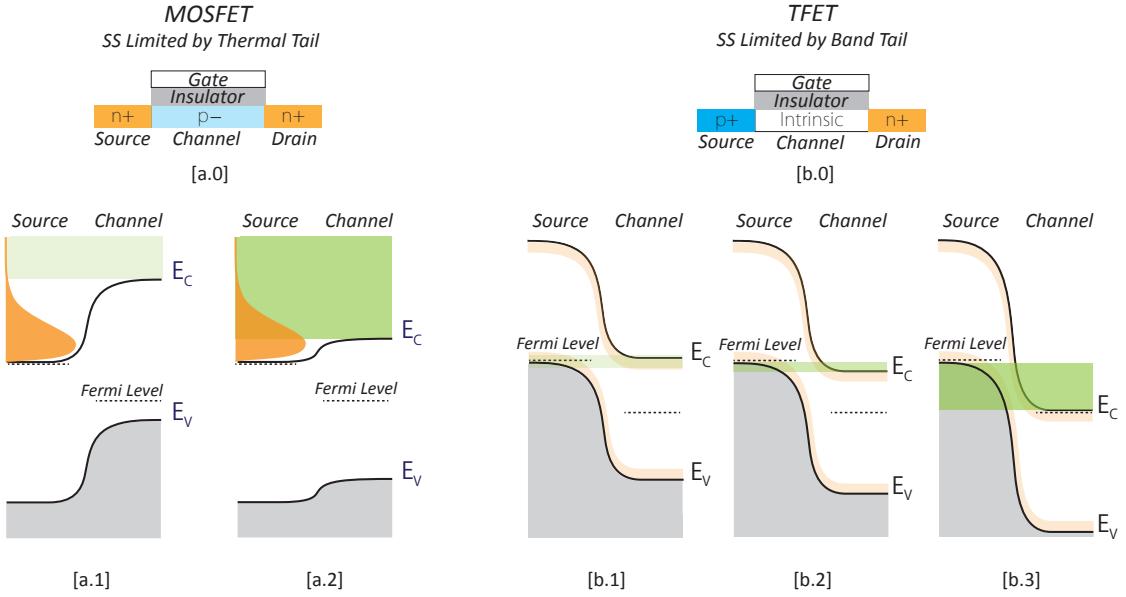


Figure 1.2: (a.0) The schematic structure of an n-channel MOSFET; the band diagrams of source and channel when the MOSFET is (a.1) OFF and (a.2) ON. The orange shapes represent free electron distributions in the source conduction band. (b.0) The schematic structure of a n-type TFET; the band-diagrams of source and channel: (b.1) the TFET has no tunnel window, however the leakage current is due to the band tail states; (b.2) the TFET has tunnel window, however due to the long tunneling distance, the tunnel current is still small; (b.3) the TFET is ON. The orange shades represent the band tail states.

1.3 Previous Works on TFETs

In term of material systems, Group IV materials such as Si⁴ and Ge⁵ exhibit smaller ON-currents resulting from their indirect band gaps and lower tunneling probability. Group III-V materials like InGaAs,⁶ InAs,⁷ and InSb⁸ have higher ON currents due to their narrower and direct band gaps. The use of staggered and broken-gap heterojunctions, such as AlGaSb/InAs,^{9,10} InAs/GaSb¹¹ and InP/GaAs,¹² boosts the ON-current by reducing the tunneling barrier. Following the successful exfoliation of graphene,¹³ other two-dimensional layered materials soon attract a lot of attention in the device community, such as metal chalcogenides,¹⁴ hexagonal boron nitride (hBN),¹⁵ black phosphorus (bP).¹⁶

Among their distinguishing and fascinating properties, their layer-dependent physical behaviors, atomic thin bodies and free of dangling bonds make them suitable materials to build TFETs. Band-to-band tunneling (BTBT) was demonstrated in dual-gated MoS₂/WSe₂ van der Waals junction by Roy et al.,¹⁷ in black phosphorus/SnSe₂ junction by Yan et al.,¹⁸ and in graphene nano-ribbon by Hamam et al.¹⁹ Sarkar et al.²⁰ first demonstrated sub-60 mV/dec in a TFET with Ge/MoS₂ tunnel junction, although the sub-60 mV/dec only occurs when the drain current density is below $10^{-4} \mu\text{A}/\mu\text{m}$. Various TFETs based on van der Waals heterojunctions of 2D layered materials have experimentally demonstrated, such as black phosphorus/MoS₂ by Xu et al.,²¹ MoS₂/WSe₂ by Roy et al.,²² and SnSe₂/WSe₂ by Yan et al.²³ However, only Yan's work reported sub-60 mV/dec at the room temperature. Recently, Li et al.²⁴ exploited the polarization in III-nitride heterojunctions such as GaN/InGaN/GaN to design TFETs.

Besides optimizing the material system, numerous device structures have been proposed recently to boost TFETs' performance. In general, TFET device structures can be divided to two categories: "lateral" TFETs vs. "vertical" TFETs. In lateral TFETs, tunneling direction is perpendicular to gate electric field and in vertical TFETs, tunneling direction and gate electric field are aligned. It is worth to note that many researchers use the word "vertical" to refer the orientation of the channel, but we suggest to use the tunneling direction relative to the gate electric field instead of the spatial orientation to classify the TFET structures. Among the lateral TFETs, high ON current of $10 \mu\text{A}/\mu\text{m}$ and subthreshold slope of 48 mV/decade have been achieved in InAs/GaAsSb/GaSb nanowire TFET by Memisevic et al.²⁵ In vertical TFETs, the tunneling occurs cross an area instead of a line in lateral TFETs, vertical TFETs can achieve high ON current of $180 \mu\text{A}/\mu\text{m}$ by Zhou et al.,¹⁰ although

the vertical TFET takes much larger footprint than Memisevic's lateral TFET. Based on those two canonical structures, many variations have been proposed. Imenabadi et al.²⁶ proposed Z-shaped TFETs to take an advantage of high ON current in vertical TFETs without increasing the device footprint. Wang et al.²⁷ and Shih et al.²⁸ developed the U-shaped TFETs to enhance ON current while suppressing leakage current. On top of the different tunneling directions, researchers often engineer the band diagram in the channel to further optimize the TFET performance. Inserting a "pocket" between the source and channel have been proposed to steepen the subthreshold slope and enhance the ON current by Huang et al.²⁹ and Long et al.,³⁰ and inserting a "pocket" between the channel and the drain to reduce the gate-drain capacitance by Kwon et al.³¹ Special Engineered band alignments in the tunneling direction by Long et al.³² and perpendicular to the tunneling direction by Zhao et al.³³ are proposed to enhance ON current, steepen subthreshold slope, and suppress leakage current. In thin body channel materials such as 2D layered materials, doping can be achieved by charge transfer or electrostatic doping from adjacent layers or plates, such as the dielectric engineered TFET³⁴ and junctionless TFET.³⁵

The performance of TFETs is often affected by several non-ideal effects. Huang et al.³⁶ looked into the reliability issue due to dielectric breakdown in TFETs. As the fundamental attribute to the subthreshold steepness, effects of band tails³⁷⁻⁴⁰ have been studied. Besides the band tail, trap assisted tunneling,⁴¹ recombination,⁴¹ and interface roughness⁴² also have significant impacts on the TFET performance. In order to mitigate those non-ideal effects, it is essential to have high quality materials and advanced fabrication techniques, as well as suitable device designs.⁴³

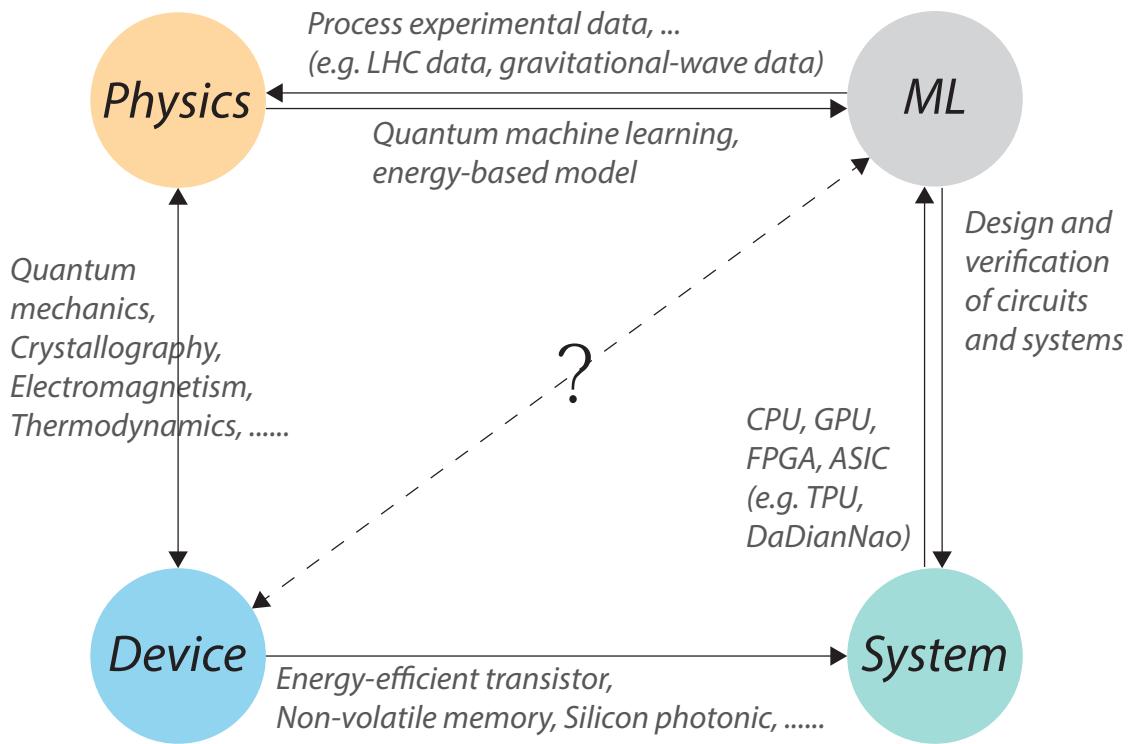


Figure 1.3: The interactions between physic, device, system and machine learning.

TFETs have been explored and evaluated in many circuit/system applications, including TCAM,⁴⁴ analog/mixed-signal circuits,^{45–47} spiking neural network,⁴⁸ active pixel sensors,⁴⁹ and GPU register file.⁵⁰

1.4 Devices and Machine Learning

The increasing computing power, algorithm advances, and the explosion of digital data fuel machine learning techniques as powerful methods to find patterns in the data. The interactions between physic, device, system and machine learning are illustrated in Fig.1.3.

As discussed above, devices are the implementation of physics, especially

condensed matter physics. On the other hand, devices are important platforms for physics researches. Recently, physics and machine learning algorithms become more and more connected. Quantum mechanism is proposed to enhance classical machine learning algorithms.^{51,52} Moreover, physical concepts such as entropy and energy have been widely used to design and reason machine learning algorithms.^{53,54} On the other hand, machine learning has been successfully applied to analyze experimental data, such as searching new particles in the Large Hadron Collider (LHC),⁵⁵ discovering gravitational waves,⁵⁶ and various topics in the condensed matter physics.^{57–63}

As one of the driving forces of machine learning popularity, device innovations keep influencing the system designs, which ultimately enable efficient hardwares for machine learning applications. For instance, FinFETs⁶⁴ have become the building blocks the latest generation of CPUs and GPUs. Non-volatile memories have promised to greatly reduce the memory access energy and delay,⁶⁵ therefore open the possibility of novel architectures for machine learning accelerators.⁶⁶ The use of photonics in networking has been considered one of many possible solutions for handling the growing demands on data centers.⁶⁷ Various architectures have been developed to accelerate machine learning algorithm, particular neural networks.^{68–70} On the other hand, machine learning algorithms have become a popular method to enable fast, accurate design and verification of electronic systems.^{71,72}

Apparently, device and machine learning researches are indirectly interacting with each other through physics and systems. However, how to apply machine learning algorithms directly to help accelerating device researches and how to design certain devices to directly solve machine learning problems re-

main open questions. In this work, we proposed a novel device for binary neural network accelerator and we developed a deep learning framework for efficient and accurate device modeling.

1.5 Brief Outline of This Work

In Chapter 2, two-dimensional heterojunction interlayer tunnel FET (Thin-TFET) is introduced. Thin-TFETs were positioned as an ultra-scaled steep transistor to offer high ON current and steep subthreshold slope. In Chapter 3, we investigated the Miller effect in vertical and lateral TFETs and discovered that vertical TFETs intrinsically have smaller Miller effect than lateral TFETs. In Chapter 4, we proposed TransiNXOR, which utilized the tunneling at both the source/channel junction and the channel/drain junction to enable exclusive not or (XNOR) logic operation in a single transistor. In Chapter 5, physics-inspired neural network (Pi-NN) is developed to learn efficient and accurate device model from experimental or simulation data.

BIBLIOGRAPHY

- ¹J. Linvill and L. Schimpf, "The design of tetrode transistor amplifiers," *Bell Labs Technical Journal*, vol. 35, no. 4, pp. 813–840, 1956.
- ²J. A. Fleming, "Instrument for converting alternating electric currents into continuous currents." Nov. 7 1905, uS Patent 803,684.
- ³S. Natarajan, M. Armstrong, M. Bost, R. Brain, M. Brazier, C.-H. Chang, V. Chikarmane, M. Childs, H. Deshpande, K. Dev *et al.*, "A 32nm logic technology featuring 2 nd-generation high-k+ metal-gate transistors, enhanced channel strain and 0.171 μm 2 sram cell size in a 291mb array," in *Electron Devices Meeting, 2008. IEDM 2008. IEEE International*. IEEE, 2008, pp. 1–3.
- ⁴L. Knoll, Q.-T. Zhao, A. Nichau, S. Trelenkamp, S. Richter, A. Schäfer, D. Es-seni, L. Selmi, K. K. Bourdelle, and S. Mantl, "Inverters with strained si nanowire complementary tunnel field-effect transistors," *IEEE electron device letters*, vol. 34, no. 6, pp. 813–815, 2013.
- ⁵S. H. Kim, H. Kam, C. Hu, and T.-J. K. Liu, "Germanium-source tunnel field effect transistors with record high i on/i off," in *VLSI Technology, 2009 Symposium on*. IEEE, 2009, pp. 178–179.
- ⁶U. E. Avci, S. Hasan, D. E. Nikonorov, R. Rios, K. Kuhn, and I. A. Young, "Understanding the feasibility of scaled iii–v tfet for logic by bridging atomistic simulations and experimental results," in *VLSI technology (VLSIT), 2012 symposium on*. IEEE, 2012, pp. 183–184.
- ⁷S. Agarwal, G. Klimeck, and M. Luisier, "Leakage-reduction design concepts for low-power vertical tunneling field-effect transistors," *IEEE Electron Device Letters*, vol. 31, no. 6, pp. 621–623, 2010.
- ⁸S. S. Sylvia, M. A. Khayer, K. Alam, and R. K. Lake, "Doping, tunnel barriers, and cold carriers in inas and insb nanowire tunnel transistors," *IEEE transactions on electron devices*, vol. 59, no. 11, pp. 2996–3001, 2012.
- ⁹Y. Lu, G. Zhou, R. Li, Q. Liu, Q. Zhang, T. Vasen, S. Doo Chae, T. Kosel, M. Wistey, H. Xing, A. Seabaugh, and P. Fay, "Performance of AlGaSb/InAs TFETs with gate electric field and tunneling direction aligned," *Electron Device Letters, IEEE*, vol. 33, no. 5, pp. 655–657, May 2012.

- ¹⁰ G. Zhou, R. Li, T. Vasen, M. Qi, S. Chae, Y. Lu, Q. Zhang, H. Zhu, J.-M. Kuo, T. Kosel *et al.*, “Novel gate-recessed vertical inas/gasb tfets with record high ion of $180 \mu\text{A}/\mu\text{m}$ at $V_{DS} = 0.5 \text{ V}$,” in *Electron Devices Meeting (IEDM), 2012 IEEE International*. IEEE, 2012, pp. 32–6.
- ¹¹ U. E. Avci and I. A. Young, “Heterojunction tfet scaling and resonant-tfet for steep subthreshold slope at sub-9nm gate-length,” in *Electron Devices Meeting (IEDM), 2013 IEEE International*. IEEE, 2013, pp. 4–3.
- ¹² B. Ganjipour, J. Wallentin, M. T. Borgstrom, L. Samuelson, and C. Thelander, “Tunnel field-effect transistors based on inp-gaas heterostructure nanowires,” *ACS nano*, vol. 6, no. 4, pp. 3109–3113, 2012.
- ¹³ A. C. Neto, F. Guinea, N. M. Peres, K. S. Novoselov, and A. K. Geim, “The electronic properties of graphene,” *Reviews of modern physics*, vol. 81, no. 1, p. 109, 2009.
- ¹⁴ Q. H. Wang, K. Kalantar-Zadeh, A. Kis, J. N. Coleman, and M. S. Strano, “Electronics and optoelectronics of two-dimensional transition metal dichalcogenides,” *Nature nanotechnology*, vol. 7, no. 11, pp. 699–712, 2012.
- ¹⁵ C. R. Dean, A. F. Young, I. Meric, C. Lee, L. Wang, S. Sorgenfrei, K. Watanabe, T. Taniguchi, P. Kim, K. L. Shepard *et al.*, “Boron nitride substrates for high-quality graphene electronics,” *Nature nanotechnology*, vol. 5, no. 10, pp. 722–726, 2010.
- ¹⁶ L. Li, Y. Yu, G. J. Ye, Q. Ge, X. Ou, H. Wu, D. Feng, X. H. Chen, and Y. Zhang, “Black phosphorus field-effect transistors,” *Nature nanotechnology*, vol. 9, no. 5, pp. 372–377, 2014.
- ¹⁷ T. Roy, M. Tosun, X. Cao, H. Fang, D.-H. Lien, P. Zhao, Y.-Z. Chen, Y.-L. Chueh, J. Guo, and A. Javey, “Dual-gated mos2/wse2 van der waals tunnel diodes and transistors,” *Acs Nano*, vol. 9, no. 2, pp. 2071–2079, 2015.
- ¹⁸ R. Yan, S. Fathipour, Y. Han, B. Song, S. Xiao, M. Li, N. Ma, V. Protasenko, D. A. Muller, D. Jena *et al.*, “Esaki diodes in van der waals heterojunctions with broken-gap energy band alignment,” *Nano letters*, vol. 15, no. 9, pp. 5791–5798, 2015.
- ¹⁹ A. M. Hamam, M. E. Schmidt, M. Muruganathan, S. Suzuki, and H. Mizuta, “Sub-10 nm graphene nano-ribbon tunnel field-effect transistor,” *Carbon*, vol. 126, pp. 588–593, 2018.

- ²⁰ D. Sarkar, X. Xie, W. Liu, W. Cao, J. Kang, Y. Gong, S. Kraemer, P. M. Ajayan, and K. Banerjee, "A subthermionic tunnel field-effect transistor with an atomically thin channel," *Nature*, vol. 526, no. 7571, pp. 91–95, 2015.
- ²¹ J. Xu, J. Jia, S. Lai, J. Ju, and S. Lee, "Tunneling field effect transistor integrated with black phosphorus-mos2 junction and ion gel dielectric," *Applied Physics Letters*, vol. 110, no. 3, p. 033103, 2017.
- ²² T. Roy, M. Tosun, M. Hettick, G. H. Ahn, C. Hu, and A. Javey, "2d-2d tunneling field-effect transistors using wse2/snse2 heterostructures," *Applied Physics Letters*, vol. 108, no. 8, p. 083111, 2016.
- ²³ X. Yan, C. Liu, C. Li, W. Bao, S. Ding, D. W. Zhang, and P. Zhou, "Tunable snse2/wse2 heterostructure tunneling field effect transistor," *Small*, vol. 13, no. 34, 2017.
- ²⁴ W. Li, S. Sharmin, H. Ilatikhameneh, R. Rahman, Y. Lu, J. Wang, X. Yan, A. Seabaugh, G. Klimeck, D. Jena *et al.*, "Polarization-engineered iii-nitride heterojunction tunnel field-effect transistors," *IEEE Journal on Exploratory Solid-State Computational Devices and Circuits*, vol. 1, pp. 28–34, 2015.
- ²⁵ E. Memisevic, J. Svensson, M. Hellenbrand, E. Lind, and L.-E. Wernersson, "Vertical inas/gaassb/gasb tunneling field-effect transistor on si with s= 48 mv/decade and i on= 10 μ A/ μ m for i off= 1 na/ μ m at v ds= 0.3 v," in *Electron Devices Meeting (IEDM), 2016 IEEE International*. IEEE, 2016, pp. 19–1.
- ²⁶ R. M. Imenabadi, M. Saremi, and W. G. Vandenberghe, "A novel pnpn-like z-shaped tunnel field-effect transistor with improved ambipolar behavior and rf performance," *IEEE Transactions on Electron Devices*, vol. 64, no. 11, pp. 4752–4758, 2017.
- ²⁷ W. Wang, P.-F. Wang, C.-M. Zhang, X. Lin, X.-Y. Liu, Q.-Q. Sun, P. Zhou, and D. W. Zhang, "Design of u-shape channel tunnel fets with sige source regions," *IEEE Transactions on Electron Devices*, vol. 61, no. 1, pp. 193–197, 2014.
- ²⁸ P.-C. Shih, W.-C. Hou, and J.-Y. Li, "A u-gate ingaas/gaassb heterojunction tfet of tunneling normal to the gate with separate control over on-and off-state current," *IEEE Electron Device Letters*, 2017.
- ²⁹ Q. Huang, R. Huang, Z. Zhan, Y. Qiu, W. Jiang, C. Wu, and Y. Wang, "A novel si tunnel fet with 36mv/dec subthreshold slope based on junction depleted-modulation through striped gate configuration," in *Electron Devices Meeting (IEDM), 2012 IEEE International*. IEEE, 2012, pp. 8–5.

- ³⁰ P. Long, J. Z. Huang, M. Povolotskyi, G. Klimeck, and M. J. Rodwell, "High-current tunneling fets with (1 1 0) orientation and a channel heterojunction," *IEEE Electron Device Letters*, vol. 37, no. 3, pp. 345–348, 2016.
- ³¹ D. W. Kwon, H. W. Kim, J. H. Kim, E. Park, J. Lee, W. Kim, S. Kim, J.-H. Lee, and B.-G. Park, "Effects of localized body doping on switching characteristics of tunnel fet inverters with vertical structures," *IEEE Transactions on Electron Devices*, vol. 64, no. 4, pp. 1799–1805, 2017.
- ³² P. Long, J. Huang, M. Povolotskyi, D. Verreck, J. Charles, T. Kubis, G. Klimeck, M. Rodwell, and B. Calhoun, "A tunnel fet design for high-current, 120 mv operation," in *Electron Devices Meeting (IEDM), 2016 IEEE International*. IEEE, 2016, pp. 30–2.
- ³³ Y. Zhao, C. Wu, Q. Huang, C. Chen, J. Zhu, L. Guo, R. Jia, Z. Lv, Y. Yang, M. Li *et al.*, "A novel tunnel fet design through adaptive bandgap engineering with constant sub-threshold slope over 5 decades of current and high ion/ioff ratio," *IEEE Electron Device Letters*, vol. 38, no. 5, pp. 540–543, 2017.
- ³⁴ H. Ilatikhameneh, T. A. Ameen, G. Klimeck, J. Appenzeller, and R. Rahman, "Dielectric engineered tunnel field-effect transistor," *IEEE Electron Device Letters*, vol. 36, no. 10, pp. 1097–1100, 2015.
- ³⁵ B. Ghosh and M. W. Akram, "Junctionless tunnel field effect transistor," *IEEE electron device letters*, vol. 34, no. 5, pp. 584–586, 2013.
- ³⁶ Q. Huang, R. Jia, J. Zhu, Z. Lv, J. Wang, C. Chen, Y. Zhao, R. Wang, W. Bu, W. Wang *et al.*, "Deep insights into dielectric breakdown in tunnel fets with awareness of reliability and performance co-optimization," in *Electron Devices Meeting (IEDM), 2016 IEEE International*. IEEE, 2016, pp. 31–5.
- ³⁷ M. A. Khayer and R. K. Lake, "Effects of band-tails on the subthreshold characteristics of nanowire band-to-band tunneling transistors," *Journal of Applied Physics*, vol. 110, no. 7, p. 074508, 2011. [Online]. Available: <http://dx.doi.org/10.1063/1.3642954>
- ³⁸ H. Zhang, W. Cao, J. Kang, and K. Banerjee, "Effect of band-tails on the subthreshold performance of 2d tunnel-fets," in *Electron Devices Meeting (IEDM), 2016 IEEE International*. IEEE, 2016, pp. 30–3.
- ³⁹ E. Memisevic, E. Lind, M. Hellenbrand, J. Svensson, and L.-E. Wernersson, "Impact of band-tails on the subthreshold swing of iii-v tunnel field-effect transistor," *IEEE Electron Device Letters*, vol. 38, no. 12, pp. 1661–1664, 2017.

- ⁴⁰ S. Sant and A. Schenk, "The effect of density-of-state tails on band-to-band tunneling: Theory and application to tunnel field effect transistors," *Journal of Applied Physics*, vol. 122, no. 13, p. 135702, 2017.
- ⁴¹ Q. Smets, A. S. Verhulst, E. Simoen, D. Gundlach, C. Richter, N. Collaert, and M. M. Heyns, "Calibration of bulk trap-assisted tunneling and shockley-read-hall currents and impact on ingaas tunnel-fets," *IEEE transactions on electron devices*, vol. 64, no. 9, pp. 3622–3626, 2017.
- ⁴² S. Sant and A. Schenk, "Modeling the effect of interface roughness on the performance of tunnel fets," *IEEE Electron Device Letters*, vol. 38, no. 2, pp. 258–261, 2017.
- ⁴³ ———, "Trap-tolerant device geometry for inas/si ptfets," *IEEE Electron Device Letters*, vol. 38, no. 10, pp. 1363–1366, 2017.
- ⁴⁴ M.-H. Tu, Y.-N. Chen, P. Su, and C.-T. Chuang, "Exploration and evaluation of tcam with hybrid tunneling fet and finfet devices for ultra-low-voltage applications," in *VLSI Technology, Systems and Application (VLSI-TSA), 2017 International Symposium on*. IEEE, 2017, pp. 1–2.
- ⁴⁵ F. Settino, M. Lanuzza, S. Strangio, F. Crupi, P. Palestri, D. Esseni, and L. Selmi, "Understanding the potential and limitations of tunnel fets for low-voltage analog/mixed-signal circuits," *IEEE Transactions on Electron Devices*, 2017.
- ⁴⁶ A. Acharya, A. B. Solanki, S. Dasgupta, and B. Anand, "Drain current saturation in line tunneling-based tfets: An analog design perspective," *IEEE Transactions on Electron Devices*, 2017.
- ⁴⁷ J. Min and P. M. Asbeck, "Compact modeling of distributed effects in 2-d vertical tunnel fets and their impact on dc and rf performances," *IEEE Journal on Exploratory Solid-State Computational Devices and Circuits*, vol. 3, pp. 18–26, 2017.
- ⁴⁸ D. Rajasekharan, T. Dutta, A. R. Trivedi, and Y. S. Chauhan, "Energy-efficient spiking neural networks based on tunnel fet," in *Emerging Electronics (ICEE), 2016 3rd International Conference on*. IEEE, 2016, pp. 1–4.
- ⁴⁹ J. Fernández-Berni, M. Niemier, X. Hu, H. Lu, W. Li, P. Fay, R. Carmona-Galán, and Á. Rodríguez-Vázquez, "Tfet-based well capacity adjustment in active pixel sensor for enhanced high dynamic range," *Electronics Letters*, vol. 53, no. 9, pp. 622–624, 2017.

- ⁵⁰ C. Xie, J. Tan, M. Chen, Y. Yi, L. Peng, and X. Fu, "Emerging technology enabled energy-efficient gpgpus register file," *Microprocessors and Microsystems*, vol. 50, pp. 175–188, 2017.
- ⁵¹ J. Biamonte, P. Wittek, N. Pancotti, P. Rebentrost, N. Wiebe, and S. Lloyd, "Quantum machine learning," *Nature*, vol. 549, pp. 195 EP –, 09 2017. [Online]. Available: <http://dx.doi.org/10.1038/nature23474>
- ⁵² T. Yoder, G. H. Low, and I. Chuang, "Quantum inference on bayesian networks," in *APS Meeting Abstracts*, 2014.
- ⁵³ C. Poultney, S. Chopra, Y. L. Cun *et al.*, "Efficient learning of sparse representations with an energy-based model," in *Advances in neural information processing systems*, 2007, pp. 1137–1144.
- ⁵⁴ D. H. Ackley, G. E. Hinton, and T. J. Sejnowski, "A learning algorithm for boltzmann machines," *Cognitive science*, vol. 9, no. 1, pp. 147–169, 1985.
- ⁵⁵ P. Baldi, P. Sadowski, and D. Whiteson, "Searching for exotic particles in high-energy physics with deep learning," *Nature communications*, vol. 5, 2014.
- ⁵⁶ B. P. Abbott, R. Abbott, T. Abbott, M. Abernathy, F. Acernese, K. Ackley, C. Adams, T. Adams, P. Addesso, R. Adhikari *et al.*, "Observation of gravitational waves from a binary black hole merger," *Physical review letters*, vol. 116, no. 6, p. 061102, 2016.
- ⁵⁷ L.-F. Arsenault, A. Lopez-Bezanilla, O. A. von Lilienfeld, and A. J. Millis, "Machine learning for many-body physics: the case of the anderson impurity model," *Physical Review B*, vol. 90, no. 15, p. 155136, 2014.
- ⁵⁸ A. G. Kusne, T. Gao, A. Mehta, L. Ke, M. C. Nguyen, K.-M. Ho, V. Antropov, C.-Z. Wang, M. J. Kramer, C. Long *et al.*, "On-the-fly machine-learning for high-throughput experiments: search for rare-earth-free permanent magnets," *Scientific reports*, vol. 4, 2014.
- ⁵⁹ S. V. Kalinin, B. G. Sumpter, and R. K. Archibald, "Big-deep-smart data in imaging for guiding materials design," *Nature materials*, vol. 14, no. 10, pp. 973–980, 2015.
- ⁶⁰ L. M. Ghiringhelli, J. Vybiral, S. V. Levchenko, C. Draxl, and M. Scheffler, "Big data of materials science: Critical role of the descriptor," *Physical review letters*, vol. 114, no. 10, p. 105503, 2015.

- ⁶¹ S. S. Schoenholz, E. D. Cubuk, D. M. Sussman, E. Kaxiras, and A. J. Liu, "A structural approach to relaxation in glassy liquids," *Nature Physics*, vol. 12, pp. 469 EP –, 02 2016. [Online]. Available: <http://dx.doi.org/10.1038/nphys3644>
- ⁶² P. Mehta and D. J. Schwab, "An exact mapping between the variational renormalization group and deep learning," *arXiv preprint arXiv:1410.3831*, 2014.
- ⁶³ J. Carrasquilla and R. G. Melko, "Machine learning phases of matter," *Nature Physics*, vol. 13, pp. 431 EP –, 02 2017. [Online]. Available: <http://dx.doi.org/10.1038/nphys4035>
- ⁶⁴ D. Bhattacharya and N. K. Jha, "Finfets: From devices to architectures," *Advances in Electronics*, vol. 2014, 2014.
- ⁶⁵ C. J. Xue, G. Sun, Y. Zhang, J. J. Yang, Y. Chen, and H. Li, "Emerging non-volatile memories: opportunities and challenges," in *Hardware/Software Codesign and System Synthesis (CODES+ ISSS), 2011 Proceedings of the 9th International Conference on*. IEEE, 2011, pp. 325–334.
- ⁶⁶ P. Chi, S. Li, C. Xu, T. Zhang, J. Zhao, Y. Liu, Y. Wang, and Y. Xie, "Prime: A novel processing-in-memory architecture for neural network computation in reram-based main memory," in *Proceedings of the 43rd International Symposium on Computer Architecture*. IEEE Press, 2016, pp. 27–39.
- ⁶⁷ Y. A. Vlasov, "Silicon cmos-integrated nano-photonics for computer and data communications beyond 100g," *IEEE Communications Magazine*, vol. 50, no. 2, 2012.
- ⁶⁸ G. Lacey, G. W. Taylor, and S. Areibi, "Deep learning on fpgas: Past, present, and future," *arXiv preprint arXiv:1602.04283*, 2016.
- ⁶⁹ N. P. Jouppi, C. Young, N. Patil, D. Patterson, G. Agrawal, R. Bajwa, S. Bates, S. Bhatia, N. Boden, A. Borchers *et al.*, "In-datacenter performance analysis of a tensor processing unit," in *Proceedings of the 44th Annual International Symposium on Computer Architecture*. ACM, 2017, pp. 1–12.
- ⁷⁰ Y. Chen, T. Luo, S. Liu, S. Zhang, L. He, J. Wang, L. Li, T. Chen, Z. Xu, N. Sun *et al.*, "Dadiannao: A machine-learning supercomputer," in *Proceedings of the 47th Annual IEEE/ACM International Symposium on Microarchitecture*. IEEE Computer Society, 2014, pp. 609–622.
- ⁷¹ W.-T. J. Chan, P.-H. Ho, A. B. Kahng, and P. Saxena, "Routability optimization

for industrial designs at sub-14nm process nodes using machine learning.” in *ISPD*, 2017, pp. 15–21.

⁷² W.-T. J. Chan, K. Y. Chung, A. B. Kahng, N. D. MacDonald, and S. Nath, “Learning-based prediction of embedded memory timing failures during initial floorplan design,” in *Design Automation Conference (ASP-DAC), 2016 21st Asia and South Pacific*. IEEE, 2016, pp. 178–185.

CHAPTER 2

PHYSICAL MODELING OF TWO-DIMENSIONAL HETEROJUNCTION INTERLAYER TUNNELING FETS (THIN-TFETS)

2.1 Introduction

The electronic integrated circuits are the hardware backbone of todays information society and the power dissipation has recently become the greatest challenge, affecting the lifetime of existing portable equipments, the sustainability of large and growing in number data centers, and the feasibility of energy autonomous systems for ambiance intelligence,^{1,2} and of sensor networks for implanted monitoring and actuation medical devices.³ While the scaling of the supply voltage, V_{DD} , is recognized as the most effective measure to reduce switching power in digital circuits, the performance loss and increased device to device variability are a serious hindrance to the V_{DD} scaling down to 0.5 V or below.

The voltage scalability of VLSI systems may be significantly improved by resorting to innovations in the transistor technology and, in this regard, the ITRS has singled out Tunnel filed effect transistors (FETs) as the most promising transistors to reduce the sub-threshold swing, SS, below the 60 mV/dec limit of MOSFETs (at room temperature), and thus to enable a further V_{DD} scaling.^{4,5} Several device architectures and materials are being investigated to develop Tunnel FETs offering both an attractive on current and a small SS, including III-V based transistors possibly employing staggered or broken bandgap heterojunctions,⁶⁻⁹ or strain engineering.¹⁰ Even if encouraging experimental results have been reported for the on-current in III-V Tunnel FETs, to achieve a sub 60

mV/dec subthreshold swing is still a real challenge in these devices, probably due to the detrimental effects of interface states.^{6,11,12} Therefore, as of today the investigation of new material systems and innovative device architectures for high performance Tunnel FETs is a timely research field in both the applied physics and the electron device community.

In such a contest, two-dimensional (2D) crystals attract increasingly more attention primarily due to their scalability, step-like density of states and absence of broken bonds at interface. They can be stacked to form a new class of tunneling transistors based on an interlayer tunneling occurring in the direction normal to the plane of the 2D materials. In fact tunneling and resonant tunneling devices have been recently proposed,¹³ as well as experimentally demonstrated for graphene-based transistors.^{14,15} Furthermore, monolayers of group-VIB transition metal dichalcogenides MX₂ (M = Mo, W; X = S, Se, Te) have recently attracted remarkable attention for their electronic and optical properties.^{16,17} Monolayers of transition-metal dichalcogenides (TMDs) have a bandgap varying from almost zero to 2 eV with a sub-nanometer thickness such that these materials can be considered approximately as two-dimensional crystals.¹⁸ The sub-nanometer thickness of TMDs can provide excellent electrostatic control in a vertically stacked heterojunction. Furthermore, the 2D nature of such materials make them essentially immune to the energy bandgap increase produced by the vertical quantization when conventional 3D semiconductors are thinned to a nanoscale thickness, and thus immune to the corresponding degradation of the tunneling current density.¹⁹ Moreover, the lack of dangling bonds at the surface of TMDs may allow for the fabrication of material stacks with low densities of interface defects,¹⁹ which is another potential advantage of TMDs materials for Tunnel FETs applications.

In this paper we propose a two-dimensional heterojunction interlayer tunneling field effect transistor (Thin-TFET) based on 2D semiconductors and develop a transport model based on the transfer-Hamiltonian method to describe the current voltage characteristics and discuss, in particular, the subthreshold swing. In Section 2.2 we first present the device concept and illustrate examples of the vertical electrostatic control, then we develop a formalism to calculate the tunneling current. Upon realizing that the subthreshold swing of the Thin-TFET is ultimately determined by the energy broadening, in Sec.2.2.3 we show how this important physical factor has been included in our calculations. In Sec.2.2.4 we address the effect of a possible misalignment between the two 2D semiconductor layers, while in Sec.2.2.5 we derive some approximated, analytical expressions for the tunneling current density, which are useful to gain insight in the transistor operation and to guide the device design. In Sec.2.3.2 we present the results of numerically calculated transfer characteristics for the Thin-TFET based on MoS₂ and WSe₂, and effects of correlation lengths, interlayer thicknesses, and energy broadening. In Sec.2.4 we discuss both n-type and p-type Thin-TFETs employing a promising material system of 2H-WSe₂ and 1T-SnSe₂, along with their capacitance model in Sec.2.4.2. The effect of a non-uniform van der Waals gap thickness and the external source and drain total access resistance are also discussed in Sec.2.4.1. Using the simulated results, we present the benchmarking results in Sec.2.4.3 and finally in Sec.2.5 we draw some concluding remarks about the modeling approach developed in the paper and about the design perspectives for the Thin-TFET.

2.2 Modeling of the Tunneling Transistor

2.2.1 Device Concept and Electrostatics

The device structure and the corresponding band diagram are sketched in Fig.2.1, where the 2D materials are assumed to be semiconductors with sizable energy bandgap, for example, transition-metal dichalcogenide (TMD) semiconductors without losing generality.^{17,20} Both the top 2D and the bottom 2D material is a monolayer and the thickness of the 2D layers is neglected in the modeling of the electrostatics.

The working principle of the tunneling transistor sketched in Fig.2.1(a) can be explained as follows. When the conduction band edge E_{CT} of the top 2D layer is higher than the valence band edge E_{VB} of the bottom 2D layer (see Fig.2.2(a)), there are no states in the top layer to which the electrons of the bottom layer can tunnel into. This corresponds to the off state of the device. When E_{CT} is pulled below E_{VB} (see Fig.2.2(b)), a tunneling window is formed and consequently an interlayer tunneling can flow from the bottom to the top 2D material. The crossing and uncrossing between the top layer conduction band and the bottom layer valence band is governed by the gate voltages and it is described by the electrostatics of the device.

To calculate the band alignment along the vertical direction of the intrinsic device in Fig.2.1 we write the Gauss law linking the sheet charge in the 2D materials to the electric fields in the surrounding insulating layers, which leads to

$$\begin{aligned} C_{TOX}V_{TOX} - C_{IOX}V_{IOX} &= e(p_T - n_T + N_D) \\ C_{BOX}V_{BOX} + C_{IOX}V_{IOX} &= e(p_B - n_B + N_A) \end{aligned} \tag{2.1}$$

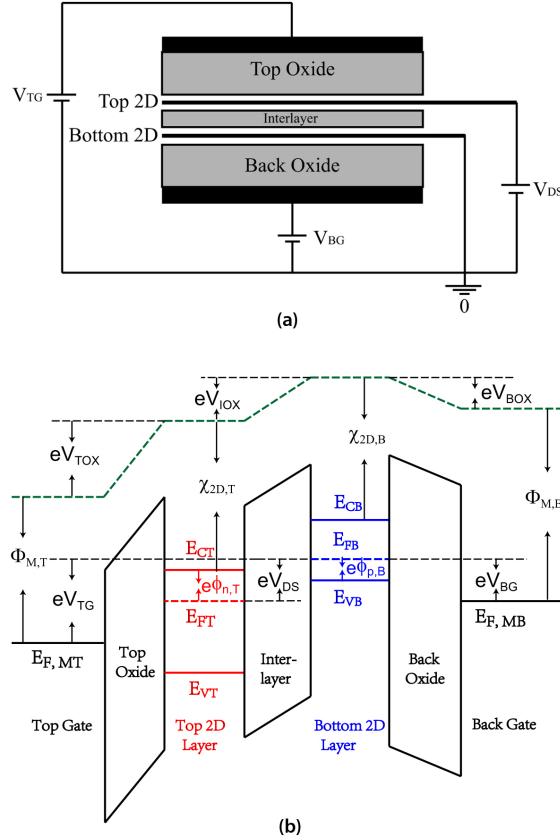


Figure 2.1: (a) Schematic device structure for the Thin-TFET, where V_{TG} , V_{BG} and V_{DS} are the top gate, bottom gate and drain to source voltages; (b) sketch of the band diagram, where $\Phi_{M,T}$, $\Phi_{M,B}$ are the work-functions and $E_{F,MT}$, $E_{F,MB}$ the Fermi levels of the metal gates, while $\chi_{2D,T}$, $\chi_{2D,B}$ are the electron affinities, E_{FT} , E_{FB} the Fermi levels, E_{CT} , E_{CB} the conduction band edges and E_{VT} , E_{VB} the valence band edges respectively in the top and bottom 2D layer. V_{TOX} , V_{IOX} and V_{BOX} are the potential drops respectively across the top oxide, interlayer and bottom oxide.

where $C_{T(I,B)OX}$ is the capacitance per unit area of top oxide (interlayer, bottom oxide) and $V_{T(I,B)OX}$ is the potential drop across top oxide (interlayer, bottom oxide). The potential drop across the oxides can be written in terms of the external voltages V_{TG} , V_{BG} , V_{DS} and of the energy $e\phi_{n,T} = E_{CT} - E_{FT}$ and $e\phi_{p,T} = E_{FB} - E_{VB}$

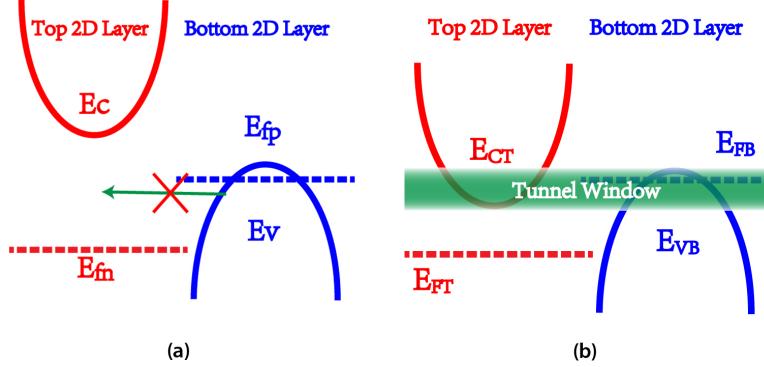


Figure 2.2: Sketch of the band alignments in a Thin-TFET between the top and bottom 2D layer in: (a) OFF state and (b) ON state.

defined in Fig.2.1(b) as

$$\begin{aligned} eV_{TOX} &= eV_{TG} + e\phi_{n,T} - eV_{DS} + \chi_{2D,T} - \Phi_{M,T} \\ eV_{BOX} &= eV_{BG} - e\phi_{p,B} + E_{GB} + \chi_{2D,B} - \Phi_{M,B} \\ eV_{IOX} &= eV_{DS} - e\phi_{p,B} - e\phi_{n,T} + E_{GB} + \chi_{2D,B} - \chi_{2D,T} \end{aligned} \quad (2.2)$$

where E_{FT}, E_{FB} are fermi levels of majority carriers in the top and bottom layer. n_T, p_T are the electron and hole concentration in the top layer, n_B, p_B the concentrations in bottom layer, $\chi_{2D,T}, \chi_{2D,B}$ are the electron affinities of the 2D materials, Φ_T, Φ_B the workfunctions of the top and back gate and E_{GB} is the energy gap in the bottom layer. Eq. 2.2 implicitly assumes that the majority carriers of the two 2D materials are at thermodynamic equilibrium with their Fermi levels, with the split of the Fermi levels set by the external voltages (i.e. $E_{FB}-E_{FT}=eV_{DS}$), and the electrostatic potential essentially constant in the 2D layers.

Since in our numerical calculations we shall employ a parabolic effective mass approximation for the energy dispersion of the 2D materials, as discussed more thoroughly in Sec.2.3, the carrier densities can be readily expressed as an

analytic function of $e\phi_{n,T}$ and $e\phi_{p,B}$ ²¹

$$n(p) = \frac{g_v m_c(m_v) k_B T}{\pi \hbar^2} \ln \left[\exp \left(-\frac{q\phi_{n,T}(\phi_{p,B})}{k_B T} \right) + 1 \right] \quad (2.3)$$

where g_v is the valley degeneracy.

When Eq.2.2 and Eq.2.3 are inserted in Eq.2.1, we obtain two algebraic equations for $\phi_{n,T}$ and $\phi_{p,B}$ that can be solved numerically and describe the electrostatics in a one dimensional section of the device.

2.2.2 Transport Model

In this section we develop a formalism to calculate the tunneling current based on the transfer-Hamiltonian method,^{22–24} as also revisited recently for resonant tunneling in graphene transistors.^{13,14,25} We start by writing the single particle elastic tunneling current as

$$I = g_v \frac{4\pi e}{\hbar} \sum_{\mathbf{k}_T, \mathbf{k}_B} |M(\mathbf{k}_T, \mathbf{k}_B)|^2 \delta(E_B(\mathbf{k}_B) - E_T(\mathbf{k}_T))(f_B - f_T) \quad (2.4)$$

where e is the elementary charge, \mathbf{k}_B , \mathbf{k}_T are the wave-vectors respectively in the bottom and top 2D material, $E_B(\mathbf{k}_B)$ $E_T(\mathbf{k}_T)$ denote the corresponding energies, f_B and f_T are the Fermi occupation functions in the bottom and top layer (depending respectively on E_{FB} and E_{FT} , see Fig.2.1), and g_v is the valley degeneracy. The matrix element $M(\mathbf{k}_T, \mathbf{k}_B)$ expresses the transfer of electrons between the two 2D layers is given by¹⁴

$$M(\mathbf{k}_T, \mathbf{k}_B) = \int_A d\mathbf{r} \int dz \psi_{T,\mathbf{k}_T}^\dagger(\mathbf{r}, z) U_{sc}(\mathbf{r}, z) \psi_{B,\mathbf{k}_B}(\mathbf{r}, z) \quad (2.5)$$

where ψ_{B,\mathbf{k}_B} (ψ_{T,\mathbf{k}_T}) is the electron wave-function in the bottom (top) 2D layer and $U_{sc}(\mathbf{r}, z)$ is the perturbation potential in the interlayer region.

Eq.2.5 acknowledges the fact that in real devices several physical mechanisms occurring in the interlayer region can result in a relaxed conservation of the in plane wave-vector \mathbf{k} in the tunneling process. We will return to the discussion of $U_{sc}(\mathbf{r}, z)$ in this section.

To proceed in the calculation of $M(\mathbf{k}_T, \mathbf{k}_B)$ we write the electron wave-function in the Bloch function form as

$$\psi_{\mathbf{k}}(\mathbf{r}, z) = \frac{1}{\sqrt{N_C}} e^{i\mathbf{k}\cdot\mathbf{r}} u_{\mathbf{k}}(\mathbf{r}, z) \quad (2.6)$$

where $u_{\mathbf{k}}(\mathbf{r}, z)$ is a periodic function of \mathbf{r} and N_C is the number of unit cells in the overlapping area A of the two 2D materials. Eq.2.6 assumes the following normalization condition:

$$\int_{\Omega_C} d\rho \int_z dz |u_{\mathbf{k}}(\rho, z)|^2 = 1 \quad (2.7)$$

where ρ is the in-plane abscissa in the unit cell area Ω_C and $A=N_C\Omega_C$.

The wave-function $\psi_{\mathbf{k}}(\mathbf{r}, z)$ is assumed to decay exponentially in the interlayer region with a decay constant κ ^{13,14} such a z dependence is absorbed in $u_{\mathbf{k}}(\mathbf{r}, z)$ and we do not need to make it explicit in our derivations. It should be noticed that absorbing the exponential decay in $u_{\mathbf{k}}(\mathbf{r}, z)$ recognizes the fact that in the interlayer region the \mathbf{r} dependence of the wave-function changes with z . In fact, as already discussed,¹³ while the $u_{\mathbf{k}}(\mathbf{r}, z)$ are localized around the basis atoms in the two 2D layers, these functions are expected to spread out while they decay in the interlayer region, so that the \mathbf{r} dependence becomes weaker when moving farther from the 2D layers.

To continue in the calculation of $M(\mathbf{k}_T, \mathbf{k}_B)$ we let the scattering potential in the interlayer region be separable in the form¹⁴

$$U_{sc}(\mathbf{r}, z) = V_B(z) F_L(\mathbf{r}) \quad (2.8)$$

where $F_L(\mathbf{r})$ is the in-plane fluctuation of the scattering potential, which is essentially responsible for the relaxation of momentum conservation in the tunneling process.

By substituting Eqs.2.6 and 2.8 in Eq.2.5 and writing $\mathbf{r}=\mathbf{r}_j+\boldsymbol{\rho}$, where \mathbf{r}_j is a direct lattice vector and $\boldsymbol{\rho}$ is the in-plane position inside each unit cell, we obtain

$$\begin{aligned} M(\mathbf{k}_T, \mathbf{k}_B) &= \frac{1}{N_C} \sum_{j=1}^{N_C} e^{i(\mathbf{k}_B - \mathbf{k}_T) \cdot \mathbf{r}_j} \int_{\Omega_C} d\boldsymbol{\rho} \int dz e^{i(\mathbf{k}_B - \mathbf{k}_T) \cdot \boldsymbol{\rho}} \times \\ &\quad \times u_{T,\mathbf{k}_T}^\dagger(\mathbf{r}_j + \boldsymbol{\rho}, z) F_L(\mathbf{r}_j + \boldsymbol{\rho}) V_B(z) u_{B,\mathbf{k}_B}(\mathbf{r}_j + \boldsymbol{\rho}, z) \end{aligned} \quad (2.9)$$

We now assume that $F_L(\mathbf{r})$ corresponds to relatively long range fluctuations so that it can be taken as approximately constant inside a unit cell, and that, furthermore, the top and bottom 2D layer have the same lattice constant, hence the Bloch functions u_{T,\mathbf{k}_T} and u_{B,\mathbf{k}_B} have the same periodicity in the \mathbf{r} plane. Moreover, for the time being we consider that the conduction band minimum in the top layer and the valence band maximum in the bottom layer are at the same point of the 2D Brillouin zone, so that $\mathbf{q} = \mathbf{k}_B - \mathbf{k}_T$ is small compared to the size of the Brillouin zone and $e^{i\mathbf{q}\cdot\boldsymbol{\rho}}$ is approximately 1.0 inside a unit cell. These considerations and approximations allow us to rewrite Eq.2.9 as

$$M(\mathbf{k}_T, \mathbf{k}_B) \simeq \frac{1}{N_C} \sum_{j=1}^{N_C} e^{i\mathbf{q}\cdot\mathbf{r}_j} F_L(\mathbf{r}_j) \int_{\Omega_C} d\boldsymbol{\rho} \int dz u_{T,\mathbf{k}_T}^\dagger(\boldsymbol{\rho}, z) V_B(z) u_{B,\mathbf{k}_B}(\boldsymbol{\rho}, z) \quad (2.10)$$

where the integral in the unit cell has been written for $\mathbf{r}_j = \mathbf{0}$ because it is independent of the unit cell.

Consistently with the assumption that \mathbf{k}_B and \mathbf{k}_T are small compared to the size of the Brillouin zone, in Eq.2.10 we neglect the \mathbf{k}_B (\mathbf{k}_T) dependence of u_{B,\mathbf{k}_B} (u_{T,\mathbf{k}_T}) and simply set $u_{T,\mathbf{k}_T}(\boldsymbol{\rho}, z) \approx u_{0T}(\boldsymbol{\rho}, z)$, $u_{B,\mathbf{k}_B}(\boldsymbol{\rho}, z) \approx u_{0B}(\boldsymbol{\rho}, z)$, where $u_{0T}(\boldsymbol{\rho}, z)$ and $u_{0B}(\boldsymbol{\rho}, z)$ are the periodic parts of the Bloch function at the band edges, which is

the simplification typically employed in the effective mass approximation approach.²¹ By recalling that the u_{0B} and u_{0T} retain the exponential decay of the wave-functions in the interlayer region with a decay constant κ , we now write

$$\int_{\Omega_C} d\rho \int dz u_{0T}^\dagger(\rho, z) V_B(z) u_{0B}(\rho, z) \simeq M_{B0} e^{-\kappa T_{IL}} \quad (2.11)$$

where T_{IL} is the interlayer thickness and M_{B0} is a \mathbf{k} independent matrix element that will remain a prefactor in the final expression for the tunneling current. Since $F_L(\mathbf{r})$ has been assumed a slowly varying function over a unit cell, then the sum over the unit cells in Eq.2.10 can be rewritten as a normalized integral over the tunneling area

$$\frac{1}{\Omega_c N_C} \sum_{j=1}^{N_C} \Omega_c e^{i\mathbf{q}\cdot\mathbf{r}_j} F_L(\mathbf{r}_j) \simeq \frac{1}{A} \int_A e^{i\mathbf{q}\cdot\mathbf{r}} F_L(\mathbf{r}) d\mathbf{r} \quad (2.12)$$

By introducing Eq.2.11 and 2.12 in Eq.2.10 we can finally express the squared matrix element as

$$|M(\mathbf{k}_T, \mathbf{k}_B)|^2 \simeq \frac{|M_{B0}|^2 S_F(\mathbf{q})}{A} e^{-2\kappa T_{IL}} \quad (2.13)$$

where $\mathbf{q} = \mathbf{k}_B - \mathbf{k}_T$ and $S_F(\mathbf{q})$ is the power spectrum of the random fluctuation described by $F_L(\mathbf{r})$, which is defined as²¹

$$S_F(\mathbf{q}) = \frac{1}{A} \left| \int_A e^{i\mathbf{q}\cdot\mathbf{r}} F_L(\mathbf{r}) d\mathbf{r} \right|^2 \quad (2.14)$$

By substituting Eq.2.13 in Eq.2.4 and then converting the sums over \mathbf{k}_B and \mathbf{k}_T to integrals we obtain

$$I = \frac{g_v e |M_{B0}|^2 A}{4\pi^3 \hbar} e^{-2\kappa T_{IL}} \int_{\mathbf{k}_T} \int_{\mathbf{k}_B} d\mathbf{k}_T d\mathbf{k}_B S_F(\mathbf{q}) \delta(E_B(\mathbf{k}_B) - E_T(\mathbf{k}_T)) (f_B - f_T) \quad (2.15)$$

Before we proceed with some important integrations of the basic model that will be discussed in Secs.2.2.3 and 2.2.4, a few comments about the results obtained so far are in order below.

According to Eq.2.15 the current is proportional to the squared matrix element $|M_{B0}|^2$ defined in Eq.2.11 and decreases exponentially with the thickness interlayer T_{IL} according to the decay constant κ of the wave-functions. Attempting to derive a quantitative expression for M_{B0} is admittedly very difficult, in fact it is difficult to determine how the periodic functions $u_{0T}(\rho, z)$ and $u_{0B}(\rho, z)$ spread out when they decay in the barrier region and, furthermore, it is not even perfectly clear what potential energy or Hamiltonian should be used to describe the barrier region itself, which is an issue already recognized and thoroughly discussed in the literature since a long time.²⁴ Our model essentially circumvents these difficulties by resorting to the semi-empirical formulation of the matrix element given by Eq.2.11, where M_{B0} is left as a parameter to be determined and discussed by comparing to experiments.

It is also worth noting that in our calculations we have not explicitly discussed the effect of spin-orbit interaction in the bandstructure of 2D materials, even if giant spin-orbit couplings have been reported in 2D transition-metal dichalcogenides.²⁶ If the energy separations between the spin-up and spin-down bands are large, then the spin degeneracy in current calculations should be one instead of two, which would affect the current magnitude but not its dependence on the gate bias. Our calculations neglected also the possible modifications of band structure in the TMD materials produced by the vertical electrical field, in fact we believe that in our device the electrical field in the 2D layers is not strong enough to make such effects significant.²⁷

The decay constant κ in the interlayer region may be estimated from the electron affinity difference between the 2D layers and the interlayer material.¹³ Moreover, according to Eq.2.15 the constant κ determines the dependence of the

current on T_{IL} , so that κ may be extracted by comparing to experiments discussing such a dependence, which, for example, have been recently reported for the interlayer tunneling current in a graphene-*h*BN system.¹⁵

As for the spectrum $S_F(\mathbf{q})$ of the scattering potential, in our calculations we utilize

$$S_F(\mathbf{q}) = \frac{\pi L_C^2}{(1 + \mathbf{q}^2 L_C^2/2)^{3/2}} \quad (2.16)$$

where L_C is the correlation length, which in our derivations has been assumed large compared to the size of a unit cell. Eq.2.16 is consistent with an exponential form for the autocorrelation function of $F_L(\mathbf{r})$,²¹ and a similar \mathbf{q} dependence has been recently employed to reproduce the experimentally observed line-width of the resonance region in graphene interlayer tunneling transistors.¹⁴ Such a functional form can be representative of phonon assisted tunneling, short-range disorder,²⁸ charged impurities²⁹ or Moiré patterns that have been observed, for instance, at the graphene-*h*BN interface.^{30–32} We will see in Sec.2.2.5 that the L_C has an influence on the gate voltage dependent current, which has a neat physical interpretation, hence a comparison to experimental data will be very informative for an estimate of L_C .

2.2.3 Effects of Energy Broadening

According to Eq.2.4 and Eq.2.15 the tunneling current is simply zero when there is no energy overlap between the conduction band in the top layer and the valence band in the bottom layer, that is for $E_{CT} > E_{VB}$. In a real device, however, the 2D materials will inevitably have phonons, disorder, host impurities in the 2D layer and be affected by the background impurities in the surrounding materi-

als, so that a finite broadening of the energy levels is expected to occur because of the statistical potential fluctuations superimposed to the ideal crystal structure.³³ The energy broadening in 3D semiconductors is known to lead to a tail of the density of states (DoS) in the gap region, that has been also observed in optical absorption measurements and denoted Urbach tail.^{34,35} It is thus expected that the finite energy broadening will be a fundamental limit to the abruptness of the turn on characteristic attainable with the devices of this work, hence it is important to include this effect in our model.

Energy broadening in the 2D systems can stem from the interaction with randomly distributed impurities and disorder in the 2D layer or in the surrounding materials,^{33,36,37} by scattering events induced by the interfaces,³⁸ as well as by other scattering sources. We recognize the fact that a detailed description of the energy broadening is exceedingly complicated due to the many-body and statistical fluctuation effects,³⁹ and thus resort to a relatively simple semi-classical treatment^{36,33}. We start by recalling that the density of states $\rho_0(E)$ for a 2D layer with no energy broadening is

$$\rho_0(E) = \frac{g_s g_v}{4\pi^2} \int_{\mathbf{k}} d\mathbf{k} \delta [E - E(\mathbf{k})] \quad (2.17)$$

where $E(\mathbf{k})$ denotes the energy relation with no broadening and g_s, g_v are spin and valley degeneracy. In the presence of a randomly fluctuating potential $V(\mathbf{r})$, instead, the DoS can be written as^{33,36}

$$\begin{aligned} \rho(E) &= \int_0^\infty dv \rho_0(v) P_v(E - v) \\ &= \frac{g_s g_v}{4\pi^2} \int_{\mathbf{k}} d\mathbf{k} \left[\int_0^\infty dv \delta [v - E(\mathbf{k})] P_v(E - v) \right] \\ &= \frac{g_s g_v}{4\pi^2} \int_{\mathbf{k}} d\mathbf{k} P_v [E - E(\mathbf{k})] \end{aligned} \quad (2.18)$$

where $P_v(v)$ is the distribution function for $V(\mathbf{r})$ (to be further discussed below),

and we have used the $\rho_0(E)$ definition in Eq.2.17 to go from the first to the second equality.

Comparing Eq.2.18 to Eq.2.17, we see that the $\rho(E)$ of the system in the presence of broadening can be calculated by substituting the Dirac function in Eq.2.17 with a finite width function $P_v(v)$, which is the distribution function of $V(\mathbf{r})$ and it is thus normalized to one.

In order to include the energy broadening in our current calculations, we rewrite the tunneling rate in Eq.2.4 as

$$\begin{aligned} \frac{1}{\tau_{\mathbf{k}_T, \mathbf{k}_B}} &= \frac{2\pi}{\hbar} |M(\mathbf{k}_T, \mathbf{k}_B)|^2 \delta [E_T(\mathbf{k}_T) - E_B(\mathbf{k}_B)] \\ &= \frac{2\pi}{\hbar} |M(\mathbf{k}_T, \mathbf{k}_B)|^2 \int_{-\infty}^{\infty} dE \delta [E - E_T(\mathbf{k}_T)] \delta [E - E_B(\mathbf{k}_B)] \end{aligned} \quad (2.19)$$

and note that, consistently with Eq.2.18, the energy broadening can be included in the current calculation by substituting $\delta[E - E(\mathbf{k})]$ with $P_v[E - E(\mathbf{k})]$. By doing so the tunneling rate becomes

$$\frac{1}{\tau_{\mathbf{k}_T, \mathbf{k}_B}} \simeq \frac{2\pi}{\hbar} |M(\mathbf{k}_T, \mathbf{k}_B)|^2 S_E(E_T(\mathbf{k}_T) - E_B(\mathbf{k}_B)) \quad (2.20)$$

where we have introduced an energy broadening spectrum S_E that is defined as

$$S_E(E_T(\mathbf{k}_T) - E_B(\mathbf{k}_B)) = \int_{-\infty}^{\infty} dE P_{vT} [E - E_T(\mathbf{k}_T)] P_{vB} [E - E_B(\mathbf{k}_B)] \quad (2.21)$$

where P_{vT} and P_{vB} is the potential distribution function due to the presence of randomly fluctuating potential $V(\mathbf{r})$ in the top and the bottom layer, respectively.

On the basis of Eq.2.20, in our model for the tunneling current we accounted for the energy broadening by using in all numerical calculations the broadening spectrum $S_E(E_T(\mathbf{k}_T) - E_B(\mathbf{k}_B))$ defined in Eq.2.21 in place of $\delta[E_T(\mathbf{k}_T) - E_B(\mathbf{k}_B)]$. More precisely we used a Gaussian potential distribution for both the top and

the bottom layer

$$P_v(E - E_{k0}) = \frac{1}{\sqrt{\pi}\sigma} e^{-(E - E_{k0})^2/\sigma^2} \quad (2.22)$$

which has been derived by Evan O.Kane for a broadening induced by randomly distributed impurities,³⁶ in which case σ can be expressed in terms of the average impurity concentration.

Quite interestingly, for the Gaussian spectrum in Eq.2.22 the overall broadening spectrum S_E defined in Eq.2.21 can be calculated analytically and reads

$$S_E(E_T(\mathbf{k}_T) - E_B(\mathbf{k}_B)) = \frac{1}{\sqrt{\pi(\sigma_T^2 + \sigma_B^2)}} e^{-(E_T(\mathbf{k}_T) - E_B(\mathbf{k}_B))^2/\sigma^2} \quad (2.23)$$

Hence also S_E has a Gaussian spectrum, where σ_T and σ_B are the broadening energies for the top and bottom 2D layer, respectively.

2.2.4 Rotational Misalignment and Tunneling Between In-equivalent Extrema

The derivations in Sec.2.2.2 assumed that there is a perfect rotational alignment between the top and the bottom layer and that the tunneling occurs between equivalent extrema in the Brillouin zone, that is tunneling from a K to a K extremum (or from K' to K' extremum). We now denote by θ the angle expressing a possible rotational misalignment between the two 2D layers (see Fig.2.3), and still assume that the top 2D crystal has the same lattice constant a_0 as the bottom 2D crystal. The principal coordinate system is taken as the crystal coordinate

system in the bottom layer, and we denote with \mathbf{r}' , \mathbf{k}' the position and wave vectors in the crystal coordinate system of the top layer (with \mathbf{r} , \mathbf{k} being the vectors in the principal coordinate system). The wave-function in the top layer has the form given in Eq.2.6 in terms of \mathbf{r}' , \mathbf{k}' , hence in order to calculate the matrix element in the principal coordinate system we start by writing $\mathbf{r}' = \hat{R}_{B \rightarrow T} \mathbf{r}$, $\mathbf{k}' = \hat{R}_{B \rightarrow T} \mathbf{k}$, where $\hat{R}_{B \rightarrow T}$ is the rotation matrix from the bottom to the top coordinate system, with $\hat{R}_{T \rightarrow B} = [\hat{R}_{B \rightarrow T}]^T$ being the matrix going from the top to the bottom coordinate system and M^T denoting the transpose of the matrix M . The rotation matrix can be written as

$$\hat{R}_{T \rightarrow B} = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \quad (2.24)$$

in terms of the rotational misalignment angle θ .

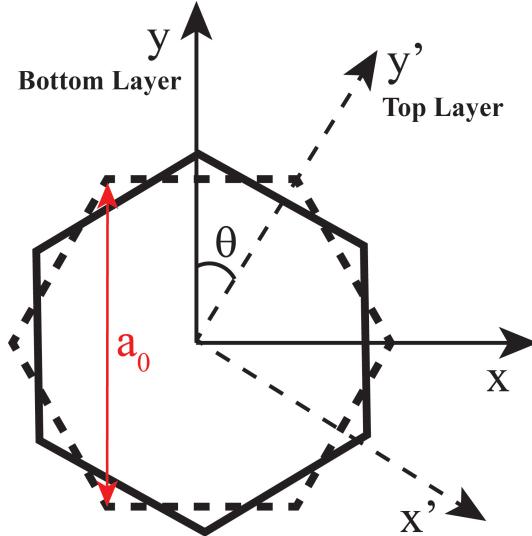


Figure 2.3: Sketch of a possible rotational misalignment between the top and bottom 2D layer, x - y is the reference coordinate for the bottom 2D layer and x' - y' is the reference coordinate for the top 2D layer. θ is the rotational misalignment angle. We assume the top layer and the bottom layer have the same lattice constant a_0 .

Consistently with Sec.2.2.2 we set $u_{T,\mathbf{k}_T}(\mathbf{r}', z) \approx u_{0T}(\mathbf{r}', z)$, $u_{B,\mathbf{k}_B}(\mathbf{r}, z) \approx u_{0B}(\mathbf{r}, z)$,

where $u_{0T}(\mathbf{r}', z)$, $u_{0B}(\mathbf{r}, z)$ are the periodic part of the Bloch function respectively at the band edge in the top and bottom layer. We then denote with \mathbf{K}_{0T} the wave-vector at the conduction band edge in the top layer (expressed in the top layer coordinate system), and with \mathbf{K}_{0B} the wave-vector at the valence band edge in the bottom layer (expressed in the principal coordinate system); the derivations in this section account for the fact that \mathbf{K}_{0T} and \mathbf{K}_{0B} may be inequivalent extrema (i.e. $\mathbf{K}_{0T} \neq \mathbf{K}_{0B}$).

By expressing \mathbf{r}' and \mathbf{k}' in the principal coordinate system we can essentially follow the derivations in Sec.2.2.2 and write the matrix element as

$$\begin{aligned} M(\mathbf{k}_T, \mathbf{k}_B) &\simeq \frac{1}{N_C} \sum_{j=1}^{N_C} e^{i(\mathbf{q} + \mathbf{Q}_D) \cdot \mathbf{r}_j} F_L(\mathbf{r}_j) \times \\ &\times \int_{\Omega_C} d\mathbf{r} \int dz u_{0T}^\dagger(\hat{R}_{B \rightarrow T}(\mathbf{r}_j + \boldsymbol{\rho}), z) V_B(z) u_{0B}(\mathbf{r}_j + \boldsymbol{\rho}, z) \end{aligned} \quad (2.25)$$

where $\mathbf{q} = (\mathbf{k}_B - \mathbf{k}_T)$ and we have introduced the vector

$$\mathbf{Q}_D = \mathbf{K}_{0B} - \hat{R}_{T \rightarrow B} \mathbf{K}_{0T} \quad (2.26)$$

Eq.2.25 is an extension of Eq.2.10 that accounts for a possible rotational misalignment between the 2D layers and describes also the tunneling between inequivalent extrema. The vector \mathbf{Q}_D is zero only for tunneling between equivalent extrema (i.e. $\mathbf{K}_{0B} = \mathbf{K}_{0T}$) and for a perfect rotational alignment (i.e. $\theta = 0$). Considering a case where all extrema are at the \mathbf{K} point, we have $|\mathbf{K}_{0B}| = |\mathbf{K}_{0T}| = 4\pi/3a_0$, then for $\mathbf{K}_{0B} = \mathbf{K}_{0T}$ the magnitude of \mathbf{Q}_D is simply given by $Q_D = (8\pi/3a_0) \sin(\theta/2)$.¹³

One significant difference in Eq.2.25 compared to Eq.2.10 is that, in the presence of rotational misalignment, the top layer Bloch function $u_{0T}(\hat{R}_{B \rightarrow T}\mathbf{r}, z)$ has a different periodicity in the principal coordinate system from the bottom layer $u_{0B}(\mathbf{r}, z)$. Consequently the integral over the unit cells of the bottom 2D layer

is not the same in all unit cells, so that the derivations going from Eq.2.10 to Eq.2.15 should be rewritten accounting for a matrix element $M_{B0,j}$ depending on the unit cell j . Such an $M_{B0,j}$ could be formally included in the calculations by defining a new scattering spectrum that includes not only the inherently random fluctuations of the potential $F_L(\mathbf{r})$, but also the cell to cell variations of the matrix element $M_{B0,j}$. A second important difference of Eq.2.25 compared to Eq.2.10 lies in the presence of \mathbf{Q}_D in the exponential term multiplying $F_L(\mathbf{r}_j)$.

For the case of tunneling between inequivalent extrema and with a negligible rotational misalignment (i.e. $\theta \approx 0$), Eq.2.26 gives $\mathbf{Q}_D = \mathbf{K}_{0B} - \mathbf{K}_{0T}$ and the current can be expressed as in Eq.2.15 but with the scattering spectrum evaluated at $|\mathbf{q} + \mathbf{Q}_D|$. Since in this case the magnitude of \mathbf{Q}_D is comparable to the size of the Brillouin zone, the tunneling between inequivalent extrema is expected to be substantially suppressed if the correlation length L_c of the scattering spectrum $S_R(\mathbf{q})$ is much larger than the lattice constant, as it has been assumed in all the derivations.

Quite interestingly, the derivations in this section suggest that a possible rotational misalignment is expected to affect the absolute value of the tunneling current but not to change significantly its dependence on the terminal voltages.

From a technological viewpoint, if the stack of the 2D materials is obtained using a dry transfer method the rotational misalignment appears inevitable.^{14,40} Experimental results have shown that, when the stack of 2D materials is obtained by growing the one material on top of the other, the top 2D and bottom 2D layer can have a fairly good angular alignment.^{41,42}

2.2.5 An Analytical Approximation for the Tunneling Current

The numerical calculations for the tunneling current obtained with the model derived in Secs.2.2.2 and 2.2.3 will be presented in Sec.2.3, while in this section we discuss an analytical, approximated expression for the tunneling current which is mainly useful to gain an insight about the main physical and material parameters affecting the current versus voltage characteristic of the Thin-TFET. In order to derive an analytical current expression we start by assuming a parabolic energy relation and write

$$E_{VB}(\mathbf{k}_B) = E_{VB} - \frac{\hbar^2 k_B^2}{2m_v} \quad E_{CT}(\mathbf{k}_T) = E_{CT} + \frac{\hbar^2 k_T^2}{2m_c} \quad (2.27)$$

where $E_{VB}(\mathbf{k}_B)$, $E_{CT}(\mathbf{k}_T)$ are the energy relation respectively in the bottom layer valence band and top layer conduction band and m_v , m_c the corresponding effective masses.

In the analytical derivations we neglect the energy broadening and start from Eq.2.15, so that the model is essentially valid only in the on-state of the device, that is for $E_{CT} < E_{VB}$.

We now focus on the integral over \mathbf{k}_B and \mathbf{k}_T in Eq.2.15 and first introduce the polar coordinates $\mathbf{k}_B=(k_B,\theta_B)$, $\mathbf{k}_T=(k_T,\theta_T)$, and then use Eq.2.27 to convert the integrals over \mathbf{k}_B , \mathbf{k}_T to integrals over respectively E_B , E_T , which leads to

$$\begin{aligned} I &\propto \int_{\mathbf{k}_T} \int_{\mathbf{k}_B} d\mathbf{k}_T d\mathbf{k}_B S_F(q) \delta(E_B(\mathbf{k}_B) - E_T(\mathbf{k}_T)) (f_B - f_T) \\ &= \frac{m_c m_v}{\hbar^4} \int_0^{2\pi} d\theta_B \int_0^{2\pi} d\theta_T \int_{E_{CT}}^{\infty} dE_T \int_{-\infty}^{E_{VB}} dE_B S_F(q) \delta(E_B - E_T) (f_B - f_T) \end{aligned} \quad (2.28)$$

where the spectrum $S_F(q)$ is given by Eq.2.16 and thus depends only on the magnitude q of $\mathbf{q}=\mathbf{k}_B-\mathbf{k}_T$. Assuming $E_{CT} < E_{VB}$, the Dirac function reduces one of the integrals over the energies and sets $E=E_B=E_T$, furthermore the magnitude

of $\mathbf{q}=\mathbf{k}_B-\mathbf{k}_T$ depends only on the angle $\theta=\theta_B-\theta_T$, so that Eq.2.28 simplifies to

$$I \propto \frac{m_c m_v (2\pi)}{\hbar^4} \int_0^{2\pi} d\theta \int_{E_{CT}}^{E_{VB}} dE S_F(q) (f_B - f_T) \quad (2.29)$$

In the on-state condition (i.e. for $E_{CT} < E_{VB}$), the zero Kelvin approximation for the Fermi-Dirac occupation functions f_B, f_T can be introduced to further simplify Eq.2.29 to

$$I \propto \frac{m_c m_v (2\pi)}{\hbar^4} \int_0^{2\pi} d\theta \int_{E_{min}}^{E_{max}} dE S_F(q) \quad (2.30)$$

where $E_{min} = \max\{E_{CT}, E_{FT}\}$, $E_{max} = \min\{E_{VB}, E_{FB}\}$ define the tunneling window $[E_{max} - E_{min}]$.

The evaluation of Eq.2.30 requires to express q as a function of the energy E inside the tunneling window and of the angle θ between \mathbf{k}_B and \mathbf{k}_T . By recalling $q^2 = k_B^2 + k_T^2 - 2k_B k_T \cos(\theta)$, we can use Eq.2.27 to write

$$q^2 = \frac{2m_v}{\hbar^2} (E_{VB} - E) + \frac{2m_c}{\hbar^2} (E - E_{CT}) - \frac{4\sqrt{m_c m_v}}{\hbar^2} \sqrt{(E_{VB} - E)(E - E_{CT})} \cos(\theta) \quad (2.31)$$

with $E = E_B = E_T$. When Eq.2.31 is substituted in the spectrum $S_F(q)$ the resulting integrals over E and θ in Eq.2.30 cannot be evaluated analytically. Therefore to proceed further we now examine the maximum value taken by q^2 . The θ value leading to the largest q^2 is $\theta = \pi$, and the resulting q^2 expression can be further maximized with respect to the energy E varying in the tunneling window. The energy leading to maximum q^2 is

$$E_M = \frac{E_{CT} + (m_c/m_v)E_{VB}}{1 + (m_c/m_v)} \quad (2.32)$$

and the corresponding q_M^2 is

$$q_M^2 = \frac{2(m_c + m_v)(E_{VB} - E_{CT})}{\hbar^2} \quad (2.33)$$

When neither the top nor the bottom layer are degenerately doped the tunneling window is given by $E_{min}=E_{CT}$ and $E_{max}=E_{VB}$, in which case the E_M defined in Eq.2.32 belongs to the tunneling window and the maximum value of q^2 is given by Eq.2.33. If either the top or the bottom layer is degenerately doped the Fermi levels become the edges of the tunneling window and the maximum value of q^2 may be smaller than in Eq.2.33.

A drastic simplification in the evaluation of Eq.2.30 is obtained for $q_M^2 \ll 1/L_c^2$, in which case Eq.2.16 returns to $S_F(q) \approx \pi L_c^2$, so that by substituting $S_F(q)$ in Eq.2.29 and then in Eq.2.15 the expression for the current simplifies to

$$I \simeq \frac{eg_v A(m_c m_v)}{\hbar^5} |M_{B0}|^2 e^{-2\kappa T_{IL}} L_c^2 (E_{max} - E_{min}) \quad (2.34)$$

where we recall that $E_{min} = \max\{E_{CT}, E_{FT}\}$, $E_{max} = \min\{E_{VB}, E_{FB}\}$ define the tunneling window.

It should be noticed that Eq.2.34 is consistent with a complete loss of momentum conservation, so that the current is simply proportional to the integral over the tunneling window of the product of the density of states in the two 2D layers. Since for a parabolic effective mass approximation the density of states is energy independent, the current turns out to be simply proportional to the width of the tunneling window. In physical terms, Eq.2.34 corresponds to a situation where the scattering produces a complete momentum randomization during the tunneling process.

As can be seen, as long as the top layer is *not* degenerate we have $E_{min}=E_{CT}$ and the tunneling window widens with the increase of the top gate voltage $V_{T,G}$, hence according to Eq.2.34 the current is expected to increase linearly with $V_{T,G}$. However, when the tunneling window increases to such an extent that q_M^2 becomes comparable to or larger than $1/L_c^2$, then part of the q values in the

integration of Eq.2.30 belong to the tail of the spectrum $S_F(q)$ defined in Eq.2.16, and so their contribution to the current becomes progressively vanishing. The corresponding physical picture is that, while the tunneling window increases, the magnitude of the wave-vectors in the two 2D layers also increases, and consequently the scattering can no longer provide momentum randomization for all the possible wave-vectors involved in the tunneling process. Under these circumstances the current is expected to first increase sub-linearly with V_{TG} and eventually saturate for large enough V_{TG} values.

2.3 Numerical Results for the Tunneling Current

2.3.1 Parabolic Band Approximation

The 2D materials used for the tunneling current calculations reported in this paper are the hexagonal monolayer MoS_2 and WTe_2 . The band structure for MoS_2 and WTe_2 have been calculated by using a density functional theory (DFT) approach,^{18,43} showing that these materials have a direct bandgap with the band edges for both the valence and the conduction band residing at the K point in the 2D Brillouin zone. Fig.2.4 shows that in a range of about 0.4 eV from the band edges the DFT results can be fitted fairly well by using an energy relation based on a simple parabolic effective mass approximation (dashed lines). Hence the parabolic effective mass approximation appears adequate for the purposes of this work, which is focussed on a device concept for extremely small supply voltages (< 0.5 V). The values for the effective masses inferred from the fitting of the DFT calculations are tabulated in Tab.2.1 together with some other material

parameters relevant for the tunneling current calculations.

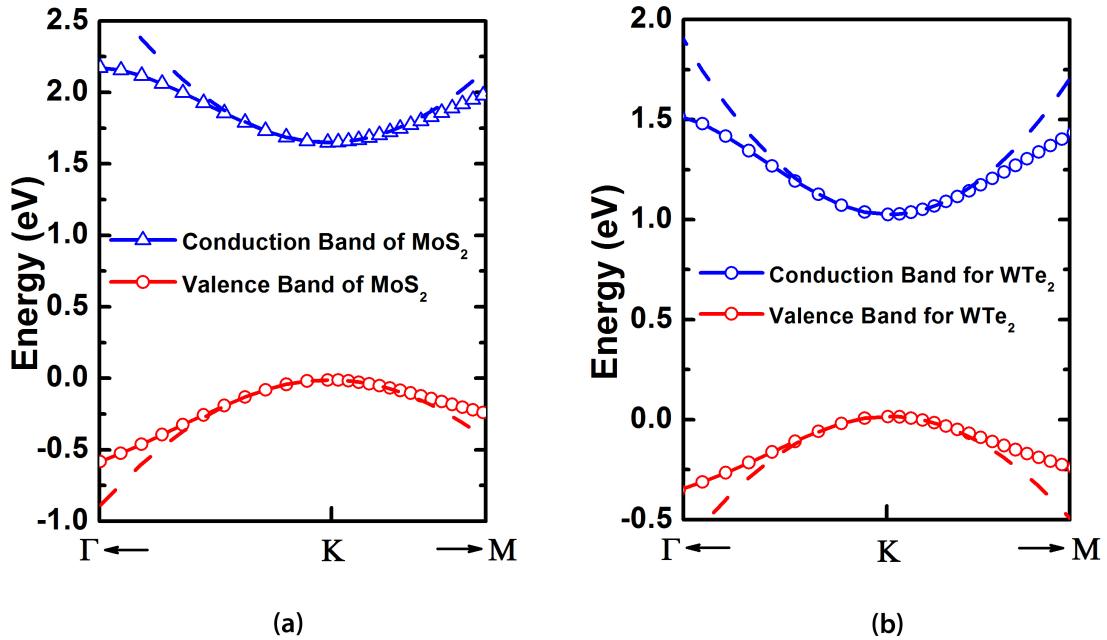


Figure 2.4: (a) Band structure for hexagonal monolayer MoS_2 and (b) hexagonal monolayer WTe_2 as obtained using DFT method described in the paper of C. Gong et.al.¹⁸ The dashed lines represent the analytical approximation obtained with a parabolic effective mass model.

2.3.2 Effects of Correlation Lengthes, Interlayer Thicknesses and Energy Broading

	Bandgap (eV)	Electron affinity (χ)	Conduction band effective mass (m_0)	Valence band effective mass (m_0)
MoS_2	1.8	4.30	0.378	0.461
WTe_2	0.9	3.65	0.235	0.319

Table 2.1: The band gaps, electron affinities and effective masses used for MoS_2 and WTe_2

In all current calculations we assume a top gate work function of 4.17 eV (Aluminium) and back gate work function of 5.17 eV (p++ Silicon) and the top and bottom oxide have an effective oxide thickness (EOT) of 1 nm (see Fig.2.1). The top 2D layer consists of hexagonal monolayer MoS₂ while the bottom 2D layer is hexagonal monolayer WTe₂. An *n*-type and *p*-type doping density of 10^{12} cm^{-2} by impurities and full ionization are assumed respectively in the top and bottom 2D layer and the relative dielectric constant of the interlayer material is set to 4.2 (e.g. boron nitride). The voltage V_{DS} between the drain and the source is set to 0.3 V and the back gate is grounded for all calculations, unless otherwise stated.

As already pointed out in Sec.2.2.2, it is very difficult to derive a quantitative expression for the tunneling matrix element $M_{B,0}$. However, the value of $M_{B,0}$ could be inferred from the experimental data. In the lack of experimental data for a vertical transistor consisting of transition-metal dichalcogenides, we have set the value of $M_{B,0}$ to be 0.01 eV in our calculations so that the resultant current density is in the same order of magnitude with the experimental value reported in the graphene/hBN system.⁴⁴

In Fig. 2.5, the results of numerical calculations are shown for the band alignment and the current density versus the top gate voltage V_{TG} . Figure 2.5(a) shows that the top gate voltage can effectively govern the band alignment in the device and, in particular, the crossing and uncrossing between the conduction band minimum E_{CT} in the top layer and the valence band maximum E_{VB} in the bottom layer, which discriminates between the on and off state of the transistor.

The I_{DS} versus V_{TG} characteristic in Fig.2.5(b) can be roughly divided into three different regions: sub-threshold region, linear region and saturation re-

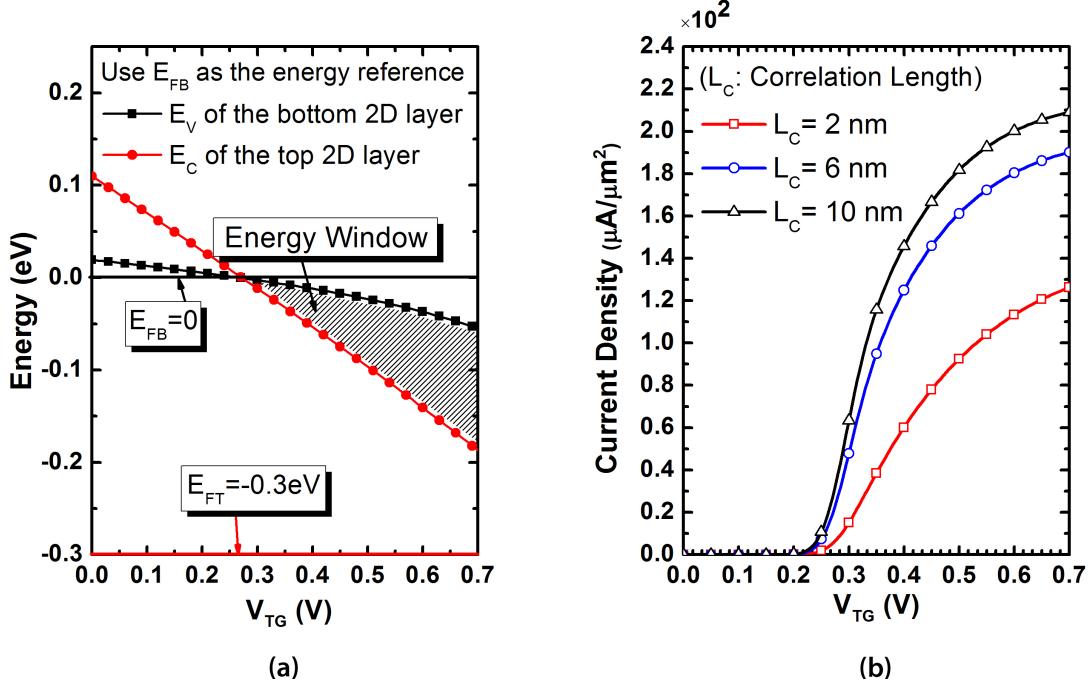


Figure 2.5: Numerical results of (a) band alignment versus the top gate voltage V_{TG} and (b) tunnel current density versus the top gate voltage V_{TG} for different values of the correlation length L_C . The parameters used in (b) are: matrix element is $M_{B0} = 0.01\text{ eV}$; decay constant of wave-function in the interlayer is $\kappa = 3.8\text{ nm}^{-1}$; energy broadening is $\sigma = 10\text{ meV}$ and interlayer thickness is $T_{IL} = 0.6\text{ nm}$ (e.g. 2 atomic layers of BN). $V_{BG} = 0$ and $V_{DS} = 0.3\text{ V}$ in both (a) and (b).

gion. The sub-threshold region corresponds to the condition $E_{CT} > E_{VB}$ (see also Fig.2.5(a)), where the very steep current dependence on V_{TG} is illustrated better in Fig.2.6 and will be discussed below.

In the second region I_{DS} exhibits an approximately linear dependence on V_{TG} , in fact the current is roughly proportional to the energy tunneling window, as discussed in Sec.2.2.5 and predicted by Eq.2.34, because the tunneling window is small enough that the condition $q_M^2 \ll 1/L_c^2$ is fulfilled. In this region I_{DS} is proportional to the long-wavelength part of scattering spectrum (i.e. small q values), hence the current increases with L_c , as expected from Eq.2.34. The

super-linear behavior of I_{DS} at small V_{TG} values observed in Fig.2.5(b) is due to the tail of the Fermi occupation function in the top layer. When V_{TG} is increased above approximately 0.5V, the current in Fig.2.5(b) enters the saturation region, where I_{DS} increasing with V_{TG} slows down because of the decay of the scattering spectrum $S_R(q)$ for q values larger than $1/L_c$ (see Eq.2.16).

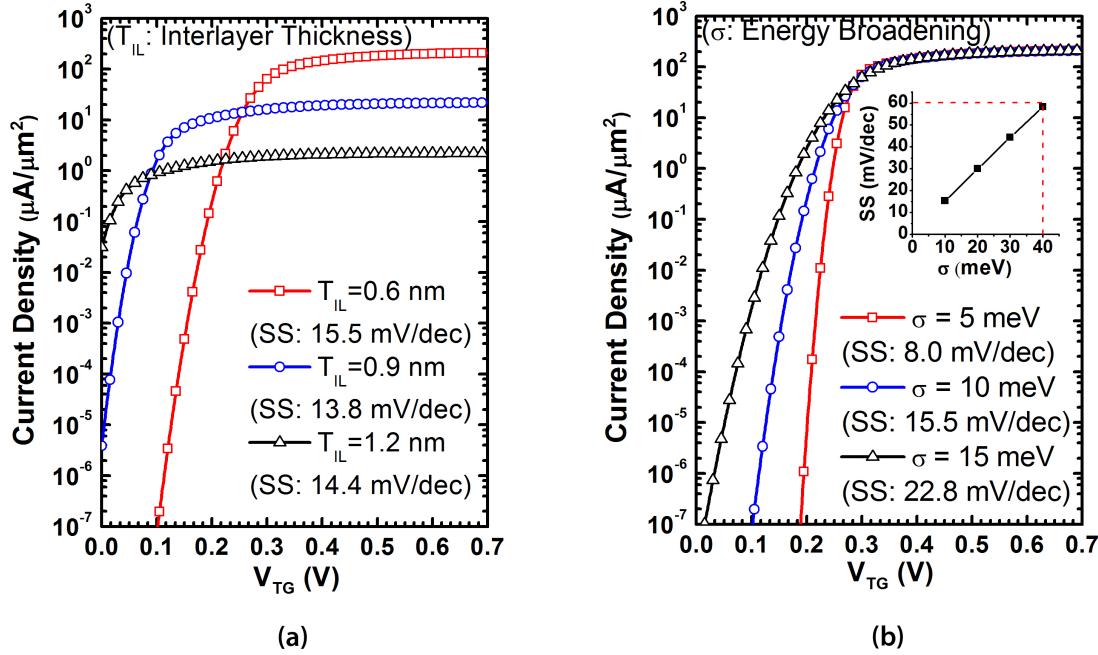


Figure 2.6: Numerical calculations for: (a) current density versus V_{TG} with several interlayer thicknesses; (b) current density versus V_{TG} with different values of energy broadening σ . The insert shows that SS increases with σ , and a SS value of 60 mV/dec corresponds to a energy broadening as high as 40 meV. The matrix element is $M_{B0} = 0.01 \text{ eV}$; the decay constant of wave-function in the interlayer is $\kappa = 3.8 \text{ nm}^{-1}$. In (a) the energy broadening is $\sigma = 10 \text{ meV}$. In (b) the interlayer thickness is $T_{IL} = 0.6 \text{ nm}$ (e.g. 2 atomic layers of BN). $V_{BG} = 0$ and $V_{TG} = 0.3 \text{ V}$ in both (a) and (b).

In Fig.2.6 we analyze the I-V curves for different interlayer thicknesses T_{IL} and broadening energies σ ; in all cases an average inverse sub-threshold slope (SS) is extracted in the I_{DS} range from 10^{-5} and $10^{-2} \mu\text{A}/\mu\text{m}^2$. Figure 2.6(a) shows that the tunneling current increases exponentially by decreasing T_{IL} , and the de-

cay constant $\kappa=3.8 \text{ nm}^{-1}$ employed in our calculations results in a dependence on T_{IL} that is roughly consistent with the dependence experimentally reported in graphene based interlayer tunneling devices.¹⁵ The threshold voltages are also shifted to lower values by increasing T_{IL} . It can be seen that the T_{IL} impact on SS is overall weak and a very steep sub-threshold region is obtained for all the T_{IL} values examined in Fig.2.6(a). This is because, in order for the Thin-TFET to obtain a small SS, it is absolutely necessary that V_{TG} has a tight control on the electrostatic potential in the top semiconductor layer, but has a negligible influence on the potential of the bottom semiconductor layer. The SS is thus insensitive to T_{IL} as long as T_{IL} does not change the control of V_{TG} on such potentials. In short, for Thin-TFETs, a larger interlayer thickness reduces substantially the current density, but does not deteriorate SS.

Figure 2.6(b) shows that according to the model employed in our calculations SS is mainly governed by the parameter σ of the energy broadening (Eq.2.22). This result is expected, as already mentioned in Sec.2.2.3, since in our model the energy broadening is the physical factor setting the minimum value for SS and the I_{DS} versus V_{TG} approaches a step-like curve when σ is zero due to the step-like DoS of these 2D semiconductors.⁴⁵ More specifically, Fig.2.6(b) shows that according to our calculations the Thin-TFET may be able to provide an SS below the 60mV/dec (i.e. the limit of conventional MOSFETs at room temperature), even for fairly large broadening energies up to about 40 meV. It is here worth noting that energy broadening and band tails have been already recognized as a fundamental limit to the SS of band-to-band tunneling transistors,⁴⁶ and are not a specific concern of the Thin-TFET. As already mentioned in Sec.2.2.3, the band tails in 3D semiconductors have been investigated by using thermal measurements and are described in terms of the so called Urbach

parameter E_0 .^{34,35} Values for E_0 comparable to the room temperature thermal energy, $k_B T \approx 26\text{ meV}$, have been reported for GaAs and InP.^{47,48} Our results suggest that energy broadening and band tails in 2D materials play a critical role in the minimum SS attainable by Thin-TFETs, and at the time of writing we are not aware of experimental data reported for band tails in monolayers of transition-metal dichalcogenides.

2.4 N-type and P-type Thin-TFETs

Out of various 2D semiconductors studied by density function theory calculations¹⁸ and experimental efforts, we chose the trigonal prismatic coordination monolayer (2H) WSe₂ and the octahedral coordination (CdI₂ crystal structure) monolayer (1T) SnSe₂ (see Fig.2.7). WSe₂/SnSe₂ stacked-monolayer heterojunction can potentially form a near broken band alignment, which reduces the voltage drop in the van der Waals gap in the on-state condition.⁶ Since there is no experimental band alignment reported for *monolayer* WSe₂ and SnSe₂, the band alignment of the WSe₂/SnSe₂ system used in this work are based on the existing experimental results of *multilayer* WSe₂ and SnSe₂,^{49–51} while their approximated effective masses are based on the DFT results of monolayer WSe₂ and SnSe₂¹⁸ (see Fig.2.7).

Following the complex band method,⁵² we assume the effective barrier height E_B of the van der Waals gap is 1 eV and the electron mass in the van der Waals gap is the free electron mass m_0 , thus the decay constant is $\kappa = \sqrt{2m_0E_B/\hbar} = 5.12\text{ nm}^{-1}$. In our model, we set the scattering correlation length L_C in $S_F(\mathbf{q})$ to $L_C=10\text{ nm}$, which is also consistent with the value employed in;¹⁴ the energy

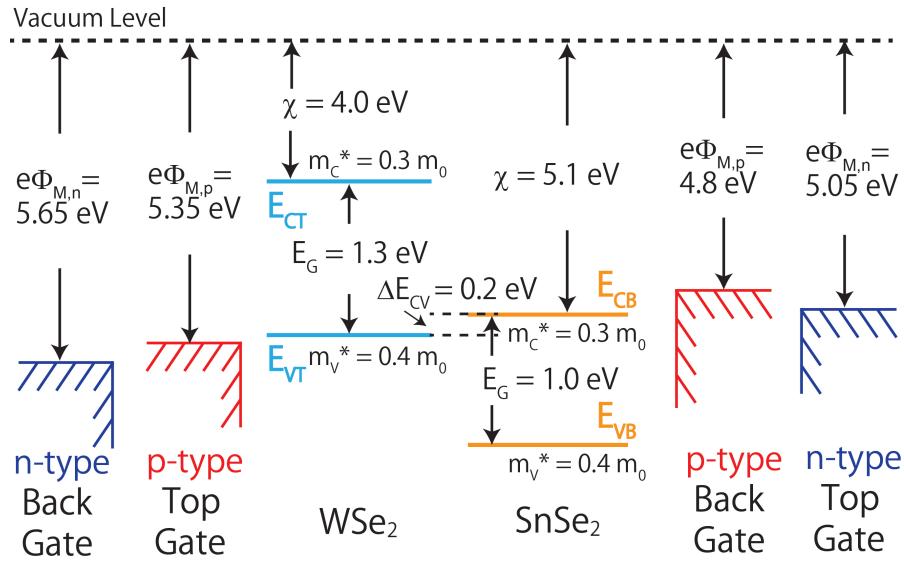


Figure 2.7: An example to realize both *n*-type and *p*-type Thin-TFETs using one pair of 2D semiconductors (2H-WSe₂ and 1T-SnSe₂) with near broken gap band alignment. For the *n*-type Thin-TFET, SnSe₂ is the top (i.e. drain) 2D layer and WSe₂ is the bottom (i.e. source) 2D layer, along with the top and back gate labeled as *n*-type in blue. While for the *p*-type Thin-TFET, WSe₂ is the top (i.e. drain) 2D layer and SnSe₂ is the bottom (i.e. source) 2D layer, along with the top and back gate labeled as *p*-type in red; Band gaps, electron affinities, effective masses are shown for WSe₂ and SnSe₂. The *n*-type and *p*-type metal work functions are tuned to give symmetric threshold voltages for the *n*-type and *p*-type Thin-TFETs.

broadening σ is set to be 10 meV.

M_{B0} in Eq.2.15 is directly related to the interlayer charge transfer time τ across the van der Waals gap, which can be written as⁵³

$$\tau^{-1} = \frac{2\pi}{\hbar} \rho |M_{B0}|^2 e^{-2\kappa T_{vdW}} S_F(q) \quad (2.35)$$

where $\rho = g_v m^* / \pi \hbar^2$ is the density of states (DOS). Recall the tunneling current can be written as:

$$J_T = \frac{g_v e |M_{B0}|^2 A}{4\pi^3 \hbar} e^{-2\kappa T_{vdW}} \times \int_{\mathbf{k}_T} \int_{\mathbf{k}_B} d\mathbf{k}_T d\mathbf{k}_B S_F(q) S_E(E_B - E_T) (f_B - f_T) \quad (2.36)$$

As can be seen from Eq.2.35 and the expression of the scattering potential spectrum $S_F(q)$ (given after Eq.2.36), due to scattering in our model, τ increases with increasing q , which is the magnitude of the wave-vector difference across the van der Waals gap defined as $q=|\mathbf{k}_T-\mathbf{k}_B|$. In a recent experiment, a charge transfer time of 25 fs has been observed across the van der Waals gap between a stacked-monolayer MoS₂/WS₂ heterostructure, which, according to Eq.2.35, gives us $M_{B0} \sim 0.02$ eV when $q=0$. We recognize that the charge transfer time might be different for different 2D heterojunctions, nevertheless, this experimentally determined charge transfer time is a reasonable value to use for the first pass estimate. Thus, we choose $M_{B0}=0.02$ eV in all following simulations.

Throughout this work, the gate length is set to be 15 nm, the back gate and source are grounded. An effective oxide thickness (EOT) of 1 nm is used for both the top and back oxide, which gives a top (back) oxide capacitance C_{TG} (C_{BG}) of 0.518 fF/ μ m. The thickness of the van der Waals gap is set to 0.35 nm, unless specified otherwise. We assume the relative dielectric constant of the van der Waals gap is 1.0, therefore the van der Waals gap capacitance C_{vdW} is 0.38 fF/ μ m. The external total access resistances are considered after the intrinsic device performance is discussed first (Figs.2.8 and 2.9).

The example material systems for *n*-type and *p*-type Thin-TFETs based on the stacked-monolayer WSe₂ and SnSe₂ are shown in Fig.2.7. The metal work functions are tuned to obtain a symmetric threshold voltage for the *n*-type and the *p*-type Thin-TFET. Figure 2.8(a) shows the band alignment versus V_{TG} . V_{TG}

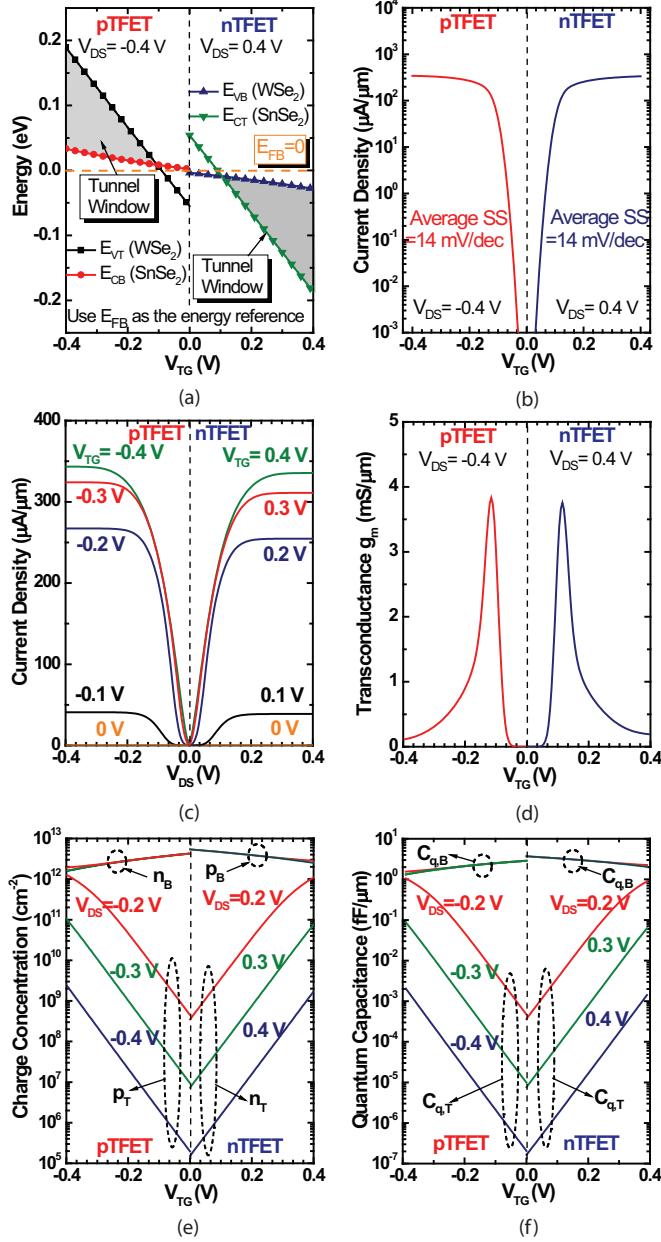


Figure 2.8: For the *n*-type and *p*-type Thin-TFETs shown in Fig. 2.7: (a) the band alignment versus V_{TG} ; (b) Current density versus V_{TG} , the average SS is calculated from $10^{-3} \mu\text{A}/\mu\text{m}$ to $10 \mu\text{A}/\mu\text{m}$; (c) the current density versus V_{DS} at various V_{TG} ; (d) the transconductance versus V_{TG} ; (e) the carrier concentration in the top and bottom 2D layers versus V_{TG} at various V_{DS} ; (f) the quantum capacitances of the top and bottom 2D layers versus V_{TG} at various V_{DS} ;

can effectively control the vertical band alignment in the device by controlling primarily the band edge of the top (i.e. drain) layer while having a weak effect on the band edge of the bottom (i.e. source) layer, so that a tunneling window is modulated. Figure.2.8(b) shows I_D versus V_{TG} transfer curves with very compelling average SS of ~ 14 mV/dec averaged from $10^{-3} \mu\text{A}/\mu\text{m}$ to $10 \mu\text{A}/\mu\text{m}$. The I_D versus V_{DS} family curves are shown in Fig.2.8(c). I_D saturates for V_{DS} when $V_{DS} > \sim 0.2$ V. The superlinear onset is also observed and the so called V_{DS} threshold voltage increases at lower V_{TG} .⁵⁴ A peak transconductances of ~ 4 mS/ μm is observed around $V_{TG}=0.12$ V (Fig.2.8(d)), which are much larger than ~ 0.8 mS/ μm reported peak transconductances of 10 nm Fin-FET.⁵⁵ In Fig.2.8(e), the top gate changes the carrier concentrations of the top 2D semiconductor much faster than of the bottom 2D semiconductor under different V_{DS} . The ability to efficiently change a hole (electron) concentration in the top 2D semiconductor while keeping a high electron (hole) concentration in the bottom 2D semiconductor is vital to achieve good electrostatics control of these Thin-TFETs. The quantum capacitance associated with the top and bottom semiconductor layers can be expressed as Eq.2.37:

$$C_{Q,T(B)} = - \left[\frac{e\partial p_{T(B)}}{\partial \phi_{p,T(B)}} + \frac{e\partial n_{T(B)}}{\partial \phi_{n,T(B)}} \right] \quad (2.37)$$

The quantum capacitances are plotted in Figure.2.8(f) under various bias conditions.

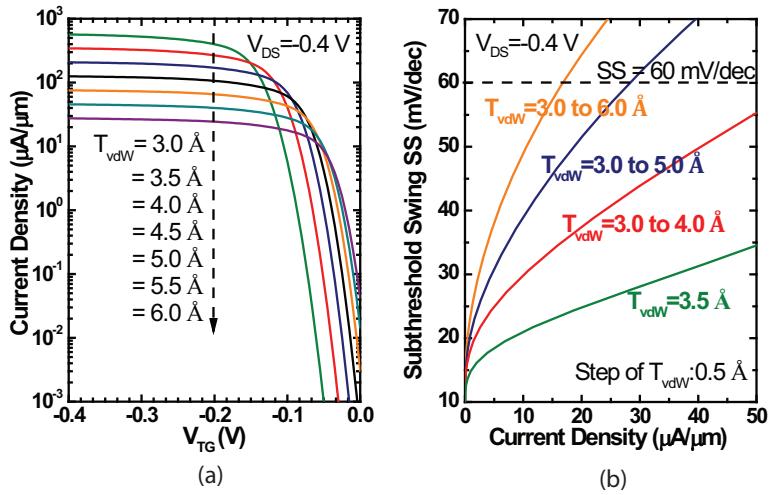


Figure 2.9: Effect of van der Waals gap thickness variation on a *p*-type Thin-TFET: (a) tunnel current density versus V_{TG} for different van der Waals gap thicknesses T_{vdW} ; (b) differential SS versus current density assuming an evenly distributed van der Waals gap thickness T_{vdW} in the specified range.

2.4.1 Effects of Non-uniform van der Waals Gap Thickness and Access Resistance

Due to the nature of van der Waals bonds, the van der Waals gap thickness is subject to intercalation of atoms/ions, interlayer rotational misalignment between 2D layers etc. For instance, in bilayer mechanically stacked Molybdenum Disulfide (MoS_2) with an interlayer twist, a maximum variation of 0.059 nm⁵⁶ was experimentally verified in the van der Waals gap thickness [22]. Surface roughening due to ripples in 2D crystals or roughness of the underlying substrates can also introduce van der Waals gap variations.⁵⁷ Meanwhile, tunneling probability is very sensitive to the tunneling distance, namely the van der Waals gap thickness in a Thin-TFET, which makes it important to investigate effects of a non-uniform van der Waals thickness. First, the Thin-TFET I-V curves are calculated by varying the van der Waals gap thickness T_{vdW} from 0.3 nm to

0.6 nm and a step of 0.05 nm (which is roughly half of the Se covalent radius⁵⁸). The results are shown in Fig.2.9(a) for a *p*-type Thin-TFET: the on current density decreases and the threshold voltage moves towards 0 when increasing the T_{vdW} . We note that, as long as the T_{vdW} is uniform, the SS remains as steep as ~ 14 mV/dec. However, for a non-uniform T_{vdW} , SS will degrade. To estimate its impact, an evenly distributed T_{vdW} over several ranges is used in the calculated differential SS shown in Fig.2.9(b). For example, for a 2D heterojunction with an evenly distributed T_{vdW} from 0.3 nm to 0.5 nm and a step of 0.05 nm, we take the corresponding I_D - V_{TG} curve for each T_{vdW} (i.e. 0.3 nm, 0.35 nm, 0.4 nm, 0.45 nm, and 0.5 nm) shown in Fig.2.9(a) and average them over the T_{vdW} range to obtain the overall I_D - V_{TG} curve for the calculation of SS. Fig.2.9(b) shows that up to 0.1 nm variation in T_{vdW} is tolerable, resulting in a sub-60 mV/dec SS over a decent current window (up to $50 \mu\text{A}/\mu\text{m}$). Depending on how Thin-TFETs are fabricated, the T_{vdW} non-uniformity may have different distributions. Our first look at its impact in this work highlights the importance to precisely control T_{vdW} .

A finite total access resistance has a critical impact on ultrascaled transistors. To date, how to minimize the total access resistance in 2D crystal based device still remains an open question. In Fig. 2.10, we show its effects on Thin-TFET by assuming several values for the total access resistance R_C . At a sufficiently high $|V_{DS}|$ of 0.4 V, maximum I_D is almost the same for a R_C of up to $320 \Omega\mu\text{m}$; a higher R_C decreases maximum I_D appreciably. Understandably, a lower R_C is necessary for a lower V_{DD} . In an ideal 2D conductor, the quantum limit of the total access resistance is inversely proportional to the square root of the carrier concentration; e.g. $\sim 52 \Omega\mu\text{m}$ for a carrier concentration of 10^{13} cm^{-2} .⁵⁹ Thus the access region of 2D semiconductors can be degenerately doped to minimize R_C .

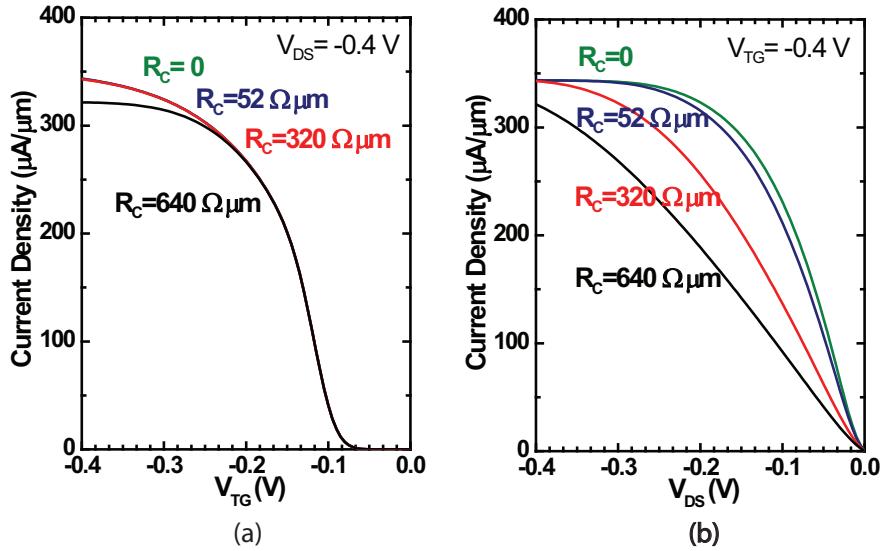


Figure 2.10: Effect of total access resistance on a *p*-type Thin-TFET: (a) I_D versus V_{TG} and (b) I_D versus V_{DS} with various total access resistance R_C values.

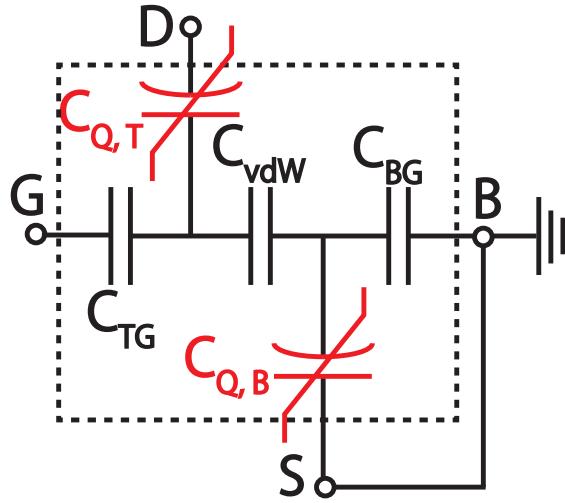


Figure 2.11: Capacitance network model of the Thin-TFET

2.4.2 Capacitance Evaluation

The gate-to-drain and gate-to-source capacitances (i.e. C_{GD} , C_{GS}) can be readily calculated from the capacitance network shown in Fig.2.11.

The quantum capacitances $C_{Q,T(B)}$ of the top (bottom) 2D semiconductor are defined in Eq.2.37 and indicated as the red non-linear capacitances in Fig.2.11. First we define C_S as:

$$1/C_S \equiv 1/C_{vdW} + 1/(C_{Q,B} + C_{BG}) \quad (2.38)$$

Then, C_{GS} and C_{GD} can be written as Eqs.2.39:

$$\begin{aligned} C_{GS} &= \frac{C_{TG}C_S}{C_{TG} + C_{Q,T} + C_S} \\ C_{GD} &= \frac{C_{TG}C_{Q,T}}{C_{TG} + C_{Q,T} + C_S} \end{aligned} \quad (2.39)$$

Due to the symmetry in these *p*-type and *n*-type Thin-TFETs as well as the similar hole and electron effective mass in these 2D crystals, we expect similar C-V characteristics for the *p*-type and *n*-type Thin-TFETs. In Fig.2.12 we plot the calculated C-V curves for the *p*-type Thin-TFETs shown in Fig.2.7. In the linear region of the I_D - V_{DS} family of curves, C_{GD} is significant, where the drain is coupled with the top gate to modulate the tunnel current. From the linear region to the saturation region, C_{GD} drops to be near zero while C_{GS} increases to its maximum. What is worthy noting is that the magnitude of a Thin-TFET capacitance is smaller than CMOS and III-V TFET benchmarked in Sec. 2.4.3 for a given gate oxide EOT thus capacitances, which stem from the serially connected capacitance components as shown in Fig.2.11. The capacitance model is useful for implementing the Thin-TFET into circuit simulations.

2.4.3 Benchmarking

The Semiconductor Research Corporation (SRC) Nanoelectronic Research Initiative (NRI) has supported research on beyond CMOS devices as reported by

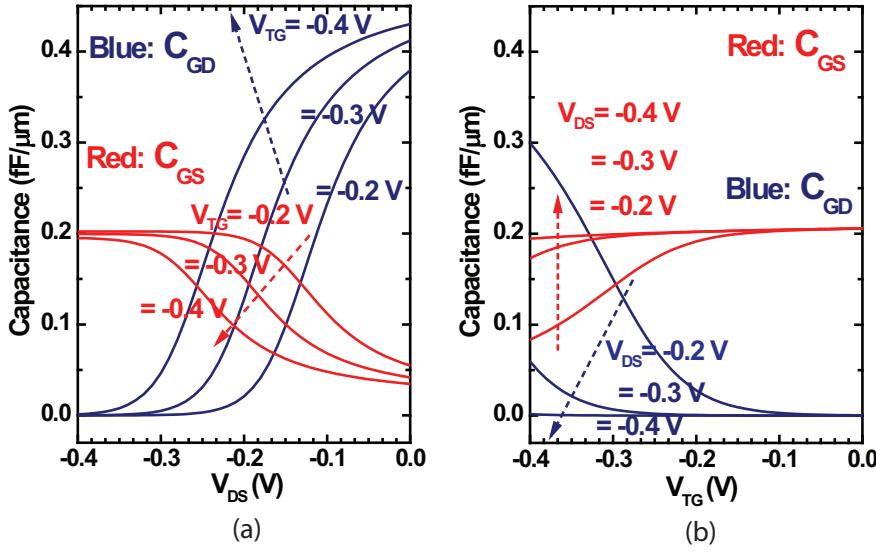


Figure 2.12: For the *p*-type Thin-TFET, (a) C_{GD} and C_{GS} versus V_{DS} at $V_{TG}=-0.2$, -0.3 , -0.4 V; (b) C_{GD} and C_{GS} versus V_{TG} at $V_{DS}=-0.2$, -0.3 , -0.4 V.

Bernstein, et al.⁶⁰ As part of the initiative, the projected performance of the beyond-CMOS devices and the CMOS of the same technology node was compared, i.e. benchmarked. The benchmarking activity has continued by Nikonov and Young^{61,62} Thin-TFET being proposed by us primarily under the support of SRC STARnet, we participated in the recent benchmarking using the Nikonov and Young (N&Y) methodology.

The N&Y methodology uses basic device performance parameters such as operating voltage ($V_{DD} = |V_{DS}|$), saturation current (I_{Dsat}), and average gate capacitance ($C_{G,avg}$), to project logic switching energy and delay. The change of the net charge under the gate ($\Delta Q = q\Delta n_s$) when V_{TG} switches from 0 to V_{DD} is the sum of the change of the net charge in the top 2D semiconductor and the bottom 2D semiconductor. The average gate capacitance ($C_{G,avg}$) is defined as $\Delta Q/V_{DD}$. Here we take the *p*-type Thin-TFET as an example, I_{Dsat} and $C_{G,avg}$ are provided in Table 2.2 for a few V_{DD} values of 0.2, 0.3, and 0.4 V and a few total

Table 2.2: Benchmarking Parameters

Parameters for Thin-TFETs with various V_{DD} and R_C						
V_{DD} (V)	0.2		0.3		0.4	
R_C ($\Omega\mu\text{m}$)	52	320	52	320	52	320
I_{Dsat} ($\mu\text{A}/\mu\text{m}$)	263	233	325	317	349	348
ΔQ ($\text{fC}/\mu\text{m}^2$)	2.34	2.80	3.33	3.72	4.30	4.47
$\Delta n_s \times 10^{12}$ ($/\text{cm}^{-2}$)	1.46	1.75	2.08	2.32	2.69	2.79
$C_{G,avg}$ ($\text{fF}/\mu\text{m}$)	0.175	0.210	0.167	0.186	0.161	0.168
Parameters for HP/LP CMOS and HetJ/HomJ TFET ⁶²						
	V_{DD} (V)	I_{Dsat} ($\mu\text{A}/\mu\text{m}$)	$C_{G,avg}$ ($\text{fF}/\mu\text{m}$)			
HP CMOS	0.73	1805	1.29			
LP CMOS	0.3	2	1.29			
HetJTFET	0.4	500	1.04			
HomJTFET	0.2	25	1.04			
Geometrical Parameters for Benchmarking						
Half-pitch (F) (nm)	EOT (nm)	Gate Length (L) (nm)	Gate Width (W) (nm)			
15	1	15	60			

access resistance R_C values of 52 and $320 \Omega\mu\text{m}$. The device parameters for High Performance (HP) CMOS, Low Power (LP) CMOS, InAs Homojunction TFET (HomJTFET) and InAs/GaSb Heterojunction TFET (HetJTFET) are taken from Ref.⁶² and we use the same geometrical parameters for all the devices as shown in Table 2.2, while neglecting the contact capacitance.

The intrinsic switching delay t_{int} and the intrinsic switching energy E_{int} are calculated by:⁶²

$$t_{int} = \frac{C_{G,avg} V_{DD}}{I_{Dsat}} \quad (2.40)$$

$$E_{int} = C_{G,avg} W V_{DD}^2$$

In Fig.2.13, we plot the projected values of t_{int} and E_{int} of the devices listed in Table 2.2.

As far as the intrinsic switching energy-delay product is concerned, the Thin-

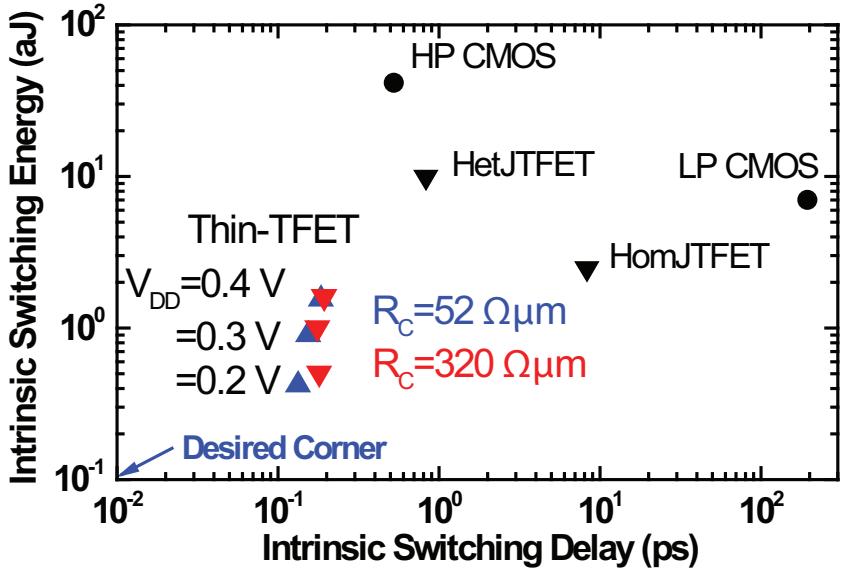


Figure 2.13: The intrinsic switching energy and delay for HP CMOS, LP CMOS, HetJTFET, HomJTFET and Thin-TFETs with $V_{DD}=0.2, 0.3, 0.4$ V and $R_C=52, 320$ $\Omega\mu\text{m}$.

TFET shows distinct energy consumption and performance advantages. For instance, Thin-TFET operation at a V_{DD} as low as 0.2 V is fast because its current is still significantly high. The most distinguishing feature of a Thin-TFET is its low intrinsic capacitance in comparison to the other devices. This advantage will be less significant when device parasitics become dominant in completed circuits.

It is observed that the Thin-TFET intrinsic switching energy-delay product moves toward the desired corner when decreasing V_{DD} from 0.4 V to 0.2 V. This is an unusual but favorable behavior for ultrascaled switches. In the case of 15 nm CMOS, I_D is roughly proportional to V_{DD} . While in the ON state of Thin-TFET, I_D has much weaker dependence on V_{TG} (see Fig. 2.10(a)) than CMOS, thus V_{DD} to I_D ratio actually decreases when scaling down V_{DD} from 0.4 V to 0.2 V. Therefore, given that $C_{G,\text{avg}}$ stays roughly the same (increasing slightly

with decreasing V_{DD}), the intrinsic switching time t_{int} slightly decreases when decreasing V_{DD} .

2.5 Discussion and Conclusions

2.5.1 Experimental Insights

Since our proposal of Thin-TFET in 2012⁶³ that is derived from our III-V TFET design,⁶ several key challenges have been identified along our pursuit in experimental demonstration of Thin-TFETs.⁶⁴ The foremost is the scarcity of electronic-grade layered materials and knowledge of their properties, in particular, the semiconductor heterojunctions with near broken gap alignment. The reasonably well-characterized material properties in the literature are largely based on bulk layered materials. An exponentially growing number of publications in the recent years on monolayer and few-layer materials are mainly theoretical calculations or based on exfoliation of naturally occurring crystals or synthesized by chemical vapor transport, which typically contains a few atomic percent of defects (impurities, vacancies etc). Both chemical vapor deposition and molecular beam epitaxy⁶⁵ are actively pursued by the community to grow electronic grade layered materials.

Besides lack of high quality layered materials and heterojunctions, the fabrication development of Thin-TFET is also challenging. It inherits all the fundamental fabrication challenges of a TFET including doping profile, alignment especially gate registry, gate dielectrics, ohmic contacts. Atomic layer deposition has been improved over years to achieve good quality gate dielectrics

on 2D crystals.⁶⁶ Using 2D dielectrics such as hexagonal boron nitride as the gate dielectrics has also been pursued.⁶⁷ Third, low resistance ohmic contacts to 2D crystal are vital to device performance. Various techniques such as external chemical doping,⁶⁸ internal chemical doping,⁶⁹ electrostatic doping such as ion doping⁷⁰ and phase-engineering from the semiconductor phase to the metallic phase of a 2D crystal,⁷¹ have been implemented to reduce the contact resistances. Furthermore, Thin-TFETs demand true precision layer number control since the properties of nearly all layered materials critically depend on the layer number when the layer number is in the range of 1-3 nm.

2.5.2 Conclusions

This paper proposed a new steep slope transistor based on the interlayer tunneling between two 2D semiconductor materials and presented a detailed model to discuss the physical mechanisms governing the device operation and to gain an insight about the tradeoffs implied in the design of the transistor.

The tunnel transistor based on 2D semiconductors has the potential for a very steep subthreshold region and the subthreshold swing is ultimately limited by the energy broadening in the two 2D materials. The energy broadening can have different physical origins such as disorder, charged impurities in the 2D layers or in the surrounding materials³⁹,³⁷ phonon scattering⁷² and microscopic roughness at interfaces.³⁸ In our calculations we accounted for the energy broadening by assuming a simple gaussian energy spectrum with no explicit reference to a specific physical mechanism. However, a more detailed and quantitative description of the energy broadening is instrumental in physical

modeling of the device and its design.

Quite interestingly, our analysis suggests that, while a possible rotational misalignment between the two 2D layers can affect the absolute value of the tunneling current, the misalignment is not expected to significantly degrade the steep subthreshold slope, which is the crucial figure of merit for a steep slope transistor.

An optimal operation of the device demands a good electrostatic control of the top gate voltage V_{TG} on the band alignments in the material stack, as shown for example in Fig.2.5(a), which may become problematic if the electric field in the interlayer is effectively screened by the high electron concentration in the top 2D layer. Consequently, since high carrier concentrations in the 2D layers are essential to reduce the layer resistivities, a tradeoff exists between the gate control and layer resistivities; as a result, doping concentrations in these 2D layers are important design parameters in addition to tuning the threshold voltage. In this respect, chemical doping of TMD materials have been recently demonstrated,^{68,73} however these doping technologies are still far less mature than they are for 3D semiconductors, and improvements in in-situ doping will be very important for optimization of the device performance. Since our model does not include the lateral transport in the 2D materials, an exploration of the above design tradeoffs goes beyond the scope of the present paper and demands the development of more complete transport models.

The transport model proposed in this work does not account for possible traps or defects assisted tunneling, which have been recently recognized as a serious hindrance to the experimental realization of Tunnel-FETs exhibiting a sub-threshold swing better than 60 mV/dec.^{11,12} A large density of states in

the gap of the 2D materials may even lead to a Fermi level pinning that would drastically degrade the gate control on the band alignment and undermine the overall device operation. In this respect, from a fundamental viewpoint the 2D crystals may offer advantages over their 3D counterparts because they are inherently free of broken/dangling bonds at the interfaces.¹⁹ However, the fabrication technologies for 2D crystals are still in an embryonal stage compared to technologies for conventional semiconductors, hence the control of defects in the 2D materials will be a challenge for the development of the proposed tunneling transistor.

The simulation results reported in this paper indicate that the newly proposed transistor based on interlayer tunneling between two 2D materials has the potential for a very steep turn-on characteristic, because the vertical stack of 2D materials having an energy gap is probably the device structure that allows for the most effective, gate controlled crossing and uncrossing between the edges of the bands involved in the tunneling process. A uniform van der Waals gap thickness and low total access resistance are vital to optimize the Thin-TFET performance. The benchmark study shows Thin-TFETs may have distinct advantages over CMOS and III-V TFETs in term of both performance and energy consumption at low supply voltages. Our modeling approach based on the Bardeen's transfer Hamiltonian is by no means a complete device model but instead a starting point to gain insight about its working principle and its design. At the present time an experimental demonstration of the device appears of crucial importance, first of all to validate the device concept, and then to help estimate the numerical value of a few parameters in the transport model that can be determined only by comparing to experiments.

BIBLIOGRAPHY

- ¹J. M. Rabaey, J. Ammer, T. Karalar, S. Li, B. Otis, M. Sheets, and T. Tuan, “Picoradios for wireless sensor networks: the next challenge in ultra-low power design,” in *Solid-State Circuits Conference, 2002. Digest of Technical Papers. ISSCC. 2002 IEEE International*, vol. 1. IEEE, 2002, pp. 200–201.
- ²R. Amirtharajah and A. P. Chandrakasan, “Self-powered signal processing using vibration-based power generation,” *Solid-State Circuits, IEEE Journal of*, vol. 33, no. 5, pp. 687–695, 1998.
- ³R. G. Dreslinski, M. Wieckowski, D. Blaauw, D. Sylvester, and T. Mudge, “Near-threshold computing: Reclaiming moore’s law through energy efficient integrated circuits,” *Proceedings of the IEEE*, vol. 98, no. 2, pp. 253–266, 2010.
- ⁴I. W. Group *et al.*, “International technology roadmap for semiconductors, 2011,” URL <http://www.itrs.net>, 2011.
- ⁵A. C. Seabaugh and Q. Zhang, “Low-voltage tunnel transistors for beyond cmos logic,” *Proceedings of the IEEE*, vol. 98, no. 12, pp. 2095–2110, 2010.
- ⁶G. Zhou, R. Li, T. Vasen, M. Qi, S. Chae, Y. Lu, Q. Zhang, H. Zhu, J.-M. Kuo, T. Kosel *et al.*, “Novel gate-recessed vertical inas/gasb tfets with record high ion of $180 \mu\text{A}/\mu\text{m}$ at $V_{DS} = 0.5 \text{ V}$,” in *Electron Devices Meeting (IEDM), 2012 IEEE International*. IEEE, 2012, pp. 32–6.
- ⁷K. Tomioka, M. Yoshimura, and T. Fukui, “Sub 60 mv/decade switch using an inas nanowire–si heterojunction and turn-on voltage shift with a pulsed doping technique,” *Nano letters*, 2013.
- ⁸L. Knoll, Q.-T. Zhao, A. Nichau, S. Trelenkamp, S. Richter, A. Schafer, D. Eseni, L. Selmi, K. K. Bourdelle, and S. Mantl, “Inverters with strained si nanowire complementary tunnel field-effect transistors,” *Electron Device Letters, IEEE*, vol. 34, no. 6, pp. 813–815, 2013.
- ⁹D. Mohata, R. Bijesh, S. Mujumdar, C. Eaton, R. Engel-Herbert, T. Mayer, V. Narayanan, J. Fastenau, D. Loubychev, A. Liu *et al.*, “Demonstration of mosfet-like on-current performance in arsenide/antimonide tunnel fets with staggered hetero-junctions for 300mv logic applications,” in *Electron Devices Meeting (IEDM), 2011 IEEE International*. IEEE, 2011, pp. 33–5.

- ¹⁰ F. Conzatti, M. Pala, D. Esseni, E. Bano, and L. Selmi, "Strain-induced performance improvements in inas nanowire tunnel fets," *Electron Devices, IEEE Transactions on*, vol. 59, no. 8, pp. 2085–2092, 2012.
- ¹¹ M. Pala and D. Esseni, "Interface traps in inas nanowire tunnel-fets and mosfets 2014;part i: Model description and single trap analysis in tunnel-fets," *Electron Devices, IEEE Transactions on*, vol. 60, no. 9, pp. 2795–2801, 2013.
- ¹² D. Esseni and M. G. Pala, "Interface traps in inas nanowire tunnel fets and mosfets—part ii: Comparative analysis and trap-induced variability," *Electron Devices, IEEE Transactions on*, vol. 60, no. 9, pp. 2802–2807, 2013.
- ¹³ R. M. Feenstra, D. Jena, and G. Gu, "Single-particle tunneling in doped graphene-insulator-graphene junctions," *Journal of Applied Physics*, vol. 111, no. 4, pp. 043711–043711, 2012.
- ¹⁴ L. Britnell, R. Gorbachev, A. Geim, L. Ponomarenko, A. Mishchenko, M. Greenaway, T. Fromhold, K. Novoselov, and L. Eaves, "Resonant tunnelling and negative differential conductance in graphene transistors," *Nature communications*, vol. 4, p. 1794, 2013.
- ¹⁵ L. Britnell, R. V. Gorbachev, R. Jalil, B. D. Belle, F. Schedin, M. I. Katsnelson, L. Eaves, S. V. Morozov, A. S. Mayorov, N. M. R. Peres, A. H. Castro Neto, J. Leist, A. K. Geim, L. A. Ponomarenko, and K. S. Novoselov, "Electron tunneling through ultrathin boron nitride crystalline barriers," *Nano Letters*, vol. 12, no. 3, pp. 1707–1710, 2012, PMID: 22380756. [Online]. Available: <http://dx.doi.org/10.1021/nl3002205>
- ¹⁶ B. Radisavljevic, A. Radenovic, J. Brivio, V. Giacometti, and A. Kis, "Single-layer mos2 transistors," *Nature nanotechnology*, vol. 6, no. 3, pp. 147–150, 2011.
- ¹⁷ Q. H. Wang, K. Kalantar-Zadeh, A. Kis, J. N. Coleman, and M. S. Strano, "Electronics and optoelectronics of two-dimensional transition metal dichalcogenides," *Nature nanotechnology*, vol. 7, no. 11, pp. 699–712, 2012.
- ¹⁸ C. Gong, H. Zhang, W. Wang, L. Colombo, R. M. Wallace, and K. Cho, "Band alignment of two-dimensional transition metal dichalcogenides: Application in tunnel field effect transistors," *Applied Physics Letters*, vol. 103, no. 5, pp. 053513–053513, 2013.
- ¹⁹ D. Jena, "Tunneling transistors based on graphene and 2-d crystals," *Proceedings of the IEEE*, vol. 101, no. 7, pp. 1585–1602, 2013.

- ²⁰ K. F. Mak, C. Lee, J. Hone, J. Shan, and T. F. Heinz, “Atomically thin mos₂: A new direct-gap semiconductor,” *Physical Review Letters*, vol. 105, no. 13, p. 136805, 2010.
- ²¹ D. Esseni, P. Palestri, and L. Selmi, *Nanoscale MOS transistors: Semi-classical transport and applications*. Cambridge University Press, 2011.
- ²² J. Bardeen, “Tunnelling from a many-particle point of view,” *Phys. Rev. Letters*, vol. 6, 1961.
- ²³ W. A. Harrison, “Tunneling from an independent-particle point of view,” *Physical Review*, vol. 123, no. 1, p. 85, 1961.
- ²⁴ C. B. Duke, *Tunneling in solids*. Academic Press New York, 1969, vol. 1999.
- ²⁵ P. Zhao, R. Feenstra, G. Gu, and D. Jena, “Symfet: A proposed symmetric graphene tunneling field-effect transistor,” *Electron Devices, IEEE Transactions on*, vol. 60, no. 3, pp. 951–957, 2013.
- ²⁶ Z. Zhu, Y. Cheng, and U. Schwingenschlögl, “Giant spin-orbit-induced spin splitting in two-dimensional transition-metal dichalcogenide semiconductors,” *Physical Review B*, vol. 84, no. 15, p. 153402, 2011.
- ²⁷ A. Ramasubramaniam, D. Naveh, and E. Towe, “Tunable band gaps in bilayer transition-metal dichalcogenides,” *Physical Review B*, vol. 84, no. 20, p. 205325, 2011.
- ²⁸ Q. Li, E. Hwang, E. Rossi, and S. D. Sarma, “Theory of 2d transport in graphene for correlated disorder,” *Physical review letters*, vol. 107, no. 15, p. 156601, 2011.
- ²⁹ J. Yan and M. S. Fuhrer, “Correlated charged impurity scattering in graphene,” *Physical Review Letters*, vol. 107, no. 20, p. 206601, 2011.
- ³⁰ M. Yankowitz, J. Xue, D. Cormode, J. D. Sanchez-Yamagishi, K. Watanabe, T. Taniguchi, P. Jarillo-Herrero, P. Jacquod, and B. J. LeRoy, “Emergence of superlattice dirac points in graphene on hexagonal boron nitride,” *Nature Physics*, vol. 8, no. 5, pp. 382–386, 2012.
- ³¹ J. Xue, J. Sanchez-Yamagishi, D. Bulmash, P. Jacquod, A. Deshpande, K. Watanabe, T. Taniguchi, P. Jarillo-Herrero, and B. J. LeRoy, “Scanning

tunnelling microscopy and spectroscopy of ultra-flat graphene on hexagonal boron nitride," *Nature materials*, vol. 10, no. 4, pp. 282–285, 2011.

³² R. Decker, Y. Wang, V. W. Brar, W. Regan, H.-Z. Tsai, Q. Wu, W. Gannett, A. Zettl, and M. F. Crommie, "Local electronic properties of graphene on a bn substrate via scanning tunneling microscopy," *Nano letters*, vol. 11, no. 6, pp. 2291–2295, 2011.

³³ P. Van Mieghem, G. Borghs, and R. Mertens, "Generalized semiclassical model for the density of states in heavily doped semiconductors," *Physical Review B*, vol. 44, no. 23, p. 12822, 1991.

³⁴ F. Urbach, "The long-wavelength edge of photographic sensitivity and of the electronic absorption of solids," *Physical Review*, vol. 92, pp. 1324–1324, 1953.

³⁵ G. Cody, "Urbach edge of crystalline and amorphous silicon: a personal review," *Journal of non-crystalline solids*, vol. 141, pp. 3–15, 1992.

³⁶ E. O. Kane, "Thomas-fermi approach to impure semiconductor band structure," *Physical Review*, vol. 131, no. 1, p. 79, 1963.

³⁷ S. D. Sarma and B. Vinter, "Thomas-fermi screening and level broadening in interacting two-dimensional electron-impurity systems," *Surface Science*, vol. 113, no. 1, pp. 176–181, 1982.

³⁸ A. Knabchen, "Self-consistent level broadening in thin films with volume and surface roughness scattering," *Journal of Physics: Condensed Matter*, vol. 7, no. 27, p. 5209, 1995.

³⁹ A. Ghazali and J. Serre, "Disorder, fluctuations and electron interactions in doped semiconductors: A multiple-scattering approach," *Solid-State Electronics*, vol. 28, no. 1, pp. 145–149, 1985.

⁴⁰ L. Britnell, R. Gorbachev, R. Jalil, B. Belle, F. Schedin, A. Mishchenko, T. Georgiou, M. Katsnelson, L. Eaves, S. Morozov *et al.*, "Field-effect tunneling transistor based on vertical graphene heterostructures," *Science*, vol. 335, no. 6071, pp. 947–950, 2012.

⁴¹ S. Tiefenbacher, C. Pettenkofer, and W. Jaegermann, "Moiré pattern in lead obtained by van der waals epitaxy of lattice mismatched ws₂/mote₂ heterointerfaces," *Surface science*, vol. 450, no. 3, pp. 181–190, 2000.

- ⁴² A. Koma, "Van der waals epitaxy for highly lattice-mismatched systems," *Journal of crystal growth*, vol. 201, pp. 236–241, 1999.
- ⁴³ G.-B. Liu, W.-Y. Shan, Y. Yao, W. Yao, and D. Xiao, "A three-band tight-binding model for monolayers of group-vib transition metal dichalcogenides," *arXiv preprint arXiv:1305.6089*, 2013.
- ⁴⁴ L. Britnell, R. V. Gorbachev, R. Jalil, B. D. Belle, F. Schedin, M. I. Katsnelson, L. Eaves, S. V. Morozov, A. S. Mayorov, N. M. Peres *et al.*, "Atomically thin boron nitride: a tunnelling barrier for graphene devices," *arXiv preprint arXiv:1202.0735*, 2012.
- ⁴⁵ S. Agarwal and E. Yablonovitch, "Pronounced effect of pn-junction dimensionality on tunnel switch sharpness," *arXiv preprint arXiv:1109.0096*, 2011.
- ⁴⁶ M. A. Khayer and R. K. Lake, "Effects of band-tails on the subthreshold characteristics of nanowire band-to-band tunneling transistors," *Journal of Applied Physics*, vol. 110, no. 7, p. 074508, 2011. [Online]. Available: <http://dx.doi.org/10.1063/1.3642954>
- ⁴⁷ J. I. Pankove, "Absorption edge of impure gallium arsenide," *Phys. Rev.*, vol. 140, pp. A2059–A2065, Dec 1965. [Online]. Available: <https://link.aps.org/doi/10.1103/PhysRev.140.A2059>
- ⁴⁸ A. V. Subashiev, O. Semyonov, Z. Chen, and S. Luryi, "Urbach tail studies by luminescence filtering in moderately doped bulk inp," *Applied Physics Letters*, vol. 97, no. 18, p. 181914, 2010. [Online]. Available: <http://dx.doi.org/10.1063/1.3510470>
- ⁴⁹ R. Schlaf, C. Pettenkofer, and W. Jaegermann, "Band lineup of a SnS₂/SnSe₂/SnS₂ semiconductor quantum well structure prepared by van der waals epitaxy," *Journal of applied physics*, vol. 85, no. 9, pp. 6550–6556, 1999.
- ⁵⁰ R. Schlaf, O. Lang, C. Pettenkofer, and W. Jaegermann, "Band lineup of layered semiconductor heterointerfaces prepared by van der waals epitaxy: Charge transfer correction term for the electron affinity rule," *Journal of applied physics*, vol. 85, no. 5, pp. 2732–2753, 1999.
- ⁵¹ L. Upadhyayula, J. Loferski, A. Wold, W. Giriat, and R. Kershaw, "Semiconducting properties of single crystals of n- and p-type tungsten diselenide (WSe₂)," *Journal of Applied Physics*, vol. 39, no. 10, pp. 4736–4740, 1968.

- ⁵² C. Sergio, Q. Gao, and R. M. Feenstra, "Theory of graphene–insulator–graphene tunnel junctions," *Journal of Vacuum Science & Technology B*, vol. 32, no. 4, p. 04E101, 2014.
- ⁵³ K. T. Lam, G. Seol, and J. Guo, "Operating principles of vertical transistors based on monolayer two-dimensional semiconductor heterojunctions," *Applied Physics Letters*, vol. 105, no. 1, p. 013112, 2014.
- ⁵⁴ L. De Michielis, L. Lattanzio, and A. M. Ionescu, "Understanding the super-linear onset of tunnel-fet output characteristic," 2012.
- ⁵⁵ B. Yu, L. Chang, S. Ahmed, H. Wang, S. Bell, C.-Y. Yang, C. Tabery, C. Ho, Q. Xiang, T.-J. King, J. Bokor, C. Hu, M.-R. Lin, and D. Kyser, "FinFET scaling to 10 nm gate length," in *Electron Devices Meeting, 2002. IEDM '02. International*, Dec 2002, pp. 251–254.
- ⁵⁶ A. M. van der Zande, J. Kunstmann, A. Chernikov, D. A. Chenet, Y. You, X. Zhang, P. Y. Huang, T. C. Berkelbach, L. Wang, F. Zhang *et al.*, "Tailoring the electronic structure in bilayer molybdenum disulfide via interlayer twist," *Nano letters*, vol. 14, no. 7, pp. 3869–3875, 2014.
- ⁵⁷ J. Brivio, D. T. L. Alexander, and A. Kis, "Ripples and layers in ultrathin mos2 membranes," *Nano Letters*, vol. 11, no. 12, pp. 5148–5153, 2011, PMID: 22010987. [Online]. Available: <http://dx.doi.org/10.1021/nl2022288>
- ⁵⁸ W. M. Haynes, *CRC handbook of chemistry and physics*. CRC press, 2012.
- ⁵⁹ D. Jena, K. Banerjee, and G. H. Xing, "2D crystal semiconductors: Intimate contacts," *Nat Mater*, vol. 13, no. 12, pp. 1076–1078, 12 2014. [Online]. Available: <http://dx.doi.org/10.1038/nmat4121>
- ⁶⁰ K. Bernstein, R. Cavin, W. Porod, A. Seabaugh, and J. Welser, "Device and architecture outlook for beyond cmos switches," *Proceedings of the IEEE*, vol. 98, no. 12, pp. 2169–2184, Dec 2010.
- ⁶¹ D. Nikonov and I. Young, "Uniform methodology for benchmarking beyond-cmos logic devices," in *Electron Devices Meeting (IEDM), 2012 IEEE International*, Dec 2012, pp. 25.4.1–25.4.4.
- ⁶² ——, "Overview of beyond-cmos devices and a uniform methodology for their benchmarking," *Proceedings of the IEEE*, vol. 101, no. 12, pp. 2498–2533, Dec 2013.

- ⁶³ “The Center for Low Energy Systems Technology (LEAST) proposal,” *led by the University of Notre Dame, submitted to SRC*, 2012.
- ⁶⁴ S. Xiao, M. Li, A. Seabaugh, D. Jena, and H. Xing, “Vertical heterojunction of mos₂ and wse₂,” in *Device Research Conference (DRC), 2014 72nd Annual*, June 2014, pp. 169–170.
- ⁶⁵ S. Vishwanath, X. Liu, S. Rouvimov, J. K. Furdyna, D. Jena, and H. G. Xing, “Molecular beam epitaxy of layered material superlattices and heterostructures,” *Bulletin of the American Physical Society*, 2014.
- ⁶⁶ L. Cheng, X. Qin, A. T. Lucero, A. Azcatl, J. Huang, R. M. Wallace, K. Cho, and J. Kim, “Atomic layer deposition of a high-k dielectric on MoS₂ using trimethylaluminum and ozone,” *ACS applied materials & interfaces*, vol. 6, no. 15, pp. 11 834–11 838, 2014.
- ⁶⁷ T. Roy, M. Tosun, J. S. Kang, A. B. Sachid, S. Desai, M. Hettick, C. C. Hu, and A. Javey, “Field-effect transistors built from all two-dimensional material components,” *ACS nano*, 2014.
- ⁶⁸ H. Fang, S. Chuang, T. C. Chang, K. Takei, T. Takahashi, and A. Javey, “High-performance single layered wse₂ p-fets with chemically doped contacts,” *Nano letters*, vol. 12, no. 7, pp. 3788–3792, 2012.
- ⁶⁹ L. Yang, K. Majumdar, Y. Du, H. Liu, H. Wu, M. Hatzistergos, P. Hung, R. Tieckelmann, W. Tsai, C. Hobbs *et al.*, “High-performance MoS₂ field-effect transistors enabled by chloride doping: Record low contact resistance (0.5 kΩ·μm) and record high drain current (460 μA/μm),” in *VLSI Technology (VLSI-Technology): Digest of Technical Papers, 2014 Symposium on*. IEEE, 2014, pp. 1–2.
- ⁷⁰ H. Xu, E. Kinder, S. Fathipour, A. Seabaugh, and S. Fullerton-Shirey, “Reconfigurable ion doping in 2h-mote₂ field-effect transistors using peo:csclo₄ electrolyte.” The 41st International Symposium on Compound Semiconductor, May 2014.
- ⁷¹ R. Kappera, D. Voiry, S. E. Yalcin, B. Branch, G. Gupta, A. D. Mohite, and M. Chhowalla, “Phase-engineered low-resistance contacts for ultrathin MoS₂ transistors,” *Nat Mater*, vol. 13, no. 12, pp. 1128–1134, 12 2014. [Online]. Available: <http://dx.doi.org/10.1038/nmat4080>
- ⁷² U. Bockelmann and G. Bastard, “Phonon scattering and energy relaxation in

two-, one-, and zero-dimensional electron gases," *Physical Review B*, vol. 42, no. 14, p. 8947, 1990.

⁷³ H. Fang, M. Tosun, G. Seol, T. C. Chang, K. Takei, J. Guo, and A. Javey, "Degenerate n-doping of few-layer transition metal dichalcogenides by potassium," *Nano letters*, vol. 13, no. 5, pp. 1991–1995, 2013.

CHAPTER 3

**COMPARATIVE STUDY OF INTRINSIC CAPACITANCES OF
THIN-TFETS AND PIN-TFET**

3.1 Enhanced Miller Effect of TFETs

Over years, most designs of TFETs fall into two categories in terms of device structures: lateral TFETs with tunneling direction perpendicular to gate electric field and vertical TFETs with tunneling direction and gate electric field aligned. For ultrascaled devices, pin-TFET based on layered 2D materials¹ is the latest breed of the lateral TFET structure, and its vertical counterpart is Two-Dimensional Heterojunction Interlayer Tunneling Field Effect Transistor (Thin-TFET).² It has been reported that pin-TFET has a much larger gate-drain capacitance (C_{GD}) compared to MOSFET, therefore exhibiting enhanced Miller effect.³ A large Miller effect leads to large voltage overshoot/undershoot and longer switching time in the transient response of a circuit. In this chapter, we show that a Thin-TFET offers smaller C_{GD} than a pin-TFET. Furthermore, this observation can be generalized that all vertical TFETs offers smaller C_{GD} than lateral TFET counterparts.

3.2 Effects of TFET Geometries: “lateral” TFETs vs. “vertical” TFETs

In a *n*-type pin-TFET structure (see Fig.3.1), a positive gate voltage induces the electron inversion in the channel. Because the source terminal charge is entirely

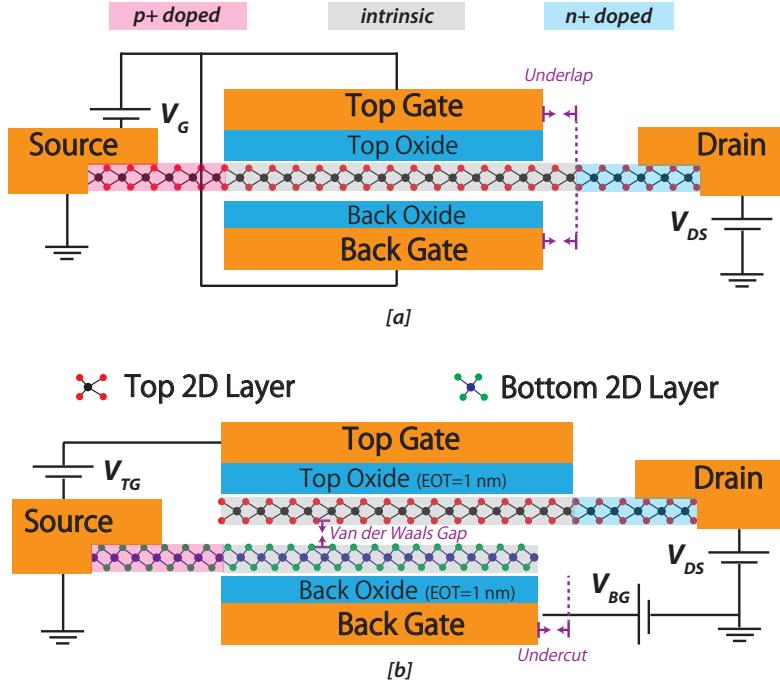


Figure 3.1: Schematic structure of (a) n-type pin-TFET and (a) n-type Thin-TFET.

composed of fixed charges in the depletion region of the tunneling junction, almost all the channel inversion charge is attributed to the drain terminal. This so-called 100/0 drain/source charge partition in a pin-TFET gives a much larger C_{GD} than in a MOSFET.³ On the contrary, in a *n*-type Thin-TFET structure, the heavily doped source is situated under the gate. In order to have efficient top gate control of the band alignment between the top layer (*i* region) and bottom layer (*p+* region) in a Thin-TFET, the top layer has to have a low carrier concentration to avoid screening out the electric field from the top gate, while the bottom layer has to have a very high carrier concentration to terminate the electric field. Because most of the electric field from top gate terminates at the source beneath the gate, the change of gate terminal charge is mainly reflected in the source instead of the drain. Therefore, a Thin-TFET intrinsically has smaller C_{GD} than a pin-TFET. Moreover, this trend is also applicable for any TFETs with vertical structures, such as the recently proposed U-shaped TFET.⁴

To better understand the effect of TFET geometries, we presented the abstract comparison between pin-TFET and Thin-TFET (see Fig.3.2). The device schematics with highlighted tunneling areas are shown in the first column. The Thin-TFET has notably larger tunneling areas than the pin-TFET, which is one of the original motivation of vertical TFETs. Larger tunneling areas can contribute to larger ON-current density.⁵ The different device geometries lead to distinguish capacitance networks shown in the second column of Fig.3.2. These 1D capacitance networks neglect the effect of parasitic capacitances and lateral junction capacitances. Because the tunneling junction between the source and channel has much smaller conductance than the forward bias junction between the drain and channel, the channel inversion charge is mostly attributed to the drain. Therefore, in the pin-TFET capacitance network model, the source and channel are approximately disconnected. Since we neglected the lateral junction capacitance between the channel and the drain, the channel capacitance is the quantum capacitance of the channel material. The Thin-TFET capacitance model has been introduced in Li et al.² It is worth to note that in the pin-TFET, both the top gate and bottom gate are biased at the same voltage. Whereas in the Thin-TFET, the bottom gate is grounded and only the top gate is biased. Because of this, the pin-TFET has better “gate control” than the Thin-TFET. To quantify the “gate control”, we define gate efficiency (GE) for TFETs as the change of energy overlap (ΔE_{OL}) over the change of gate voltage (ΔV_G) times unit charge. Gate efficiency tells how much the tunneling window changes by changing certain amount of gate voltage. The gate efficiencies of pin-TFET and Thin-TFET are shown in the fourth column of Fig.3.2. From the equations, the gate efficiency of pin-TFET is roughly twice of Thin-TFET, given the following assumptions: 1) the bottom 2D layer in Thin-TFET is heavily doped, namely $C_{Q,B}$ is

much larger than $C_{Interlayer}$; 2) $C_{Interlayer}$ is almost the same as C_{TG} in Thin-TFET; 3) C_{TG} in Thin-TFET is the same as C_G in pin-TFET. It is advantageous for pin-TFET in term of gate efficiency since its lateral geometries allows the top and bottom gates control the channel potential together. While for Thin-TFET, potential difference is required between the top and bottom gates, therefore the top layer 2D is essential controlled by the only one gate. The expressions for gate-drain capacitance (C_{GD}) of pin-TFET and Thin-TFET are listed in third column of Fig.3.2. If we use the same assumptions above, C_{GD} of Thin-TFET is roughly half of C_{GD} of pin-TFET.

In a Thin-TFET, an undercut (labeled in Fig.3.1) is necessary to achieve a small SS.⁶ Because of the absence of $p+$ source in the undercut region, the top gate in the undercut region is strongly coupled with the drain. Therefore, the undercut region in Thin-TFET will increase C_{GD} . On the other hand, an underlap (labeled in Fig.3.1) in pin-TFET is used to address the ambipolar problem in TFETs.⁷ For a n-type TFET, when applying more negative gate voltage, the valance band in the channel will move above the conduction band in the drain, therefore forming a tunneling junction between the channel and the drain. It is the so-called ambipolar effect of TFETs, which is undesired under the regular circuit design (we will discuss how to use this ambipolar effect to create more interesting device behavior in Chapter 4). By leaving a certain channel region close to the drain (underlap) not controlled by the gate, the possible tunneling current through the channel and drain junction is suppress, and the ambipolar effect is mitigated. In term of C_{GD} , the underlap in pin-TFET can be viewed as "decoupling" between the channel and the drain. Therefore the underlap in pin-TFET helps to decrease C_{GD} . We will show how undercut and underlap affects the capacitances in Thin-TFET and pin-TFET respectively in the later sections.

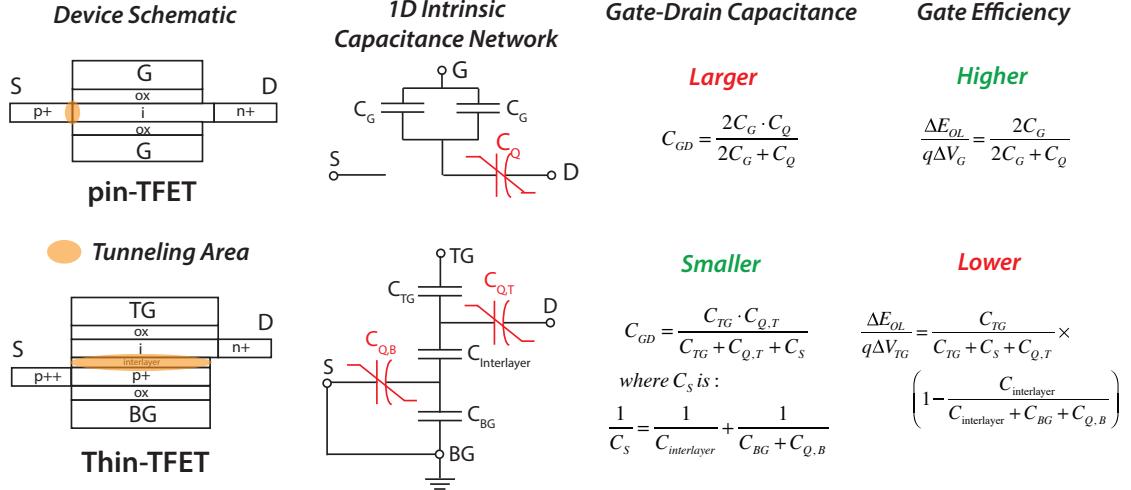


Figure 3.2: Comparison between pin-TFET and Thin-TFET: (Column 1) Device schematics with the tunneling area highlighted. Thin-TFET has a larger tunneling area, which can potentially render a higher tunnel current; (Column 2) 1D intrinsic capacitance networks, where C_G is the gate capacitance of the pin-TFET, C_Q is the quantum capacitance of the channel material in the pin-TFET, $C_{T(B)G}$ is the top (bottom) gate capacitance of the Thin-TFET, $C_{Q,T(B)}$ is the quantum capacitance of the top (bottom) material, and $C_{Interlayer}$ is the interlayer capacitance between the top and bottom materials in the Thin-TFET; (Column 3 & 4) analytical expressions for C_{GD} and gate efficiency.

3.3 Numerical Simulations of C-V Curves

3.3.1 Simulation Methods

The numerical simulations of C-V curves of pin-TFET and Thin-TFET are based on solving 2D Poisson equation using *LENO*.⁸ In TFETs, the tunnel junction has the highest resistance in the normal operation conditions. Therefore, it is reasonable to assume all the potential difference between the source and drain voltage drops across the tunnel junction, and the Quasi-Fermi levels are flat elsewhere. The capacitances are computed via the following steps:

1. By solving 2D Poisson equation, obtain the electric fields in the gate oxides

at different biases;

2. Using the electric fields in the gate oxides, obtain the area charge densities in the gate terminals (Q_G) at different biases;
3. Calculate the gate capacitance (C_{GG}), gate-to-source capacitance (C_{GS}), and gate-to-drain capacitance (C_{GD}) using Equation 3.1.

$$\begin{aligned} C_{GG} &= \frac{\Delta Q_G}{\Delta V_G} \\ C_{GD} &= \frac{\Delta Q_G}{\Delta V_D} \\ C_{GS} &= \frac{\Delta Q_G}{\Delta V_S} \end{aligned} \quad (3.1)$$

It is worth noting that for pin-TFET, Q_G is the total charge on both the top and bottom gate terminals, while for Thin-TFET, Q_G is just the charge on the top gate terminal and the corresponding V_G is the top gate voltage since Thin-TFET is gated by the top gate only.

This capacitance model is known as the quasi-static quasi-equilibrium model. When deriving the capacitances only from the Poisson equations, we made two approximations:

1. Quasi-static approximation, also known as the low frequency approximation, where the finite charging time for the inversion layer is ignored.
2. Quasi-equilibrium approximation, also known as the low current approximation, where the change of charges in the channel due to the injected tunnel current from the source is ignored.

The material and device parameters of the pin-TFET and the Thin-TFET used in the simulations are listed in Table 3.3.1.

Material system for <i>n</i> -type pin-TFETs								
	Bandgap (eV)	Electron affinity (χ) (eV)	m_c^* (m_0)	m_v^* (m_0)				
WTe ₂	0.75	4.05	0.37	0.3				
Φ_M (eV)	4.13							
Lead region doping level: $N_{D(A)} = 2 \times 10^{12} \text{ cm}^{-2}$								
Material system for <i>n</i> -type Thin-TFETs								
	Top 2D layer		Bottom 2D layer					
Materials	SnSe ₂		WSe ₂					
	Bandgap (eV)	Electron affinity (χ) (eV)	m_c^* (m_0)	m_v^* (m_0)				
WSe ₂	1.3	4.0	0.3	0.4				
SnSe ₂	0.9	5.1	0.3	0.4				
	$\Phi_{M,T}$ (eV)	$\Phi_{M,B}$ (eV)	$N_{D,B}(\text{cm}^{-2})$	$N_{A,B}(\text{cm}^{-2})$				
<i>n</i> -Thin-TFET	5.20	5.95	0	7×10^{13}				
Top layer lead region doping level: $N_{D(A)} = 2 \times 10^{12} \text{ cm}^{-2}$								
Valley degeneracy for WTe ₂ , WSe ₂ , SnSe ₂ : 2								
Simulated Device Parameters for Thin-TFETs and pin-TFETs								
Gate length (nm) (nm)	S/D length (nm)	interlayer thickness (Thin-TFET) (nm)		Gate EOT (nm)				
10	5	0.35		1				

Table 3.1: The material and device parameters of pin-TFETs and Thin-TFETs in the simulation.

3.3.2 Simulation Results with Different Undercut/Underlap Lengths

The simulation results of bias dependent C_{GG} , C_{GD} and C_{GS} of pin-TFETs and Thin-TFETs with different undercut/underlap lengths are shown in Fig.3.3. In Fig.3.3(a), the black curves are C_{GD} versus the gate voltage V_G (for Thin-TFET,

$V_G=V_{TG}$) at different V_{DS} for pin-TFETs, and the red curves are for Thin-TFETs. For both pin-TFETs and Thin-TFETs, C_{GD} increases with decreasing V_G . This trend can be understood using the simple capacitance network in Fig.3.2. When applying more positive gate voltage, the quantum capacitance (C_Q for pin-TFET, $C_{Q,T}$ for Thin-TFET) increases due to the increasing free carrier concentration in the channel, which increases C_{GD} . Moreover, the C_{GD} decreases with increasing V_{DS} . Similar with the explanation above, more positive V_{DS} means the potential difference between the channel potential to the drain potential will decrease, which leads to lower free carrier concentration in the channel and sequentially smaller quantum capacitance (C_Q for pin-TFET, $C_{Q,T}$ for Thin-TFET).

The C_{GD} values of Thin-TFETs are roughly half of the ones of pin-TFETs as discussed in Section 3.2. A smaller C_{GD} means less severe Miller effect. With longer underlap region, C_{GD} of pin-TFETs decrease; while with longer undercut region, C_{GD} of Thin-TFETs slightly increase. Figure.3.3(d) shows the trend of C_{GD} . As for C_{GS} shown in Fig.3.3(b), because in Thin-TFET, the gate is more “coupled” with the source, the C_{GS} of Thin-TFETs are significant larger than the C_{GS} of pin-TFETs. Due to the so called 100/0 drain/source charge partition in pin-TFET, the C_{GS} of pin-TFETs are almost zero. Precisely the same reason gives Thin-TFETs a relatively smaller C_{GD} and larger C_{GS} compared to pin-TFETs: the gate has stronger “coupling” with the source.

During the switching of the devices in a inverter, C_{GG} of the devices are the input capacitance, which affects the energy dissipation. A smaller C_{GG} means less charges need to be moved in order to switch a inverter. Therefore, we also compared C_{GG} of pin-TFETs and Thin-TFETs. Shown in Fig.3.3(c), Thin-TFETs have smaller C_{GG} compared to pin-TFET, which indicates Thin-TFETs are more

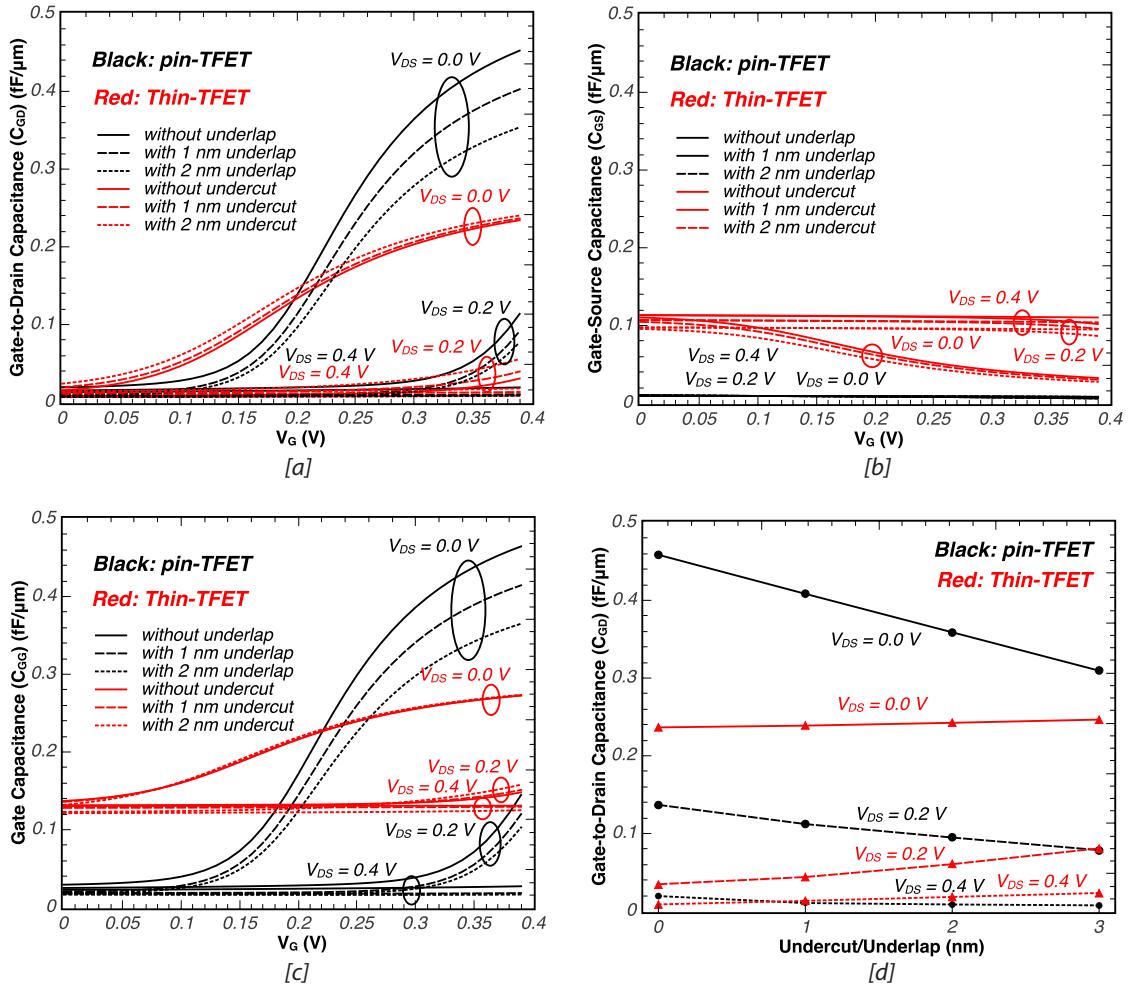


Figure 3.3: (a) C_{GD} versus V_G at different V_{DS} for both pin-TFETs (black lines) and Thin-TFETs with different underlap and undercut lengths respectively; (b) C_{GS} versus V_G at different V_{DS} or both pin-TFETs (black lines) and Thin-TFETs with different underlap and undercut lengths respectively; (c) C_{GG} versus V_G at different V_{DS} or both pin-TFETs (black lines) and Thin-TFETs with different underlap and undercut lengths respectively; (d) C_{GD} versus underlap/undercut length at different V_{DS} and $V_G = 0.4V$.

energy efficient than pin-TFETs. We will discuss this in details in the next section.

Besides the capacitances, we also compared the gate efficiencies of pin-TFETs and Thin-TFETs, and their non-linear onset effect in the output characteristics. As discussed in Section 3.2, the gate-efficiency (GE) of pin-TFETs is roughly

twice higher than Thin-TFETs. In Fig.3.4(a), the black line is the GE of the pin-TFET and the red line is the GE of the Thin-TFET. The GE of the Thin-TFET is measured at the center of the channel, therefore it is independent of the undercut effect. The solid lines in Fig.3.4(a) are the GE right at the threshold voltage, namely when the valence band edge of the source is aligned with the conduction edge of the channel. The dash lines are the average GE when changing V_G from 0 to 0.4 V. The simulation results of GE are in agreement with the analysis in Section 3.2. Another common phenomenon of TFETs is the so-called non-linear or super-linear onset of tunnel-FET output characteristic.⁹ The non-linear onset of tunnel-FET output characteristic happens when V_{DS} is small, I_D is exponentially dependent on V_{DS} instead of linearly dependent on V_{DS} . On the other hand, the non-linear onset can be viewed as higher threshold voltages at smaller V_{DS} . In Fig.3.4(b), we show the threshold voltage versus the drain voltage (V_{DS}) for both pin-TFET and Thin-TFET. The threshold voltages staying almost constant at higher V_{DS} for both pin-TFET and Thin-TFET. At lower V_{DS} , pin-TFET's threshold voltage doesn't increase as much as Thin-TFET's, which indicated that pin-TFETs has less non-linear onset in the output characteristics when compared with Thin-TFETs.

Since the undercut in Thin-TFETs increases C_{GD} , we investigated what is the minimal necessary undercut length. The ideal case is that the whole channel in Thin-TFETs has the same threshold voltage. However, because the influence of the drain, the threshold voltage is lower at the drain-side edge (see Fig.3.5(a)). The uniformity of threshold voltages in Thin-TFETs lead to detriment sub-threshold slope, which has been discussed in Section 2.4.1. In Fig.3.5(a), the red line is the threshold voltage at the center of the channel (indicated in Fig.3.5(a)). The dash lines are the threshold voltage at the drain-side edge (indicated in

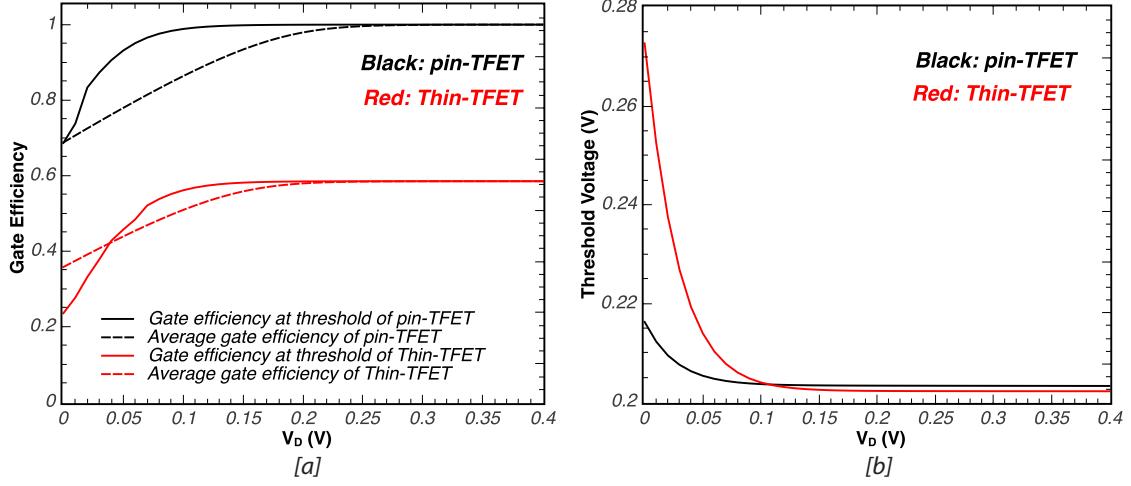


Figure 3.4: (a) Gate efficiencies versus the drain voltages V_{DS} for both pin-TFETs and Thin-TFETs, the solid lines are the gate efficiency at the threshold while the dash lines are the average gate efficiency when swiping V_G from 0 to 0.4 V; (b) the threshold voltages versus the drain voltages V_{DS} for both pin-TFETs and Thin-TFETs, the increasing threshold voltages at smaller V_{DS} lead to the non-linear onset in the output characteristics.

Fig.3.5(a)). When there is no undercut, the threshold voltages at the drain-side edge are significantly smaller than the one in the center of the center, which would deteriorate the sub-threshold slop. To balance between the sub-threshold slop and the C_{GD} , we use undercut equals 1 nm in the following simulation.

3.4 Complimentary TFET Inverters

In order to evaluate the impact of C_{GD} in Thin-TFETs and pin-TFETs based circuits, material parameters have been chosen to render symmetric behaviors in both the *p*-type and *n*-type devices. The Thin-TFET used here has 1 nm undercut and the pin-TFET has 1 nm underlap. The complementary TFET (CTFET) inverter is shown in Fig.3.6. We can write its charge conservation equation (see Equation 3.2):

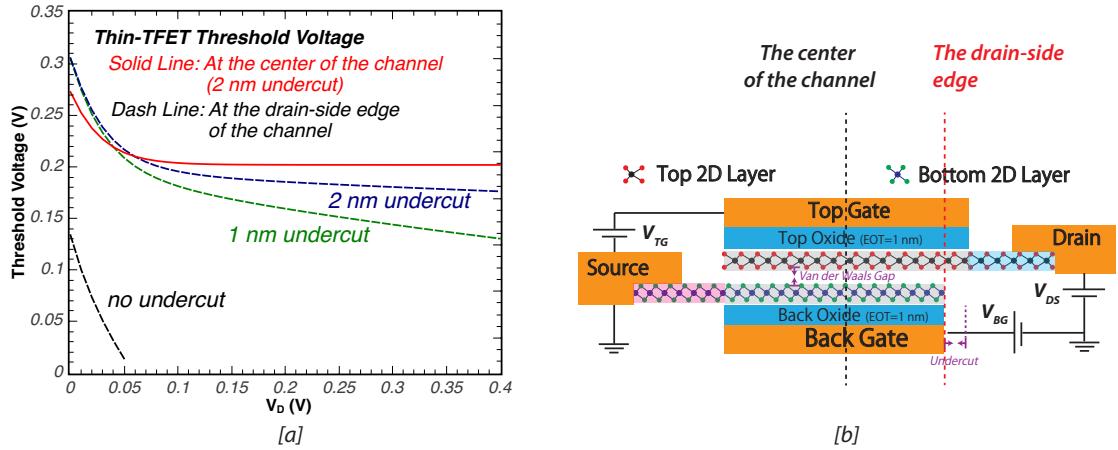


Figure 3.5: (a) The threshold voltages versus the drain voltages for Thin-TFETs with different undercut lengths. The red solid line is the threshold voltages computed at the center of the channel (shown in (b)), the dash lines are the threshold voltages computed at the drain-side edge (shown in (b)). The differences between the dash lines and the red solid line indicate the non-uniformity of the threshold voltages along the channel of Thin-TFETs.

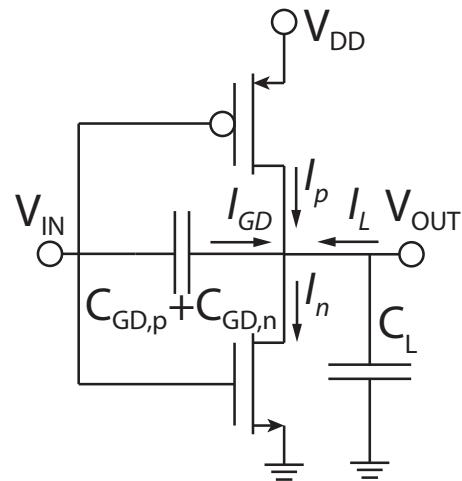


Figure 3.6: The schematic layout of the complementary TFET (CTFET) inverter. $C_{GD,n}$ and $C_{GD,p}$ are the gate-to-drain capacitance of the n-type and p-type TFETs. C_L is the load capacitance.

$$C_L \frac{dV_{OUT}}{dt} = (C_{GD,n} + C_{GD,p}) \frac{d(V_{IN} - V_{OUT})}{dt} + I_p - I_n \quad (3.2)$$

For a given sequence of $V_{IN}(t)$, we can define the right hand side as $f(t, V_{OUT}(t), V_{IN}(t))$. Then we can solve for $V_{OUT}(t)$ using the backward Euler Method:

Result: The output voltage V_{OUT} at each time step t

for each time step t_k **do**

Initialization: $i = 1$ and $V_{OUT}^i(t_k) = V_{OUT}(t_{k-1})$;

while δ doesn't meet convergence condition **do**

$V_{OUT}^{i+1}(t_k) = V_{OUT}(t_{k-1}) + \Delta t \times f(t_k, V_{OUT}^i(t_k), V_{IN}(t_k))$;

$\delta = V_{OUT}^{i+1}(t_k) - (V_{OUT}(t_{k-1}) + \Delta t \times f(t_k, V_{OUT}^{i+1}(t_k), V_{IN}(t_k)))$;

end

$V_{OUT}(t_k) = V_{OUT}^{i+1}(t_k)$;

end

Algorithm 1: The backward Euler method used to compute the transient response of CTFET inverters

Ignoring the static energy dissipation of CTFET inverters due to the leakage current, the dynamic energy dissipation can be written as:

$$E_{Dynamic} = V_{DD} \int_{V_{OUT}=0 \rightarrow V_{DD}} I_n dt + V_{DD} \int_{V_{OUT}=V_{DD} \rightarrow 0} I_p dt + E_{shortcircuit} \quad (3.3)$$

where $E_{shortcircuit}$ is the short circuit current if the n -type and p -type TFETs are ON simultaneously. Since we designed the threshold voltage $V_{th,n}$ of the n -type TFET and $V_{th,p}$ of the p -type TFET such that $|V_{th,n}| + |V_{th,p}| > V_{DD}$, the short circuit

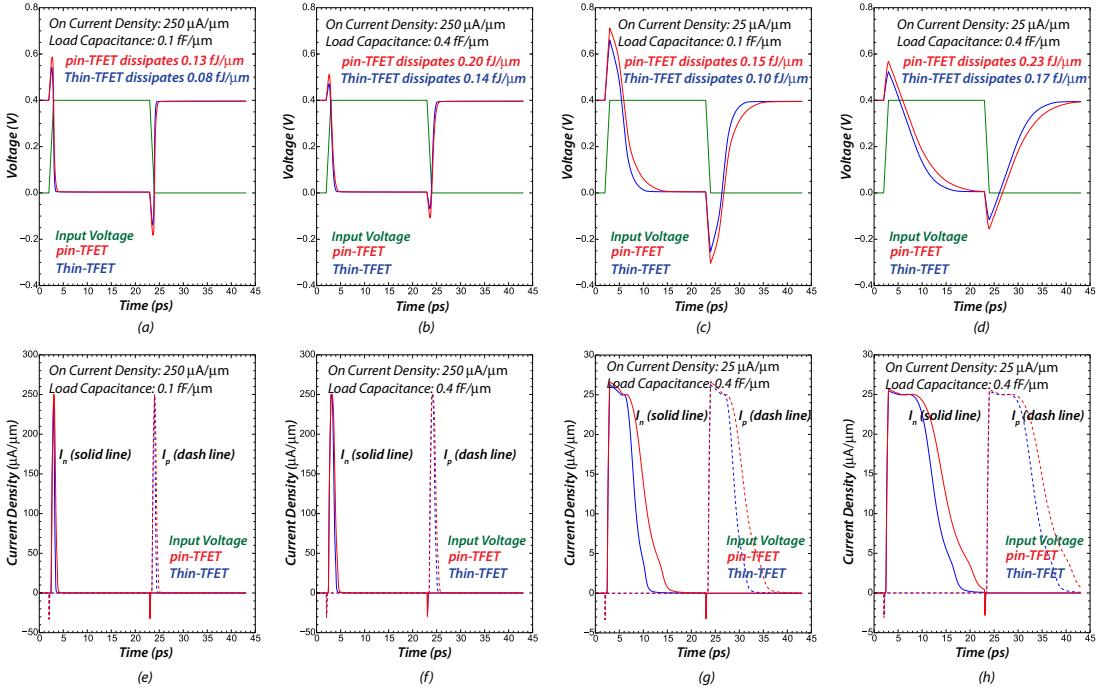


Figure 3.7: (a)-(d) the input and output voltages of pin-TFETs and Thin-TFETs based CFET inverter versus time with different ON current density and load capacitance; (e)-(h) the current density of pin-TFETs and Thin-TFETs in the inverters versus time with different On current density and load capacitance.

current is eliminated. Therefore we can ignore $E_{shortcircuit}$ when computing the energy dissipation.

The transient response of the CTFET inverters is shown in Fig.3.7. The green lines in Fig.3.7(a)-(d) are the input voltages. In order to clearly illustrate the output voltage behaviors, we use the square wave as the input voltages. In practice, the input voltage will be the output voltage of the previous stage. The output voltage of pin-TFETs and Thin-TFETs are shown in red line. First, both of them have the significant overshoot/undershoot. The overshoot/undershoot happens when there is a substantial large capacitance directly connecting the input and output, namely C_{GD} . When the input voltage ramping up from low to high, the capacitance between the input and output will attempt to keep the poten-

tial difference between the input and the output, which leads to a displacement current flow from the input to the output via C_{GD} and start to charge C_L . Consequently, the output voltage has been pushed over V_{DD} before the ON current of the *n*-type TFET pulls down the output voltage. The undershoot voltage can be explained in a similar way. The overshoot/undershoot also lead to propagation delays, which measures the time delays between the signal edge (i.e. $V_{DD}/2$) of the input voltages and of the output voltages. In the Fig.3.7(a)-(d), we show that Thin-TFET based CTFET inverters has smaller overshoot/undershoot voltage when comparing to pin-TFET based CTFET inverters since Thin-TFET has smaller C_{GD} . Besides C_{GD} , a smaller C_L and higher ON current density can also help to reduce the overshoot/undershoot voltage. Since the energy dissipation is determined by the integral of current over time, the areas under the curve in Fig.3.7(e)-(h) are proportional to the energy dissipation. Comparing Thin-TFETs and pin-TFETs, Thin-TFET based CTFET inverter can save around 30% energy dissipation and has shorter propagation delay due to smaller C_{GD} .

3.5 Conclusion

Due to its vertical stacking structure, a Thin-TFET intrinsically has much smaller C_{GD} than a pin-TFET. With mitigated Miller effect, Thin-TFET inverters can save around 30% power dissipation then pin-TFET inverters. This finding will help to guide future designs of TFET structures.

BIBLIOGRAPHY

- ¹ H. Ilatikhameneh, Y. Tan, B. Novakovic, G. Klimeck, R. Rahman, and J. Appenzeller, "Tunnel field-effect transistors in 2-d transition metal dichalcogenide materials," *IEEE Journal on Exploratory Solid-State Computational Devices and Circuits*, vol. 1, pp. 12–18, Dec 2015.
- ² M. O. Li, D. Esseni, J. J. Nahas, D. Jena, and H. G. Xing, "Two-dimensional heterojunction interlayer tunneling field effect transistors (thin-tfets)," *IEEE Journal of the Electron Devices Society*, vol. 3, no. 3, pp. 200–207, 2015.
- ³ S. Mookerjea, R. Krishnan, S. Datta, and V. Narayanan, "On enhanced miller capacitance effect in interband tunnel transistors," *IEEE Electron Device Letters*, vol. 30, no. 10, pp. 1102–1104, 2009.
- ⁴ W. Li, H. Liu, S. Wang, and S. Chen, "Reduced miller capacitance in u-shaped channel tunneling fet by introducing heterogeneous gate dielectric," *IEEE Electron Device Letters*, vol. 38, no. 3, pp. 403–406, 2017.
- ⁵ G. Zhou, R. Li, T. Vasen, M. Qi, S. Chae, Y. Lu, Q. Zhang, H. Zhu, J.-M. Kuo, T. Kosel *et al.*, "Novel gate-recessed vertical inas/gasb tfets with record high i on of 180 μ A/ μ m at v ds= 0.5 v," in *Electron Devices Meeting (IEDM), 2012 IEEE International*. IEEE, 2012, pp. 32–6.
- ⁶ Y. Lu, G. Zhou, R. Li, Q. Liu, Q. Zhang, T. Vasen, S. Doo Chae, T. Kosel, M. Wistey, H. Xing, A. Seabaugh, and P. Fay, "Performance of AlGaSb/InAs TFETs with gate electric field and tunneling direction aligned," *Electron Device Letters, IEEE*, vol. 33, no. 5, pp. 655–657, May 2012.
- ⁷ T. Krishnamohan, D. Kim, S. Raghunathan, and K. Saraswat, "Double-gate strained-ge heterostructure tunneling fet (tfet) with record high drive currents and 60mv/dec subthreshold slope," in *2008 IEEE International Electron Devices Meeting*, Dec 2008, pp. 1–3.
- ⁸ M. Li, "Leno device simulator," https://github.com/Oscarlight/leno/tree/master/Leno_beta1.5, 2015.
- ⁹ L. De Michielis, L. Lattanzio, and A. M. Ionescu, "Understanding the superlinear onset of tunnel-fet output characteristic," 2012.

CHAPTER 4

**XNOR-ENABLED TRANSISTOR (TRANSIXNOR) FOR BINARIZED
NEURAL NETWORK ACCELERATOR**

4.1 Introduction

In recent years, deep neural networks (DNNs) have become an important type of machine learning algorithms, and achieved substantial improvements in a wide range of tasks including object recognition in images,^{1,2} speech recognition,³ machine translation,⁴ image generation⁵ and game plays.^{6,7}

However, the state-of-art DNNs have a lot of parameters and expensive computational cost. Especially for mobile and embedded systems, the size of the model and the energy consumption during inference are crucial. Numerous researches have been conducted to provide efficient hardware designs for DNNs.⁸ Since the data intensive nature of deep learning, data movement becomes the speed bottleneck and dominate the energy consumption. The concept known as *processing-in-memory* aims at bringing the memory closer to the computation. Among various non-volatile memories (NVM) based architectures, the resistive RAM (RRAM) crossbar array allows computing the analog matrix-vector multiplication in the constant time $O(1)$, therefore provides massive acceleration of forward and backward pass of DNNs with reduced power consumption and increased integration density.⁹ In order to accelerate the weight update in RRAM crossbars, Gokmen et al.¹⁰ proposed to significantly simplify the multiplication operation itself by using stochastic computing technique, thus achieving the $O(1)$ time complexity for the weight update cycle of the training algorithm. However, processing fixed point numbers with RRAMs has several drawbacks

including resistance variations, stuck-at faults and AD/DA overheads.¹¹ Chen et al.¹¹ developed a weight-memristor mapping algorithm based on bipartite matching and the self-healing capability of neural networks to improve the precision.

On the other hand, the energy and area costs of computation are reduced rapidly by decreasing the number of bits used to represent the weight and the activation.¹² The recently introduced binarized neural network (BNNs) with binary weights and activations¹³⁻¹⁷ turns the most computationally expensive convolutions into bitwise operations, as well as dramatically shrinks the model size. Many efforts have been made to design specific hardware to accelerate BNNs. Zhong et al.¹⁸ and Umuroglu et al.¹⁹ contribute to building fast and flexible FPGA accelerators for BNNs. Tutu et al.²⁰ built the specialization tier for BNNs in the *Celerity* chip. The performances of BNNs in different hardware platforms such as FPGA, CPU, GPU, and ASIC were compared by Eriko et al.²¹

The RRAM crossbar architectures mentioned above can further support BNNs.²²⁻²⁴ From one standpoint, the *processing-in-memory* capability of RRAM crossbars can achieve faster and more energy efficient implementation of BNNs along with smaller chip areas; from the other standpoint, BNNs use the single-bit RRAM devices, which can tolerate more variation and be more reliable than the multi-bit RRAM devices. Moreover, compared to the multi-bit RRAM crossbar, the single-bit RRAM crossbar is more energy efficient during computation and required no AD/DA overhead.

At the heart of deep neural networks is general matrix-matrix multiplication (GEMM) or general matrix-vector multiplication (GEMV). Both the forward and backward pass of the fully-connected layer and the convolution layer can be

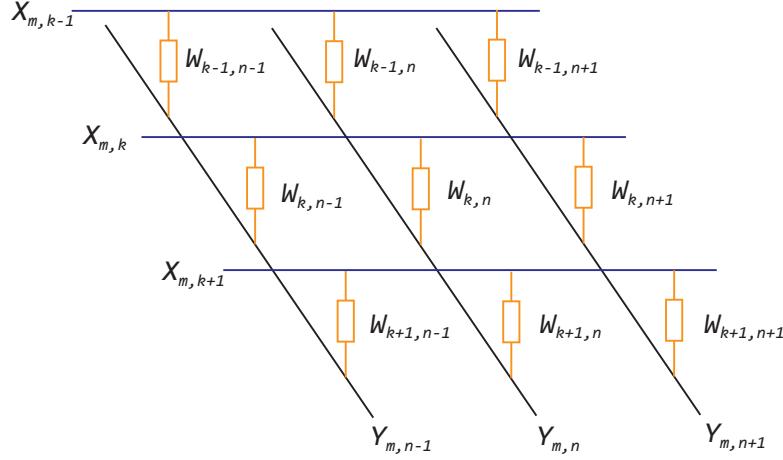


Figure 4.1: The RRAM crossbar architecture. X is the input voltages signals, W is the weight matrix whose elements are the RRAM conductivities, and Y is the output current signals. The relationship of X , W , Y is shown in Eq.4.1

built on GEMM or GEMV. In a RRAM crossbar architecture,²³ parallel GEMV is performed by using the conductivities of each RRAM devices as the weight matrix W , the input voltage signals as the input vector X and the output current signal as the output vector Y (shown in Fig.4.1).

The relationship between the input vectors X and output vectors Y can be expressed as in Eq.4.1:

$$Y_{m,n} = \sum_k X_{m,k} \cdot W_{k,n} \quad (4.1)$$

When the input vector X and weight matrix W are binary (i.e. $X_{m,k}, W_{k,n} \in \{0, 1\}$), the multiplication in Equation 4.1 is replaced by XNOR (noted as \otimes). However, XNOR can not be implemented directly with RRAM. Therefore XNOR is implemented as Eq.4.2:

$$Y_{m,n} = \sum_k X_{m,k} \otimes W_{k,n} = \sum_k X_{m,k} \cdot W_{k,n} + \overline{X_{m,k}} \cdot \overline{W_{k,n}} \quad (4.2)$$

Curious readers may wonder why binary multiplication is equivalent to XNOR. In the BNNs,^{13–15} the binarization is done by constraining the variable to +1 and -1 (instead of 1 and 0). If we define +1 as *true* and -1 as *false*, it was obvious that the binary multiplication with +1 and -1 is equivalent to XNOR as shown in Eq.4.3.

$$\begin{aligned} 1 \cdot 1 &= 1 \otimes 1 = 1 \\ -1 \cdot 1 &= -1 \otimes 1 = -1 \\ 1 \cdot -1 &= 1 \otimes -1 = -1 \\ -1 \cdot -1 &= -1 \otimes -1 = 1 \end{aligned} \quad (4.3)$$

In the circuit, we normally use 1 and 0 instead of 1 and -1. Fortunately, these two representations are interchangeable through Eq.4.4.

$$2 \sum_i^N a_i \otimes b_i - N = \sum_i^N c_i \otimes d_i \quad (4.4)$$

$a_i, b_i \in \{0, 1\}, c_i, d_i \in \{-1, 1\}$

We will keep the convention in circuit design by using 1 and 0, but follow the binary multiplication rule of -1 and 1, which is the XNOR operation.

Since the RRAM crossbar architecture doesn't natively support XNOR, we need in total 2 multiplications, 1 addition, and 2 bit complements. In RRAM

crossbar, the multiplications and additions can be done in parallel, therefore there is no extra time consumption. From Eq.4.2, it may seem like implementing XNOR required twice as many as RRAMs in the crossbar comparing with its multi-bit counterpart. Note that, in order to represent negative weights in the multi-bit RRAM, each weight is implemented by a pair of RRAM devices.²⁵ Therefore there is no extra area penalty. Taking the bit complement, however, may become a potential overhead.

What if we can use a single device to compute XNOR? It will lead to around 50% area saving and potentially faster and more energy efficient implementation. Due to the channel to drain tunneling, TFETs are known to have ambipolar behavior, which is considered undesirable in logic circuits.²⁶⁻³¹ However, the ambipolar behavior can enable interesting logic operations, such as exclusive or (XOR) and its complement XNOR. In this letter, we utilize the ambipolar behavior in TFETs to propose a novel device, TransiXNOR, a dual-gate XNOR-enable transistor based on Zener tunneling. Eventually, we propose to integrate a non-volatile memory, for instance RRAM, and use it as the building block to create a new crossbar architecture to compute binary GEMV by utilizing the unique NXOR functionality of TransiNORs.

4.2 Dual-gated XNOR-enable Transistor: TransiXNOR

4.2.1 Device Working Principle

The schematic structure of TransiXNOR is shown in Fig.4.2(a). The structure resembles a double-gated tunnel field effect transistor (DG-TFET). But there are

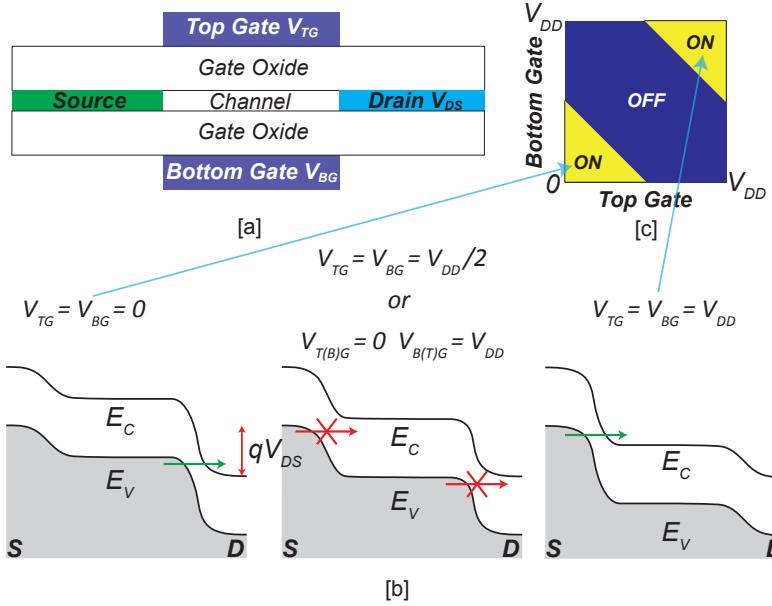


Figure 4.2: (a) The schematic structure of TransiXNOR; (b) the band diagrams at different gate bias conditions of TransiXNOR when V_{DS} equals V_{DD} : (left) the channel/drain tunnel junction is ON when both V_{TG} and V_{BG} are 0; (right) the source/channel tunnel junction is ON when both V_{TG} and V_{BG} are V_{DD} ; (middle) both the channel/drain tunnel junction and source/channel tunnel junction are OFF at the bias conditions such as both V_{TG} and V_{BG} are $V_{DD}/2$, or one gate is V_{DD} and the other is 0; (c) The schematic mapping of transiXNOR ON/OFF states at different V_{TG} and V_{BG} when V_{DS} is V_{DD} , which resembles XNOR logic.

three major differences:

1. The top gate and bottom gate are biased independently;
2. The channel has to be thin enough such that the top gate and bottom gate control the same conducting channel;
3. The tunneling current plane is alternated between the source/channel junction and channel/drain junction.

The working principle of TransiXNOR is shown in Fig.4.2(b): when both V_{TG} and V_{BG} are zero biases and V_{DS} biased at V_{DD} , the channel is electrostatically p -doped such that the valence band edge of the channel is above the conduction

band edge of the drain. Therefore, the channel/drain tunnel junction is ON. On the other hand, when both V_{TG} and V_{BG} are biased to V_{DD} and V_{DS} biased at V_{DD} , the channel is gated to be n -type such that the conduction band edge of the channel is below the valence band edge of the source. Therefore, the source/channel tunnel junction is ON. However, when both V_{TG} and V_{BG} are biased at $V_{DD}/2$, or one gate at V_{DD} and the other gate at 0, both the source/channel junction and channel/drain channel are OFF.

Therefore, if we map the transiXNOR ON/OFF states with respect to V_{TG} and V_{BG} at $V_{DS} = V_{DD}$ (shown in Fig.4.2(b)), transiXNOR is ON only when V_{TG} and V_{BG} are either both low or both high, and OFF otherwise. This behavior is precisely XNOR logic.

4.2.2 Simulation Approach

To demonstrate the concept of TransiXNOR, we choose 2 quintuple-layer (2QL) Bi_2Se_3 ³² as an example channel material. The simulated device structure is shown in Fig.4.2(a). Following the Bi_2Se_3 TFET simulation by Zhang et al.,²⁸ 2QL Bi_2Se_3 channel material is used with a thickness of 1.4 nm and a relative static dielectric constant ϵ_r of 100 (from bulk Bi_2Se_3 ³³). The bandgap of 2QL Bi_2Se_3 is 0.252 eV and its electron/hole effective mass is $0.124/2.23 \text{ m}_0$.^{28,32} The source is p -doped with a fixed negative charge concentration of $5.5 \times 10^{13} \text{ cm}^{-2}$ and the drain is n -doped with a fixed positive charge concentration of $3.8 \times 10^{12} \text{ cm}^{-2}$. We define the source degeneracy $\Delta E_S = E_V - E_{FS}$, and drain degeneracy $\Delta E_D = E_{FD} - E_C$. The gate length is 18 nm and the source/drain region is 10 nm. The top and bottom gate oxides are both 1.1 nm HfO_2 with ϵ_r of 25. The

workfunction of the gate metals is 0.215 eV above the conduction edge E_C of Bi_2Se_3 . Ballistic transport is solved self-consistently with the 2D Poisson equation, within the Non-Equilibrium Greens function (NEGF) formalism, using the NanoTCAD ViDES simulation environment.³⁴

4.2.3 Results and discussion

The transport characteristics and the corresponding band diagrams and current spectra are shown in Fig.4.3. As discussed in the Section 4.2.1, when $V_{DS} = 0.2$ V and $V_{BG} = 0.2$ V, the I_{DS} vs. V_{TG} curve resembles a *n*-type TFETs (shown in Fig.4.3(a.1)). When both V_{TG} and V_{BG} biased at 0.2 V, the valence band edge E_V in the source is above the conduction band edge E_C in the channel and the tunneling happens at the source/channel junction (shown in Fig.4.3(a.2)). On the other hand, when $V_{DS} = 0.2$ V and $V_{BG} = 0.1$ V, the I_{DS} vs. V_{TG} curve resembles a *p*-type TFETs (see Fig.4.3(c.1)). When both V_{TG} and V_{BG} biased at 0 V, E_V in the channel is above E_C in the drain and tunneling happens at the channel/drain junction (shown in Fig.4.3(c.2)). Therefore, the transiXNOR is ON when both V_{TG} and V_{BG} are biased at 0 V or 0.2 V. When $V_{DS} = 0.2$ V and $V_{BG} = 0.1$ V, the I_{DS} vs. V_{TG} curve shows ambipolar operation (shown in Fig.4.3(b.1)). The asymmetry between the *n* and *p* branch is due to the difference in the conduction/valence band effective masses of 2QL Bi_2Se_3 . The drain current is minimal when both V_{TG} and V_{BG} are biased at 0.1 V, or one gate at 0.2 V and the other gate at 0, since there is no tunneling window at both source/channel junction and channel/drain junction (shown in Fig.4.3(b.1)).

The key to TransiXNOR design is to manage the three major components

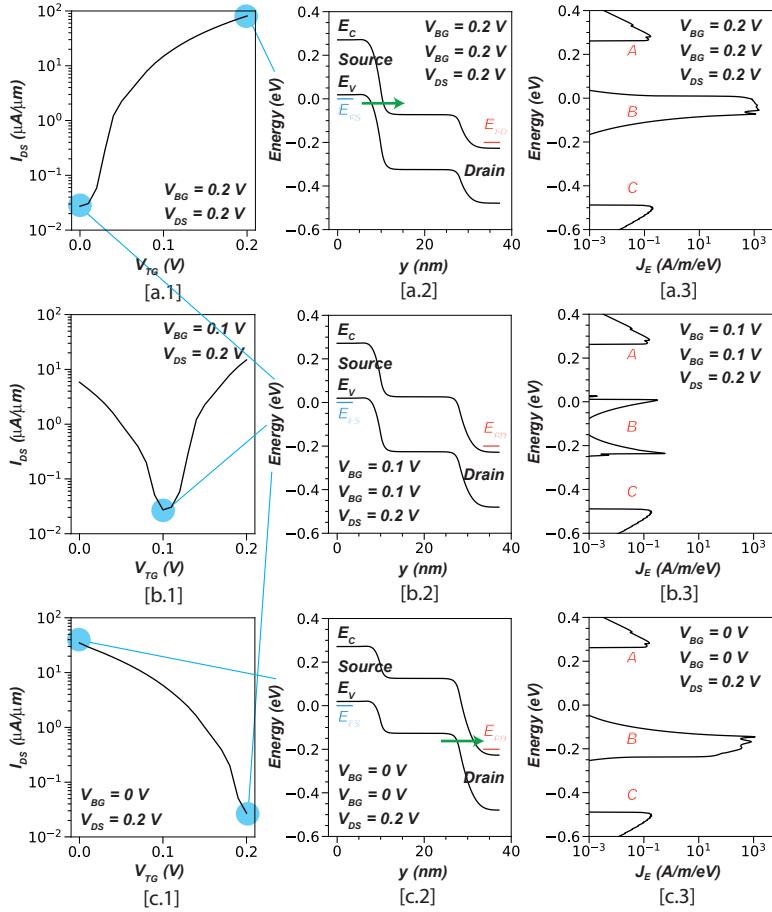


Figure 4.3: (a.1) The I-V curve of the drain current I_{DS} versus V_{TG} when $V_{BG} = 0$ V and $V_{DS} = 0.2$ V; (a.2) The band diagram and (a.3) the current spectrum when $V_{DS} = 0.2$ V and both $V_{TG} = V_{BG} = 0$ V. (b.1) The I-V curve of the drain current I_{DS} versus V_{TG} when $V_{BG} = 0.1$ V and $V_{DS} = 0.2$ V; (b.2) The band diagram and (b.3) the current spectrum when $V_{DS} = 0.2$ V and both $V_{TG} = V_{BG} = 0.1$ V. (c.1) The I-V curve of the drain current I_{DS} versus V_{TG} when $V_{BG} = 0.2$ V and $V_{DS} = 0.2$ V; (c.2) The band diagram and (c.3) the current spectrum when $V_{DS} = 0.2$ V and both $V_{TG} = V_{BG} = 0.2$ V.

of the drain current: A) electron thermionic current, B) inter-band tunneling current, C) hole thermionic current. In the OFF state (shown in Fig.4.3(b.3)), sufficient high doping levels in both source and drain are required in order to reduce the thermionic current (A and C). However if the doping levels were too high, the tunneling energy window ($\Delta E_S + \Delta E_D + qV_{DS}$) became larger than the channel bandgap $E_{G,Channel}$ and TransiXNOR can not be turned off. A suf-

ficient long gate length is necessary to reduce the direct inter-band tunneling from source to drain. The workfunction of the gate metals is designed to minimize the drain current when both V_{TG} and V_{BG} are biased at $V_{DD}/2$ or one gate at V_{DD} and the other at 0. In the ON state, the inter-band tunneling current (B) peaks at either the source/channel junction or the channel/drain junction (shown in Fig.4.3(a.3, c.3)). The V_{DD} has to be large enough so that both the source/channel and the channel/drain junction can be turned ON, while V_{DD} has to be smaller than the channel bandgap so that the tunneling energy window ($\Delta E_S + \Delta E_D + qV_{DS}$) is smaller than $E_{G,Channel}$ when $V_{DS} = V_{DD}$. Therefore, V_{DD} should be chosen to be slightly smaller than $E_{G,Channel} - \Delta E_S - \Delta E_D$.

The output characteristics and the corresponding band diagrams and current spectra are shown in Fig.4.4. When $V_{BG} = 0.2$ V, the output characteristics of TransiXNOR resemble an ordinary *n*-type TFET as shown in Fig.4.4(a.1). Comparing the band diagrams at the same gate biases but different V_{DS} in Fig.4.4(a.2) and Fig.4.3(a.2), the tunneling energy window of the source/channel junction remains virtually unchanged, therefore its inter-band tunneling current stays almost the same when changing V_{DS} from 0.2 V to 0.1 V (shown in Fig.4.4(a.3)). On the other hand, when $V_{BG} = 0$ V, the channel/drain junction is used as the tunnel junction, V_{DS} inevitably affects the tunneling energy window of the channel/drain tunnel junction. Therefore, when $V_{BG} = 0$ V, the output characteristics of TransiXNOR resemble *p*-type tunnel diodes. Comparing the band diagrams at the same gate biases but different V_{DS} in Fig.4.4(b.2) and Fig.4.3(c.2), the tunneling energy window of the channel/drain junction decreases when changing V_{DS} from 0.2 V to 0.1 V, so does its inter-band tunneling current (shown in Fig.4.4(b.3)).

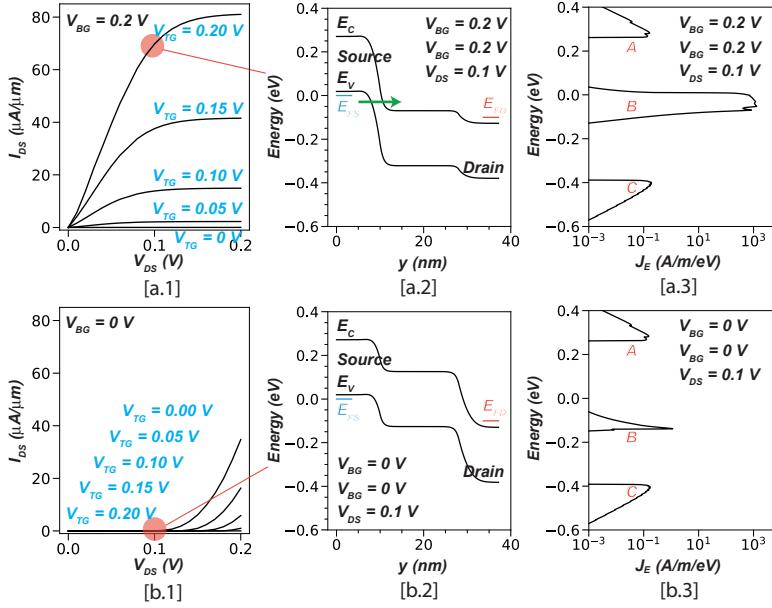


Figure 4.4: (a.1) The family characteristic of TransiXNOR with various V_{TG} at $V_{BG} = 0.2$ V; (a.2) The band diagram and (a.3) the current spectrum when $V_{DS} = 0.1$ V and both $V_{TG} = V_{BG} = 0.2$ V. (b.1) The family characteristic of TransiXNOR with various V_{TG} at $V_{BG} = 0$ V; (b.2) The band diagram and (b.3) the current spectrum when $V_{DS} = 0.1$ V and both $V_{TG} = V_{BG} = 0$ V.

A grid of drain current maps with different V_{TG} and V_{BG} at different V_{DS} is shown in Fig.4.5. Because of the diode-like output characteristics stemming from the channel/drain tunnel junction, TransiXNOR establishes XNOR behavior when V_{DS} is larger than $V_{DD}/2$ but AND behavior instead when V_{DS} is smaller than $V_{DD}/2$.

4.3 TransiXNOR Crossbar Architecture for Binary matrix-vector Multiplication

We utilize both the TransiXNOR and the Resistive RAM (RRAM) to build a XNOR cell shown in Fig.4.6.

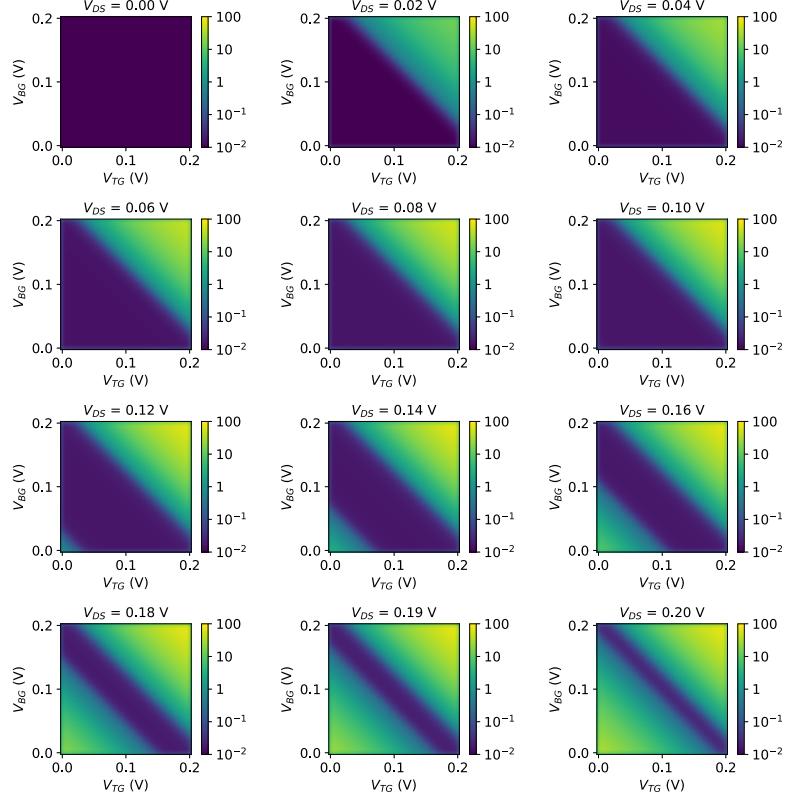


Figure 4.5: A grid of 2D mappings of I_{DS} along both V_{TG} and V_{BG} axes at different V_{DS} . The coloring represents the current density in logarithm. When V_{DS} is larger than 0.1 V (half V_{DD}), the transiXNOR resembles XNOR logic; and when V_{DS} is smaller than 0.1 V, the transiXNOR resemble AND logic.

In the cell, the bit line and work line are used to write to the RRAM. The RRAM is in series with a regular resistor R . The RRAM is set to either low resistance state R_L or high resistance state R_H . After mapping each element $W_{k,n}$ of the 2D binary weight matrix W to either R_L or R_H state of each RRAM, the word line is set floating, and the bit line is set to ground. During the computing, the source line is set to V_{DD} . The resistance R of the regular resistor is designed such as the bottom gate voltage of TransiXNOR is close to zero when RRAM is at the low resistance state R_L and close to V_{DD} when RRAM at the high resistance state R_H . Given the RRAM ON/OFF resistance ratio k (i.e. R_H/R_L), the optimal R to maximize the margin between high and low bottom gate voltage is $\sqrt{k}R_L$,

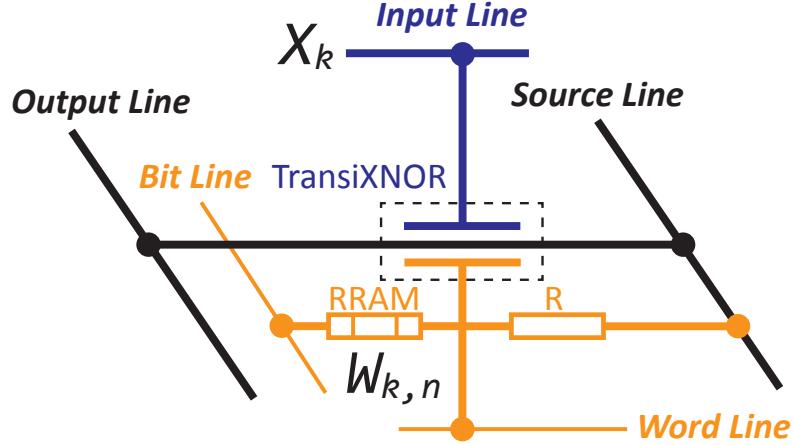


Figure 4.6: The XNOR cell built with TransiXNOR and RRAM. The bit line and work line are used to write to the RRAM. The RRAM is in series with a regular resistor. After writing each element $W_{k,n}$ of the 2D binary weight matrix W to the each RRAM, the word line is set floating, and the bit line is set to ground. During the computing, the source line is set to V_{DD} , and each element X_k of the input vector X is set through each input line in parallel. The current entering the output line represent the XNOR result of $W_{k,n}$ and X_k .

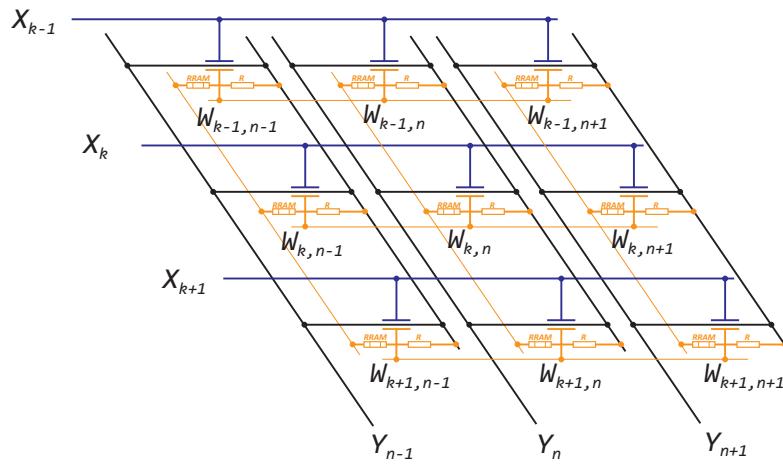


Figure 4.7: The XNOR array to compute $Y = W \times X$ in the constant time.

namely \sqrt{k} times of the low resistance R_L . When ratio k are 10^2 , 10^3 , and 10^4 , the high bottom gate voltage are $0.90V_{DD}$, $0.97V_{DD}$, and $0.99V_{DD}$ respectively; the low bottom gate voltage are $0.10V_{DD}$, $0.03V_{DD}$, and $0.01V_{DD}$ respectively. Since the ON/OFF resistance ratio of RRAM can be larger than 10^4 ,³⁵ the bottom gate voltage has large enough margin between high and low voltage to correctly perform the XNOR logic. Each element X_k of the input vector X is set as the voltage signal on each input line in parallel. The current entering the output line represent the XNOR result of $W_{k,n}$ and X_k . Putting the XNOR cell shown in Fig.4.6 into an array, we got the XNOR array (shown in Fig.4.7) computing binary GEMV in the constant time. The current emerging from Y_n is the sum of the currents through each XNOR cell along the output line due to Kirchoff's law. Therefore, Y , which is $W \times X$, can be read out in parallel from the output lines.

4.4 Conclusion

In this chapter, we proposed a XNOR-enabled transistor: TransiXNOR. TransiXNOR is based on a double gate lateral TFET structure but it uses not only the source/channel junction as a tunnel junction, but also the channel/drain junction. This unique dual junctions enable TransiXNOR to be ON when and only when the top and bottom gate voltage are both high or both low.

Since binary multiplication with -1 and 1 is equivalent to XNOR operation. We proposed a TransiXNOR cell with integrated RRAM to compute the binary product between the memory state in RRAM and input voltage signal. By integrating the TransiXNOR cells into a crossbar architecture, we could compute

the binary GEMV in the constant time, which can be used to greatly accelerate binarized neural network.

BIBLIOGRAPHY

- ¹ Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.
- ² K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
- ³ D. Amodei, S. Ananthanarayanan, R. Anubhai, J. Bai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, Q. Cheng, G. Chen *et al.*, "Deep speech 2: End-to-end speech recognition in english and mandarin," in *International Conference on Machine Learning*, 2016, pp. 173–182.
- ⁴ D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.
- ⁵ I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.
- ⁶ V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller, "Playing atari with deep reinforcement learning," *arXiv preprint arXiv:1312.5602*, 2013.
- ⁷ S. Singh, A. Okun, and A. Jackson, "Artificial intelligence: Learning to play go from scratch," *Nature*, vol. 550, no. 7676, p. 550336a, 2017.
- ⁸ V. Sze, Y.-H. Chen, T.-J. Yang, and J. Emer, "Efficient processing of deep neural networks: A tutorial and survey," *arXiv preprint arXiv:1703.09039*, 2017.
- ⁹ Y. Wang, L. Xia, M. Cheng, T. Tang, B. Li, and H. Yang, "Rram based learning acceleration," in *Compliers, Architectures, and Sythesis of Embedded Systems (CASES), 2016 International Conference on*. IEEE, 2016, pp. 1–2.
- ¹⁰ T. Gokmen and Y. Vlasov, "Acceleration of deep neural network training with resistive cross-point devices," *arXiv preprint arXiv:1603.07341*, 2016.
- ¹¹ L. Chen, J. Li, Y. Chen, Q. Deng, J. Shen, X. Liang, and L. Jiang, "Accelerator-friendly neural-network training: Learning variations and defects in rram crossbar," in *2017 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, 2017, pp. 19–24.

- ¹² M. Horowitz, “1.1 computing’s energy problem (and what we can do about it),” in *Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2014 IEEE International*. IEEE, 2014, pp. 10–14.
- ¹³ M. Courbariaux, Y. Bengio, and J.-P. David, “Binaryconnect: Training deep neural networks with binary weights during propagations,” in *Advances in Neural Information Processing Systems*, 2015, pp. 3123–3131.
- ¹⁴ I. Hubara, M. Courbariaux, D. Soudry, R. El-Yaniv, and Y. Bengio, “Binarized neural networks,” in *Advances in neural information processing systems*, 2016, pp. 4107–4115.
- ¹⁵ M. Rastegari, V. Ordonez, J. Redmon, and A. Farhadi, “Xnor-net: Imagenet classification using binary convolutional neural networks,” in *European Conference on Computer Vision*. Springer, 2016, pp. 525–542.
- ¹⁶ S. Zhou, Y. Wu, Z. Ni, X. Zhou, H. Wen, and Y. Zou, “Dorefa-net: Training low bitwidth convolutional neural networks with low bitwidth gradients,” *arXiv preprint arXiv:1606.06160*, 2016.
- ¹⁷ W. Tang, G. Hua, and L. Wang, “How to train a compact binary neural network with high accuracy?” in *AAAI*, 2017, pp. 2625–2631.
- ¹⁸ R. Zhao, W. Song, W. Zhang, T. Xing, J.-H. Lin, M. Srivastava, R. Gupta, and Z. Zhang, “Accelerating Binarized Convolutional Neural Networks with Software-Programmable FPGAs,” *Int'l Symp. on Field-Programmable Gate Arrays (FPGA)*, Feb 2017.
- ¹⁹ Y. Umuroglu, N. J. Fraser, G. Gambardella, M. Blott, P. Leong, M. Jahre, and K. Vissers, “Finn: A framework for fast, scalable binarized neural network inference,” in *Proceedings of the 2017 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays*, ser. FPGA ’17. New York, NY, USA: ACM, 2017, pp. 65–74. [Online]. Available: <http://doi.acm.org/10.1145/3020078.3021744>
- ²⁰ T. Ajayi, K. Al-Hawaj, A. Amarnath, S. Dai, S. Davidson, P. Gao, G. Liu, A. Lotfi, J. Puscar, A. Rao *et al.*, “Celerity: An open-source risc-v tiered accelerator fabric,” in *Symp. on High Performance Chips (Hot Chips)*, 2017.
- ²¹ E. Nurvitadhi, D. Sheffield, J. Sim, A. Mishra, G. Venkatesh, and D. Marr, “Accelerating binarized neural networks: Comparison of fpga, cpu, gpu, and asic,” in *Field-Programmable Technology (FPT), 2016 International Conference on*. IEEE, 2016, pp. 77–84.

- ²² L. Ni, Z. Liu, H. Yu, and R. Joshi, "An energy-efficient digital reram-crossbar based cnn with bitwise parallelism," *IEEE Journal on Exploratory Solid-State Computational Devices and Circuits*, 2017.
- ²³ T. Tang, L. Xia, B. Li, Y. Wang, and H. Yang, "Binary convolutional neural network on rram," in *Design Automation Conference (ASP-DAC), 2017 22nd Asia and South Pacific*. IEEE, 2017, pp. 782–787.
- ²⁴ S. Yu, Z. Li, P. Y. Chen, H. Wu, B. Gao, D. Wang, W. Wu, and H. Qian, "Binary neural network with 16 mb rram macro chip for classification and online training," in *2016 IEEE International Electron Devices Meeting (IEDM)*, Dec 2016, pp. 16.2.1–16.2.4.
- ²⁵ M. Prezioso, F. Merrikh-Bayat, B. Chakrabarti, and D. Strukov, "Rram-based hardware implementations of artificial neural networks: progress update and challenges ahead," in *Proc. of SPIE Vol*, vol. 9749, 2016, pp. 974 918–1.
- ²⁶ U. E. Avci, R. Rios, K. Kuhn, and I. A. Young, "Comparison of performance, switching energy and process variations for the tfet and mosfet in logic," in *VLSI Technology (VLSIT), 2011 Symposium on*. IEEE, 2011, pp. 124–125.
- ²⁷ T. Krishnamohan, D. Kim, S. Raghunathan, and K. Saraswat, "Double-gate strained-ge heterostructure tunneling fet (tfet) with record high drive currents and 60mv/dec subthreshold slope," in *2008 IEEE International Electron Devices Meeting*, Dec 2008, pp. 1–3.
- ²⁸ Q. Zhang, G. Iannaccone, and G. Fiori, "Two-dimensional tunnel transistors based on bi₂se₃ thin film," *IEEE Electron Device Letters*, vol. 35, no. 1, pp. 129–131, 2014.
- ²⁹ C. Anghel, A. Gupta, A. Amara, A. Vladimirescu *et al.*, "30-nm tunnel fet with improved performance and reduced ambipolar current," *IEEE Transactions on Electron Devices*, vol. 58, no. 6, pp. 1649–1654, 2011.
- ³⁰ S. Sahay and M. J. Kumar, "Controlling the drain side tunneling width to reduce ambipolar current in tunnel fets using heterodielectric box," *IEEE Transactions on Electron Devices*, vol. 62, no. 11, pp. 3882–3886, 2015.
- ³¹ J. Wu and Y. Taur, "Reduction of tfet off-current and subthreshold swing by lightly doped drain," *IEEE Transactions on Electron Devices*, vol. 63, no. 8, pp. 3342–3345, 2016.

³² Y. Zhang, K. He, C.-Z. Chang, C.-L. Song, L.-L. Wang, X. Chen, J.-F. Jia, Z. Fang, X. Dai, W.-Y. Shan *et al.*, “Crossover of the three-dimensional topological insulator bi₂se₃ to the two-dimensional limit,” *Nature Physics*, vol. 6, no. 8, pp. 584–588, 2010.

³³ H. Köhler and C. R. Becker, “Optically active lattice vibrations in bi₂se₃,” *physica status solidi (b)*, vol. 61, no. 2, pp. 533–537, 1974.

³⁴ G. Fiori and G. Iannaccone, “Nanotcad vides,” May 2016 [Online].

³⁵ S. H. Jo, T. Kumar, C. Zitlaw, and H. Nazarian, “Self-limited rram with on/off resistance ratio amplification,” in *VLSI Technology (VLSI Technology), 2015 Symposium on*. IEEE, 2015, pp. T128–T129.

CHAPTER 5

ARTIFICIAL NEURAL NETWORKS (ANNS) FOR DEVICE COMPACT MODELING

5.1 Introduction

Device compact modeling bridges device researches to their applications, allowing circuit level simulations before the hardwares are production-ready. The predominant compact models are physics-based,^{1,2} where fundamental device physics are used as the building blocks, and empirical equations are hand-crafted to modify and merge physical expressions into smooth analytical functions. However developing a high-quality physics-based compact models is very expensive and time-consuming. In order to quickly incorporate new generations of devices into circuit simulations, data-oriented modeling methods are developed to circumvent the detailed physics, focusing on delivering numerically stable and computationally efficient models directly from the device data.

Table look-up models are current widely-used data-oriented models. Artificial neural networks (also rebranded as “deep learning”) has also raised a lot of interests.^{3–6} Comparing table look-up models and artificial neural networks (ANNs), in theory, the neural network model performs better on the following three aspects.

1. Scalability: In order to achieve certain level of accuracy, the table lookup model needs a large amount of data, and the space complexity increases exponentially with increasing dimensions. In contrast, the neural network model is lightweight and scalable.

2. Generalization: The table lookup model has poor generalization performance. The polynomial fitting used in the table lookup model often has high out-of-sample errors. In contrast, by using correct learning algorithms, neural network model can be well generalized, which make it more robust against noises.
3. Smoothness: An ideal compact model needs to be infinitely differentiable. The table lookup model is not infinitely differentiable due to the nature of polynomial fitting, while using higher order polynomial fitting will improve the smoothness, and it is at the expense of computation efficiency. Therefore, the table lookup model is not possible to be both smooth and computationally efficient. In contrast, the neural network model is guaranteed to be infinitely differentiable.

Despite of all the theoretical benefits of a neural network, a fundamental question arise: *Can neural network model very small current in the deep sub-threshold region or around V_{DS} equals zero?*. Gradient-based learning in a neural network relies on the gradients of the loss function. A common loss function is the mean squared error (MSE): $1/N * \sum_i (pred_i - label_i)^2$, where N is the number of examples, $label_i$ and $pred_i$ are the true value and the neural network output of each example i . Its partial gradient with respect to $pred_i$ scales linearly with the value of $pred_i$. In the context of device modeling, the value of $pred_i$ (i.e. current density) varies over 8 orders of magnitudes. Assuming the max value of $pred_i$ has been normalized to 1. For a very small current value ($< 10^{-6}$), its partial gradient is too small to have significant impacts on the training. Even worse is that most ANNs compact modeling frameworks^{3,4,6} used the vanilla form of feed-forward neural networks known as multi-layer perceptions (MLPs) with hyperbolic tangent ($tanh$) activation functions. In this neural network architecture, nothing

stops its prediction value to oscillate around zero when the label value is very small. It leads to the unphysical behaviors in both the I_D - V_{DS} and I_D - V_{GS} curves.⁷ This is a fundamental limitation of MLPs with \tanh activation functions and the MSE loss function.

Learned from the groundbreaking successes of deep learning in image classification⁸ and speech recognition,⁹ it is important of leveraging invariants in the problem to design structured neural network architectures. In the previous work,⁷ we encoded the fundamental device physics into the new neural network architecture: Physics-based neural networks (Pi-NN) to overcome the fundamental limitation of MLPs. As far as we are aware, Pi-NN is the only neural network based compact modeling framework can accurately model the deep sub-threshold region. The major criticism of the previous work is that the current needs to multiply a scalar function in the form of $\exp(-a(V_G + b))$ for better deep sub-threshold modeling. Besides complicating the modeling process by introducing more hyper-parameters, this pre-processing method failed to improve deep sub-threshold modeling when the drain voltage affecting the threshold voltage (e.g. Drain-induced barrier lowering (DIBL) effect¹).

In this work, we redesigned the loss function of Pi-NN to successfully eliminate the need of this tricky pre-processing step in the original work,⁷ laying the groundwork for the new compact modeling framework: Pi-NN. We discussed how Pi-NN utilized the invariants of device physics in the section 5.3. In the section 5.5, we proposed the new reweighted L1 loss function for efficient training. The Pi-NN framework has been used to generate compact models from experimental data of a Gallium Nitride (GaN) High Electron Mobility Transistor (HEMT),¹⁰ and theoretical simulated data of Two-dimensional Het-

erostructure Interlayer Tunneling Field Effect Transistors (Thin-TFETs).¹¹ The proposed Pi-NN framework is 1) the *only* framework that can generate an accurate and smooth transistor compact model in all operation regimes; and 2) an *generic* framework such that it can be used to model very different devices (e.g. Tunnel FETs, HEMTs, and other exotic transistors) without any device-specific modification and preprocessing.

5.2 Previous Works

There have been several recent advances aimed at developing neural network based compact models. Wang and Zhang et al.^{5,12,13} proposed to use neural networks for RF and microwave design. In the proposed knowledge-based neural models,⁵ the neural network structure embedded the empirical or semi-analytical functions as the activation functions, and the problem dependent boundary functions as the “boundary layer”. This neural network structure combines the empirical functions usually valid only in a certain region of the parameter space. However, this knowledge-based neural models relies on the selected empirical functions which may be different for different devices. Therefore it may not be a generic framework. Moreover, the quality of these models are largely dependent on the quality of the selected empirical functions. And those practical empirical functions may not be available for emerging devices. Xu and Root et al.^{14–16} developed the commercialized framework NeuroFET inside Agilent IC-CAP. However, this framework works poorly in the very small current region (e.g. deep sub-threshold region) as discussed in the Section 5.1. This limitation was also recognized by Zhang et al.⁶ Zhang et al. enhanced the ANN model accuracy with data preprocessing, which transfers (V_{GS} , V_{DS} , I_{DS})

to $(V_{GS}, \log(V_{DS}), \log(I_{DS}))$. However, this data preprocessing method has three limitations: 1) An assumption of this method is the linear $I_{DS} - V_{DS}$ dependence, but exponential $I_{DS} - V_{DS}$ dependence is also very common due to short channel effects such as DIBL; 2) Since the model has to exchange source and drain when a negative V_{DS} is given, the model won't be able to guarantee smooth derivatives across V_{DS} equals zero; 3) For V_{DS} equals zero, either an exact zero current needs to be assigned or a smooth function needs to be implemented to guarantee an exact zero current. Also, one additional training data point close to zero is required for improving the model accuracy in the linear region.

All of these limitations stem from the ignorance of the device physics in the MLP network structure. Regardless of the preprocessing, MLPs still treats V_{DS} and V_{GS} inputs interchangeable even though they are responsible for two very different physical effects in the device. The more graceful solution is to incorporate these intrinsic structures in the input space into the neural network architecture. The two key contributions of this work are 1) the Pi-NN architecture, which is structured according to the invariants in the fundamental device physics, and 2) the reweighted L1 loss function, which ensures accurate modeling in all device operation region without the need of tricky preprocessing steps. Combining the Pi-NN architecture and the reweighted L1 loss function, the new Pi-NN framework can 1) eliminate the need of the preprocessing in Zhang et al.,⁶ 2) generate accurate and smooth compact models in all operation region; 3) easily adapt to different devices.

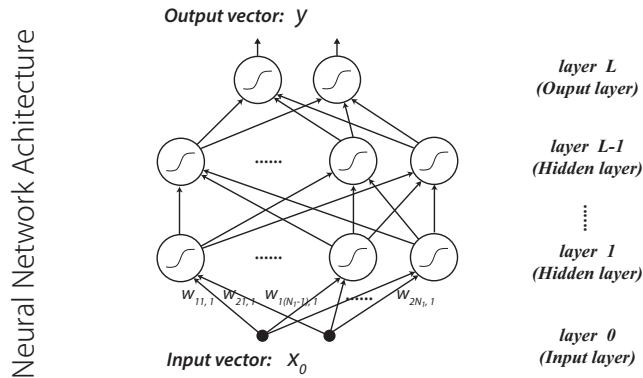
5.2.1 Low Current Regime Challenge

When the V_{DS} is close to zero or the device is in subthreshold region, the current is very small comparing to the ON current. Modeling the output with very large range (around 8 order of magnitudes) is challenging for neural networks. Moreover, I-V relationships are also very different in the low current regime for the gate voltage or the drain voltage. I_{DS} is exponentially dependent on V_{GS} while linearly dependent of V_{DS} in the low current regime. Therefore, there are the following challenges when approximating with the neural network:

1. For common loss functions, such as L2 loss, the gradient decreases with smaller outputs. Therefore, neural networks failed to accurately model the low current regime;
2. It is hard to model exact zero in MLP. However, when V_{DS} equals zero, the output should be exact zero;
3. Along different input features V_{GS} and V_{DS} , the relationships are vastly different (one is exponential and another is linear).

We further illustrate these challenges by using the MLP neural network to generate a compact model for the DC I-V curves of the Thin-TFET.¹¹ Thin-TFET structure is shown in Fig.3.1(b). The training data are simulated for the top gate voltage (V_{TG}) from 0 to 0.4 V and the drain-source voltage (V_{DS}) from -0.1 to 0.4 V with an uniform step of 0.01 V, while the test data are for V_{TG} from 0.005 to 0.405 V and V_{DS} from -0.095 to 0.405 V with an uniform step of 0.01 V. The MLP neural network architecture and its well-established learning algorithms are shown in Fig.5.1.

Previous works:
Multilayer Perceptron (MLP) Neural Network



Symbols, Constraint and Parameters

Symbols:

- l : Layer Index
- N_l : No. of neurons in the l^{th} layer
- x_l : The output vector of the l^{th} layer
- z_l : The intermediate vector of the l^{th} layer
- σ : The entrywise product of two vectors
- L : No. of layers in the network
- y : The output vector of the network
- x_0 : The input vector of the network

Parameters: W_l : the weight matrix, whose element $w_{ji,l}$ connects the i^{th} neuron of the $(l-1)^{\text{th}}$ layer to the j^{th} neuron of the l^{th} layer
 b_l : the bias vector, whose element $b_{i,l}$ is the bias of the i^{th} neuron of the l^{th} layer

Feedforward
Function

```
for each layer l in [1, 2, ..., L] :
     $z_l = W_l x_{l-1} + b_l$ 
     $x_l = \sigma(z_l)$ 
     $y = x_L$ 
```

$$E = \frac{1}{2} \|y - t\|^2 \quad \text{where } t \text{ is the desired output vector of } y$$

Error Backpropagation

```
(* For the output layer *)
 $\frac{\partial E}{\partial x_L} = \frac{\partial E}{\partial y_L}$ 
(* For the hidden layers *)
for each layer l in [L, L-1, ..., 1] :
     $\frac{\partial E}{\partial z_l} = \frac{\partial x_l}{\partial z_l} \circ \frac{\partial E}{\partial x_l}$ 
     $\frac{\partial E}{\partial x_{l-1}} = (W_l)^T \frac{\partial E}{\partial z_l}$ 
```

Figure 5.1: The Multiplayer Perception (MLP) neural network model

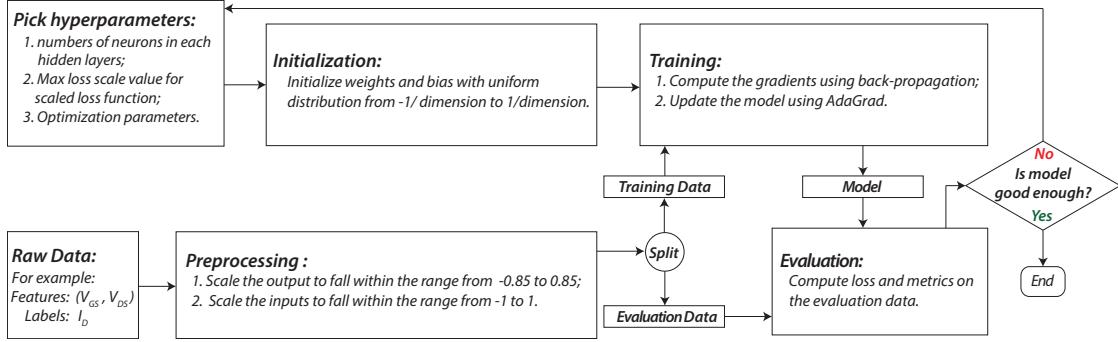


Figure 5.2: A training procedure for Artificial Neural Network (ANN) device compact modeling.

We follow the training procedure shown in Fig.5.2. After some initial training, we choose to use MLP neural networks with two hidden layers and defined its hyper-parameter as (i, j) , where i is the number of neurons in the first hidden layer and j is the number of neurons in the second hidden layer. Each neuron uses the hyperbolic tangent function $\tanh(x) = (e^x - e^{-x}) / (e^x + e^{-x})$ as the activation function. By choosing the hyper-parameter (i, j) to be $(5, 5)$, $(7, 7)$ and $(9, 9)$, these three MLP neural networks were trained for 5 million epochs. Using the loss function defined in Fig. 3, the root-mean-squared (R.M.S) deviations for training data and test data are plotted in Fig.5.3(a). The test errors are used to evaluate the generalization ability of the model, namely how the model fit the unseen data. As shown in Fig. 4(a), the test errors stay close to the training errors, which indicated a good generalization. We choose to plot the I-V curves modeled by the MLP neural network with 7 \tanh neurons in the first and second hidden layers as shown in Fig.5.3(b), which gives a neural network with 15 neurons and 85 parameters in total. Figure 5.3(c-f) show the I-V curves generated by the MLP neural network compact model along with the training data and the test data. Good fitting in the linear scale is achieved for both the I_{DS} - V_{DS} and the I_{DS} - V_{TG} curves. However, if we zoom in the region near $V_{DS} = 0$, I_{DS} is not zero when V_{DS} is zero, indicating the I_{DS} - V_{DS} relationship is unphysical

around $V_{DS} = 0$ (see Fig. 4(e) and the inset). Moreover, the I_{DS} - V_{TG} relationship is also unphysical in the sub-threshold region (shown in Fig.5.3(f)). The fundamental reason of these unphysical behaviors is that the MLP neural network has no knowledge of the device physics; therefore the fitting is no longer physical when I_D is very small. In order to eliminate these unphysical behaviors, we have to design a neural network with *a priori* knowledge of the fundamental device physics.

5.3 The Idea of Pi-NN: Structured Physical System

Structured models make independent assumptions to limit the size of the configuration set. Structures introduce invariance in the system. Invariance serves as the prior knowledge. Prior knowledge profoundly influences the effectiveness of learning. For example, the convolutional neural network (CNN)¹⁷ incorporates spatial invariance, and the recurrent neural network (RNN)¹⁸ incorporates natural ordering.

When comes to device modeling, we first note that the inputs V_{DS} and V_{TG} are related to two different physical effects: V_{DS} drives the current through the device while V_{TG} controls the channel potential profile to change the magnitude of the current. Therefore, V_{DS} and V_{TG} should be fed to two different neural networks as shown in Fig.5.4. According to the fundamental device physics, we know I_{DS} - V_{DS} curves have a linear region at small V_{DS} and a saturation region at large V_{DS} . This behavior is similar to a *tanh* function. This indicates V_{DS} should be fed into a neural network with *tanh* activation functions (*tanh* subnet). To ensure I_{DS} equals zero when V_{DS} equals zero, all the *tanh* neurons in the *tanh*

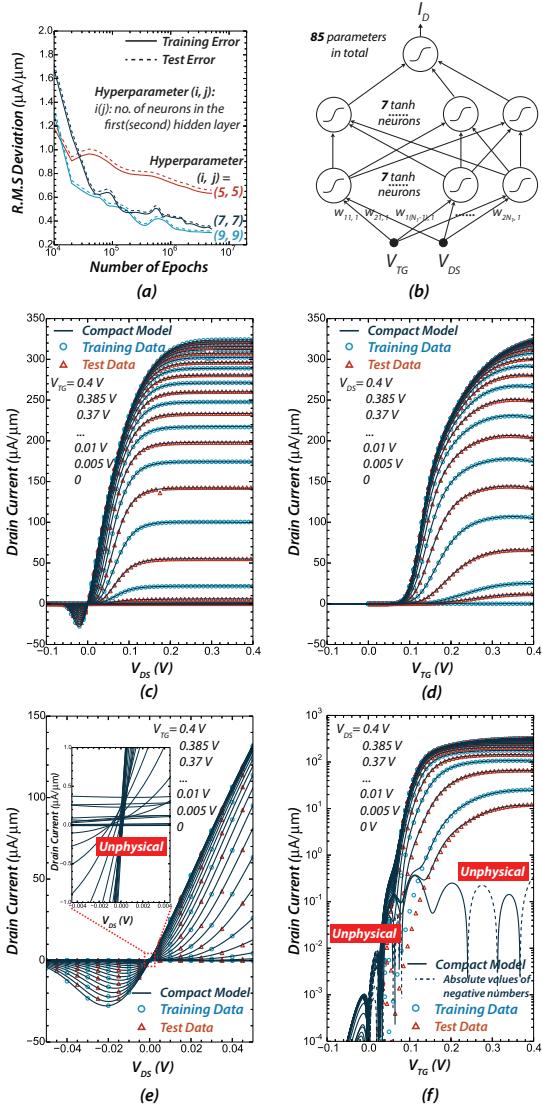


Figure 5.3: The compact model of the n-type Thin-TFET derived based on the MLP neural network widely used in previous works,^{3–5} (a) the training errors and test errors for a variety of hyper-parameters; (b) the MLP neural network with 7 *tanh* neurons in the first and second hidden layers. From (c) to (f), the I-V curves generated by the MLP neural network shown in (b) are plotted along with the training data and the test data: (c) I_{DS} versus V_{DS} at different V_{TG} ; (d) I_{DS} versus V_{TG} at different V_{DS} in linear scale; (e) I_{DS} versus V_{DS} at different V_{TG} around $V_{DS} = 0 \text{ V}$, the embedded plot shows unphysical I_{DS} - V_{DS} relationships around V_{DS} equals 0; (f) I_{DS} versus V_{TG} at different V_{DS} in semi-log scale, unphysical oscillation of I_{DS} around zero appears in the sub-threshold region and when $V_{DS} = 0 \text{ V}$.

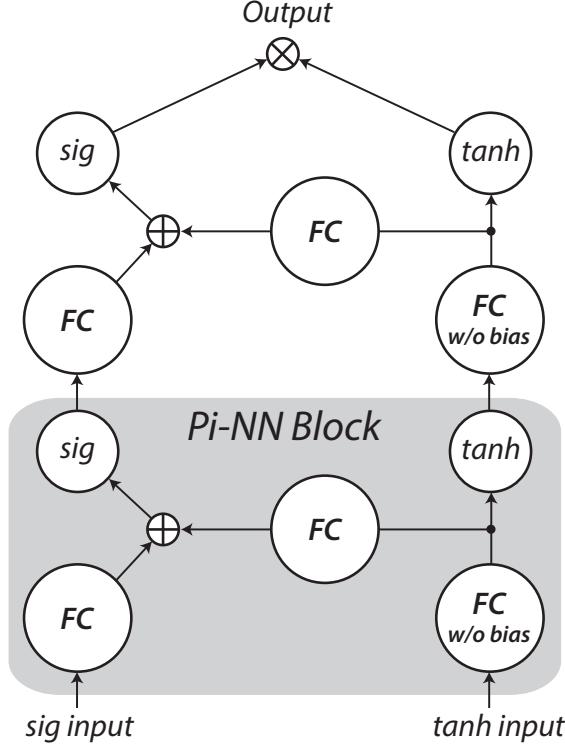


Figure 5.4: The architecture of Pi-NN. The shaded area indicates a Pi-NN block, which is the building block of Pi-NN network.

subnet must have no bias terms. On the other hand, the I_{DS} - V_{TG} curves have an exponential turn-on in the sub-threshold region and then become a polynomial in the ON region. This is best simulated as a *sigmoid* function $\text{sig}(x)=1/(1+e^{-x})$. Therefore, V_{TG} is fed into a neural network with *sigmoid* activation functions (*sig* subnet). It should be noted that we assumed gate leakage current is negligible, so V_{TG} would not change the sign of I_{DS} . The final drain current is the entrywise product of the outputs of the *tanh* subnet and the *sig* subnet. This entrywise product reflects the control of V_{TG} on the drain current driven by V_{DS} . In addition, V_{DS} can affect the channel potential profile controlled by V_{TG} due to various non-ideal effects such as the short channel effects. Therefore weighted connections are added between each layer in the *tanh* subnet and its corresponding layer in the *sig* subnet. By embedding the above device physics in a neural

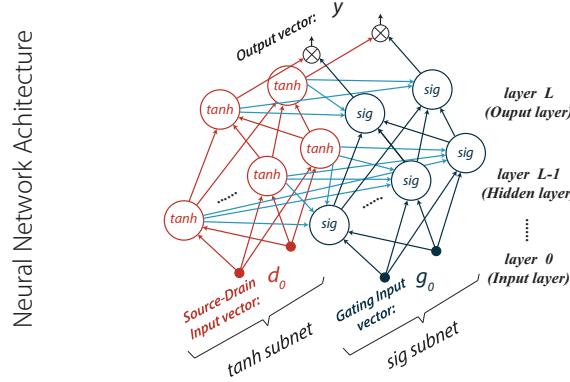
network structure, we arrive at the Physics-Inspired Neural Network (Pi-NN). The Pi-NN architecture and its pseudo-codes for the feed-forward and error back-propagation algorithms are shown in Fig.5.5.

5.4 Adjoint Sensitivity Network

The neural network sensitivity analysis is to find the network output sensitivities with respect to variations in the inputs of multilayer feedforward neural networks with differentiable activation function.^{19,20} This analysis has been used to understand variable contributions in the neural network²¹ and visualize the importance of inputs with respect to classification decision.^{22,23} Recently, the sensitivity analysis also provides the gradient information to fool the deep neural networks to produce high confidence classifications of unrecognizable images.²⁴

In the context of device modeling, unlike the application of sensitivity analysis mentioned above, we would like to train the origin network with the output sensitivity with respect to the inputs. Therefore, we create a new network called adjoint sensitivity neural network (adjoint network).²⁵ Shown in Fig.5.6, the adjoint network share the same parameters (weights) with the origin one, and the take the activations (i.e. the output of the activation function) from each layer in the origin network as the input. The adjoint network has the same input dimension as the output of the original network, and the same output dimension as the input of the original network. The input to the adjoint network is a vector with only one non-zero element (the selector). If the i^{th} element of the selector is non-zero, the output of the adjoint network is gradient of i^{th} element of output with respect to the input vector. Here the adjoint sensitivity network mainly

This work:
Physics-Inspired Neural Network (Pi-NN)



Neural Network Architecture
Symbols, Constraint and Parameters

- Symbols:**
- l : Layer Index
 - L : No. of layers in the network
 - $N_l^{(S)}$: No. of neurons in the l^{th} layer of the **tanh(sig) subnet**
 - y : The output vector of the network
 - $d_o(g_o)$: The input vector of the **tanh(sig) subnet**
 - $d_l(g_l)$: The output vector of the l^{th} layer in the **tanh(sig) subnet**
 - z_l : The intermediate vector of the l^{th} layer in the **tanh(sig) subnet**
 - \circ : The entrywise product of two vectors

Constraint: $N_l^T = N_l^S$ (The output layers of the two subnets have same no. of neurons)

Parameters:

- $W_l^{(S)}$: the weight matrix for the **tanh(sig) subnet**, whose element $w_{ji,l}^{(S)}$ connects the i^{th} neuron of the $(l-1)^{\text{th}}$ layer to the j^{th} neuron of the l^{th} layer
- W_l^P : the weight matrix for the peephole connections from the **tanh subnet** to the **sig subnet**, whose element $w_{ji,l}^P$ connects the i^{th} neuron of the l^{th} layer in the **tanh subnet** to the j^{th} neuron of the l^{th} layer in the **sig subnet**
- b_l^S : the bias vector, whose element $b_{ii,l}^S$ is the bias of the i^{th} neuron of the l^{th} layer in the **sig subnet**

(Note: There is no bias in the tanh neurons of the tanh subnet)

Feedforward
Loss Function
Error Backpropagation

```
for each layer l in [1, 2, ..., L] :
     $z_l^T = W_l^T d_{l-1}$  (* tanh subnet *)
     $d_l = \tanh(z_l^T)$  (* tanh subnet *)
     $z_l^S = W_l^P d_l + W_l^S g_{l-1} + b_l^S$ 
    (* sig subnet & peephole connections *)
     $g_l = \text{sig}(z_l^S)$  (* tanh subnet *)
     $y = d_l \circ g_l$ 
```

$$E = \frac{1}{2} \|y - t\|^2 \quad \text{where } t \text{ is the desired output vector of } y$$

```
(* For the output layer *)
 $\frac{\partial E}{\partial g_l} = d_l \circ \frac{\partial E}{\partial y}, \frac{\partial E}{\partial d_l} = g_l \circ \frac{\partial E}{\partial y}$ 
(* For the hidden layers *)
for each layer l in [L, L-1, ..., 1] :
     $\frac{\partial E}{\partial z_l^S} = g_l \circ (1 - g_l) \circ \frac{\partial E}{\partial g_l}$  (* sig subnet *)
     $\frac{\partial E}{\partial g_{l-1}} = (W_l^S)^T \frac{\partial E}{\partial z_l^S}$  (* sig subnet *)
     $\frac{\partial E}{\partial z_{j,i}^T} = (1 - d_l \circ d_l) \circ \left( \frac{\partial E}{\partial d_l} + (W_l^P)^T \frac{\partial E}{\partial z_l^T} \right)$ 
    (* tanh subnet & peephole connections *)
     $\frac{\partial E}{\partial d_{l-1}} = (W_l^T)^T \frac{\partial E}{\partial z_l^T}$  (* tanh subnet *)
```

Figure 5.5: The Physics-Inspired Neural Network (Pi-NN) model.

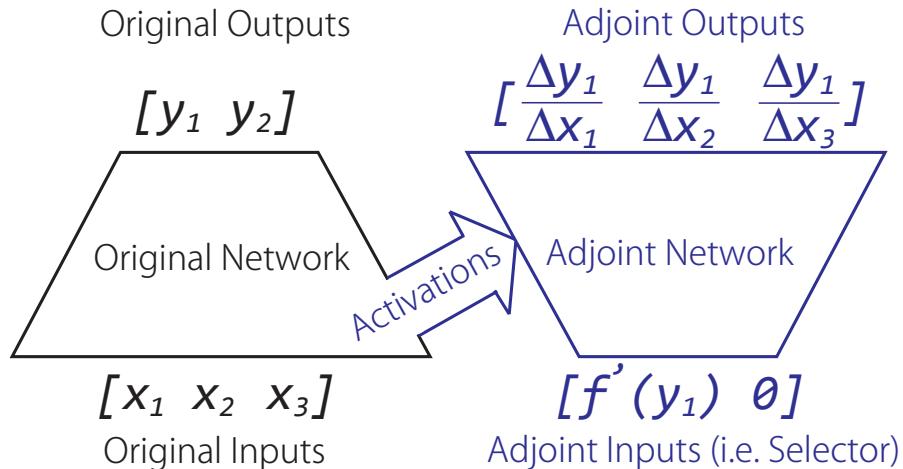


Figure 5.6:

serves two applications:

1. *Use the adjoint sensitivity network to train the parameters of the original neural network:* In circuit network analyzer measurements, we are able to obtain gradient responses of the outputs with respect to the inputs. For example, a capacitance between two terminals is the partial derivative of the charge from one terminal with respect to the voltage on the other terminal (shown in Fig.3.1). Training an adjoint sensitivity network on the capacitance data will, at the same time, generate the original neural network, which outputs the terminal charges at different terminal voltage. More generally, the adjoint sensitivity network trains on the Jacobian matrix of a vector function, and its trained parameters are shared with the original neural network, which approximates the vector function.
2. *Train the parameters of the original neural networks and use the adjoint sensitivity to output the first-order partial derivative between the outputs and inputs (i.e. Jacobian matrix):* In I-V modeling, the first-order partial derivatives of the drain current with respect to the gate and drain voltage are the transcon-

ductance and output conductance respectively. Also monotonicity of the model in interpolation and extrapolations is an important model verification metric for some devices. The adjoint sensitivity network can be used to verify the monotonicity in interpolation and extrapolations.

The adjoint sensitivity network can be constructed from the origin neural network layer-by-layer (shown in Fig.5.7). First, we define a sensitivity vector β :

$$\begin{aligned}
\beta_{qi}^l &= \frac{\partial y_q}{\partial \gamma_i^l} \\
&= \sum_{k=1}^{N_{l+1}} \frac{\partial y_q}{\partial \gamma_k^{l+1}} \frac{\partial \gamma_k^{l+1}}{\partial \gamma_i^l} \\
&= \sum_{k=1}^{N_{l+1}} \frac{\partial y_q}{\partial \gamma_k^{l+1}} \frac{\partial \gamma_k^{l+1}}{\partial x_i^l} \frac{\partial x_i^l}{\partial \gamma_i^l} \\
&= x_i^l(1 - x_i^l) \sum_{k=1}^{N_{l+1}} \beta_{qk}^{l+1} W_{ki}^l
\end{aligned} \tag{5.1}$$

For a MLP with sigmoid activation function in Fig.5.7(a), y_q is q^{th} element of the output vector; γ_i^l is the i^{th} element of the output vector of FC layer in l^{th} layer; N_l is the number of neurons in l^{th} layer; x_i^l is the i^{th} element of the output vector of sigmoid functions (i.e. activation) in l^{th} layer; and W_{ki}^l is the weight connecting i^{th} element of the input vector to k^{th} element of the output vector in l^{th} layer. We can turn Eq.5.1 into a computation graph, resulting in Fig.5.7(a). For the last layer L in the origin layer (i.e. the first layer in the adjoint layer):

$$\beta_{qi}^L = y_q(1 - y_q) \text{ iff } q = i \text{ otherwise } 0 \tag{5.2}$$

Therefore the input vector of the adjoint network only has one non-zero el-

ement, and if the q^{th} element of the input vector is non-zero then the output vector is the gradient of q^{th} element of output with respect to the input vector.

As for the Pi-NN block and its adjoint network shown in Fig.5.7(b), there are two input vectors and two output vectors in the adjoint block. The output vector of Pi-NN is the entrywise product of the output vectors S^L and T^L from the last Pi-NN block. Therefore, we define two sensitivity vector β and α defined in Eq.5.3 and Eq.5.4.

$$\beta_{qi}^l = \frac{\partial y_q}{\partial \gamma_i^l} = S_i^l(1 - S_i^l) \sum_{k=1}^{N_{l+1}^S} \beta_{qk}^{l+1} [W_S]_{ki}^{l+1} \quad (5.3)$$

$$\alpha_{qj}^l = \frac{\partial y_q}{\partial \epsilon_j^l} = (1 - (T_j^l)^2) \sum_{k=1}^{N_{l+1}^T} \alpha_{qk}^{l+1} [W_T]_{kj}^{l+1} + \sum_{i=1}^{N_{l+1}^I} \beta_{qi}^l [W_I]_{ij}^{l+1} \quad (5.4)$$

As shown in Fig.5.7(b), y_q is q^{th} element of the output vector; γ_i^l is the i^{th} element of the output vector of FC layer in l^{th} layer of *sig* subnet; ϵ_j^l is the j^{th} element of the output vector of FC layer in l^{th} layer of *tanh* subnet; $N_l^{S(T)}$ is the number of neurons in l^{th} layer of *sig* (*tanh*) subnet; $S(T)_{i(j)}^l$ is the $i(j)^{th}$ element of the output vector of sigmoid (*tanh*) functions (i.e. activation) in l^{th} layer of *sig* (*tanh*) subnet; and $[W_S]_{ki}^l$ is the weight connecting i^{th} element of the input vector to k^{th} element of the output vector in l^{th} layer of *sig* subnet, similarly $[W_T]_{kj}^l$ in l^{th} layer of *tanh* subnet, and $[W_T]_{kj}^l$ in l^{th} layer between the *sig* subnet and *tanh* subnet. We can turn Eq.5.3 and Eq.5.4 into a computation graph as well, resulting in Fig.5.7(b). For the last layer L in the origin layer (i.e. the first layer in the adjoint layer):

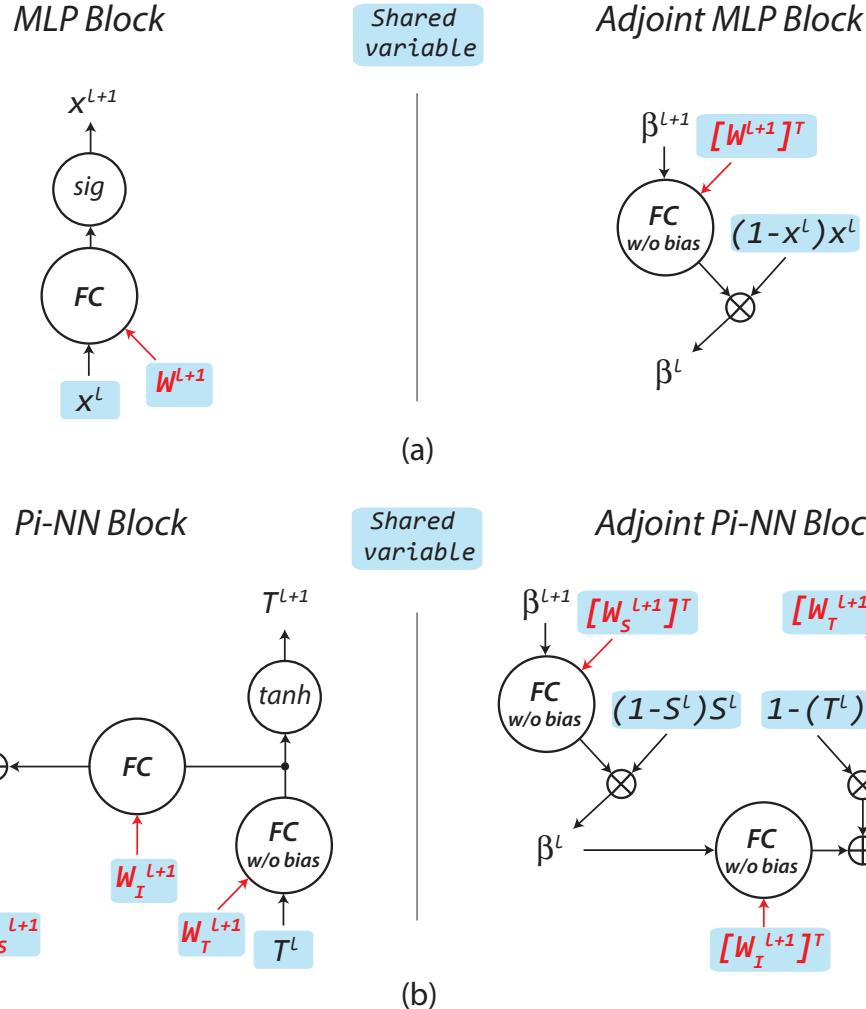


Figure 5.7: (a) The adjoint network of a fully connected (FC) layer with sigmoid activation functions, where $\beta = \nabla_y y$, y is the output of FC layer, and y is the outputs of the neural network; (b) The adjoint network of a Pi-NN block, where $\beta = \nabla_y y$ and $\alpha = \nabla_\delta y$, y is the output of FC layer in *sig* subnet, δ is the output of FC in *tanh* subnet, and y is the outputs of the neural network.

$$\begin{aligned} \beta_{qi}^L &= T_i^L (1 - S_i^L) S_i^L \quad \text{iff } q = i \text{ otherwise } 0 \\ a_{qj}^L &= S_i^L (1 - (T_i^L)^2) \quad \text{iff } q = j \text{ otherwise } 0 \end{aligned} \tag{5.5}$$

After we construct the computation graph for both the original and the adjoint network, we can utilize the “automation” differentiation function in modern deep learning libraries such as Tensorflow and Caffe2 to generate the com-

putation graphs for back-propagation and updating.

5.5 Weighted L1 Loss Function

Even though Pi-NN has been structured for device modeling, a suitable optimization algorithm is still needed to train Pi-NN properly. Unlike other machine learning tasks, the output of the model, namely the current density, varies over 8 order of magnitudes and precise modeling in all output range is required. To meet this demand, instead of usual mean square error loss function (L2 loss), we proposed to use L1 loss function, and re-weight the element-wise loss to give significant large gradient signals when the output value is extremely small. This new loss function is named “weighted L1 loss function”. Figure 5.8 illustrates how the weighted L1 loss function is computed.

The target value has a large range over 8 order of magnitudes illustrated in Fig.5.8(a). We would like to assign higher weight to the loss with small target value (shown in Fig.5.8(b)). The weight is computed as in Eq.5.6:

$$weight_i = \frac{Max(|target|)}{|target_i|} \quad (5.6)$$

If the absolute value of $target_i$ is 10^9 times smaller than the maximum value among the absolute values in $target$, its weight will be 10^9 . This weight value for extremely small value could so large that trap the optimizer into some bad local minimal. Therefore, we apply a hand-tuned “max loss scale” to limit the the maximum scale of the weight (shown in Fig.5.8(c)). The scaled weight is computed using Eq.5.7:

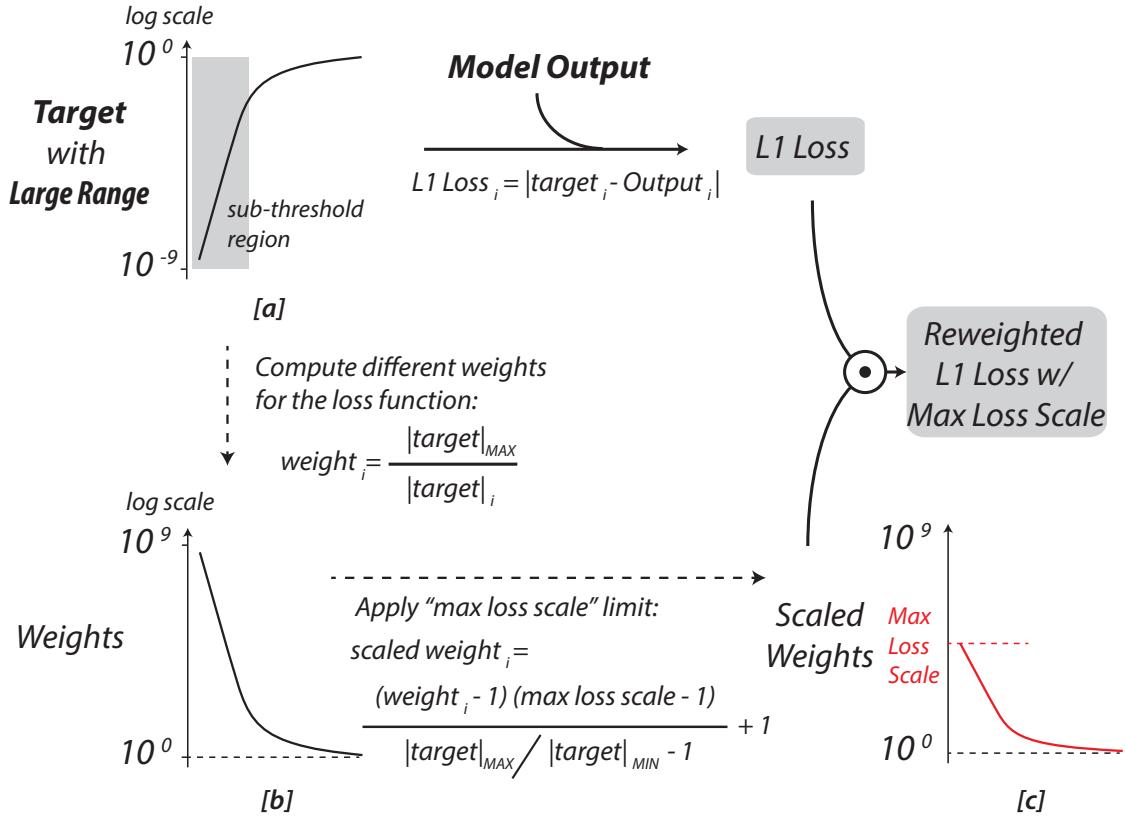


Figure 5.8: Construction of the weighted L1 loss with max scale loss limit.

$$\text{scaled weight}_i = (\text{weight}_i - 1) \frac{\text{max loss scale} - 1}{\text{Max}(\text{weight}_i) - 1} + 1 \quad (5.7)$$

where the $\text{Max}(\text{weight}_i) = \text{Max}(|\text{target}|) / \text{Min}|\text{target}|$. Finally we compute the entrywise product of the L1 loss and the scaled weight to get the weighted L1 loss function:

$$\text{weighted L1 loss}_i = \text{scaled weight}_i \odot |\text{target}_i - \text{output}_i| \quad (5.8)$$

Since L1 loss is a non-smooth function, it is considered to be “unstable” because it tends to “jump around” the solution. This “instability” property is due to that the gradient of L1 loss stays constant no matter how close to the solu-

tion. In practice, we find this instability property actually helps to prevent the optimizer getting stuck in the sub-optimal local minima. On the other hand, in order to achieve a stable solution, we use AdaGrad algorithm,²⁶ where its learning rate increases with accumulation of the squared gradients. Therefore the final solution will be stable due to decreasing learning rate.

In order to evaluation the performance with different *max loss scale*, we use two metrics: the first one is the “weighted L1 metric”. Unlike the weighted L1 loss used for training, there is no “max loss scale” applied to weighted L1 metric, namely defined in Eq.5.9:

$$\text{weighted L1 metric}_i = \text{weight}_i \odot |\text{target}_i - \text{output}_i| \quad (5.9)$$

Weighted L1 metric provides an consistent measure of model performance for different max loss scale values. Since the very small target values in the training data receive very large weight, weighted L1 metric is usually dominated by the errors coming from the targets with small values.

5.6 Experiments

5.6.1 Modeling of GaN HEMT

To demonstrate the ability of Pi-NN to accurately model I-V characteristics, we trained Pi-NN on the experimental measurement data of a promising high-power, high frequency device: Gallium Nitride (GaN) HEMT. The details of the device is discussed by Schuette et al.¹⁰ Its I-V characteristics are plotted in

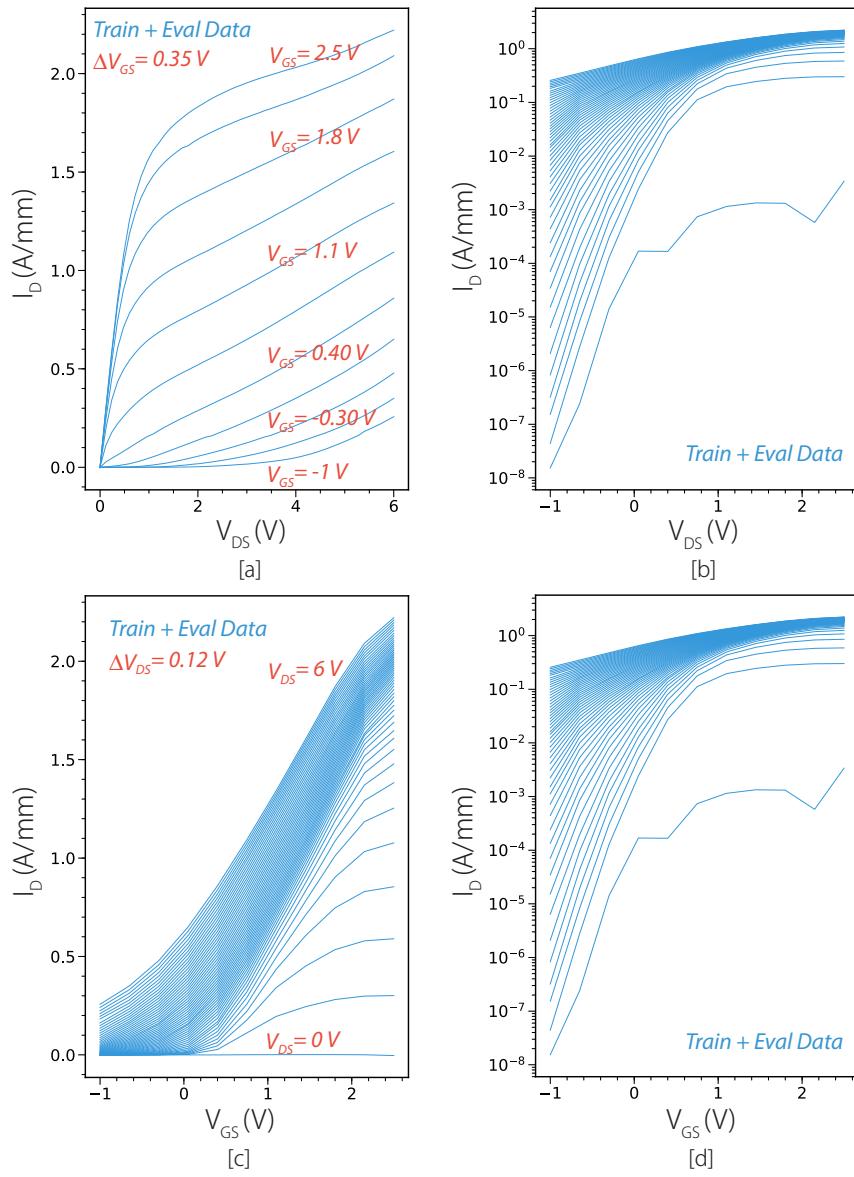


Figure 5.9: The I-V characteristics of a GaN HEMT: (a) I_{DS} versus V_{DS} at different V_{GS} in the linear scale and (b) in the log scale; (c) I_{DS} versus V_{GS} at different V_{DS} in the linear scale and (d) in the log scale.

Fig.5.9.

Throughout this experiment, We randomly left 20% of data as the evaluation set. We also fixed the Pi-NN structure to have two layers of Pi-NN blocks as

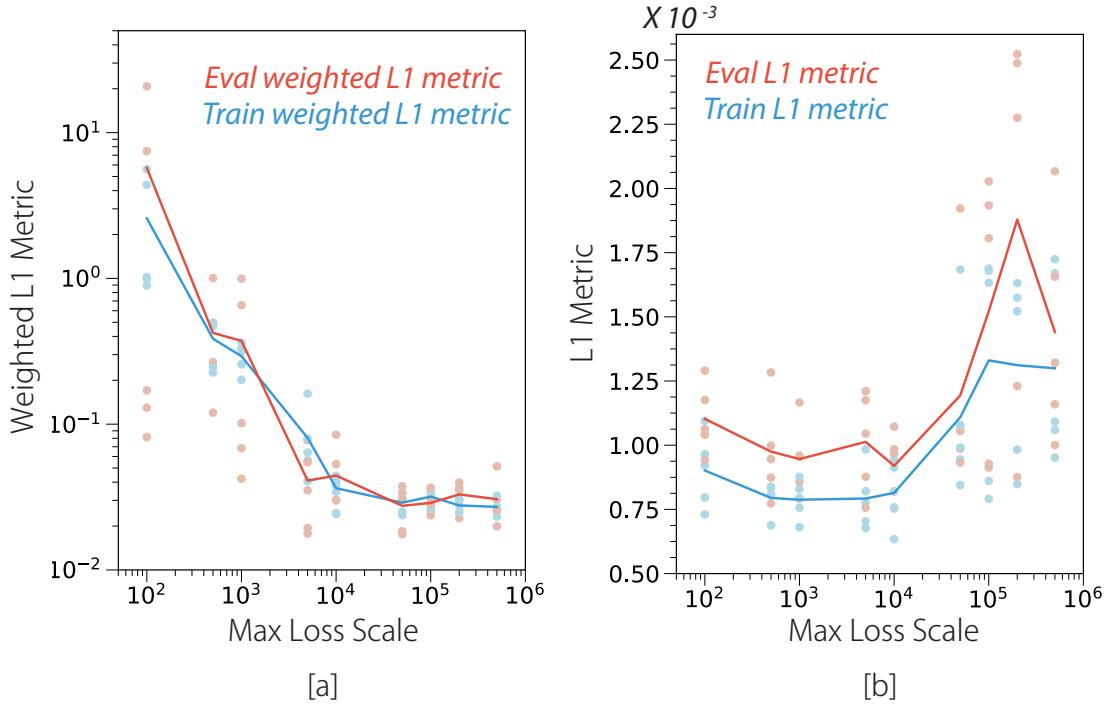


Figure 5.10: (a) The weighted L1 metric versus max loss scale. Each red circle represents the training weighted L1 metric, and each blue circle represents the evaluation weighted L1 metric. The blue and red line are the average value of each runs. (b) The L1 metric versus max loss scale. Each red circle represents the training L1 metric, and each blue circle represents the evaluation L1 metric. The blue and red line are the average value of each runs.

shown in Fig.5.4. Both the FC layers in the first Pi-NN block have 16 neurons and both the FC layers in the second Pi-NN block have 1 neuron. To select the max scale loss and base learning rate for optimization, we first fixed the base learning rate of AdGrad to 0.1 (with $\epsilon = 0.0001$), and varied the max scale loss from 10^2 to 5×10^5 . Since the stochastic nature of the optimization, for each max scale loss value, we repeated the training 5 times, and each model were trained for 10^6 epochs.

When increasing max loss scale, the weighted L1 metric value decreases as shown in Fig.5.10(a), while the L1 metric value increases as shown in Fig.5.10(b). As discussed at the end of Section 5.5, lower weighted L1 metrics indicates bet-

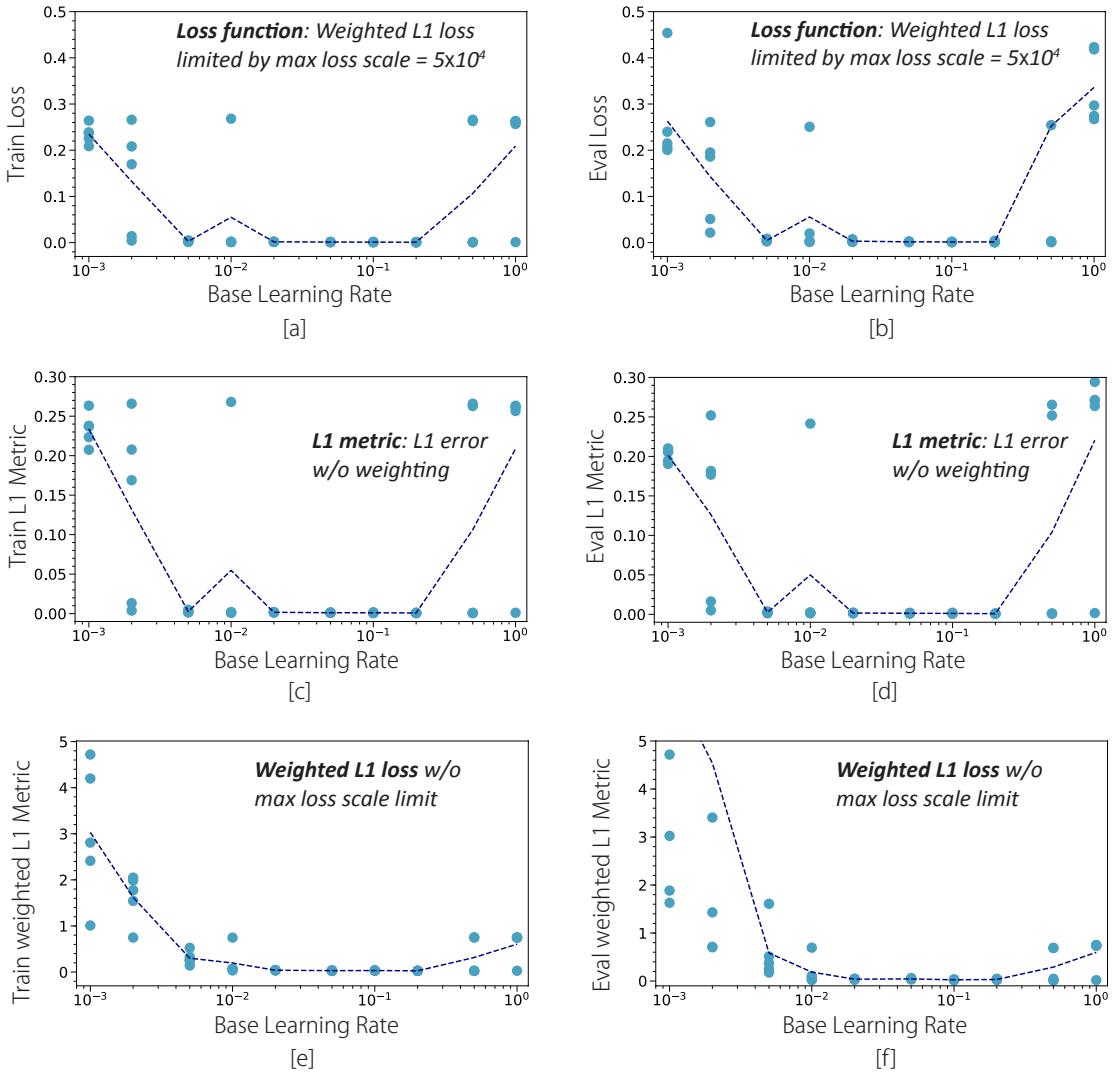


Figure 5.11: Each blue circuit represents one run, and the dash line is the average value of multiple runs. (a) The train weighted L1 losses (with max loss scale limit) versus different base learning rates; (b) The evaluation weighted L1 losses (with max loss scale limit) versus different base learning rates; (c) The train L1 metric versus different base learning rates; (b) The evaluation L1 metric versus different base learning rates; (a) The train weighted L1 metric (without max loss scale limit) versus different base learning rates; (b) The evaluation weighted L1 metric (without max loss scale limit) versus different base learning rates.

ter small value region accuracy, and lower L1 metrics indicates better large value region accuracy. Therefore, we picked 5×10^4 as our max loss scale. After determining the max loss scale, we varied the base learning rate shown in Fig.5.11.

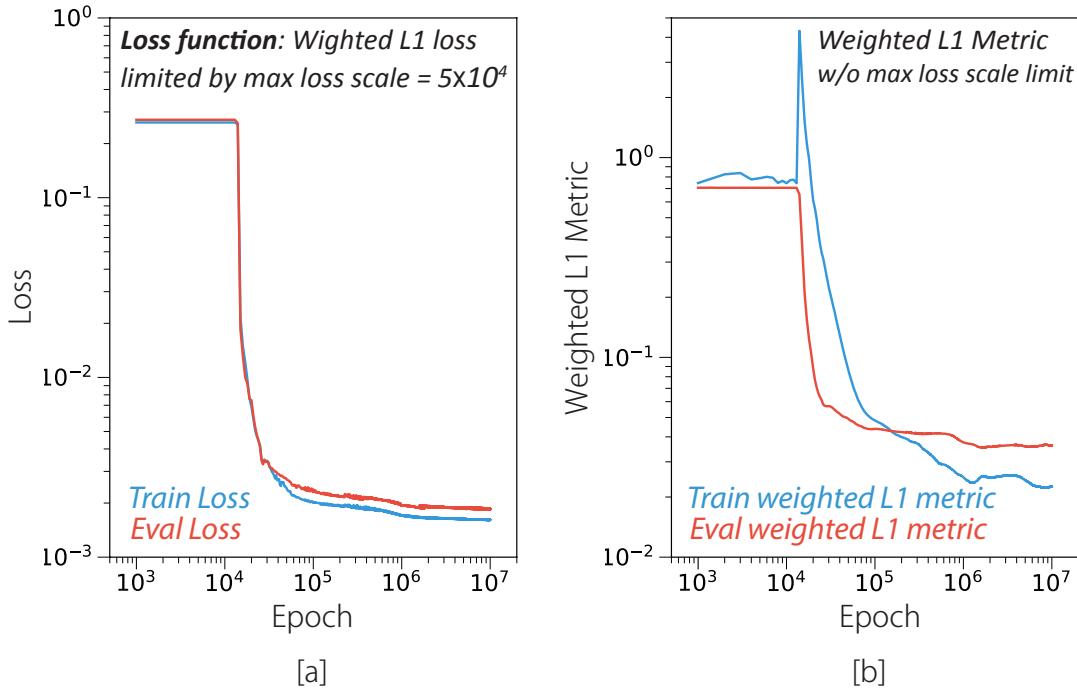


Figure 5.12: (a) The training/evaluation loss (weighted L1 loss with max loss scale limit) versus epochs; (b) the training/evaluation weighted L1 metric (without max loss scale limit) versus epochs.

Judged Fig.5.11, if the base learning is either too big or too small, the model performance becomes bad and unstable. We chose the base learning rate equals 0.1 since it was able to achieve good and stable optimization result in both weighted L1 metrics and L1 metrics.

With max loss scale equals 5×10^4 and base learning rate equals 0.1, we plotted the weighted L1 loss (with max loss scale limit) and weighted L1 metric (without max loss scale limit) at different epochs in Fig.5.12.

As shown in Fig.5.12(a), the optimizer first got stuck in a local minima before found its way to further optimize the model. The evaluation loss stays close with the train loss, which indicate good generalization. Figure 5.12(b) plots the weighted L1 metric without max loss scale limits. It is worth to note that the

training weighted L1 metric had a huge spike before decreasing. It means, if we would train with the weighted L1 metric without the max scale loss, we would never be able to get out of the local minima. This observation proves the necessity of having the max scale loss limit on the weighted L1 loss.

The model output along with the training and evaluation data are shown in Fig.5.13. As shown, the model has excellent agreement with both the training and evaluation data. Note that this GaN HEMT has a relatively severe DIBL effect. The threshold voltage is also controlled by the drain voltage. So at small V_{GS} , the I_{DS} is not only exponentially dependent on V_{GS} , but also exponentially dependent on V_{DS} . This DIBL effect is challenging to model with traditional hand-crafted compact model. Pi-NN, on the other hand, is able to model this complicate dependence very well.

The interpolation and extrapolation abilities of a device model are also crucial. In Fig.5.14, When extending V_{DS} to 80% beyond the training V_{DS} range and V_{TG} to +/- 40% beyond the training V_{GS} range, the model has shown no abnormal behavior.

Another important metric of a GaN HEMT model is monotonicity first-order derivative. An abnormal oscillation in the device model will lead to artificial high-frequency oscillations in the circuit simulation. We used the adjoint network introduced in Section 5.4 to plot out both the transconductance and output conductance (shown in Fig.5.15).

Both the transconductance and output conductance are positive and smooth throughout the whole voltage range. The red arrow line in Fig.5.15(a) indicates the peak transconductance voltage shifts at different drain voltages, which can

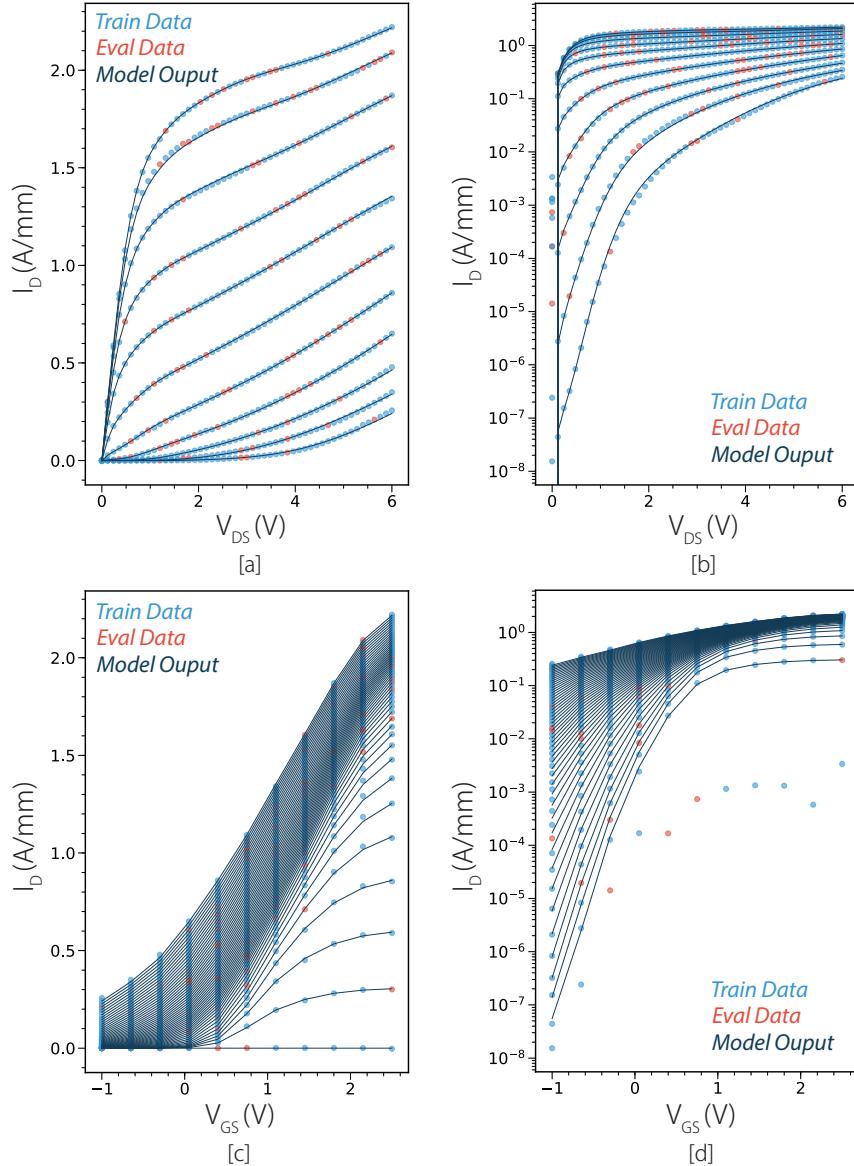


Figure 5.13: The I-V curves generated by the Pi-NN model are plotted along with the training data (blue circles) and the evaluation data (red circles): (a) I_{DS} versus V_{DS} at different V_{GS} in the linear scale and (b) in the log scale; (c) I_{DS} versus V_{GS} at different V_{DS} in the linear scale and (d) in the log scale.

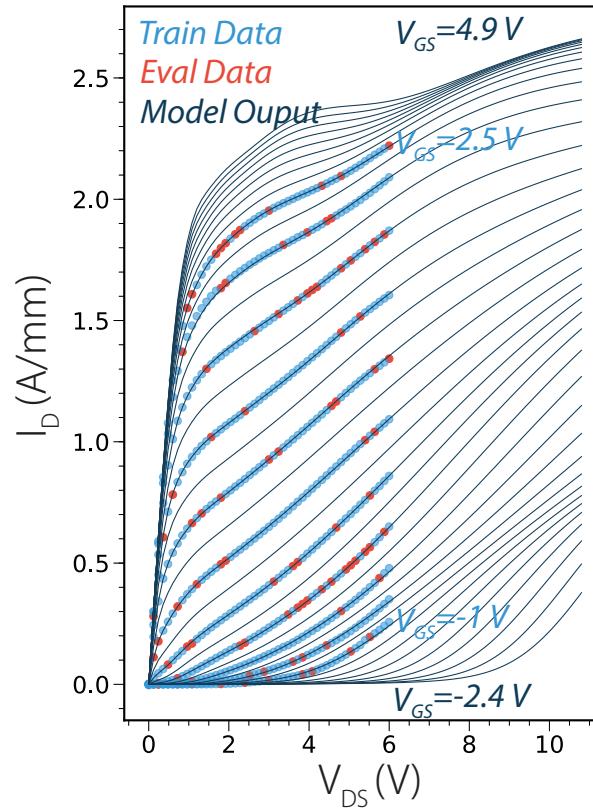


Figure 5.14: The I-V curves generated by the Pi-NN model are plotted along with the training data (blue circles) and the evaluation data (red circles) for I_{DS} versus V_{DS} at different V_{GS} in the linear scale. V_{DS} for the model are extended to 80% beyond the training V_{DS} range and V_{GS} extended to +/- 40% beyond the training V_{GS} range.

be explained by the combination of self-heating effect and DIBL effect. Overall, Pi-NN was able to generate accurate, smooth, and computation efficient device compact model.

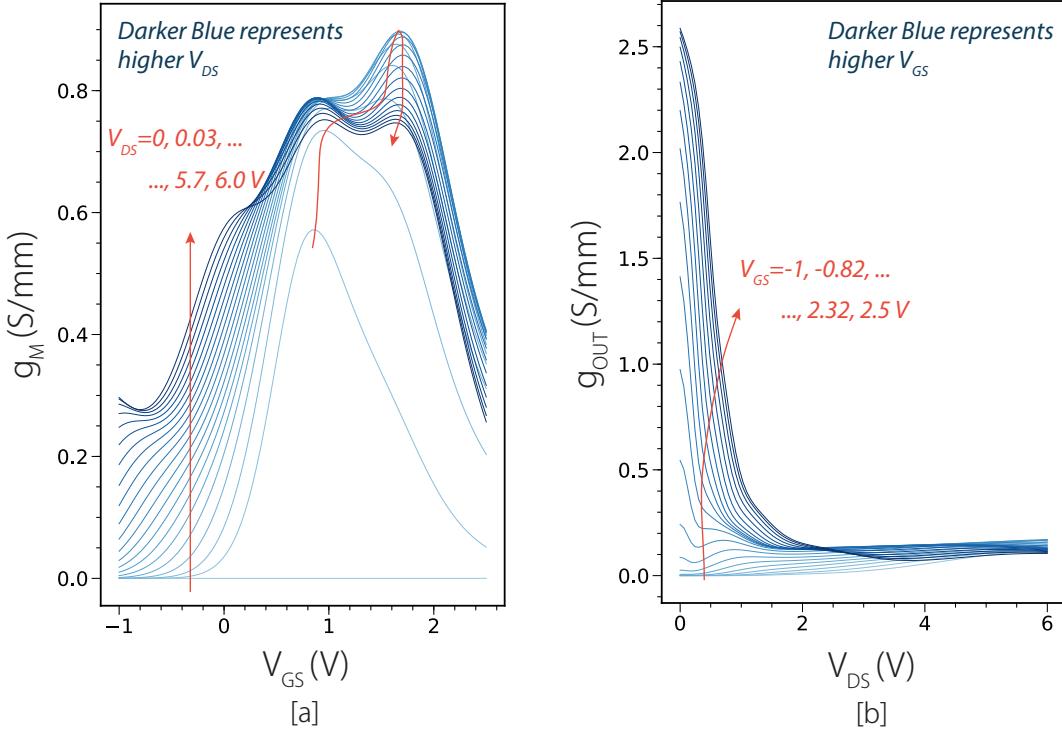


Figure 5.15: (a) The partial derivatives of the drain current with respect to the gate voltage (transconductance) versus gate voltage at different drain voltages; (b) the partial derivatives of the drain current with respect to the drain voltage (output conductance) versus the drain voltage at different gate voltages. The red arrow line in (a) indicates the peak transconductance voltage shifts at different drain voltages, which can be explained by the combination of self-heating effect and DIBL effect.

5.6.2 Modeling of Thin-TFET

We also tested the Pi-NN model on the Thin-TFETs simulation data. After initial training, we chose to use Pi-NNs with one hidden layer and define the hyper-parameter as (m, n) , where m is the number of the \tanh neurons in the hidden layer and n is the number of the sigmoid neurons in the same hidden layer. The test errors stay close to the training errors as shown in Fig. (a), which indicates good generalization. Balancing between model complexity and accuracy, we

chose the model with the hyper-parameter (2, 3) as shown in Fig.5.16(b), which give a small Pi-NN model with only 7 neurons and 20 parameters in total. Excellent modeling is demonstrated in both the ON region (shown in Fig. 6(c, d)) and the sub-threshold region (shown in Fig. 6(f)). The I_{DS} - V_{DS} relationship around V_{DS} equals zero is shown in Fig.5.16(e). All the unphysical behaviors that appeared in the MLP neural network model (shown in Fig.5.3) have been eliminated. Moreover, thanks to the embedded device physics, the Pi-NN requires much less parameters than the MLP neural network, which results in a smaller, more efficient compact model.

5.7 Conclusion

Motivated by the need of high-quality compact models for emerging devices, we have proposed a novel neural network: Pi-NN, along with weighted L1 loss function for device compact modeling. With fundamental device physics incorporated, the Pi-NN method can produce accurate, smooth and computational efficient transistor models with good generalization ability. GaN HEMT and Thin-TFET are presented as examples to illustrate the capabilities of Pi-NN. The adjoint network of Pi-NN have also been developed to model the differential information in the device measurements. Finally, the Pi-NN framework has been implemented in Caffe2, which can be readily integrated on commercial measurement and modeling systems.

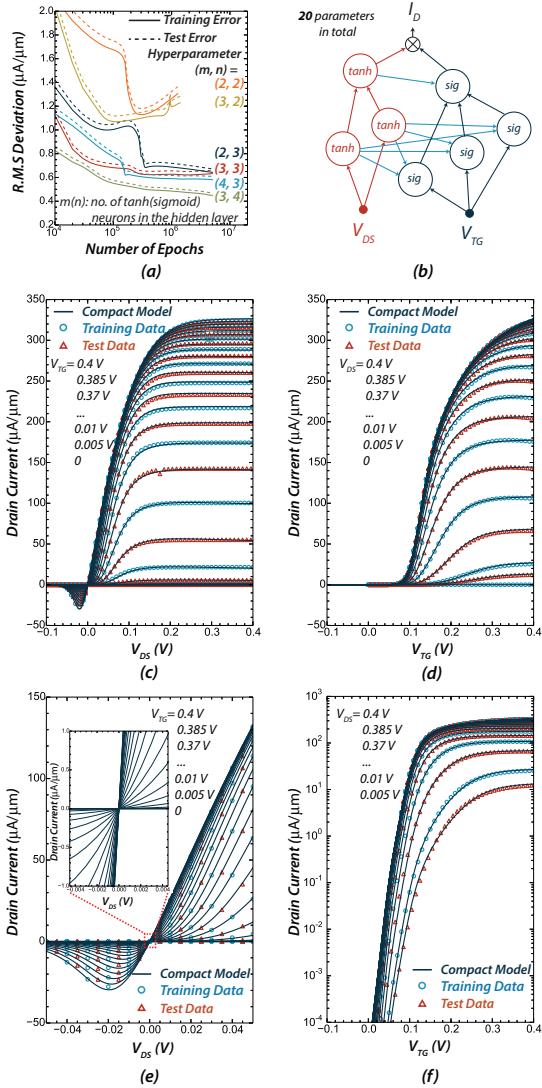


Figure 5.16: For the Pi-NN developed in this work, (a) the training errors and test errors for a variety of hyper-parameters. (b) the Pi-NN model with 2 *tanh* neurons and 3 *sigmoid* neurons in the hidden layer. From (c) to (f), the I-V curves generated by the Pi-NN model shown in (b) are plotted along with the training data and the test data: (c) I_{DS} versus V_{DS} at different V_{TG} ; (d) I_{DS} vs. V_{TG} at different V_{DS} in linear scale; (e) I_{DS} vs. V_{DS} at different V_{TG} around $V_{DS} = 0$, the embedded plot shows well-behaved I_{DS} - V_{DS} relationship around $V_{DS} = 0$; (f) I_{DS} vs. V_{TG} at different V_{DS} in semi-log scale, good fitting is achieved in the sub-threshold region. All the unphysical behaviors of the MLP neural network (shown in Fig.5.3) are eliminated, and the size of the neural network is largely reduced.

BIBLIOGRAPHY

- ¹ Y. Tsividis and C. McAndrew, *Operation and Modeling of the MOS Transistor*. Oxford Univ. Press, 2011.
- ² Y. S. Chauhan, S. Venugopalan, M. A. Karim, S. Khandelwal, N. Paydavosi, P. Thakur, A. M. Niknejad, and C. C. Hu, "Bsimindustry standard compact mosfet models," in *ESSCIRC (ESSCIRC), 2012 Proceedings of the*. IEEE, 2012, pp. 30–33.
- ³ J. Xu and D. E. Root, "Advances in artificial neural network models of active devices," in *Numerical Electromagnetic and Multiphysics Modeling and Optimization (NEMO), 2015 IEEE MTT-S International Conference on*. IEEE, 2015, pp. 1–3.
- ⁴ H. B. Hammouda, M. Mhiri, Z. Gafsi, and K. Besbes, "Neural-based models of semiconductor devices for spice simulator," *American Journal of Applied Sciences*, vol. 5, no. 4, pp. 385–391, 2008.
- ⁵ F. Wang and Q.-J. Zhang, "Knowledge-based neural models for microwave design," *IEEE Transactions on Microwave Theory and Techniques*, vol. 45, no. 12, pp. 2333–2343, 1997.
- ⁶ L. Zhang and M. Chan, "Artificial neural network design for compact modeling of generic transistors," *Journal of Computational Electronics*, pp. 1–8, 2017.
- ⁷ M. Li, O. İrsoy, C. Cardie, and H. G. Xing, "Physics-inspired neural networks for efficient device compact modeling," *IEEE Journal on Exploratory Solid-State Computational Devices and Circuits*, vol. 2, pp. 44–49, 2016.
- ⁸ A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.
- ⁹ G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.
- ¹⁰ M. L. Schuette, A. Ketterson, B. Song, E. Beam, T.-M. Chou, M. Pilla, H.-Q. Tserng, X. Gao, S. Guo, P. J. Fay *et al.*, "Gate-recessed integrated e/d gan hemt

- technology with $f_t/f \max; 300$ ghz," *IEEE Electron Device Letters*, vol. 34, no. 6, pp. 741–743, 2013.
- ¹¹ M. O. Li, D. Esseni, J. J. Nahas, D. Jena, and H. G. Xing, "Two-dimensional heterojunction interlayer tunneling field effect transistors (thin-tfets)," *IEEE Journal of the Electron Devices Society*, vol. 3, no. 3, pp. 200–207, 2015.
- ¹² Q.-j. Zhang and K. C. Gupta, *Neural networks for RF and microwave design (Book+ Neuromodeler Disk)*. Artech House, Inc., 2000.
- ¹³ Q.-J. Zhang, K. C. Gupta, and V. K. Devabhaktuni, "Artificial neural networks for rf and microwave design-from theory to practice," *IEEE transactions on microwave theory and techniques*, vol. 51, no. 4, pp. 1339–1350, 2003.
- ¹⁴ J. Xu, D. Gunyan, M. Iwamoto, A. Cognata, and D. E. Root, "Measurement-based non-quasi-static large-signal fet model using artificial neural networks," in *Microwave Symposium Digest, 2006. IEEE MTT-S International*. IEEE, 2006, pp. 469–472.
- ¹⁵ D. E. Root, J. Xu, J. Horn, and M. Iwamoto, "The large-signal model: theoretical foundations, practical considerations, and recent trends," *Nonlinear Transistor Model Parameter Extraction Technique*, pp. 123–170, 2011.
- ¹⁶ D. E. Root, "Future device modeling trends," *IEEE Microwave Magazine*, vol. 13, no. 7, pp. 45–59, 2012.
- ¹⁷ Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.
- ¹⁸ J. L. Elman, "Finding structure in time," *Cognitive science*, vol. 14, no. 2, pp. 179–211, 1990.
- ¹⁹ S. Hashem, "Sensitivity analysis for feedforward artificial neural networks with differentiable activation functions," in *Neural Networks, 1992. IJCNN., International Joint Conference on*, vol. 1. IEEE, 1992, pp. 419–424.
- ²⁰ L. Fu and T. Chen, "Sensitivity analysis for input vector in multilayer feedforward neural networks," in *Neural Networks, 1993., IEEE International Conference on*. IEEE, 1993, pp. 215–218.

- ²¹J. D. Olden and D. A. Jackson, “Illuminating the black box: a randomization approach for understanding variable contributions in artificial neural networks,” *Ecological modelling*, vol. 154, no. 1, pp. 135–150, 2002.
- ²²K. Simonyan, A. Vedaldi, and A. Zisserman, “Deep inside convolutional networks: Visualising image classification models and saliency maps,” *arXiv preprint arXiv:1312.6034*, 2013.
- ²³W. Samek, A. Binder, G. Montavon, S. Lapuschkin, and K.-R. Müller, “Evaluating the visualization of what a deep neural network has learned,” *IEEE transactions on neural networks and learning systems*, 2017.
- ²⁴A. Nguyen, J. Yosinski, and J. Clune, “Deep neural networks are easily fooled: High confidence predictions for unrecognizable images,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 427–436.
- ²⁵J. Xu, M. C. Yagoub, R. Ding, and Q. J. Zhang, “Exact adjoint sensitivity analysis for neural-based microwave modeling and design,” *IEEE Transactions on Microwave Theory and Techniques*, vol. 51, no. 1, pp. 226–237, 2003.
- ²⁶J. Duchi, E. Hazan, and Y. Singer, “Adaptive subgradient methods for online learning and stochastic optimization,” *Journal of Machine Learning Research*, vol. 12, no. Jul, pp. 2121–2159, 2011.

CHAPTER 6

FUTURE WORKS

6.1 Non-ideal effects in Thin-TFETs

Many non-ideal effects in Thin-TFETs need to be studied. In experiments, it is usually hard to achieve monolayer top and bottom 2D materials. Therefore it is important to study the effect of the top and bottom 2D layer thickness. We found if there is no chemical doping in both the top and bottom 2D layers, increasing layer thickness of either top layer or bottom layer will result in less steep subthreshold slope. Chemical doping in the bottom layer can prevent subthreshold slope degradation even with increasing bottom layer thickness.

The permittivity of the van der Waals gap directly affects the gate efficiency of Thin-TFETs. Therefore higher permittivity leads to less steep subthreshold slope and lower ON current. The interfacial trap density (D_{it}) also has significant impact on Thin-TFET. Since Thin-TFET is made of layers of 2D materials, it is also interesting to study which location of D_{it} has the most profound influence on the device performance. When D_{it} is unavoidable, we could design a D_{it} tolerant device structure to move critical regions away from D_{it} . Moreover, trap-assist tunneling (TAT) and ShockleyReadHall (SRH) recombination are known to limit the TFETs' performance. The in-depth studies of TAT and SRH recombination in Thin-TFETs are still undergoing efforts.

In the experiments, the access region between the channel to the contacts often plays a important role in Thin-TFET performance. As shown in the optical image in Fig.6.1,¹ the access region between the overlapping SnSe₂/WSe₂

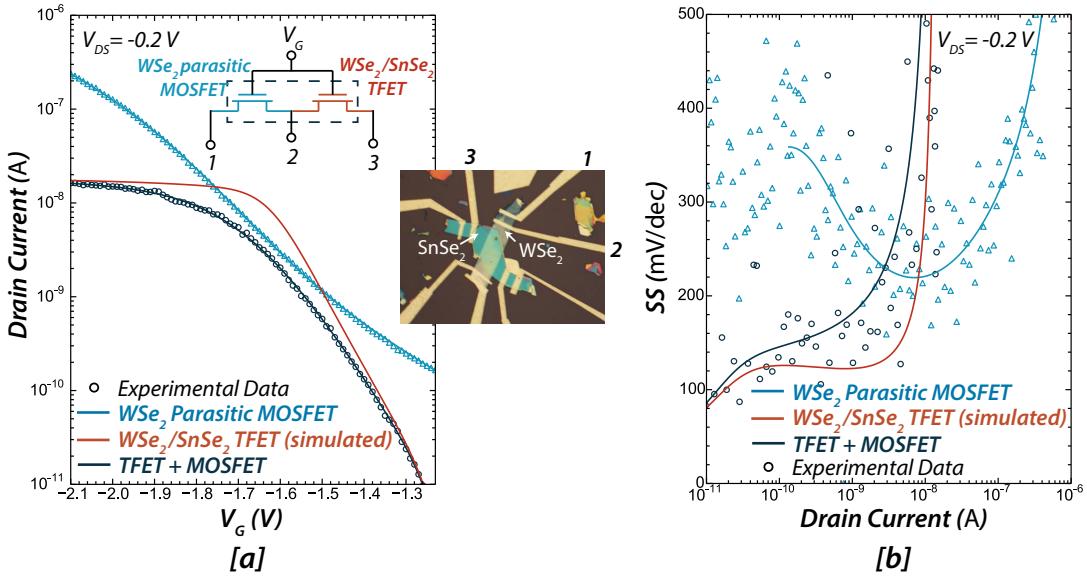


Figure 6.1: (a) I_D - V_G curves of the measured WSe_2 parasitic MOSFET, the $\text{WSe}_2/\text{SnSe}_2$ Thin-TFET (TFET + MOSFET), and the intrinsic $\text{WSe}_2/\text{SnSe}_2$ TFET, the insets show the optical image of the device and the equivalent circuit with the parasitic MOSFET; (b) the corresponding SS curves for the parasitic MOSFET, the $\text{WSe}_2/\text{SnSe}_2$ Thin-TFET (TFET + MOSFET), and the intrinsic TFET.

junction and WSe_2 contacts becomes a lateral parasitic WSe_2 MOSFET. This parasitic MOSFET will limit the subthreshold steepness of Thin-TFET (as shown in Fig.6.1(a-b)). To eliminate this parasitic MOSFET, the access region has to be heavily doped either chemically or electrostatically.

Which material systems are the best to realize Thin-TFET also remains an open question. Black-phosphorous, due to its small electron affinity, can be excellent *p*-type layer in Thin-TFETs. Incorporating Black-phosphorous in Thin-TFET is under active researches in our group.

6.2 Experimental Demonstration of TransiXNOR

In order to realize TransiXNOR, the channel material is the key. Unlike normal TFETs, where the tunneling junction is only between the source and channel, both the source/channel junction and the channel/drain junction will be served as tunnel junctions. When the tunnel junction is at the source and channel interface, the electrons tunnel from the source valence band to the channel conduction band; when the tunnel junction is at the channel and drain interface, the electrons tunnel from the channel valence band to the drain conduction band. Therefore, both the conduction and valence band of the channel material involve in the tunneling process in TransiXNOR. The first requirement of the channel material is the bandgap. The bandgap of the channel material has to be larger than V_{DD} in order to have low leakage in the OFF state, and the bandgap has to be close to V_{DD} in order to have a high ON current.

The second requirement of the channel material is the thickness. For TransiXNOR to work, the two gates of TransiXNOR have to control the same channel. If the two gates controlled their individual channels under the gates, the device will behave like the two TFETs in parallel, thus the XNOR behavior cannot be achieved. On the other hand, for 3D materials, scaling down to ultra-thin body will increase its bandgap due to the quantization effect. Using 2D layered materials can achieve atomically thin bodies with reasonable bandgaps. However, finding the right 2D layered materials with the suitable bandgap, doping the source/drain region in 2D layer materials, and integrating 2D layers materials in CMOS compatible system are all challenging. One possibility is to use 2D layered materials in the channel region, and use 3D materials in the source/drain region. Therefore, not only doping the source/drain region is no

longer a problem, but we can design staggered/broken band alignments in both the source/channel and channel/drain junction to boost the ON current. However, how to build edge contact 3D/2D junction is still under research.²

6.3 Adjoint Network as Regularization in Pi-NN

Due to its structured architecture and embedded device physics, Pi-NN has shown very good generalization capability without explicit regularization. However, for device modeling, we sometimes want to enforce a certain set of rules beyond the region that training data are available. For example, we would like to enforce the model to be monotonic increasing in the region that experimental measurement are either not available or impossible. Co-training with the adjoint network gives us access to the output sensitivities with respect to the inputs. Therefore, we can add a regularization term using the output sensitivities to penalize negative first derivative. Moreover, the adjoint network also gives us access to the sensitivity of each activation in the model. Since the activation contributing little to the final output may cause small oscillations,³ we could prune out the corresponding neurons during training.

Pi-NN framework is readily applicable to TransiXNOR and other emerging devices. Initial integration of Pi-NN model and the circuit simulator has started. The inputs from circuit/system designers will help Pi-NN become the potential new paradigm of device compact modeling.

BIBLIOGRAPHY

- ¹ M. O. Li, R. Yan, D. Jena, and H. G. Xing, “Two-dimensional heterojunction interlayer tunnel fet (thin-tfet): From theory to applications,” in *Electron Devices Meeting (IEDM), 2016 IEEE International*. IEEE, 2016, pp. 19–2.
- ² A. Allain, J. Kang, K. Banerjee, and A. Kis, “Electrical contacts to two-dimensional semiconductors,” *Nature materials*, vol. 14, no. 12, pp. 1195–1205, 2015.
- ³ F. Girosi, M. Jones, and T. Poggio, “Regularization theory and neural networks architectures,” *Neural computation*, vol. 7, no. 2, pp. 219–269, 1995.