Integrated Phase-Change Optoelectronics for Energy-Efficient Computing

Yuhan He

Corpus Christi College

University of Oxford

Supervisor: Prof. Harish Bhaskaran

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

Doctor of Philosophy

Trinity 2025

This thesis is dedicated to
my father and mother
for their continued love and support

Acknowledgements

Earning a DPhil is a journey that feels both too long and too short. Looking back, there have been so many small, life-changing moments that shaped who I am today. Time and again, I realize that we are too arrogant to admit how little we truly know about the world and ourselves.

First of all, I would like to express my deepest thanks to my supervisor Prof. Harish Bhaskaran for accepting me twice to join the group. I don't know how he managed to build such a family-like group with diversity and autonomy. His own curiosity and drive for lifelong learning and his insights about creativity and decision making have largely changed my view of the world and life. He has shown me the power of clarity, precision and simplicity in academic communication, and has taught me not only to excel in my field, but also to stay true to my passions, to take initiative and to embrace responsibility with confidence. I have seen how my potential has been unlocked during the past four years, which is not possible without all his trust, support, and sincere guidance.

I would like to thank my industrial supervisor, Dr. Francesca Parmigiani, for her unsparing support. Without the Microsoft studentship and the rare extension award, I would not afford to study at Oxford. The precious opportunity to visit Microsoft Research Cambridge offered me insight from an industry perspective.

I would like to thank Nikolaos Farmakidis, who is my co-supervisor within the group. I learned most of my experiments and simulation skills from him, which he taught me hand in hand. He has set a strong example for how to explore my curiosity, stay organized, and manage my priorities.

I would like to thank Mengyun Wang for her endless care, support, and guidance from the first day I arrived at Oxford. She is always a good mentor and has taught me to be professional while taking care of my emotions. I would like to thank Samarth Aggarwal, who guided me and took me under his wings from the beginning. And I would like to thank June Sang Lee for always being patient to listen to my small ideas and thoughts, and coaching me on how to design experiments thoughtfully and effectively. I also would like to thank Dr. Bowei Dong, the deep thinker who inspired so many insightful discussions.

Thanks to all the past and present group members of the ANE group: Prof. Zengguang Cheng, who introduced and guided me to the group; Xuan Li, for all the long chats about life, research and career; Yi Zhang, for all the moments we fight alongside one another to grow up as independent researchers; Prof. Carlos Rios Ocampo, "Things take time" has been healing me for a long time; and Yu Shu, Mingde Du, Guoce Yang, Utku Emre Ali, Wen Zhou, Angel Ortega, Håvard Hem Toftevaag, Kairan Huang, Jock Dearlove, Tara Milne, Maxine Ong, Sijing Zhong, Serena Sabnani, Sharon Oregel-Chaumont, Peiman Hosseini, Johannes Feldmann,

Nathan Youngblood, James Tan, Eugene Soh, Sirui Liu, Zhiyun Xu, Zhongyu Tang, Barry Hao Yu, Yuhang Lee, Tangsheng Cheng, Ruveyda Taşkın, Sergei Zotkin, Antonios Mikallou, Andy Moskalenko and Luci Bywater, for all the support, help, and memorable group moments.

I also want to thank my college, Corpus Christi Collge, for providing a supportive academic environment and generous financial assistance throughout my studies, especially during the pandemic period, with special thanks to the support from my college advisor Prof. Peter Nellist and Dr. Christopher Patrick, academic registrar Rachel Clifford and Aoife Walsh.

Thanks to all my friends at Oxford: Yumeng Yin, for her daily support, encouragement and treasurable piano days; Junjie Liu, for his selfless and patient sharing of knowledge and guidance on cleanroom practices and research experience; and Yang Lu, Xinya Niu, Ruiying Shu, Minyi Zhang, Yitao Zhu, Jin Xu, Jiale Fu, Zhixin Chen, Lina Chen, Kun Peng, Biyang Wang, to name a few, for all the academic discussions and valuable time spent in badminton and cooking. I would like to thank the people behind the scene who provided support for my research, Paul Pattinson and Radka Chakalova for cleanroom maintenance, and Paul Warren for IT support. I also would like to thank the RisingWISE program for inspiring me through networking with enterprising women in industry and academia.

Special thanks to Prof. Huanglong Li, my undergraduate mentor, for his constant help and support. He introduced me to the world of academic research and always inspired me to do genuine, meaningful research. Thanks to Prof. Huaqiang Wu, who guided me to understand the industry side of the research and the importance of communication in research. I also would like to thank Prof. Xudong Huang, Prof. Kaiyu Cui, and Prof. Yang Li, who introduced me to the world of photonics.

I would like to thank my old friends in China and the US, especially Yuanmuliu Hu, Yaxin Xiao, Wanwei Ren, and Qian Lin, for all their love and support and for shining in their own lives and career paths, encouraging me along the way. I also would like to thank the Grad Lounge podcast for its insightful sharing on academic life and the Jungian Psychological Types framework for supporting my self-exploration. They all inspired me to walk through the dark moments.

Special thanks to my partner Fengxiang Bai, the gift of my Oxford life, who completed me and taught me self-compassion. I treasure all our discussion about life and research, all the delicious meals we cooked together and all the happy or bitter moments during our exploration of the world.

Finally, I would like to thank my parents for their endless love and support. And special thanks to myself, for never losing hope in engaging in life and always fighting back to rebuild my life.

Yuhan He
Oxford, May 30 2025

Abstract

Digital electronics have dominated modern computing for decades. However, the emergence of artificial intelligence applications drives a growing demand for enhanced computing performance, and motivates a shift toward a new design paradigm to tackle the data-transferring bottleneck in conventional computing architecture. In-memory computing, featuring co-located memory and processing units, stands out as a compelling solution, with non-volatile devices serving as critical components.

Meanwhile, the advancement of silicon photonics has led to renewed attention on photonic computing. Combining the high-bandwidth and low-latency advantages of photonics with the maturity of electronics, the electro-optical in-memory computing architecture opens up new opportunities for future energy-efficient computing systems. The key challenge lies in developing energy-efficient, scalable non-volatile optoelectronics devices, while circumventing the intrinsic size mismatching problem between photonics and electronics.

The content of the thesis encompasses the design, simulations, fabrication process, experimental measurements, and analysis results of the development of phase-change optoelectronic devices, the core representative of non-volatile optoelectronics. Close attention is devoted to devices based on silicon-on-insulator (SOI) platforms, given their excellent integration potential with CMOS-based electronics.

The first part of the thesis is devoted to the development of plasmonic phase-change devices on a SOI platform. With an optimized corrugated grating design to reduce insertion loss, such devices have achieved a more than three-fold improvement in coupling efficiency compared to their Si_3N_4 counterparts. A self-aligned fabrication process is further exploited to mitigate the alignment challenge. Moreover, binary and multilevel programmability have been experimentally demonstrated in these devices with an ultra-low minimum optical switching energy of 3.82 pJ.

Next, the implementation of non-plasmonic integrated electro-optic phase-change devices is presented. By structuring the phase-change material as a nanoscale constriction, this design allows programmability and readability of phase-change materials in both electrical and optical domains without placing constraints on the metallic layers, thus revealing significant potential for CMOS-compatible scaling. Phase-change constriction devices have demonstrated sub-10 pJ electrical switching energy and a high electro-optical modulation efficiency of 0.15 nJ/dB. Further, in-plane photosensitivity of the device has been characterized, opening new avenues for constructing energy-efficient electro-optical systems that incorporate both in-memory programming and sensing abilities.

Finally, system integration of electro-optical computing units has been assessed by exploiting a foundry-fabricated integrated photonic chip. The performance of

integrated volatile modulators and photodetectors is compared to that of integrated phase-change material-based non-volatile optoelectronics using experimental characterization results.

Contents

List of Figures

List of Tables

List of Abbreviations

ADC . . . . . Analogue-to-digital converter

AFM . . . . . Atomic force microscope

AI . . . . . Artificial Intelligence

ALD . . . . . Atomic layer deposition

BPD . . . . . Balanced photodetector

CAD . . . . . Computer-Aided Design

CIFAR-10 . . Canadian Institute For Advanced Research-10 database

CMOS . . . . . Complementary metal-oxide-semiconductor

CNN . . . . . Convolutional neural network

CW . . . . . Continuous wave

DAC . . . . . Digital-to-analogue converter

DMD . . . . . Digital micromirror device

EAM . . . . . Electro-absorption modulator

EBL . . . . . Electron beam lithography

EDFA . . . . . Erbium-doped fiber amplifier

EOM . . . . . Electro-optical modulator

FET . . . . . Field-effect Transistor

FLOPS . . . . Floating-point operations per second

FPGA . . . . . Field programmable gate array

GST . . . . . $Ge_2Sb_2Te_5$ phase change alloy

I/O . . . . . Input-Output

ITO . . . . . Indium Tin Oxide

MAC . . . . . Multiply and accumulate

MIM . . . . . Metal-insulator-metal structure

MMI . . . . . Multi-mode interferometer

MNIST . . . . Modified National Institute of Standards and Technology database

MRR . . . . . Microring resonator

MVM . . . . . Matrix-vector multiplication

MZI . . . . . Mach-Zender interferometer

ONN . . . . . Optical neural networks

PD . . . . . Photodetector

PIC . . . . . Photonic integrated circuits

RIE . . . . . Reactive ion etching

sccm . . . . . standard cubic centimeter per minute

SEM . . . . . Scanning electron microscope

Si . . . . . Silicon

SiN . . . . . Silicon Nitride

SiO_2 . . . . . Silicon Dioxide

SLM . . . . . Spatial light modulator

SNN . . . . . Spiking neural networks

SOI . . . . . Silicon on insulator

TIA . . . . . Trans-impedance amplifier

TIR . . . . . total internal reflection

TOPS . . . . . Tera-operations per second

VCSEL . . . . Vertical cavity surface emitting laser

VOA . . . . . Variable optical attenuators

WDM . . . . . Wavelength division multiplexing

1 Introduction

Contents

1.1 Motivation

1.1.1 In-memory Computing: The Trend

Since the first general-purpose digital computer, ENIAC, came out in 1945 [1], the von Neumann architecture [2] has dominated modern computing for several decades. It makes use of data transferring between the central processing unit (CPU) and the memory unit to process input data and generate output results. Thus, the bandwidth and latency of the data transfer impose an upper bound on the achievable computing performance. Billions of electronic unit cells, transistors, have been squeezed onto the chip scale to implement the central processing unit with a minimum feature size close to 10 nm [3], but their speed and energy performance cannot still meet the requirements of the era of "big data" or artificial intelligence (AI) [Figure 1.1]. The approaching scaling-down limit and the intrinsic bottleneck

of von Neumann architecture on data transferring call for a new design paradigm of future computing.

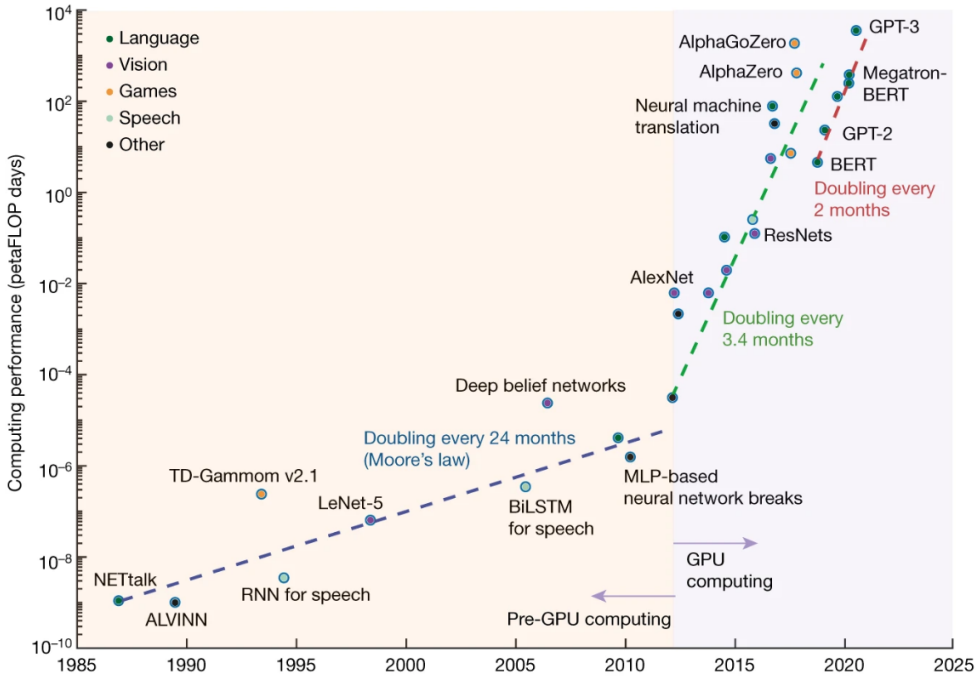

Figure 1.1: Computing power demands over the past four decades. Reprinted from [4].

In-memory computing, featuring co-located memory and processing units, provides a promising way to tackle the data transferring bottleneck. Extensive research has been conducted to build electrical in-memory computing systems[5, 6]. The key components, memristors [7, 8], are a type of memory device that retain their states (e.g. distinct resistance or transmittance levels) after programming without consuming additional energy, rendering them uniquely suited for low-power computing applications.

Given their compatibility with CMOS integration, wafer-scale electrical in-memory computing chips have been demonstrated [9, 10], offering advantages in both computing density and energy efficiency over CPU and GPU [5]. However, the operation speed of electronics (< 5 GHz) are limited by capacitive delays and power dissipation, restricting future performance advancements in computing efficiency and latency (lower-bounded by the inverse of the bandwidth, i.e. 200 ps) [11].

With recent advances in silicon photonics, photonic in-memory computing processors are redefining the boundaries of classical computing [11–13]. Photonics provide more degrees of freedom, such as wavelength [14, 15] and polarization [16], thus enabling high-bandwidth multidimensional information processing within memory. Moreover, given the absence of parasitic capacitance, photonics allow for lower computing latency and thus higher computing speed and energy efficiency [11–13]. However, since the development of integrated all-optical systems remains in its infancy, extra electro-optical conversions are still required for current photonic systems to be compatible with digital general-purpose modern computing systems.

1.1.2 Electro-optical In-memory Computing: The Challenge

Electro-optical in-memory computing architectures [Figure 1.2], which combine the advantages of both electrical and photonic domains, are preferred for future high-performance and energy-efficient computing systems.

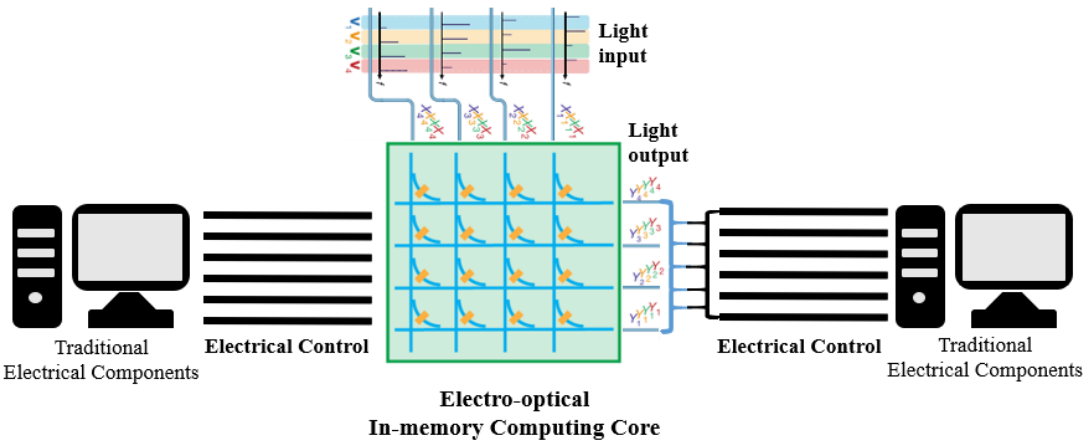

Figure 1.2: An integrated framework for electro-optical computing systems. Based on [14].

To build electro-optical computing architectures, electro-optic memristors [7] are attracting attention because they enable programming and readout of the

memory in both electrical and optical domains, i.e. providing dual electrical-optical functionality, bridging the efficiency of electronics and the bandwidth of photonics without requiring extra optical to electrical (OE) and electrical to optical (EO) conversions.

Phase-change materials, and notably the commonly employed alloy $Ge_2Sb_2Te_5$ (GST), are exploited to modulate and store information in the optical transmissivity and electrical resistivity of different states (amorphous and crystalline), thus have been widely used as both optical memristors [17–23] and electrical memristors [10, 24–26] in computing systems. However, it has been challenging to implement phase-change electro-optic memristors in an integrated system because the mechanisms for optical and electrical programming are different [Figure 1.3]. Evanescent coupling-based optical programming [18] requires adequate interaction length, while electrical programming prefers small confined cells for better Joule heating [24]. Such size mismatching between the electrical and optical domains poses hurdles in achieving both good heat confinement for electrical programming or sensing, and sufficient light-matter interaction for optical programming or sensing.

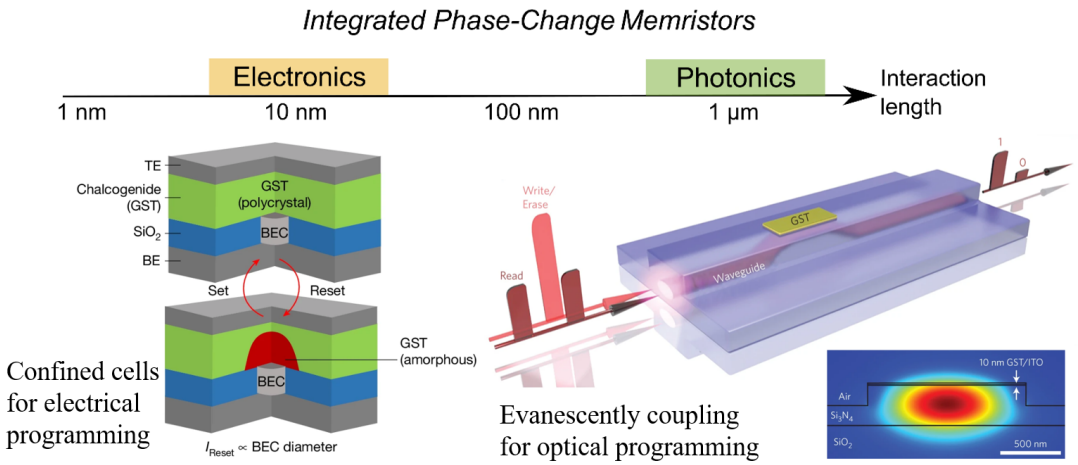

Figure 1.3: Interaction length for integrated electrical and optical phase-change memristors. TE: top electrode; BE: bottom electrode; BEC: bottom electrode contact with the BEC diameter $\leq 20\ nm$. Based on [18, 27, 28].

This thesis aims to provide experimental implementations to tackle this problem, which offer the potential for fully integrated energy-efficient electro-optical computing systems that combine in-memory programming and sensing capabilities.

1.2 Outline and Objectives

1.2.1 Objectives

This thesis addresses the challenges of building energy-efficient electro-optical computing systems on a SOI platform. This is achieved by first implementing plasmonic phase-change devices on a SOI platform. With a corrugated grating design, plasmonic phase-change devices with high mode-coupling efficiency ($-3.57\ dB$) and ultra-low optical switching energy (3.82 pJ) are experimentally demonstrated.

Next, a novel non-plasmonic electro-optic memristor is proposed. By structuring the phase-change material as a bow-tie constriction, the electrically generated heat profile of the device is geometrically confined and combined with photonics. This design allows programmability and readability of phase-change materials in both electrical and optical domains without placing constraints on the metallic layers, and achieves high electro-optical modulation efficiency. As far as is known, it is the first experimental demonstration of integrated electro-optic memristors with dual electro-optical functionality that allows for CMOS-compatible materials and fabrication, facilitating scalable integration. Meanwhile, in-plane photosensitivity of the device is explored, enabling a versatile electro-optical unit cell combining both programming and sensing abilities.

Moreover, system integration of electro-optical computing units is explored, exploiting a foundry-fabricated integrated photonic chip. The design of the photonic chip is discussed alongside experimentally measured results.

1.2.2 Outline

The content of this thesis is outlined as follows:

- **Chapter 2** A comprehensive literature review has been presented covering state-of-the-art implementations for integrated electrical and photonic in-memory computing systems. The advantages and drawbacks of different technical routes in relation to benchmark metrics are discussed. Further, advances in volatile and nonvolatile integrated electro-optical components within computing systems are reviewed.

- **Chapter 3** An introduction to the theoretical background of integrated photonics, along with discussions of the fabrication process and experimental setups, is presented.

- **Chapter 4** Experimental development and demonstration of plasmonic grating nanogap phase-change devices on SOI platforms are discussed in this chapter. A self-aligned fabrication flow is demonstrated, followed by discussion of the coupling efficiency and switching behaviors of the device.

- **Chapter 5** Experimental development and demonstration of non-plasmonic phase-change electro-optic memristors exploiting a constriction structure. The switching behavior of the devices, along with in-plane photosensitivity, is discussed.

- **Chapter 6** Discussions of the system design for electro-optical computing. This chapter presents the design and characterization of foundry-fabricated devices, together with a brief introduction to the development of a transceiver system based on an RFSoC FPGA platform.

- **Chapter 7**: Discussions of future work and the potential impact of the work demonstrated in the thesis.

1.3 Key Findings and New Science

Key results of this work relate to the development of two types of electro-optic devices based on phase-change materials, $Ge_2Sb_2Te_5$ in particular.

The key results and new science include the following :

- A self-aligned fabrication process to implement plasmonic phase-change devices has been developed.

- Plasmonic phase-change devices with corrugated grating designs have been implemented on a SOI platform, demonstrating 3× improved mode-coupling efficiency and 4× lower optical switching energy for $Ge_2Sb_2Te_5$ compared to previous implementations on a Si_3N_4 platform.

- An integrated non-plasmonic electro-optic memristor with dual electro-optical functionality and potential for CMOS-compatible scaling has been proposed. The device has demonstrated low switching energy and high electro-optical modulation efficiency, as well as photodetection capabilities.

Moreover, this project has left new tools developed by the author, as described in Chapter 6, which will be useful for current and future members of the Advanced Nanoscale Engineering group at the University of Oxford, where this thesis was conceived. The new tools include:

- Designed passive and active components of photonic circuits on a foundry run chip with high-speed integrated photodetectors and electro-absorption modulators (EAMs).

- Developed the transmitter and receiver system based on RFSoC FPGA.

1.4 Statement of Originality

This thesis entails the original work performed by the author during her time as a DPhil student at the University of Oxford. All the results and experiments presented here have been obtained by the author solely. Where relevant, the collaborators' contributions have been acknowledged. Grammar and language proofreading has been assisted by AI-based lauguage models, ChatGPT and Writefull. The author takes full responsibility for the final content.

2

Integrated Phase-Change Optoelectronics for Computing: A Literature Review

Contents

To describe the background of electro-optical in-memory computing, this chapter first reviews the mainstream technical routes for electrical in-memory computing and photonic in-memory computing, respectively, followed by a comparison of figures of merit. The last section provides a brief introduction of current implementations of electro-optical components, which form the foundation of the experimental work in this thesis.

2.1 Technical Routes for Electrical In-memory Computing

Modern computer architecture comprises a memory hierarchy (Figure 2.1) to reduce memory access latency, with high-speed, low-latency memory components placed close to the processing unit [29], such as transistor-based static random access memories (SRAMs) and dynamic random access memories (DRAMs).

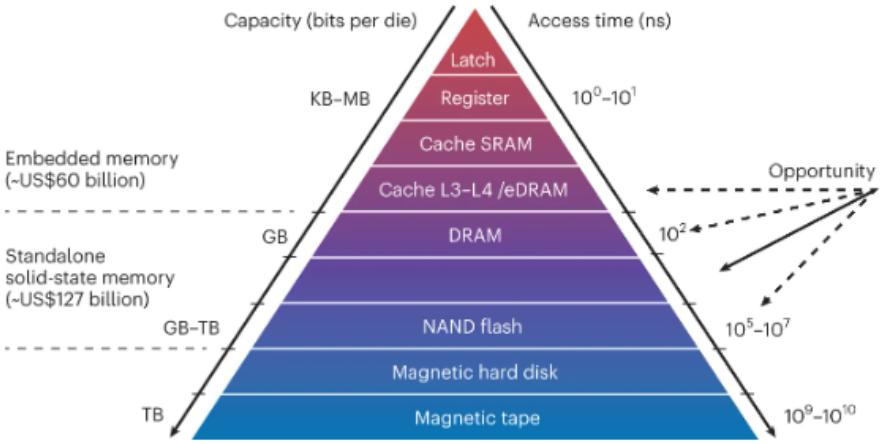

Figure 2.1: Memory hierarchy. Reprinted from [29].

Fast as they are, transistor-based memory units require a constant power supply and dynamic refreshing operations, leading to extra power consumption. Alternatively, nonvolatile memristive devices [Figure 2.2(a)] emerge as a power-efficient solution for memory implementations, i.e. non-volatile memories (NVM). They can be switched between high and low resistances and retain their resistance after the supply power is retrieved, combining the memory function and programmability [8].

Moreover, memristive devices provide an energy-efficient way to implement matrix-vector multiplication (MVM)-based computing applications. Instead of exploiting tens of transistors to implement one multiplication operation, the output current of one memristive device intrinsically calculates the multiplication results of the input voltage and the device conductance according to Ohm's law ($I = V \times G$).

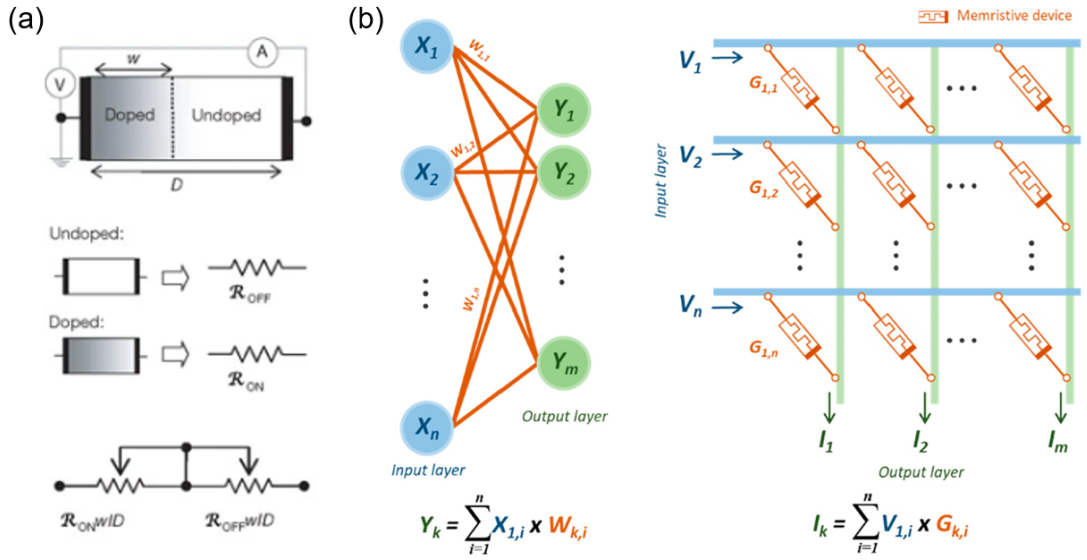

Figure 2.2: Schematics for memristive devices and arrays. (a) Simplified equivalent circuit for a resistive switching-based memristive device. Reprinted from [30]. (b) Implementing matrix-vector multiplication with a memristive device array. Reprinted from [7].

Similarly, arrays of such devices perform MVM in a single step, which are critical operations for neural network-related computing tasks, with the conductivity of memristive devices acting as configurable weight elements (G_{ij} and W_{ij} in Figure 2.2(b)).

This section first introduces representative mechanisms of electrical memristive devices, followed by a survey of the state-of-the-art electrical in-memory computing system implementations. Finally, a comparison of key performance metrics is presented.

2.1.1 Memristive Devices

Memristive devices are mostly based on a metal/insulator/metal nanocell structure [8]. Although broad material options have been explored, such as carbon nanotubes[31, 32], two-dimensional materials [33] or polymers [34, 35], this review focuses on devices made of phase-change, metal-oxide, magnetic or ferroelectric

materials [Figure 2.3], which are technical routes with relatively mature applications.

Resistive random access memory (RRAM)

In 2008, Hewlett Packard Labs reported the first explicit implementation of memristors based on TiO_x devices [30], after Leon Chua proposed the concept in 1971 [36]. Such metal oxide-based resistive switching devices are also referred to as resistive random access memories (RRAMs) [Figure 2.3(a)]. When voltage is applied across the metal electrodes, the electrical field drives the metallic ions from the highly diffusible metal electrode or the oxygen ions from the metal-oxide insulator layer to move, forming doped regions and changing device resistance. The dopant distribution, and thus the device resistance, are retained after removing the external voltage, and they are recoverable when reverse bias is applied [36].

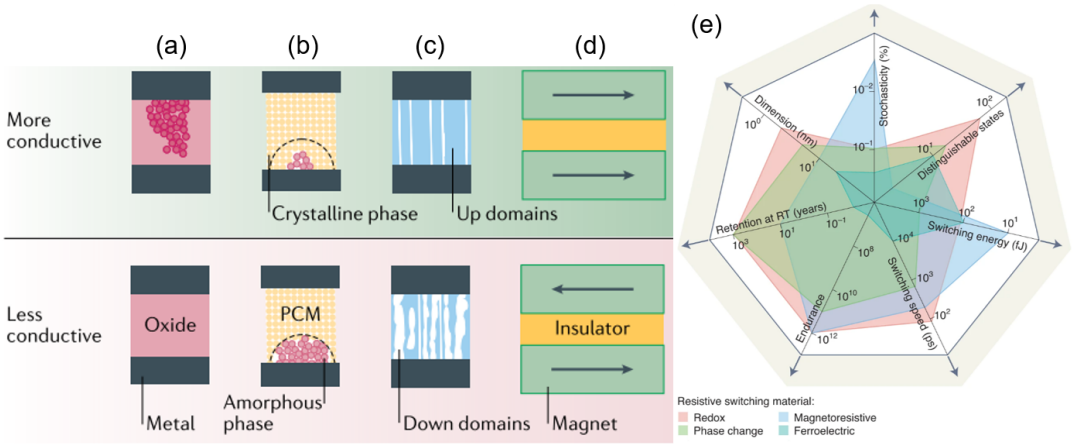

Figure 2.3: Different implementations of memristive devices. (a)-(d) represent metal-oxide, phase-change, ferroelectric and magnetic materials-based memristive devices, respectively. Reprinted from [37]. (e) Benchmark metrics for the four technical routes. Reprinted from [38].

The design of RRAMs provides good CMOS compatibility and high scalability, featuring fast switching speed (<10 ns), large high resistance and low resistance (HRS/LRS) ratios (>100, up to 2,048 conductance levels [39]) and low switching energy (<0.1 pJ) [8]. However, thermal fluctuation and the stochasticity of the

nanofilament formation and rupture processes induce variations in switching voltages and state resistance, limiting the endurance of the devices. The most repeated endurances are between 10^6 and 10^7 cycles [8].

Phase-change memory (PCM)

Chalcogenide materials, such as $Ge_2Sb_2Te_5$, exhibit different resistances in their various phases (crystalline and amorphous) [Figure 2.3(b)] and have been commercially used as phase-change memories (PCMs) [40, 41]. PCMs undergo phase transitions as a result of Joule heating and threshold switching [24, 42].

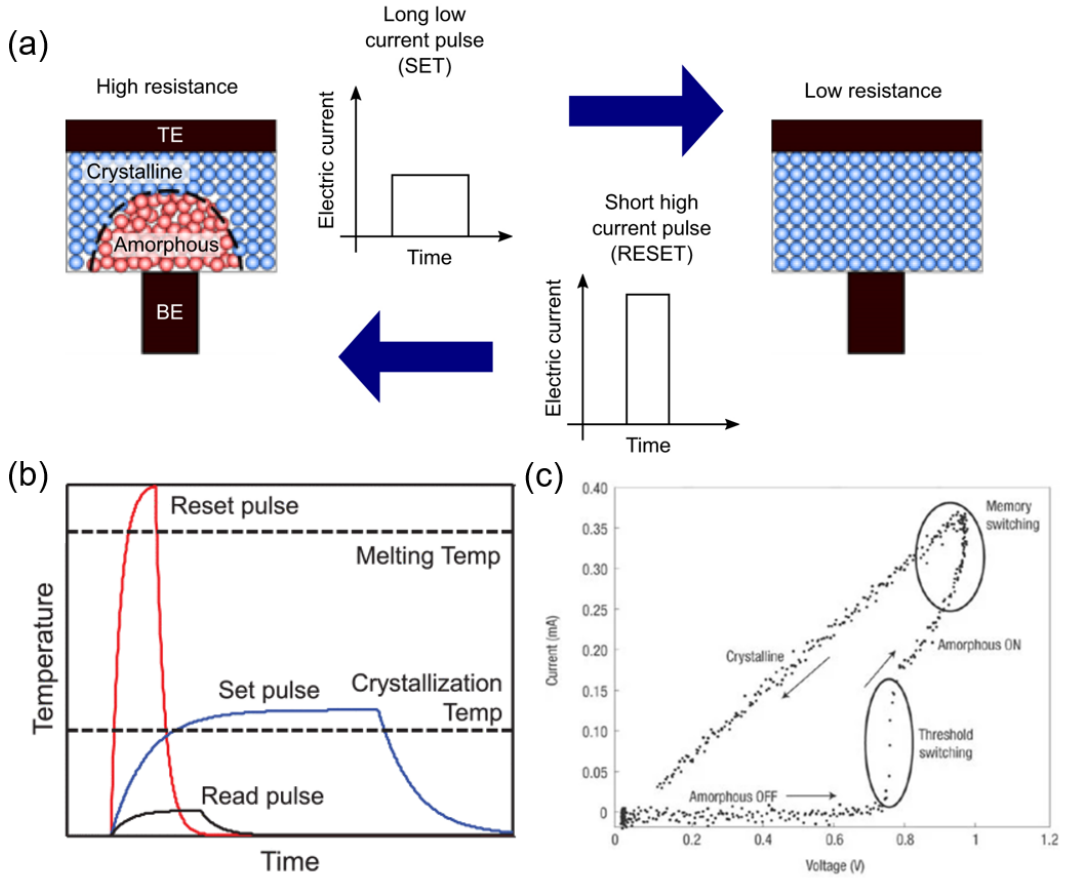

Figure 2.4: Switching mechanism of PCMs. (a) Device schematics under SET and RESET operations. TE: top electrode, BE: bottom electrode. Reprinted from [43]. (b) Temperature profile of the PCM devices under set, reset and read pulses. Reprinted from [42]. (c) Typical current-voltage curve of PCMs initially in an amorphous state. Reprinted from [24].

A short high-power pulse melts the crystalline material (high absorption, low

resistance) and rapidly cools it to a disordered, amorphous state (high transmission, high resistance). On the other hand, a moderately long pulse restructures the amorphous material to the crystalline state [Figure 2.4(a)-(b)]. The amorphization process begins with threshold switching, where the device resistance drops abruptly when the applied voltage is higher than a threshold. Afterwards, Joule heating dominates to further heat the material above the melting temperature [24] [Figure 2.4(c)]. During crystallization, many nuclei (small crystalline regions) firstly form then grow slowly to achieve full crystallization for nucleation-dominated material GST, requiring sufficient pulse duration for crystal lattice reconstruction [24, 44].

PCMs offer high scalability (cell size $< 10\ nm$), low programming voltage ($< 3\ V$), and large HRS/LRS ratios (>100). Yet, it requires longer writing pulses (50-100 ns) compared to RRAMs, limited by the slow recrystallization process [8]. PCMs demonstrate endurance up to 10^{12} [45]. Similar to RRAMs, PCMs suffer from inherent variability of switching voltages and state resistance, owning to atomic rearrangements and structural relaxation [8].

Ferroelectric tunnel junction (FTJ)

Placing a thin (a few nanometers) ferroelectric insulator (such as $BaTiO_3$ [46, 47]) between two electrodes forms a ferroelectric tunnel junction (FTJ) [Figure 2.3(c)]. The out-of-plane resistance of such devices is dependent on the orientations of the crystalline unit cells (polarized ferroelectric domains) within the ferroelectric insulator, tuning the tunneling barrier height and tunneling probability. With a ferroelectric layer of a few nanometers, quantum-mechanical tunneling becomes possible, where electrons tunnel through the ferroelectric layer with energy lower than the barrier height. The ferroelectric polarization of the domains can be tuned by an external electrical field and remains fixed after the field is removed [48].

Such devices, based on the quantum tunneling effect, demonstrate low write energy (0.1 pJ) and provide long retention time(∼10 years). Their experimental

performance is limited by the relatively lower switching speed (10 ns [38]), limited endurance ($\sim 10^7$) and HRS/LRS resistance ratios (<100) [8].

To clarify, FTJ is different from ferroelectric random access memory (FeRAM) and ferroelectric field-effect transistor (FeFET). The three-terminal device FeFET is formed by replacing the channel material or the gate insulator layer of a transistor with ferroelectric material to tune the channel conductivity in a non-volatile manner, while FeRAM is not a memristive device [8]. The dielectric capacitor layer of DRAM is replaced with a ferroelectric layer in FeRAM to provide non-volatile data storage.

Magnetic random access memory (MRAM)

Magnetic random access memories, or MRAMs, are widely used as disk memories. Two magnetic layers separated by an insulator form a magnetic tunnel junction (MTJ) [Figure 2.3(d)]. One of the magnetic layers has a pinned magnetic state, while the state of the other layer is free to be changed by external electrical stresses, enabling the out-of-plane resistance of the MTJ to be reconfigurable.

As a standalone commercial memory, MRAMs provide excellent endurance ($\sim 10^{15}$) and robustness to thermal disturbance, at the cost of high writing energy, low HRS/LRS resistance ratio (~ 2) and complex stacks for scalability [8].

Table 2.1: Performances of commercial stand-alone memories in 2021. Based on [8, 45].

Metrics	DRAM	FeRAM	PCM	RRAM	MRAM
Cell area	6 to $8F^2$	6 to $30F^2$	4/4L F^2	6 to 30 F^2	6 to 30 F^2
Bits per die	16 GB	8 Mb	**256 Gb**	8Mb	1Gb
Retention	50 ms	>10 years	>10 years	>10 years	>10 years
Endurance	$\sim$ **10^{15}**	$\sim$ **10^{15}**	$\sim 10^{12}$ [45]	$\sim 10^7$	$\sim$ **10^{15}**
Read/Write time	$\sim$ 10 ns	10 to 100 ns	10 to 100 ns	$\sim$ 100 ns	$\sim$ 10 ns
HRS/LRS ratio	-	<100	**>100**	**>100**	~ 2
Cell energy	$\sim$ **10 fJ**	$\sim$ 0.1 pJ	$\sim$ 10 pJ	~0.1 pJ	$\sim$ 0.1 pJ
2021 price	\$ 0.50/Gb	>\$ 1000/Gb	$\leq$ \$ **0.30/Gb**	$\sim$ \$ 1000/Gb	\$ 40 to 70/Gb

A comparison of the merit figures for these devices is summarized in Figure 2.3(e) and Table 2.1 ("L" is the number of layers in a three-dimensional configuration, and

"F" represents the minimum lithography feature size) [8]. In summary, RRAM and PCM devices exhibit the best overall performance, featuring high HRS/LRS ratio for multilevel programmability with the highest number of conductance levels and the largest integration capacity (bits per die), respectively. MRAMs support the best endurance and lowest stochasticity, but are limited in multilevel programmability; while the emerging FTJ devices is promising for ultra-low energy switching, but still requires future efforts to optimize device endurance.

2.1.2 System Implementations

Different designs of NVM-based weight cells have been explored for MVM operations. Instead of exploiting a single memristive device (1R) as weight elements, connecting one transistor with one passive memristive device to form the 1T1R design is most commonly used [Figure 2.5(a)]. The transistor serves as a selector to reduce the sneak path current and provide more precise control of the conductance of the memristive device [5, 6]. Multiple memristive devices can also be grouped to form one weight element, enabling fine tuning of the device precision [49] and allowing the weights to take on negative values [Figure 2.5(b)].

Given high compatibility with CMOS technologies and multilevel programmability, system implementations [Figure 2.5(c)] based on PCM and RRAM devices are the most explored.

IBM research groups scale up PCM devices to 34-tile [50] and 64-core [10] systems with on-chip digital operations and on-chip communications, supporting up to 35 million configurable weights. End-to-end inference tasks, such as speech transcription based on Transformer [51] or recurrent neural network transducer (RNNT) [52] and imagine classification based on ResNet-9 [53], are demonstrated on-chip with near software-equivalent accuracy, encoding 8-bit input/output precision through pulse-modulated durations. Up to an energy efficiency of 9.76 TOPS/W and a throughput of 63.1 TOPS/s were achieved for MVM operations. Energy

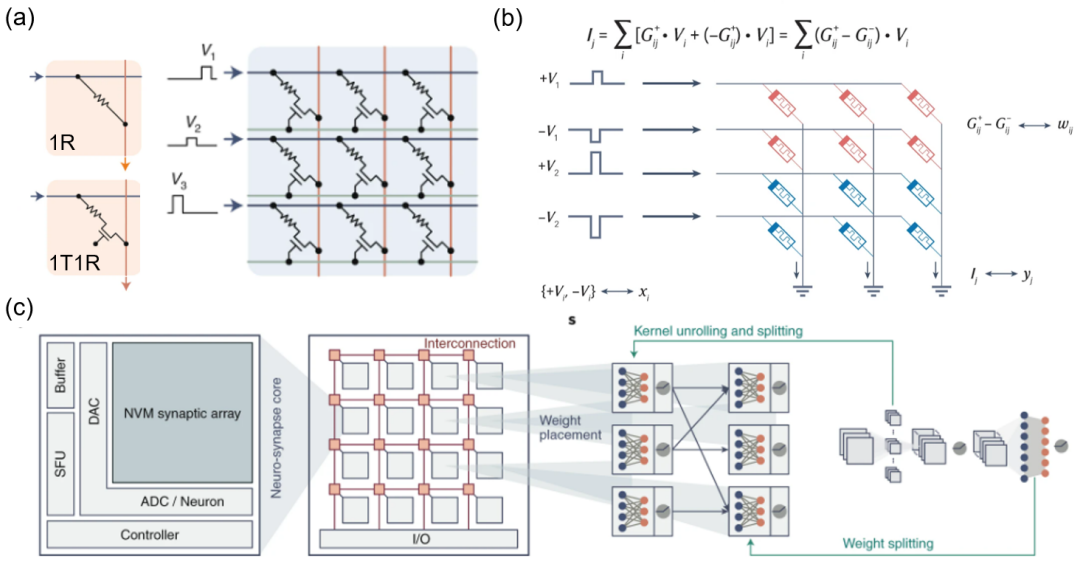

Figure 2.5: Memristive arrays. (a) NVM-based weight cell configurations. Adapted from [5]. (b) Differential configurations for implementing weights. Adapted from [6]. (c) Generic NVM-based computing architecture for neural network applications. SFU: special function units. Reprinted from [5].

efficiency can be further improved with lower communication cost, greater tile usage, and more efficient digital peripheral circuits [50].

Various research groups in both industry and academia have investigated large-scale implementations of RRAM-based computing systems, with subareas spanning from unit cell configuration, delicate peripheral circuit design (ADC, DAC, control units, interconnects) to architecture and algorithm optimizations [5, 6]. The first fully hardware-implementated RRAM convolutional neural network (CNN) system was proposed in 2020 [54]. The system integrated eight 2kb RRAM arrays to support a 5-layer CNN model for MNIST digit image recognition [55] and provided software-comparable accuracy with advantages in energy efficiency and latency compared to GPU implementations. Numerous practical neural network applications have been demonstrated since then. Recent progress has led to RRAM integration up to 4 MB capacity [56] for larger neural network models, such as ResNet-20 and ResNet-50 [57], achieving an energy efficiency of 51.4 TOPS/W and total latency

of 472.71 μs for CIFAR-10 [58] image classifications under an 8-bit input–weight precision configuration.

Early implementations of MRAM-based in-memory computing have also been proposed [59, 60], enabling one-shot binary MVM operations with dimensions up to 128×128 and demonstrating classification accuracies equivalent to digital computations based on the CIFAR-10 and MNIST datasets. The challenges for their future scaling up include accumulated analogue noise and limited resistance-state contrast.

According to the available literature, there are not yet large-scale implementations of FTJ-based in-memory computing due to the lack of device stability regarding conductance states and endurance [6]. Yet, FeFET-based devices have emerged as energy-efficient candidates to implement dynamic reconfigurable field programmable gate arrays [61] and hybrid analog-digital in-memory computing systems [62].

2.1.3 Figures of Merit

From a system-level perspective, there are some general metrics that are widely used to compare different computing platforms [5, 6].

The first and most commonly used one is computing density, which describes how many Tera-Operations are processed Per Second in a unit area with the unit TOPS/mm^2. The second is power efficiency with the TOPS/W unit, or thermal design power (TDP) for products in the industry. There are also other metrics, such as precision (bits) for input/weight/output and latency (ns/ps), which are critical to the performance of practical applications.

Table 2.2 provides a comparison of the above technical routes. Volatile SRAM-based implementations are included because multi-bit SRAMs can also serve as analog in-memory computing unit cells in a crossbar array architecture. When

executing MVM operations, continuous supply voltages provided by the peripheral circuits are required to retain the SRAM states [63, 64].

The computing density and energy efficiency indicated in the brackets are normalized to a 1-bit input-weight precision configuration for fair comparison. The capacity refers to the number of in-memory computing unit s (PCM, RRAM, MRAM and SRAM cells) integrated in a single chip. Here, PCM-based technology provides the highest integration capacity (35 Mb [50]), while SRAM-based implementation offers the best energy efficiency and computing density with smaller integration capacity, given its compatibility with the most advanced technology node (4 nm [64]) and higher computing precision. Nevertheless, it is worth noting that even with a less advanced technology node, RRAM-based implementation has already provided high energy efficiency comparable to SRAM implementations, showing great potential for future energy-efficient computing applications.

Table 2.2: State-of-the-art inference demonstrations with in-memory computing

Device	PCM	RRAM	MRAM	SRAM
CMOS technology	14 nm	22 nm	22 | 28 nm	16 | **4 nm**
Capacity	**35 Mb** | 4.2 Mb	4 MB | 8 Mb	128 kb | 4 kb	4.5 Mb | 54 kb
Operation speed	100 | 400 MHz	200 | - MHz	- | 11.1 MHz	0.2 | **1.49** GHz
Input/weight/output Precision (bit)	8/analog/8	8/8/(24|19)	1/1/4	4/4/8 | 12/12/36
Energy efficiency (TOPS/W)	12.4 | 9.76 (396.8) | (312.3)	51.4 | 21.6 (**3290**) | (1382)	5.1 | 405	121 | 42.4 (1936) | (**6106**)
Computing density (TOPS/mm^2)	- | 1.55 - | (49.60)	0.284 | 0.01 (18.18) | (0.64)	0.758 | 4.43	2.67 | 33.3 (42.72) | (**4795**)
Reference	[50] | [10]	[56] | [57]	[59] | [60]	[63] | [64]

There is not yet enough data in the literature to compare the latency for different technical routes. Two related metrics are commonly referenced in the literature: the first metric is access time for multiply-and-accumulate (MAC) operations under certain precision configurations, with example values being 14.4 ns for an 8/8/19-bit RRAM configuration [57] and 6.6 ns for an 8/8/22-bit SRAM configuration [65]. The other is the inference latency for specific applications ranging from several to hundreds of μs, which depends on neural network structures, such as 1.52 μs for the

PCM-based ResNet-9 network [10] and 472.7 μs for larger RRAM-based ResNet-20 network [56]. Non-volatile implementations have the potential to provide lower latency in real applications compared to SRAM-based volatile implementations, as the former does not require off-chip weight buffers to hold and continuously refresh network weights [10].

2.2 Technical Routes for Photonic In-memory Computing

Although breakthroughs [5] have been observed on the electronic side owing to its high compatibility with the current manufacturing process, the inherent advantages [11, 66] of photonics such as bandwidth (parallelism) and low latency, which are much sought after in data-center applications, herald new opportunities for photonic computing.

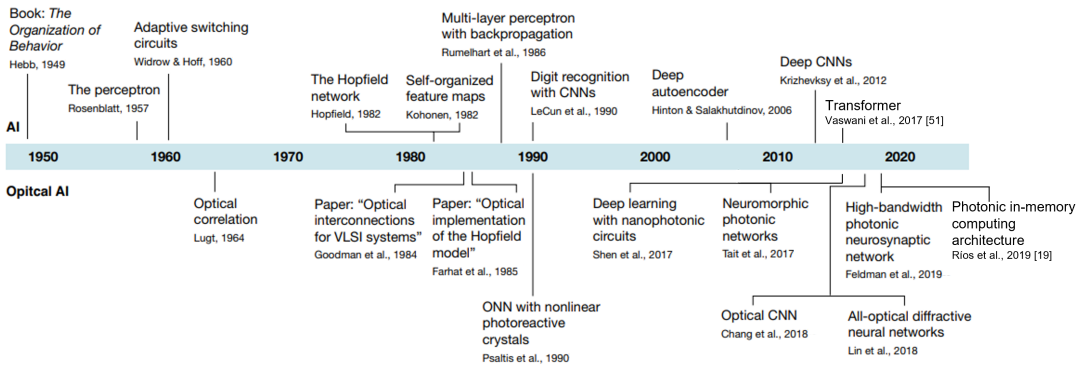

Figure 2.6: Timeline of advancements in optical and photonic in-memory computing. Adapted from [67] and modified to include recent developments [19, 51].

A timeline [67] of the advancements in AI-related optical and photonic implementations is provided in Figure 2.6. The history of optical computing can be traced back to the last century [68, 69]. However, on-chip computing with light or photonic computing has emerged only in recent decades, thanks to the rapidly developing silicon manufacturing technology and the challenges posed by the growing demand of neural network implementation in the so-called information age. Numerous reviews

[12, 13, 67, 70–72] have provided insight into the development of this field, and this section examines several mainstream technical routes for integrated photonic computing, analyzing their potential applications and current limitations with a summary of metrics comparison.

2.2.1 Interferometric Mesh Architectures

After several years of exploration into non-von Neumann computing schemes in the electronics domain for artificial neural network applications, the photonic neural network (ONN) architecture proposed by Shen *et al.* [73] has sparked renewed interest in photonic computing.

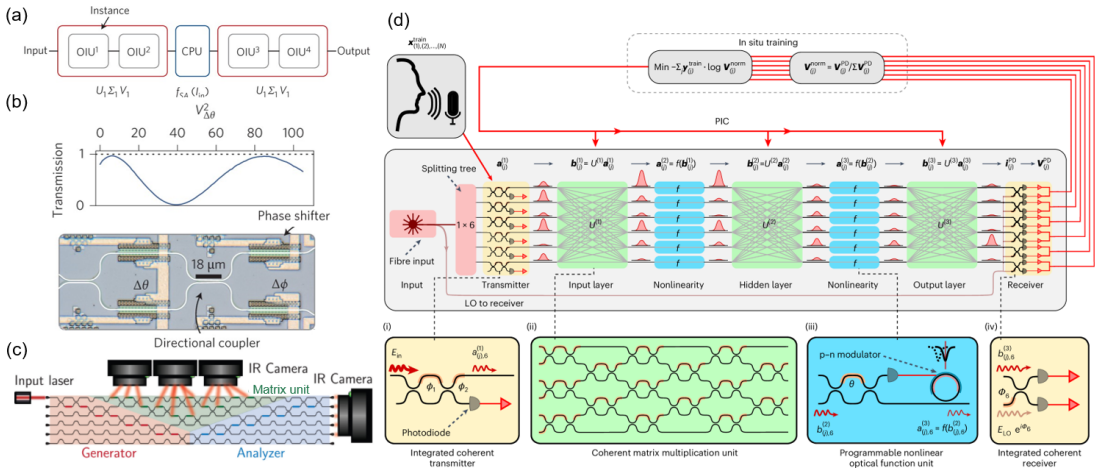

Figure 2.7: Photonic neural network schematic based on an interferometric mesh architecture. (a) Workflow of the proposed optical neural network (ONN). OIU: optical inference unit. Reprinted from [73]. (b) Schematic of a phase shifter and its transmission curve. Adapted from [73]. (c) A layout for implementing matrix-vector multiplication with in-situ back-propagation and IR monitors. Adapted from [74]. (d) Photonic integrated circuit (PIC) architecture for an end-to-end, coherent optical processor. Reprinted from [75].

They chose a deep learning network as their target, multilayer perception (MLP) in specific, realized photonic linear matrix multiplications and simulated the nonlinear functions. By integrating 56 programmable Mach-Zehnder interferometers (MZIs) on-chip to implement the interlayer weights of a 3-layer network, they completed the vowel recognition task with reasonable accuracy.

The ONN architecture is illustrated in Figure 2.7(a)-(c). Real-valued matrices are decomposed into the product of two unitary matrices (U and V) and one diagonal matrix (Σ) using singular value decomposition (SVD)[76], and each optical interference unit (OIU) can be used to implement one unitary matrix V or the multiplication result $U\Sigma$. The key elements of the OIU are universal linear optics [77] based on MZIs. By electrically tuning the thermo-optic phase shifters, the refractive index of the MZI waveguides changes, providing a phase shift accordingly. The internal phase shifter ($\Delta\theta$) defines the splitting ratio of the MZI while the external phase shifter ($\Delta\phi$) modulates the differential output phase, enabling arbitrary multiplication of unitary matrix [Figure 2.7(b)].

Given good compatibility with silicon photonics foundry processes, larger-scale hardware implementations of ONN up to 64×64 have been realized in industry [78], together with progress in the algorithm and architecture levels. The in-situ back-propagation method has been experimentally demonstrated in 2023 [74, 79], with dedicated phase control assisted by real-time infrared (IR) camera monitoring [Figure 2.7(c)]. This design enables both training and inference functions of neural networks. The architecture is further improved in 2024 [75], with advances in the perturb-based forward only training method and on-chip microring modulator-based programmable nonlinear activation units, demonstrating single-chip implementation of a fully integrated optical deep neural network [Figure 2.7(d)].

In addition to neural network applications, generic programmable photonics could also be implemented making use of this design approach with MZIs and phase shifters, analogous to a field-programmable gate array (FPGA) in electronics. Instead of utilizing lookup tables as basic cells, the fundamental units of these optical FPGAs are optical gates (Figure 2.8), as stated in the review by Bogaerts *et al.* [80]. These 2×2 optical gates modulate both the output splitting ratio and the relative phase of the light wave through phase shifters or tunable couplers, functioning as 'bar' states, 'cross' states or mixed partial states [Figure 2.8(b)-(e)].

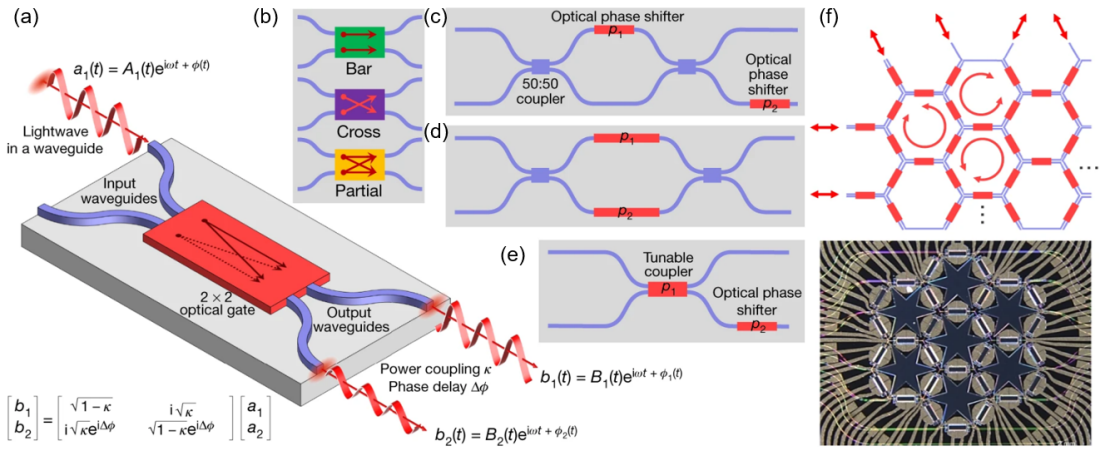

Figure 2.8: Forward-only and recirculating meshes. Adapted from [80]. (a) Mechanism of universal 2×2 optical gates. (b) Schematic for the function of the gates, when tuned between 'bar' and 'cross' states. (c)-(e) Different designs for the gates. (f) Recirculating meshes in hexagonal geometries.

Apart from forward-only meshes, recirculating routing function could also be realized when applying such a modulating method in waveguide loops [Figure 2.8(f)]. These general-purpose routing designs can be reconfigured as different photonic units (e.g. ring modulators) for photonic computing systems, and are promising for future communications, sensing, and broadband analog signal processing [80–82].

The main drawbacks for these programmable MZI designs are their relatively large footprints ($\sim$ 100 μm length or more for one MZI) and thermal crosstalk, or the challenge of achieving thermal stabilization for phase control. Extensive efforts have been made to explore the self-calibration scheme [83–85] to mitigate this difficulty.

2.2.2 Microring Weight Banks

Programmable MZIs are not the initial technical route for implementing photonic computing on an integrated platform. Earlier designs focused more on programmable microring resonators (MRRs), which serve as filter banks to modulate multi-wavelength signals [86–90].

The MRR weight-bank-based network node [86] was proposed in 2014 as an on-chip interconnection protocol [Figure 2.9(a)]. The optical signal is first coupled to a balanced photodetector (BPD) pair via two MRR weight banks, which shape excitatory (positive) and inhibitory (negative) inputs, respectively. Then the photodetectors convert the optical signals into an electrical weighted sum, modulating the excitable laser neuron for threshold detection and spike generation. The output of the laser neuron will be wavelength-division multiplexed (WDM) and broadcast to other nodes, which completes one the so-called weight-and-broadcast loop.

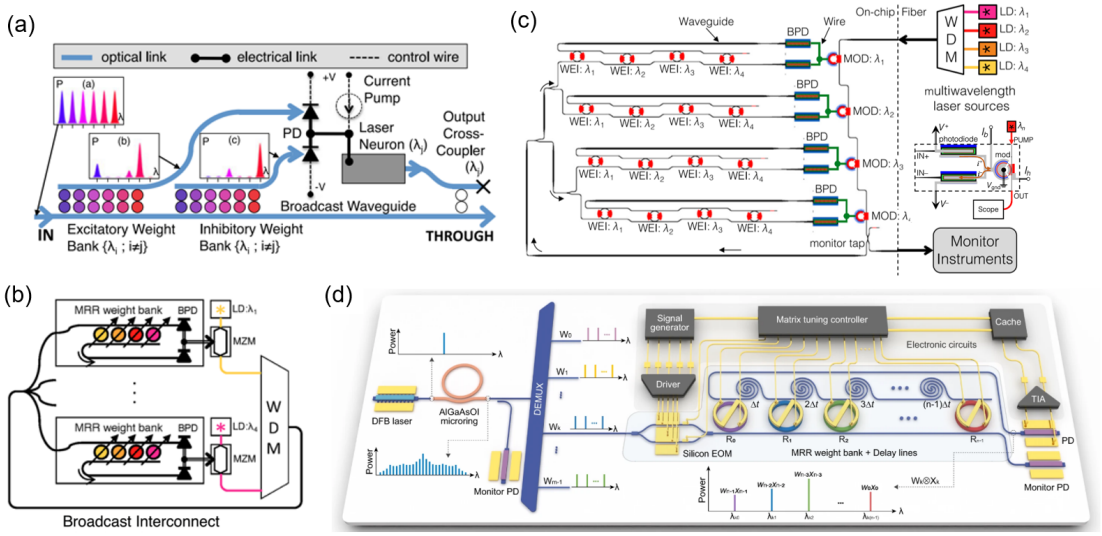

Figure 2.9: MRR weight bank-based Broadcast-and-Weight architecture. (a) Broadcast-and-weight protocol node. PD: photodetector. Reprinted from [86] (b) Experimental schematic. MRR: microring resonator; BPD: balanced photodetector; LD: laser diode; MZM: Mach-Zehnder modulator; WDM: wavelength-division multiplexer. Reprinted from [91] (c) Neural network schematic. Adapted from [92]. (d) Neural network schematic. Adapted from [93].

This protocol was further improved and experimentally implemented [91, 94]. As shown in Figure 2.9(b), both through and drop ports of the MRRs are used during the weighting process, thus only one weight bank is required for one node. The output current of the BPD is converted to electrical voltage, acting as the modulation input for Mach-Zehnder modulators (MZM) to realize nonlinearity.

A 2-node network has been experimentally demonstrated, and the simulation of a 24-node network has shown a 294× speedup for solving differential equation problems over a CPU benchmark. In 2019, a silicon photonic modulator neuron was proposed and experimentally implemented [92] by replacing the aforementioned MZM with a ring modulator as in Figure 2.9(c). This design provides more freedom, utilizing both thermal (heater current bias I_h) and electrical (modulator current bias I_b) tuning to modify the transfer function of the ring modulator and further integrating the BPD on the chip. The electrical tuning influences the depth and quality factor of the resonance peak, while the thermal tuning produces peak shifts to maintain the desired wavelength of interest. Combined thermal and electrical parameters enable reconfigurable nonlinearity.

The MRR weight bank has been further combined with a microcomb laser source [Figure 2.9(d)] to build an integrated processing unit [93]. By programming the MRR weight banks as 2×2 kernel matrices, the processing unit implements CNNs for image edge detection and digit recognition applications. With a dedicated calibration procedure and time-wavelength multiplexing, the processing unit achieves a weight precision of 9 bits and a computing density of 1.04 TOPS mm^{-2} (17 *Gbaud* modulation rate with a photonic core footprint of ~ 0.131 mm^2), yet the nonlinear activation function is implemented on a digital computer.

This kind of MRR weight bank-based designs provides a smaller footprint compared to MZI-based technical routes, yet large-scale implementations have not been materialized, owning to the interbank thermal crosstalk and the sensitivity of the rings that caused weight inaccuracy and poor environmental robustness [91, 92].

2.2.3 Phase-change Photonic Crossbar Arrays

Another design based on phase-change photonics [14, 21, 95–97] shows greater potential for future scaling. This flexible technical route enables versatile network

architectures.

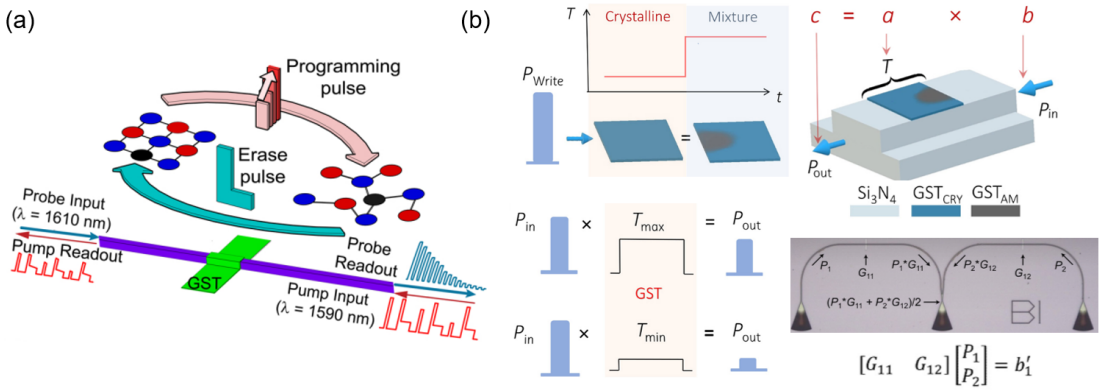

Figure 2.10: Phase-change materials-based photonic devices. (a) Pump-and-probe scheme for programming phase-change photonic devices. The pump input programs the material states while the continuous wave (CW) probe light readout the material states without changing the states. Reprinted from [20]. (b) Phase-change photonic devices for multiply-and-accumulate (MAC) operations. Adapted from [19].

Thin film phase-change materials are deposited above [18] or within [98] nanophotonic structures, such as waveguides, to form phase-change photonic devices. Optical modes of the nanophotonic structures couple evanescently to phase-change materials. Therefore, the state of the material, together with the transmittance of the device, can be modulated with engineered optical pulses [Figure 2.10(a)]. Taking the input optical power ("a") and the transmittance of the device ("b") as two factors, and the output power as the result ("c"), phase-change photonic devices function as multiplication operators while optical combiners offer an accumulation function [Figure 2.10(b)]. MVM operations with 5-bit precision have been experimentally demonstrated [14, 20].

Device-level photonic MAC operations are further expanded to matrix multiplications with a crossbar array configuration [14]. Combining with a multi-wavelength frequency-comb source [99, 100] and a dedicated directional coupler design for light splitting, phase-change photonic crossbar arrays materialize the inherent advantage of photonics in wavelength parallelization to provide high computing density and bandwidth. In this work, input information is programmed as signal amplitude in

different wavelength channels by variable optical attenuators (VOA), then couples into the crossbar core via wavelength division multiplexers (WDMs) and broadband 3D couplers [101], where the multiplications between vectors and matrix occur [Figure 2.11(a)]. Image processing and digit recognition tasks were demonstrated in this platform using CNN and it reached a speed of two tera-multiply-accumulate operations per second (2 TMAC/s or 4 TOPS).

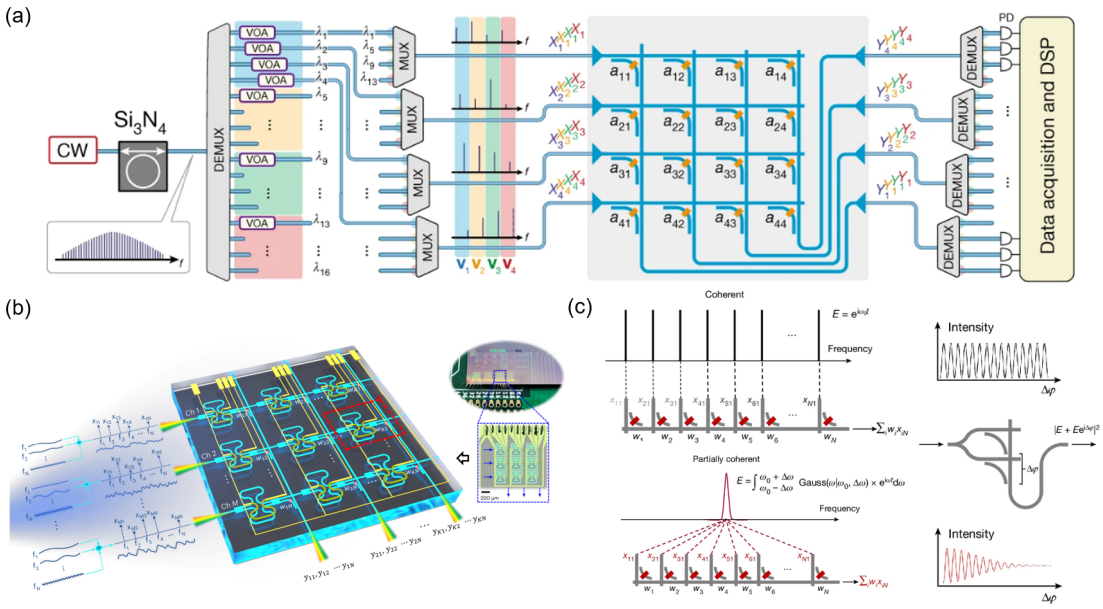

Figure 2.11: Phase-change materials-based photonic crossbar arrays. (a) Wavelength-multiplexing with an integrated phase-change photonic computing core. Adapted from [14]. (b) Photonic computing core with wavelength-multiplexing and RF modulation. Adapted from [96]. (c) Incoherent scheme for the photonic crossbar array. Adapted from [97].

The bandwidth advantage of photonics has been further exploited by Dong *et al.* [96], which adds a third dimension of freedom through radio-frequency modulation of photonic signals [Figure 2.11(b)] and demonstrates a two-orders-of-magnitude higher parallelism. Moreover, with light sources of reduced degree of coherence [Figure 2.11(c)], the interference effects between different physical channels of the crossbar array can be reduced, further enhancing the computing bandwidth by leveraging parallelism from both space-multiplexing and wavelength-multiplexing [97].

The crossbar architecture provides the highest computing density for photonic computing among all the above-mentioned work, attributing to the small unit cell size and the freedom to exploit more degrees of parallelism for higher computing bandwidth. Yet, the challenge in scaling up such crossbar architecture lies in the limited accuracy of the directional coupler splitting ratio. The larger the array size, the smaller the minimum coupling ratio, thus the system is more sensitive to fabrication variations [14, 102]. The power loss of the directional coupler also begins to dominate the system loss as the array size increases. The scaling potential of this architecture has been discussed in Li *et al.*, and calculated to support a maximum matrix dimensions of 494×494 [95], bounded by the sensitivity of the photodetector.

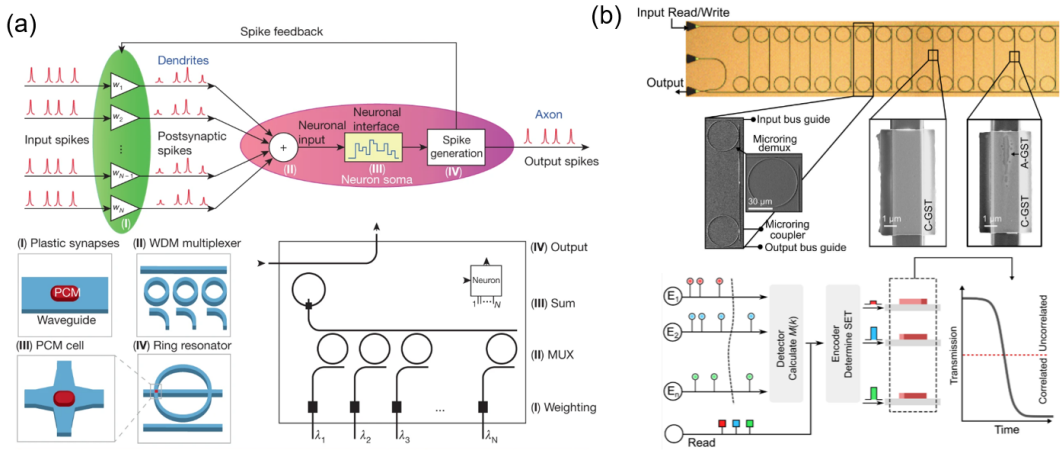

Figure 2.12: Phase-change materials-based MRR architectures. (a) Phase-change photonic devices for spiking neural network implementation. Adapted from [21]. (b) Phase-change photonic devices for correlation detection. Adapted from [22].

Note that along with the crossbar array architecture, phase-change materials have also been applied to other photonic computing architectures. In addition to functioning as synapses to weight the input signal, phase-change photonic devices further realize summation and threshold detection when combined with MRR structures [21]. The input spikes are implemented with optical pulses at different wavelengths created from continuous-wave lasers. When the combined power of the

input spikes is higher than the switching threshold of the phase-change cell on the sum-up resonator and switches its states, the probe pulse will be transmitted past the ring resonator and generate an output spike. Depending on how the output of this sum-up resonator is handled, as feedback spike or input to the next layer, both unsupervised learning spiking neural network (SNN) and forward neural network are feasible in this configuration [Figure 2.12]. The summation and threshold detection feature of the phase transition has been later exploited for correlation detection applications (i.e., identifying temporal correlations between data streams) [22], providing high-speed and high-bandwidth advantages compared to conventional electronic approaches. Early implementations of phase-change material-based phase-shifters have also been proposed for MZI mesh and routing applications [103–105], which will be discussed in Section 2.3.2.

2.2.4 Diffractive Neural Network

Diffractive neural networks also reveal versatility in their applications [106–108]. This free space design is similar to previous implementations based on nonlinear crystal holography [109], which have the much-sought ability to implement millions of pixel-level weights.

As illustrated in Figure 2.13(a), the amplitude or/and phase of the input wave are modulated by the complex-valued transmission or reflection coefficient of pixels in a diffractive layer, akin to the weight and bias modulating the input in a neural network. This kind of architecture can be used to perform both classification and imaging tasks while the coefficients were trained in silico (i.e. computer-based simulations) using error backpropagation [110]. In the earlier design, the diffractive layers were 3D-printed before the experiments and fixed afterwards, and the detector arrays were placed on the output plane to receive classification or imaging results [106].

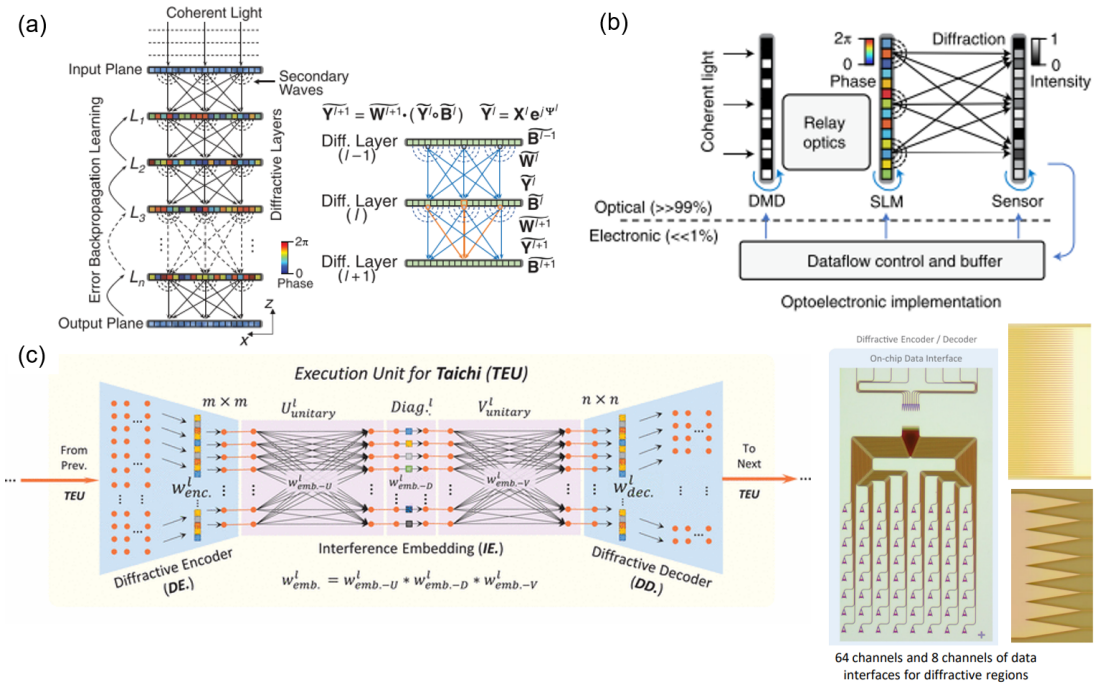

Figure 2.13: Schematics for Diffractive Neural Networks (DNN). (a) The first prototype of all-optical DNN. Adapted from [106]. (b) Reconfigurable diffractive processing unit. Reprinted from [108]. (c) On-chip diffractive units. Adapted from [111].

This diffractive neural network(D^2NN) architecture is further expanded to the Fourier space by incorporating a dual $2f$ optical system [107]. With working wavelengths within the visible range, the Fourier-space diffractive deep neural network (F-D^2NN) provides good performance for video processing tasks such as detection of salient objects. Moreover, by adding physical nonlinear layers (photorefractive crystal (SBN: 60)) to perform activation functions, higher accuracy in classification tasks and better saliency detection performance have been achieved compared to the real-space D^2NN architecture.

One drawback of the previous designs is that the diffractive layer is fixed after 3D printing, limiting the flexibility of the diffractive neural network. Later works have offered an alternative method to enable reconfigurability within the diffractive architecture [108]. Exploiting the digital micromirror device (DMD) as the input encoder and the phase spatial light modulator (SLM) as the diffractive weight modulator, the renewed architecture in Figure 2.13(b) has been shown to be

programmable for versatile neural network structures, including the weight-sharing convolutional neural network (diffractive network in network or D-NIN-1) and the diffractive recurrent neural network (D-RNN). With further adaptive training algorithm that takes the experimental output of the layers into consideration during the training process, this architecture offers outstanding experimental accuracies for image classification and video recognition benchmark tasks, though the in-silico training process is quite demanding for such a large scale neural network. The above-mentioned diffractive optical analog computing layers are subsequently combined with electronic analog computing units, providing an end-to-end all-analog solution for vision tasks [112].

Recent studies have also explored on-chip implementations of diffractive units [111, 113]. A theoretical work proposed a diffractive graph neural network (DGNN) to handle graph-structured data, leveraging metalines (slot arrays)-based [114, 115] diffractive photonic computing units (DPUs). The on-chip diffractive photonic computing units have since been experimentally implemented, serving as diffractive encoders and decoders in a distributed computing architecture [Figure 2.13(c)] for performing classification and versatile content generation tasks [111]. It is worth noting that though this work claims ultra-high energy efficiency as 160-TOPS/W, most of the operation counts are attributed to the fixed diffractive encoding and decoding units (12.59 MFLOPs), while the critical MAC computing operations (512 FLOPs) performed in the programmable MZI units contribute only a small portion. To enable a fair comparison, the performance analysis in the benchmark section (Section 2.2.6) will focus solely on the programmable MAC computing operations.

2.2.5 Emerging Routes

There are also other emerging technical routes for implementing photonic computing that cannot be classified as any of the above-mentioned architectures.

One major technical route is based on time-multiplexed photonic matrix-matrix multipliers [102, 116, 117]. Instead of fixing matrix weights on the photonic hardware and modulating optical input signals as input vectors, the time-multiplexed photonic multiplier computes vector-matrix multiplications on the fly by encoding both inputs and weights in time as optical signals [Figure 2.14(a)-(b)]. The input signal is optically fanned out as several copies to be coherently detected alongside the weight signal by homodyne detectors, providing multiplication products. This time-multiplexed design circumvents the calibration and device stability challenge associated with other analog fixed-weight methods and makes use of time-integration detection to bridge the gap in operating speed between photonics and electronics.

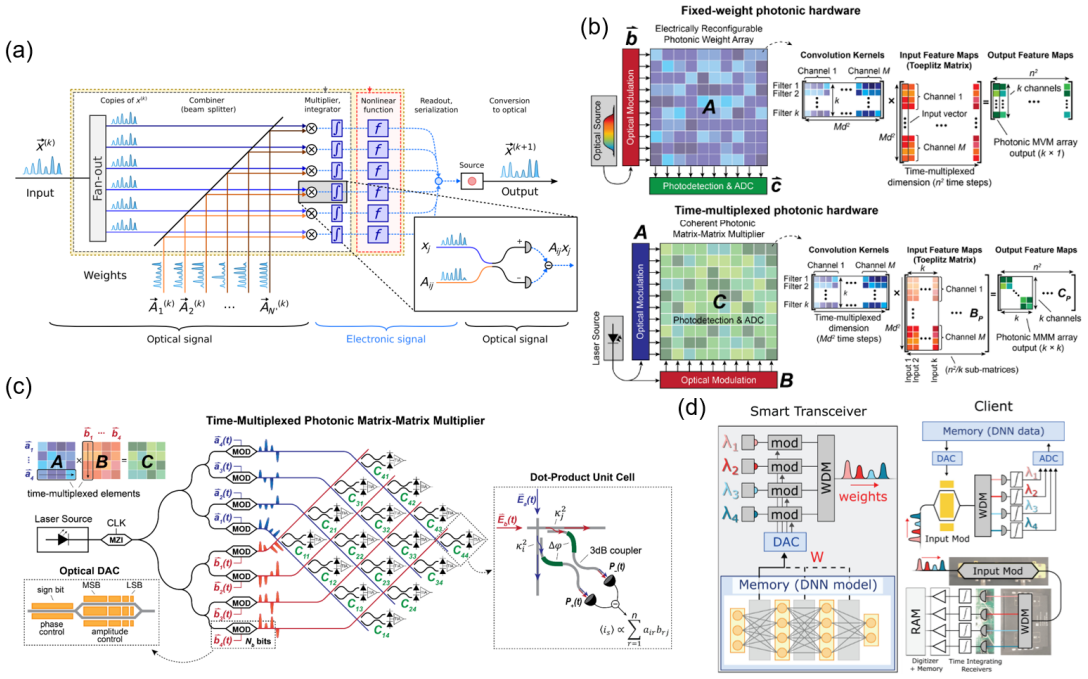

Figure 2.14: Schematics for time-multiplexed photonic computing. (a) Schematic of a single layer of homodyne optical neural network. Reprinted from [118]. (b) Comparison between fixed-weight photonic hardware and time-multiplexed photonic hardware. Adapted from [102]. (c) Proposed layout for integrated time-multiplexed photonic matrix-matrix multiplier. Reprinted from [102]. (d) Matrix-vector multiplication based on time-integrating receivers for edge computing. Adapted from [119].

This time-multiplexed architecture has been implemented in free-space systems

exploiting vertical-cavity surface-emitting laser (VCSEL) arrays [116] as coherent sources. Initial explorations with on-chip systems have also been proposed [102]. By replacing beam splitters with 3 dB directional couplers and employing the crossbar array architecture [Figure 2.14(c)], integrated unit cells have been experimentally demonstrated. However, this design requires an additional pair of photodetectors for coherent detection at each crossing of the crossbar array, which are implemented by free-space image sensors. The reliance on free-space components limits the scalability of the integrated system [102, 117]. There are also scalability challenges that arise from optical losses, phase noise, and the accuracy of the coupler splitting ratio or power distribution [117].

A comparison of energy and latency performance between the time-multiplexed architecture and the fixed-weight architecture has been discussed in [102]. When the input matrix size exceeds the crossbar size, time-multiplexed architectures achieve lower latency and higher energy efficiency compared to fixed-weight architectures, attributable to the absence of weight matrix reprogramming. Taking advantage of time-integrated detections, the time-multiplexed architecture enables a theoretical energy efficiency of ~ 10 fJ/MAC or > 300 TOPS/W for a 64×64 crossbar array operating at $f = 12\ GHz$ and 5-bit precision [102], while the experimental prototype operates at 1 kHz with a precision less than 4 bits [102].

Such time-multiplexed architectures can be further extended to support edge computing applications [119]. In a photonic edge computing architecture called Netcast, the cloud server optically encodes the weights of deep neural networks in time and wavelength, and periodically broadcasts the optical weight signals to edge devices ('clients') through smart transceivers. Local clients apply input activation data to related weight signals and perform time-integration operations for each frequency to calculate the final MAC results [Figure 2.14(d)]. This approach shifts the resource-hungry weight-training process to the cloud and enables ultra-efficient inference applications (<1 photon per multiply) at the edge.

Apart from architectures exploiting time-multiplexing and frequency-multiplexing, emerging integrated implementations to leverage synthetic frequency (synthetic frequency modulation enabled by acousto-optic modulators [120] or ring modulators [121]), multimode [23, 122], and polarization dimensions [16, 123, 124] continue to push the bandwidth limits of photonic computing.

With respect to applications, other than neural networks, optimization problems are another area in which photonic computing has been extensively investigated. Ising machine, the physical system that makes use of the well-known mathematical model in statistical mechanics, provides a judicious method for solving optimization problems. The Hamiltonian of a 2-dimensional N-spin Ising model without an external field could be described as $H = -\sum_{ij}^{N} J_{ij}\sigma_i\sigma_j$, where J_{ij} is the coupling coefficient between spin i and j, while σ_i and σ_j indicate their states in +1 or -1 [125]. A large number of optimization problems can be converted into finding the solution to this kind of equation [126]. When implementing the Ising model in a physical system with customized coupling coefficients, the minimum energy state or the ground state of the system intrinsically provides the solution to the target optimization problem. In recent years, both fiber-based [125, 127, 128] and on-chip [129–131] implementations have been proposed to solve optimization problems such as MAX-CUT problems, demonstrating systems with up to 2000 nodes. Higher node counts have been achieved with SLM-based free-space implementations, supporting 4×10^4 all-to-all coupling spins [132, 133].

In addition to the Ising machine, physics solvers based on metasurfaces [134], inverse design [135, 136], and reservoir computing [137, 138] are also under extensive explorations but are beyond the scope of this review.

2.2.6 Figures of Merit

Similarly to the general metrics used for comparing different electrical in-memory computing platforms, four commonly used metrics are taken into consideration to

benchmark the main technical routes discussed above for implementing integrated photonic computing platforms [13, 89, 139]: computing density (TOPS/mm^2), energy efficiency (TOPS/W), latency (ns or ps) and throughput (TOPS).

Here, computing density is the key metric to quantify the trade-off between the relatively large footprint associated with photonic devices, as compared to electronic counterparts, and their inherent parallelism advantage (higher TOPS). Latency and throughput are the two metrics that have the largest impact on the performance of computing systems in practical applications.

Unlike electrical in-memory computing, where wafer-scale and full-system implementations are ready, the implementations of integrated photonic in-memory computing systems are still limited within small-scale arrays (up to 6×6 in academia [75] and 128×128 in industry [140]). Thus, the performance benchmarking is more of scaling performance estimation based on device performance. A comparison of the aforementioned metrics between different photonic and electronic computing systems is provided in Figure 2.15 [13].

The integrated photonic computing system is considered to have M input and N output physical channels, Q parallel input vectors for each channel, working at an input sampling rate f with fixed 4-bit precision [Figure 2.15(a)-(b)]. Thus, throughput and latency can be defined as the following: $Throughput = 2fMNQ$ and $Latency = L_{eff}/c$ where L_{eff} represents the effective optical path length and c is the speed of light in vacuum [13]. Without capacitive delay, which undermines the performance of electrical in-memory computing systems, MZI, MRR, and PCM crossbar array-based implementations all provide sub-ns latency. Experimentally, 410 ps latency has been achieved for photonic neural network implementations, consisting of 3 layers of linear multiplications performed by a MZI mesh-based 6×6 weight matrix and electro-optical nonlinear operations [75].

For moderate scaling ($M = N = 10$, $Q = 4$ and $f = 10\ GSa\ s^{-1}$), the throughput of the photonic computing system is limited by the small scale (M,

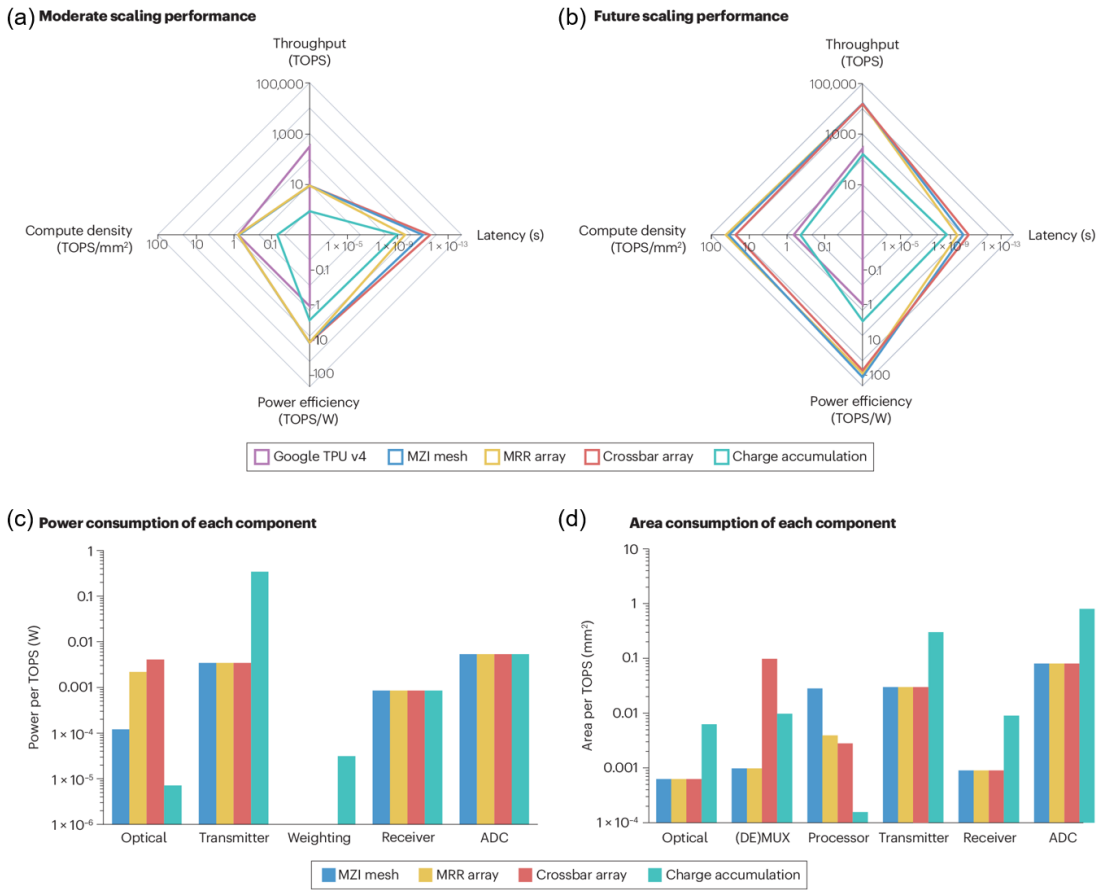

Figure 2.15: Scaling performance comparison between different technical routes for integrated photonic computing. Adapted from [13]. (a) Figure of merit at moderate scaling: $M = N = 10$, $Q = 4$ and $f = 10\ GSa\ s^{-1}$. (b) Figure of merit at future scaling: $M = N = 100$, $Q = 16$ and $f = 50\ GSa\ s^{-1}$. Here the latency is calculated for each MVM operation, and charge accumulation represents the time-integration technical route. (c) Power consumption and (d) Area consumption of each component at future scaling. Optical refers to the optical power required to drive photonic processing and the area for optical sources.

N, and Q) of the computing core. Nevertheless, the power efficiency of photonic implementations has already surpassed the benchmark electrical chip, Google TPU v4 [141, 142], with a comparable computing density. With future scaling up of photonic computing cores to $M = N = 100$, $Q = 16$ and $f = 50\ GSa\ s^{-1}$, there is potential to further improve energy efficiency and computing density by one order of magnitude. Importantly, electrical components (transmitter, weight holding, ADC and receiver) contribute a larger portion to power and area breakdowns of the photonic computing system than optical components (optical sources, (DE)MUX

and processor) [Figure 2.15(c)-(d)]. Analog-digital conversions, which dominate the power and area consumption of electrical peripheral circuits are thus the bottleneck for both electrical [6] and photonic in-memory computing systems.

It is worth noting that the performance comparisons discussed in this section and Section 2.1.3 are mainly about the inference of neural network applications, or more specifically the programmable MAC computing operations. The other important phase, training or learning phase of neural network applications, still lacks a clear benchmark for comparing different distinct training methods. Moreover, for most photonic computing implementations, the training phase still relies on digital computers (in silico). From a device perspective, the refreshing rate and programming energy are two typical metrics to describe the performance for the training phase. Less explored metrics, such as cascadability, are also important for future scaling and practical applications.

2.3 The Need for Integrated Optoelectronics

Photonic computing systems offer new opportunities to enhance computing performance, particularly in terms of low latency, high bandwidth, and energy-efficient passive transmission. Notwithstanding, optoelectronics are still indispensable for practical photonic computing applications, since nonlinear functions and general-purpose computing for data storage, retrieval, weight training, and system control are still dominated by conventional digital computers. The performance of such OE/EO conversion components (e.g. bandwidth, conversion efficiency, precisions) predominantly determines the performance (e.g. speed, energy efficiency, signal-to-noise ratio) of the overall photonic computing system.

This section will focus on reviewing key optoelectronic components of the integrated photonic computing system, especially on a silicon platform, and analyzing the challenges and possibilities for future optimization.

2.3.1 Volatile Optoelectronics

One type of prevalent optoelectronic device for electrical-to-optical conversion is the integrated optical modulator, which has been commonly used in technical routes based on coherent nanophotonics and microring weight banks [73, 75, 92, 93]. With these devices, electrical signals modulate the amplitude, phase, or polarization of input light signals, and in some cases the weights of the neural network.

An early review has surveyed silicon-based optical modulators [143]. Given the absence of second-order nonlinearities in centrosymmetric silicon crystals [144], silicon-based electro-optical modulators primarily employed either slow thermal-optic effect [145] or carrier-related principles, such as carrier accumulation, carrier injection, and carrier depletion. Among them, an operating speed exceeding 40 $Gbit\ s^{-1}$ has been demonstrated for carrier-depletion-based modulators [146]. Alternatively, heterogeneous integration with other materials pushes forward the performance of silicon modulators. Commercialized SiGe-based electro-absorption modulators (EAM) have exploited the configurable electro-absorption coefficient of germanium (Ge) and achieved a 3dB bandwidth beyond 50 GHz [147]. Moreover, recent progress in the integration of silicon platforms with emerging materials, such as lithium niobate, offers promising metrics, supporting $> 100\ Gbit\ s^{-1}$ data rates, 70 GHz bandwidth, and 2.5-dB insertion loss [148, 149].

In addition to modulators, optoelectronic components are also critical for post-processing. Instead of end-to-end systems processing optical signals only, most of the technical routes mentioned above rely on bench-top photodetectors to convert the output light intensity into electrical signals for detection. On-chip photodetectors provide a better signal-to-noise ratio, and thus lower energy consumption for integrated photonic computing architecture.

The widely used on-chip photodetectors are based on waveguide-integrated germanium photodiodes [75, 92, 150]. Emerging routes include heterogeneous

integration with 2D [151] or quasi-2D (e.g. tellurium [152]) materials, providing high-speed (>40 GHz), ultra-wide bandwidth (from visible to short-wave infrared) detection.

2.3.2 Phase-Change Optoelectronics

Combined with photonic structures, non-volatile materials, such as phase-change materials, enable energy-efficient implementations and novel applications for optoelectronic devices.

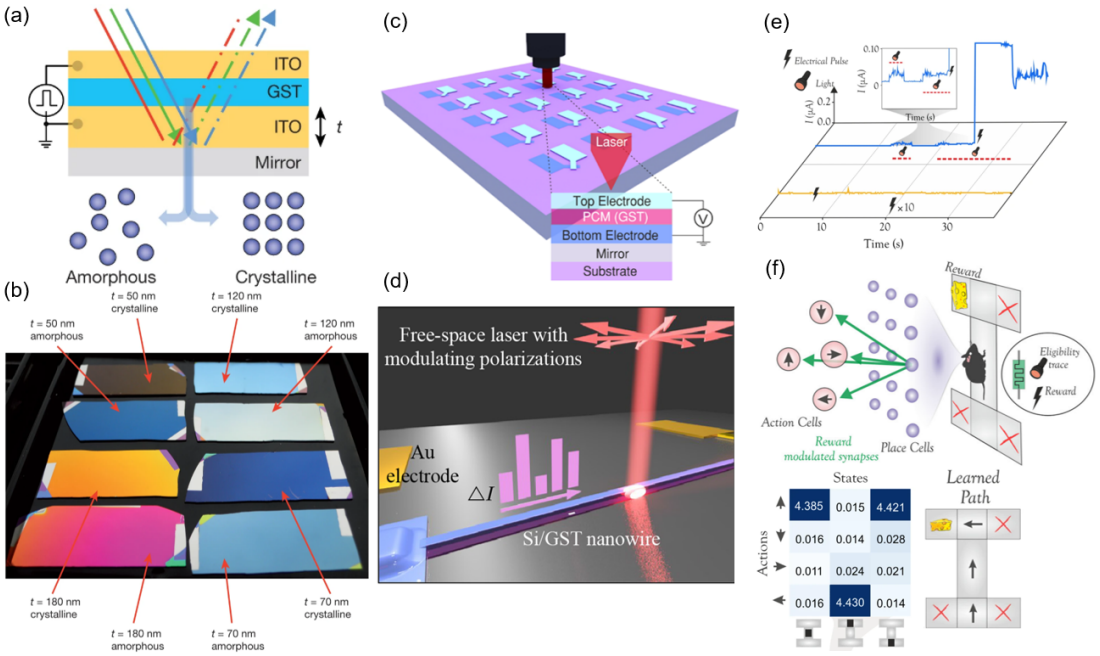

Figure 2.16: Phase-change material-based free-space optoelectronics. (a)-(b) A layered structure comprising thin-film materials, ITO/GST/ITO, with different GST thicknesses (t) and states showing different reflected colors. Adapted from [153]. (c) Crossbar phase-change devices for photo-response measurements. Reprinted from [154]. (d) Device schematic for polarization-selective switching. Reprinted from [16]. (e)-(f) Mixed-mode behavior of a phase-change device for solving navigation tasks. Without light illumination, electrical pulses cannot trigger a switching event. The rodent in the maze learns the correct route from the initial position to the cheese reward through reinforcement learning, with phase-change devices emulating the learned weights (in μS). Adapted from [155].

Early implementations are mostly based on devices with free-space light configurations. Electrical pulses applied to electrodes via electrical probes [153] or thermal heaters [156] induce phase transitions of the thin-film phase change material, thus

modulating the reflection or transmission of the devices [Figure 2.16(a)]. Such structures offer reconfigurable display colors [Figure 2.16(b)], and provide an effective approach to implement the reconfigurable coating layer for smart windows, which can adjust the reflection of solar radiation in different seasons while maintaining visible transmission [156]. When designed as microstructures, namely active metasurfaces, phase-change materials can be electrically programmed for versatile functions, such as tunable filters, lenses, and beam-steering devices [157].

Similarly, light illumination on phase-change devices can also modulate their electrical properties [Figure 2.16 (c)-(d)]. Apart from phase-dependent photo-detection behaviors [154], where the electrical current detects the light intensity, multiple dimensions of light properties, such as polarization [16], can also be reflected in the electrical readouts of phase change devices. Combining the two modulation factors, light illumination and electrical pulses, phase change devices can be configured as more complicated bio-inspired components, such as multi-factor learning units [155]. With light and electrical triggers indicating eligibility and rewards in a reinforcement learning process, coexistence of both triggers programs the phase-change devices to dedicated weights which reveal the correct solution in a navigation task [Figure 2.16 (e)-(f)].

On the contrary, it is only recently that integrated electro-optical phase-change devices have been proposed [103–105, 158–162], given the challenge of size mismatch between integrated electrical and optical phase-change devices as discussed in Section 1.1.2.

An approach to addressing this problem is to enhance light-matter interactions at the nanometer scale, such as exploiting plasmonic structures [158, 159]. By placing the gold electrodes closely (with a nanogap of tens of nanometers) on a Si_3N_4 photonic platform, the strong field confinement of the metal-insulator-metal (MIM) structure enables dual electrical-optical functionality, i.e. programming and reading of the devices in both electrical and optical domains [Figure 2.17 (a)-(b)].

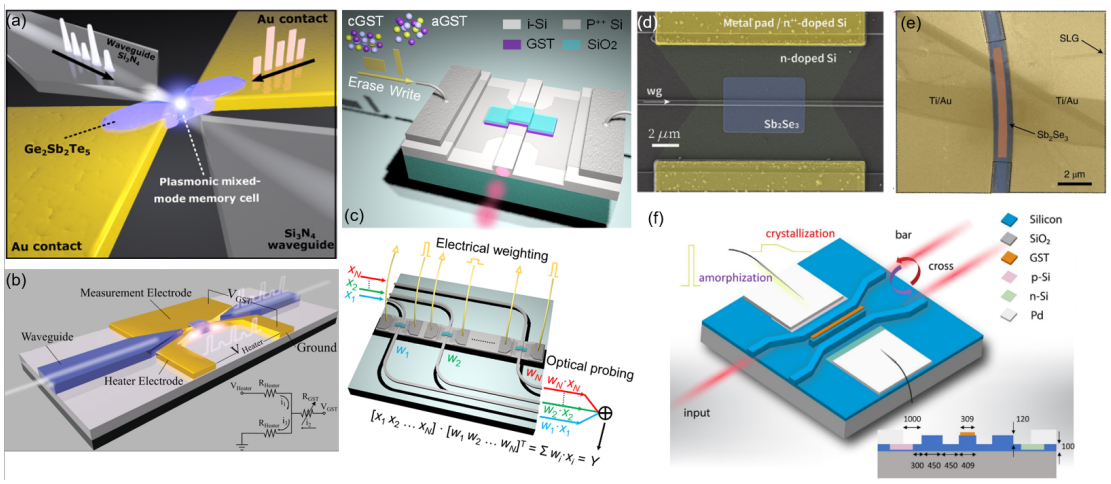

Figure 2.17: Integrated Phase-Change Optoelectronics. (a) Schematic for plasmonic nanogap phase-change devices. Adapted from [158]. (b) Schematic for plasmonic phase-change nanoheater devices. The electrical heating pathway and the readout pathway are separated to achieve larger switching volume. Reprinted from [159]. (c) Micro-heater based photonic-electronic dot-product engine. Adapted from [160]. (d) Sb_2Se_3-based phase-change phase shifter using a doped-silicon micro-heater. Reprinted from [105]. (e) Graphene-assisted phase-change phase shifter. Reprinted from [162] (f) 2×2 programmable unit based on GST. Reprinted from [103].

This design has demonstrated energy consumption in the range of tens of picojoules for both electrical and optical programming at a highly compact device scale (active area of 100 $nm \times 50$ nm). This is a very promising approach to construct unit cells for energy-efficient electro-optical computing systems, yet the high insertion loss ($\sim$ 10 dB) and stringent requirements for fabrication and alignment process pose challenges for future scaling.

Alternatively, another recently emerged approach combines external heaters with silicon waveguides [Figure 2.17 (c)-(e)]. Dedicated heater designs exploit various materials such as doped silicon [103–105, 160, 161] and graphene [162] to provide uniform Joule heating for phase change materials deposited above, allowing effective electrical switching for large volumes. This approach shows advantages in scalability. It can be CMOS compatible and scaled up to implement practical computing tasks such as imaging processing and recognition [160]. On the other hand, dual electrical-optical functionality has not been achieved in these devices

because their electrical readouts are independent of the phase-change material state. These devices also require larger footprints (several to hundreds of μm per unit cell) and higher electrical switching energy (several to tens of nJ) [103–105, 160–162]. Aside from acting as weight banks in a photonic computing core, heater-based phase-change photonic devices can also function as non-volatile tunable couplers and phase shifters [Figure 2.17 (f)], paving the way for energy-efficient low-loss optical interconnect systems [103, 105, 163, 164].

3

Techniques and Methods

Contents

This chapter introduces the theory, general fabrication process, and measurement setups for photonic phase-change devices.

3.1 Basic Photonic Structures

A typical silicon photonic phase-change device consists of several commonly employed elements, including grating couplers to couple light from off-chip fibers to on-chip waveguides, dielectric waveguides to propagate light signals, and dielectric or plasmonic waveguides for interaction with phase-change materials to modulate light signals. This section introduces the theory behind these components.

3.1.1 Dielectric Waveguides

The function of dielectric waveguides in photonics is similar to that of electrical wires in integrated electrical circuits. Waveguides connect different functional units, with their material composition and geometry defining the light propagation property. The theoretical concepts presented in this section have been derived from [165].

The working principle of waveguides is similar to that of optical fibers, with both components guiding light through total internal reflection (TIR). Waveguides are usually made up of two types of materials, namely the core (with a refractive index of n_1) and the cladding (n_2 and n_3 or more), where the refractive index of the core is higher than the cladding ($n_1 > n_2, n_3$).

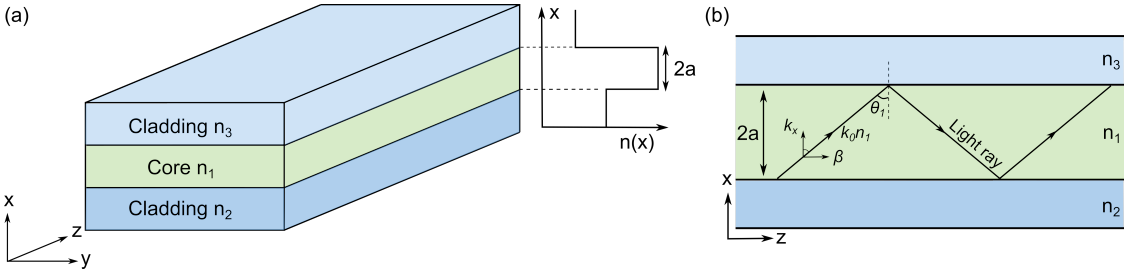

Figure 3.1: Schematic of a planar waveguide, extending infinitely in $\hat{z}$ and $\hat{y}$ direction and confined in $\hat{x}$ direction with a thickness of 2a. (a) Schematic of a planar waveguide. Inset: index profile (n(x)) of the different dielectric layers. (b) Light rays in the planar waveguide, indicating the propagation angle θ_1 and propagation constant β.

Firstly, a simple planar waveguide is considered [Figure 3.1(a)]. The electric ($\mathbf{E}$) and magnetic ($\mathbf{H}$) fields of propagating modes in the waveguide can be derived from Maxwell's equations. For a source-free, linear, isotropic dielectric waveguide, Maxwell's equations can be written as follows:

$$\begin{aligned} \nabla \times \mathbf{E} &= -\mu_0 \mu \frac{\partial \mathbf{H}}{\partial t} \\ \nabla \times \mathbf{H} &= \varepsilon_0 \varepsilon \frac{\partial \mathbf{E}}{\partial t} \\ \nabla \cdot \mathbf{E} &= 0 \\ \nabla \cdot \mathbf{H} &= 0 \end{aligned} \tag{3.1}$$

where μ_o and ε_0 represent the magnetic permeability and electric permittivity in vacuum, respectively. Relative permeability $\mu = 1$ in the non-magnetic dielectric medium, and relative permittivity ε, also known as the dielectric constant, can be expressed as $\varepsilon = n_i^2$ $(i = 1, 2, 3)$ for lossless dielectrics.

Assuming that the light wave, with time-harmonic fields, propagates in the z-direction, a plane wave solution for the fields can be expressed as the following form:

$$\begin{aligned} \mathbf{E}(x, y, z, t) &= \mathbf{E}(x, y)e^{j(\omega t - \beta z)} \\ \mathbf{H}(x, y, z, t) &= \mathbf{H}(x, y)e^{j(\omega t - \beta z)} \end{aligned} \tag{3.2}$$

where $\beta = 2\pi n_{eff}/\lambda = 2\pi n_1 sin\theta_1/\lambda$ is the propagation constant of the wave, with θ_1 being the light propagation angle with respect to the x axis [Figure 3.1(b)] and λ the wavelength in the vacuum.

Since the planar waveguide structure is infinite in the $\hat{y}$ direction and is only confined in the $\hat{x}$ direction, the electrical field is invariant in $\hat{y}$, i.e. $\partial\mathbf{E}/\partial y = 0$ and $\partial\mathbf{H}/\partial y = 0$. Applying this condition to equations (3.1)-(3.2), the components of the electric and magnetic field can be obtained as follows:

$$\begin{aligned} j\beta E_y &= -j\omega\mu_0 H_x & j\beta H_y &= j\omega\varepsilon_0 n_i^2 E_x \\ -j\beta E_x - \frac{\partial E_z}{\partial x} &= -j\omega\mu_0 H_y & -j\beta H_x - \frac{\partial H_z}{\partial x} &= j\omega\varepsilon_0 n_i^2 E_y \\ \frac{\partial E_y}{\partial x} &= -j\omega\mu_0 H_z & \frac{\partial H_y}{\partial x} &= j\omega\varepsilon_0 n_i^2 E_z \end{aligned} \tag{3.3}$$

The equations can be further rearranged:

$$\begin{aligned} (k^2 - \beta^2)E_x &= -j\beta\frac{\partial E_z}{\partial x} & (k^2 - \beta^2)H_x &= -j\beta\frac{\partial H_z}{\partial x} \\ (k^2 - \beta^2)E_y &= j\omega\mu_0\frac{\partial H_z}{\partial x} & (k^2 - \beta^2)H_y &= -j\omega\varepsilon_0 n_i^2\frac{\partial E_z}{\partial x} \end{aligned} \tag{3.4}$$

where $k^2 = \omega^2\mu_0\varepsilon_0 n_i^2 = k_0^2 n_i^2$ $(i = 1, 2, 3, k_0 = 2\pi/\lambda)$.

The wave equations of two independent electromagnetic modes of particular interest can be derived from equations (3.3)-(3.4), denoted as transverse electric (TE) mode and transverse magnetic (TM) mode, respectively:

$$\frac{\partial^2 E_y}{\partial x^2} + (k_0^2 n_i^2 - \beta^2) E_y = 0 \ (TE \ mode) \tag{3.5}$$

where $E_z = 0$, thus $E_x = H_y = 0$.

$$\frac{\partial^2 H_y}{\partial x^2} + (k_0^2 n_i^2 - \beta^2) H_y = 0 \ (TM \ mode) \tag{3.6}$$

where $H_z = 0$, thus $H_x = E_y = 0$.

To support guided modes, for which the fields are confined in the core and exponentially decay in the cladding, the condition $k_0^2 n_1^2 > \beta^2 = k_0^2 n_{eff}^2 > k_0^2 n_i^2$ $(i = 2, 3)$ must be fulfilled. In addition, the boundary conditions require that both the electric field $(E_y(x))$ and its derivative $(\partial E_y(x)/\partial x)$ are continuous at the boundaries $(x = \pm a)$, providing discrete solutions for the effective index (n_{eff}). Under these conditions, equations (3.5)-(3.6) can be solved to compute the field distributions in the waveguide.

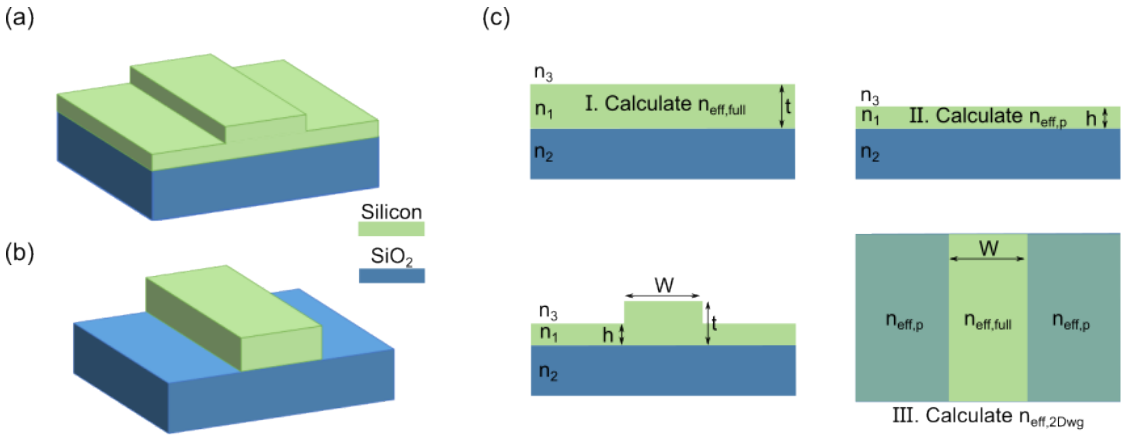

Figure 3.2: Geometries of 2D waveguides based on silicon-on-insulator substrates. (a) Rib waveguides (with a partially-etched silicon layer). (b) Ridge/slab waveguides. (c) Using effective index method (EIM) to decouple the 2D geometry of a rib waveguide into horizontal planar waveguides (steps I and II) and one vertical planar waveguide (step III). The effective indices calculated in the horizontal planar waveguides are used as the updated refractive indices to calculate the effective index of the vertical planar waveguide in step III, providing an approximate value for the effective index of the rib-waveguide.

Confining the modes in both $\hat{x}$ and $\hat{y}$ forms 2D waveguide geometries. The most commonly used are rib and ridge/slab waveguides [Figure 3.2(a)-(b)]. Specifi-

cally, waveguides based on silicon-on-insulator substrates have been used in this thesis.

With the refractive index and thus the electrical fields varying along both directions, the analytical solutions mentioned above are no longer applicable. An approximate solution can be obtained from the effective index method (EIM)[166]. By decoupling the two-dimensional geometry into several vertical and horizontal planar waveguides, the effective index can be calculated first in one direction and used as an updated refractive index for the next calculation [Figure 3.2(c)]. However, this approximate method can become complicated and inaccurate with complex geometries.

Therefore, numerical methods are exploited to calculate the mode profile in this thesis, using an *Ansys Lumerical* or *COMSOL multiphysics* mode solver.

3.1.2 Grating Couplers

Edge couplers and grating couplers [167] are two commonly used forms to couple light from optical fibers to on-chip waveguides. Edge coupling provides high coupling efficiency and broadband bandwidth; however, the accessibility of the devices is physically limited to the edge of the chip. To facilitate access in experiments, grating couplers are employed in this thesis.

Figure 3.3(a) illustrates the cross-sectional schematic of a grating coupler. It consists of a tailored periodic array of partially etched grating structures to diffract the off-plane light to the waveguide propagation direction, or vice versa. According to the Bragg condition [168], the following phase matching equation must be satisfied to add the diffracted light ($n_i k_0 sin\theta_i$) constructively along the waveguide propagation direction ($\beta = 2\pi n_{eff}/\lambda$):

$$n_i k_0 sin\theta_i + \frac{2\pi m}{p} = \beta \tag{3.7}$$

where m is the diffraction order.

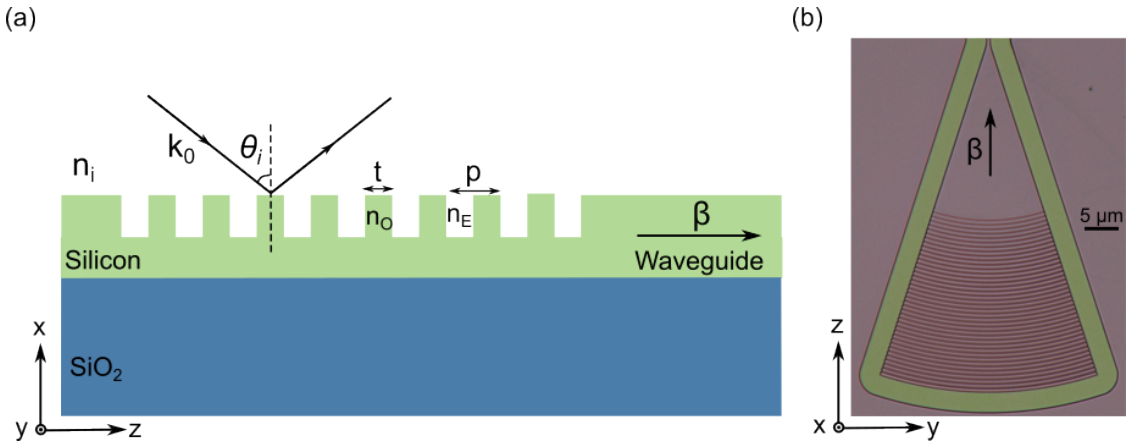

Figure 3.3: Schematic of grating couplers. (a) Cross-sectional schematic of a uniform grating coupler. θ_i: incident angle; n_O, n_E: effective index of the uneteched tooth and etched trench, respectively. (b) Top view optical microscope image of a patterned apodised grating coupler. The fill factor transitions from a lower value at the bottom to a higher value at the top along the $\hat{z}$ axis.

Thus, considering air as the top cladding layer($n_i = 1$), and taking $m = 1$ (first diffraction order), the period of the grating (p) is given by

$$p = \frac{\lambda}{n_{eff} - sin\theta_i} \tag{3.8}$$

where $n_{eff} = ff \cdot n_O + (1 - ff) \cdot n_E$ (fill factor (ff)=t/p) for uniform gratings.

In this thesis, apodised grating couplers [169] (specifically grating couplers with gradually varying fill factors) have been used to improve coupling efficiency [Figure 3.3(b)], and the incident angle is chosen as an optimum angle of $\theta_i = 8°$ [166]. For future work, emerging topological light-guiding technologies [170] and 3D couplers [101] can be further exploited to enhance coupling efficiency and expand coupling bandwidth.

3.1.3 Plasmonic Waveguides

As discussed in Section 1.1.2 and Section 2.3.2, the size of dielectric waveguides (hundreds of nanometers to several micrometers) is larger than that of electronic devices. To address the size-mismatch challenge for integrated optoelectronics, plasmonic waveguides provide a compact solution by enabling highly confined

modes [158, 159]. In this section, the operating principles of plasmonic waveguides are introduced, with the theoretical concepts derived from [171].

The optical properties of metals are largely determined by the behavior of their free electrons. With an external field $\mathbf{E}$, free electrons can be treated as oscillators driven by the electric field and damped through collisions. The equation of motion for free electrons can be expressed as:

$$m\frac{\partial^2 \mathbf{r}}{\partial t^2} + m\gamma\frac{\partial \mathbf{r}}{\partial t} = -e\mathbf{E} \tag{3.9}$$

where e and m are the charge and the effective optical mass of each electron, respectively, and $\gamma = 1/\tau$ (typically ~ 100 THz at room temperature) is the collision frequency with τ known as the relaxation time of the free electron gas.

Assuming a time-harmonic driving field $\mathbf{E}(t) = E_0 e^{-i\omega t}$, a time-harmonic solution $\mathbf{r}(t)$ to describe the displacement of the electrons can be solved as:

$$\mathbf{r}(t) = \frac{e/m}{\omega^2 + i\gamma\omega}\mathbf{E}_0 e^{-i\omega t} \tag{3.10}$$

Taking into account the density of the electrons (N), the macroscopic polarization $\mathbf{P}$ is then given by

$$\mathbf{P} = -Ne\mathbf{r} = -\frac{Ne^2/m}{\omega^2 + i\gamma\omega}\mathbf{E} \tag{3.11}$$

Meanwhile, the electric displacement field ($\mathbf{D}$) in linear, isotropic media can be described as:

$$\mathbf{D} = \varepsilon_0\mathbf{E} + \mathbf{P} = \varepsilon_0\varepsilon\mathbf{E} \tag{3.12}$$

Combining equations 3.11 and 3.12, the frequency-dependent dielectric function (also known as the Drude model) to describe the dispersive properties of metals is obtained as:

$$\varepsilon(\omega) = 1 - \frac{Ne^2/m\varepsilon_0}{\omega^2 + i\gamma\omega} = 1 - \frac{\omega_p^2}{\omega^2 + i\gamma\omega} \tag{3.13}$$

where $\omega_p = \sqrt{Ne^2/\varepsilon_0 m}$ is the plasma frequency. The complex dielectric function $\varepsilon(\omega) = \varepsilon_1(\omega) + i\varepsilon_2(\omega)$ can be further rearranged to derive the real and imaginary components:

$$\begin{aligned} \varepsilon_1(\omega) &= 1 - \frac{\omega_p^2}{\omega^2 + \gamma^2} \\ \varepsilon_2(\omega) &= \frac{\omega_p^2}{\omega\gamma + \omega^3/\gamma} \end{aligned} \tag{3.14}$$

Considering a metal-dielectric interface, the electromagnetic fields can be coupled to the oscillations of conduction electrons and excite electromagnetic surface waves, also known as surface plasmon polaritons (SPP), which propagate along the interface and are evanescently confined in the perpendicular direction. The simple geometry of a flat interface between a dielectric half-space ($x > 0$) and a metal half-space ($x < 0$) is shown in Figure 3.4(a).

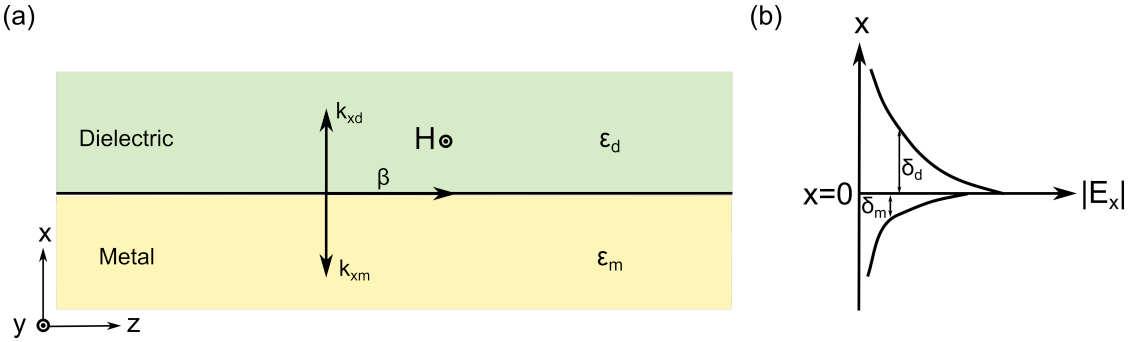

Figure 3.4: Surface plasmon polaritons (SPP) propagates along a metal-dielectric interface. (a) Geometry of a metal-dielectric interface. The SPP surface wave propagates along the $\hat{z}$ direction. k_{xd} and k_{xm} are x components of the wave vector in the dielectric and metal space, respectively. The TM mode exists with the magnetic field $\mathbf{H}$ in the $\hat{y}$ direction. (b) Electrical field distribution along the x axis of a TM mode.

The electric and magnetic field distribution of both planes in the form of $\mathbf{E}(x, y, z) = \mathbf{E}(y)e^{\pm k_x x}e^{-j\beta z}$ and $\mathbf{H}(x, y, z) = \mathbf{H}(y)e^{\pm k_x x}e^{-j\beta z}$ can be solved following the same procedure as in Section 3.1.1, using equation sets 3.3 and 3.6. The TM

mode profile is graphically shown in Figure 3.4(b), with the dispersion relation obtained as :

$$\beta = k_0\sqrt{\frac{\varepsilon_m\varepsilon_d}{\varepsilon_m+\varepsilon_d}} = \frac{\omega}{c}\sqrt{\frac{\varepsilon_m\varepsilon_d}{\varepsilon_m+\varepsilon_d}} \qquad (\frac{k_{xd}}{k_{xm}} = -\frac{\varepsilon_d}{\varepsilon_m}) \tag{3.15}$$

The field decays exponentially away from the interface, with decay lengths $\delta_d = 1/|k_{xd}|$ ($\sim \lambda/2$) and $\delta_m = 1/|k_{xm}|$ in the dielectric and metal domains, respectively. The propagation (attenuation) length of the mode is defined as $\delta_z = (2\beta'')^{-1}$ with β'' being the imaginary component of the propagation constant. For plasmonic modes, long propagation length and strong field confinements are preferred, thus δ_z/δ_d is the key metric for plasmonic devices.

The above metal-dielectric interface can be extended to form infinite multilayer metal-dielectric-metal waveguides [Figure 3.5(a)] where the above analysis is still applicable. The modes of the two interfaces interact when the width of the core material (2a) is smaller than the decay length (δ_d). In this thesis, the plasmonic waveguide structure is designed with more complex geometries, thus numerical simulations are exploited to solve the field distribution. Simulated mode profiles of a metal-air-metal structure on a SiO_2 substrate are presented in Figure 3.5(b), where smaller dielectric gap provides stronger field enhancement.

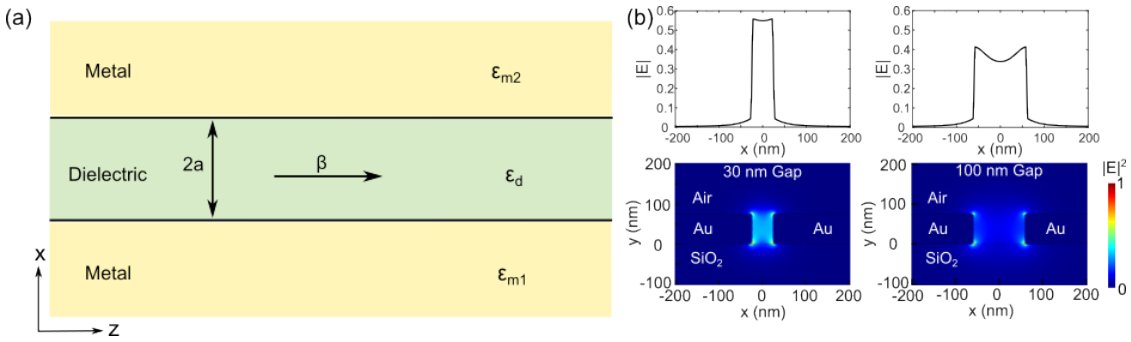

Figure 3.5: Plasmonic modes of metal-dielectric-metal waveguides. (a) Geometry of an infinite multilayer metal-dielectric-metal waveguide. (b) Simulated mode profiles (via a Lumerical mode solver) for plasmonic nanogap structures on a SiO_2 substrate. The thickness of the Au is 75 nm.

3.2 Fabrication Techniques

In this thesis, device fabrication starts with diced $10 \times 10\ mm^2$ chips on silicon-on-insulator (SOI) substrates. Specifically, the substrate consists of a 220-nm thick silicon layer above a 3-µm buried oxide.

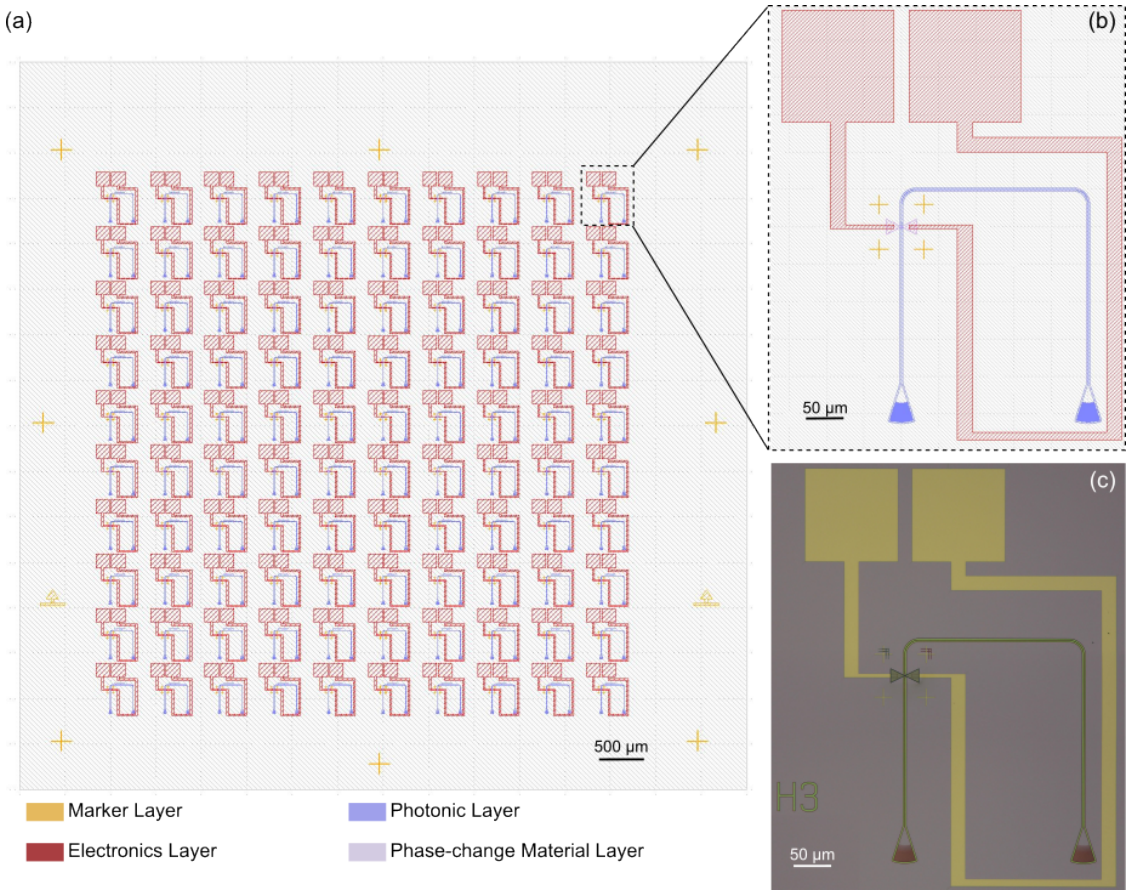

Figure 3.6: Overview of the layout used for fabrication. (a) Overview of the layout used for fabrication. Scale bar: 500 μm. (b) Zoom-in layout for a device. Scale bar: 50 μm. (c) Optical microscope image of a device. Scale bar: 50 μm.

The phase-change optoelectronic device fabrication includes steps corresponding to four main layers [Figure 3.6]: alignment markers, electrical circuits (contact electrodes), nanophotonic circuits (waveguides and grating couplers) and phase-change materials. For each layer, the layout is first designed using python with the open source gdshelpers package, then exported as a GDS-II CAD file for later electron beam lithography (EBL) fabrication steps.

The EBL process exploited in this thesis has been optimized by previous group members. The general fabrication steps for each lithography layer are as follows:

(1) **Clean and preheating.** Clean the chip by sequential immersion in acetone and isopropyl alcohol (IPA) for 5 minutes each. Ultrasonic agitation is applied for the first layer but is omitted for the subsequent layers to avoid damaging the fabricated patterns. The chip is then dried using a nitrogen gas flow. Afterwards, the chip is preheated on the hot plate and then treated with a low-power oxygen plasma to improve adhesion during photoresist coating.

(2) **Photoresist spin-coating.** EBL photoresist is spun on the chip, followed by a soft-bake step. The choice of photoresist depends on the resolution requirements and related deposition/etching steps for each layer. A single layer positive resist AR-P 6200 (CSAR 62) is used for most layers, requiring 45 second spin coating at 4000 rpm and then 3 minutes of soft baking at 150 °C. Bilayer positive resist PMMA (polymethyl methacrylate) is used for phase-change materials deposition and detailed recipes will be discussed in Section 3.2.4.

(3) **EBL exposure.** The spin-coated chip is then exposed using a 50 kV EBL system (JEOL JBX5500) at 100 pA with a dose of 180 $\mu C/cm^2$ for CSAR 62 or 700 $\mu C/cm^2$ for PMMA.

(4) **Development.** The exposed chip is then developed following a development process in accordance with the specific photoresist. Chips coated with CSAR-62 are developed in AR 600-546, methyl isobutyl ketone (MIBK) and IPA in sequence for 30 seconds, 15 seconds, and 15 seconds, respectively. The developed chip is then dried with a nitrogen gas blow.

(5) After inspection using an optical microscope, the dried chips are ready for the corresponding deposition or etching steps for each layer.

The design and fabrication considerations for each layer are discussed in the following subsections.

3.2.1 Alignment markers and exposure resolution

To improve the alignment accuracy of different layers, a two-step alignment process is followed with two types of alignment markers: global markers for chip-level alignment and local markers for device-level alignment.

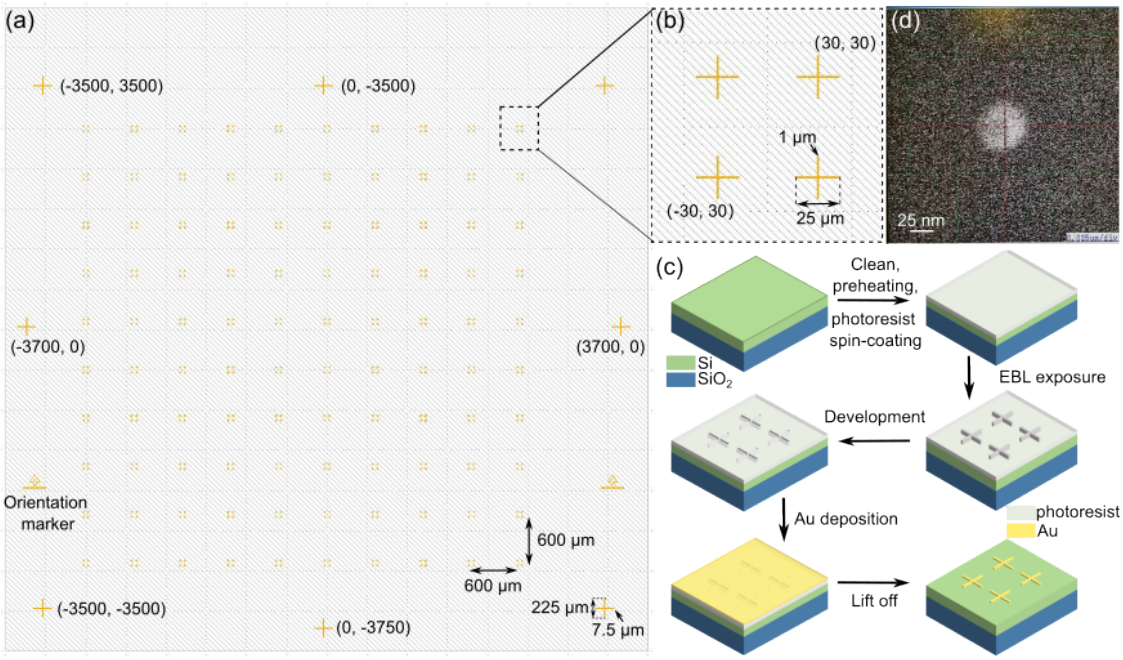

Figure 3.7: The layout of alignment markers. The units for the marker coordinates are μm. (a) The layout of global markers for chip-level alignment. The markers are designed with a length of 225 μm and a width of 7.5 μm. (b) The layout of local markers for device-level alignment. The markers are designed with a length of 25 μm and a width of 1 μm. (c) The process flow for depositing gold alignment markers. (d) SEM image for a contamination dot burnt by the focused electron beam.

Figure 3.7(a)-(b) present the layout of these two types of alignment markers. During the chip-level alignment step, a pair of global markers is selected for detection to update the chip position in the EBL coordinate system and to correct rotational

deviation. The device-level alignment is then performed for each device by detecting one of the local markers. The positions of the local markers are designed to fit within the same writing field ($100 \times 100\ \mu m^2$) of the critical device feature, such as plasmonic nanogaps and phase-change material constrictions, to avoid stage movements between marker detection and pattern exposure.

The fabrication process flow of the alignment markers is illustrated in Figure 3.7(c). After the aforementioned EBL exposure and development process, a 5 nm Cr adhesion layer and 75 nm Au are deposited on the chip by thermal evaporation. The chip then undergoes overnight lift-off in the heated remover (MICROPOSIT remover 1165).

It is worth noting that during the later phase of the DPhil, the EBL auto-focus system was malfunctioning, thus a manual focusing process was required. After the normal calibration process, the focus and astigmatism of the electron beam were manually adjusted, and contamination dots were utilized for final optimization before exposure. By focusing the electron beam on the chip slightly away from the device area for around 10 s, a bright round contamination dot can be created if the calibration condition is well optimized. An example dot with a radius of 25 nm is shown in Figure 3.7 (d), indicating a good exposure resolution.

3.2.2 Photonic Layer

To fabricate the layer of photonic circuits (waveguides, grating couplers, mode converters, and crossing structures), a reactive ion etching (RIE) step is required after the EBL exposure and development steps.

The fabrication process flow of the photonic layer is illustrated in Figure 3.8. During the RIE step, the 220-nm silicon layer is half-etched with a patterned photoresist (CSAR-62) mask. A mixture of 40 sccm CH_3 and 15 sccm SF_6 is used with 100 W RF power for 70 seconds to achieve an etch depth of 110 nm. The remaining photoresist is then removed by immersing the chip in the heated remover

(MICROPOSIT remover 1165) for ∼ 10 min. Finally, the chip is cleaned with acetone and IPA, and then dried with a nitrogen gas blow.

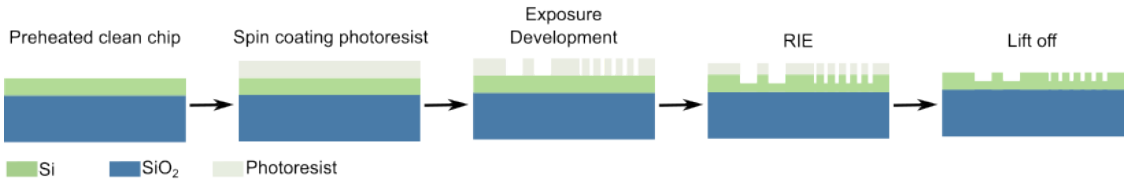

Figure 3.8: Fabrication process of the photonic layer.

It is worth noting that the photoresist mask is also etched during the RIE process, depending on the selective ratio. Thus, the thickness of the photoresist should also be taken into consideration when choosing appropriate photoresist for longer etching time. For fully etching steps exploited in the fabrication process of plasmonic devices (Sections 4.2 and 4.3.3), double layer CSAR 62 is used as the photoresist mask to ensure that the mask remains functional for the entire etching process. During the spin-coating process, the first layer CSAR is spin-coated and soft-baked for 1 minute before the spin-coating and 3-minute soft baking of the second layer. Other steps remain the same for the EBL exposure and development process.

3.2.3 Electronics Layer

The fabrication process of the electronics layer is similar to that of the alignment markers (Figures 3.7(c) and 3.9). A thermal evaporation step is exploited to deposit gold contacts or pads after the EBL exposure and development process, followed by a lift-off step.

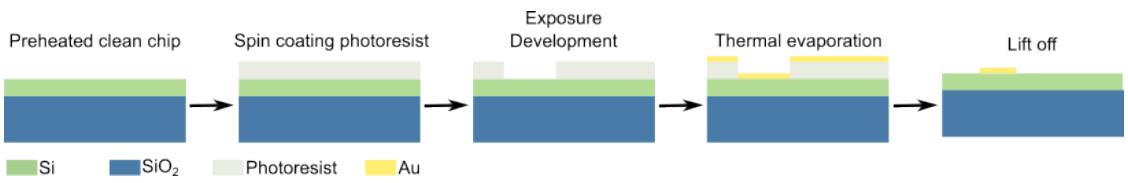

Figure 3.9: Fabrication process of the electronics layer.

The fabrication sequence of the photonic layer and the electronics layer depends on the specific requirements of different projects. For the plasmonic project (Section

4.2), the fabrication process of the Au contacts is combined with that of the alignment markers, circumventing unnecessary alignment deviations. Moreover, to fulfill the high-resolution requirements of plasmonic contacts while balancing the time efficiency of patterning masks for large gold pads (over $100 \times 100\ \mu m^2$), the process indicated in Figure 3.9 is carried out twice to fabricate the small contacts and large gold pads separately. The photoresist mask for gold pads is defined through EBL exposure at 10 nA with a larger writing field of $1000 \times 1000\ \mu m^2$. The deposition thickness of the gold pads is designed to be thicker (220 nm after a 5 nm Cr adhesion layer) to ensure good contact. Accordingly, a double-layer photoresist is used during the fabrication process to facilitate lift-off. The fabrication process for the waveguides takes place afterward.

For constriction devices (Section 5.3.2), the fabrication process for the electronics layer is performed after the photonic layer, with an electrical isolation layer of AlO_x deposited between the two layers. The AlO_x layer is deposited via atomic layer deposition (ALD, Savannah S200) without photoresist masks. Two precursors, $Al(CH_3)_3$(g) and H_2O(g), are alternatively pumped as 15 ms pulses with 20-s intervals at 150 °C. 50 cycles of such alternative pumping are exploited to achieve a deposition thickness of $\sim 5\ nm$.

3.2.4 Phase-change material layer

As explained in Section 2.1 and Section 2.2.3, the states of the phase-change material is closely related to the temperature change. Thus, the layer of phase-change material is always fabricated as the last step to avoid unintended temperature perturbation and preserve the material properties before measurements.

Figure 3.10 illustrates the fabrication process of the phase-change material layer. The phase-change material is deposited via RF magnetron sputtering, defined by a bilayer PMMA photoresist mask. Both layers of the PMMA photoresist are spin-coated at 6000 rpm for 1 minute and then soft-baked at 180 °C for 10 minutes.

After EBL exposure, the chip is developed in a solution with a 1:3 ratio of MIBK and IPA for 1 minute and IPA for 30 seconds, then dried with a nitrogen gas blow. The resolution difference between the two PMMA layers creates an undercut (reverse T-shape) photoresist profile, which facilitates clean material removal during lift-off.

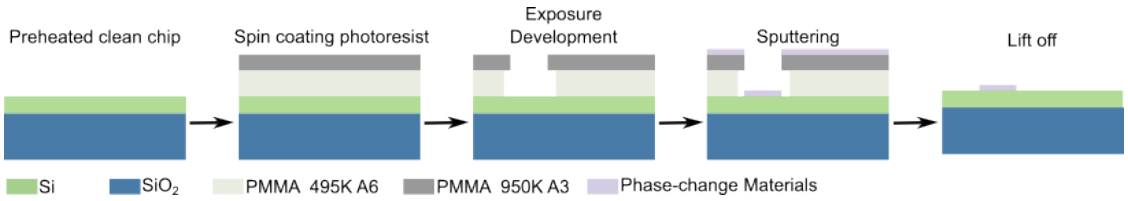

Figure 3.10: Fabrication process of the phase-change materials layer. PMMA: polymethyl methacrylate; 495K and 950K are the molecular weight of the polymeric material; A6 and A3 represent 6% and 3% dilution of PMMA in Anisole, respectively.

Two RF sputtering systems, from Nordiko and AJA International Inc, have been used during the DPhil period. The former system with a direct deposition source is preferred for the high aspect ratio structures exploited in the plasmonic project. However, it broke down. Therefore, most of the sputtering process for this thesis has been carried out with the AJA system. The AJA system is equipped with confocal sources, thus during the sputtering process, the sample is fixed towards the outer edge of the sample holder close to the source to improve the working angle.

A light cleaning step using RF bias plasma (100 W RF for 1 min at 10 $mTorr$ chamber pressure), also known as back-sputtering, is performed before sputtering to ensure good contact between the deposited material and the substrate. A phase-change material, specifically $Ge_2Sb_2Te_5$, is deposited with 30 W RF power at 3 mTorr chamber pressure. Next, a capping layer of $ZnS{-}SiO_2$ is deposited to avoid oxidation of the phase-change material, with 50 W RF power at 3 mTorr chamber pressure. Afterwards, the deposited sample is lifted off by immersion in heated acetone for around 3 hours.

3.3 Experimental Setups

This section introduces the sample stage and two experimental setups customized for optical programming and electrical programming, respectively. The details of the key equipment are also added to each section.

3.3.1 The sample stage

An optical image of the sample stage is presented in Figure 3.11. Electrical signals are sent and received through high-speed RF probes (Model 40A, GGB) contacting the pads of the devices, while optical signals are coupled to the grating couplers of the device through multi-channel fiber arrays.

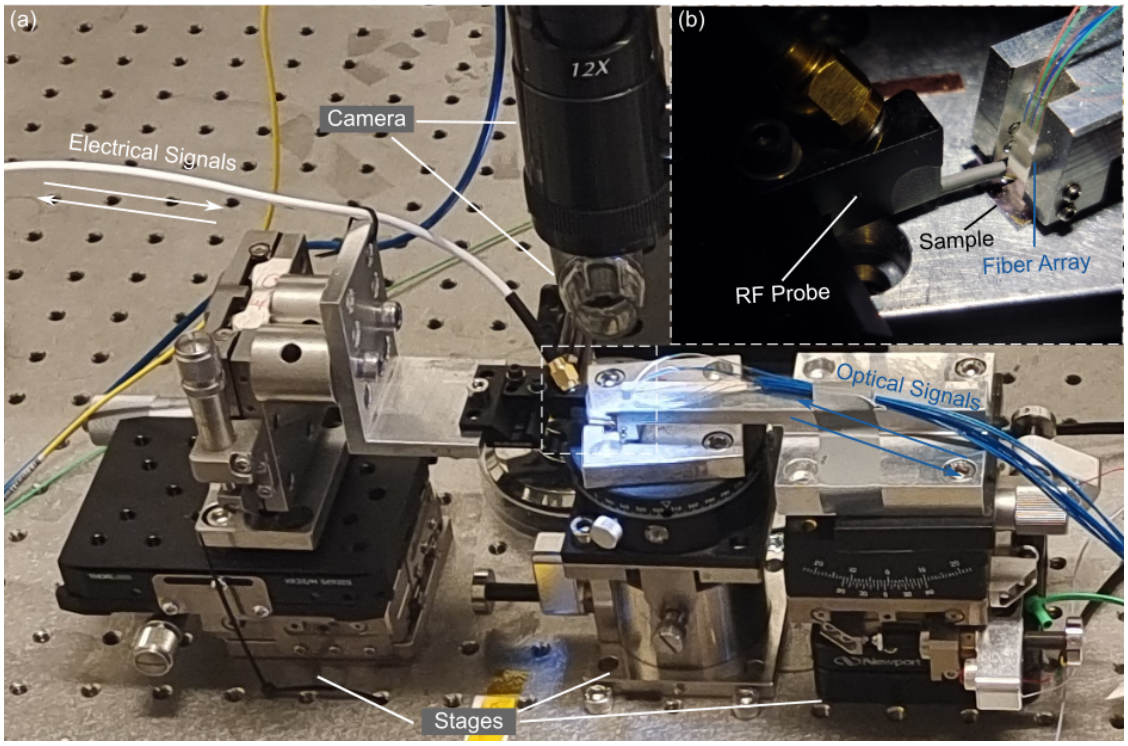

Figure 3.11: Optical images of the sample stage (a) with a zoom-in image (b) of the RF probes and fiber arrays.

Three positioning stages (from Thorlabs and OptoSigma) are used to align the RF probe, sample, and fiber arrays in the X, Y, and Z (height) directions with micrometer precision. Rotation stages are mounted on the translation stages to

correct rotational misalignment between the sample and the fiber array. A lateral camera and an overhead microscope system (a CCD camera from Thorlabs with a zoom lens system from Navitar) are used to provide visual feedback during the alignment process.

3.3.2 Optical measurement setup

Figure 3.12 illustrates the pump-probe setup [18] for optical programming with both electrical and optical readouts.

For simple transmission measurements, only the probe line (blue pathway) is required, with a probe laser, a polarizer before the device, a photodetector, and the data acquisition equipment.

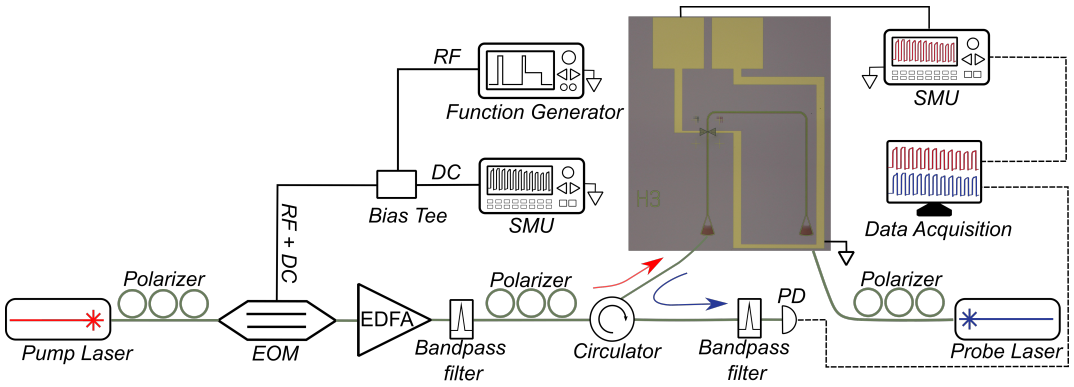

Figure 3.12: Optical switching setup with both electrical and optical readouts. SMU: source measure unit; PD: photodetector; EOM: electro-optical modulator; EDFA: erbium-doped fiber amplifier. *Inset*: optical image of a device (green: waveguide layer; gold: electrical pads).

To add programming functionality, a pump line (red pathway) is required. The pump laser was modulated with a function generator controlled electro-optical modulator (EOM) to send programming pulses. Pump pulses were amplified by an erbium-doped fiber amplifier (EDFA) before being coupled to the device. A probe laser was employed to monitor the device transmission continuously. Circulators and filters were used to separate the pump and probe signals and filter out the EDFA noise after amplification. The optical signals were detected by photodetectors

then collected by data acquisition (DAQ) units. The electrical voltage and current are supplied and measured by a source meter unit (SMU), then collected by the DAQ units.

Lasers

A continuous wave (CW) tunable laser, Santec TSL-570, has been employed as the probe laser to monitor the broadband response of the photonic devices in the range of 1500-1630 nm (C-L band). The Santec laser supports a fast wavelength sweep speed of up to 100 nm/s and can be tuned by LabVIEW programs developed by previous group members. An L-band tunable laser (N7711A, Keysight) has been exploited as the pump laser, providing maximum output power $\geq +13\ dBm$ ($20\ mW$).

Photodetectors

Different types of photodetectors have been used to meet different response and bandwidth requirements. Model 2011 and 2053 from New Focus are InGaAs fiber-optic receivers with a detection range of 900-1700 nm. They provide bandwidths of 200 kHz and 10 MHz, respectively, with a maximum conversion gain of $18.8 \times 10^6\ V/W$ (peak responsivity of $1.0\ A/W$).

To detect high-speed dynamic response, photodetectors with higher sampling rates are required. 125 MHz (Model 1811, New Focus) and 1 GHz (Model 1611, New Focus) photodetectors have been exploited in high-speed experiments in Section 5.4.4 and 6.1.1, together with an oscilloscope (TDS7404, Tektronix) of 4 GHz bandwidth to collect data. These two types of photodetectors support a wider bandwidth at the cost of a lower maximum conversion gain ($4 \times 10^4\ V/W$ and $700\ V/W$, respectively), with the same peak responsivity. Therefore, to detect small transmission changes, amplifiers might be required in certain experiments (such as in Section 5.4.4).

Source meter units (SMUs)

A dual channel source (Model 2614B, Keithley) has been exploited to supply the voltage and measure the current response of the devices. This type of SMU supports high-accuracy output voltage ($\leq 5\ \mu V,\ \leq 50\ \mu V,\ \leq 500\ \mu V$ in the range of 200 mV, 2 V, 20 V, respectively) and high-resolution current measurements ($\leq 100\ fA$, $\leq 1\ pA$, $\leq 10\ pA$ and $\leq 100\ pA$ in the range of 100 nA, 1 μA, 10 μA and 100 μA). However, the unit requires a measure setting time in the order of 100 μs, thus the capability for high-speed measurements is limited.

Data acquisition (DAQ) units

A 14-bit multi-channel data acquisition unit from National Instruments (NI USB-6009) has been used to collect and continuously display low-frequency signals from the photodetectors and the SMU, using LabVIEW programs. This unit can also be used as a bias voltage source to supply DC voltage up to 5 V (200 mA).

Function generators

Arbitrary function generators (AFG3102C and AFG3152C from Tektronix, with 100-MHz and 150-MHz bandwidths, respectively) drive the EOM to generate customized pump pulses. The pump pulse shapes are programmed by LabVIEW programs.

For high-speed eye diagram acquisition (Section 6.1.1), a 3-GHz pulse generator (8133A, Agilent) has been exploited to generate pseudo-random binary sequences.

EOM

The pump laser is modulated by a lithium niobate electro-optical modulator (2623NA, Lucent Technologies), supporting $V_\pi < 5\ V$, insertion loss of 3.7 dB and modulation bandwidths up to 10 GHz within the C-band and L-band wavelength ranges. A bias tee (ZFBT-4R2GW+, Mini-Circuits) combines the electrical pump pulse generated by the function generator and a DC bias voltage generated from the SMU or DAQ to form control signals for the EOM. The DC bias is set to the

corresponding drive voltage at which the EOM provides minimum transmission, so that the output transmission can be tuned at the largest extinction ratio with the minimum electrical voltage.

EDFA

A high-gain (29 dB amplification for small input signal of -6 dBm) erbium-doped fiber amplifier (AEDFA-CL-23, Amonics), operating in the C-band and L-band wavelength ranges, has been exploited to amplify the EOM-modulated optical signal to sufficient power levels for programming phase-change materials. The amplification of the EDFA is programmable by tuning the amplifier current.

For input signal levels smaller than $-6\ dBm$ (Section 5.4.4), a pre-amplifier (AEDFA-PA-35-B-FA, Amonics) is required, providing 35 dB optical gain for input signal levels starting from $-40\ dBm$.

Notably, the EDFAs provide broadband gains, thus both the signal at the working wavelength and the out-of-band noise are amplified. Therefore, tunable bandpass filters (OTF-320, Santec) are required to filter out the out-of-band noise prior to the device.

Polarizers, bandpass filters and optical circulators

Given that grating couplers and EOMs are polarization-sensitive components, in-line polarizers are necessary before the EOM and the device to maximize the transmission. The in-line polarizer consists of a quarter-wave plate, a half-wave plate, and another quarter-wave plate to generate arbitrary polarizations, exploiting stress-induced birefringence in a single-mode fiber.

A three-port optical circulator (6015-3-APC, Thorlabs) has been employed to guide the probe light and the pump light to travel in only one direction (clockwise) with minimal loss ($\leq 1.0\ dB$). Another tunable bandpass filter (OTF-320, Santec) has been used before detecting the probe light to filter out the stray light from the high-energy pump light.

3.3.3 Electrical measurement setup

Figure 3.13 illustrates the setup [158] for electrical programming with both electrical and optical readouts.

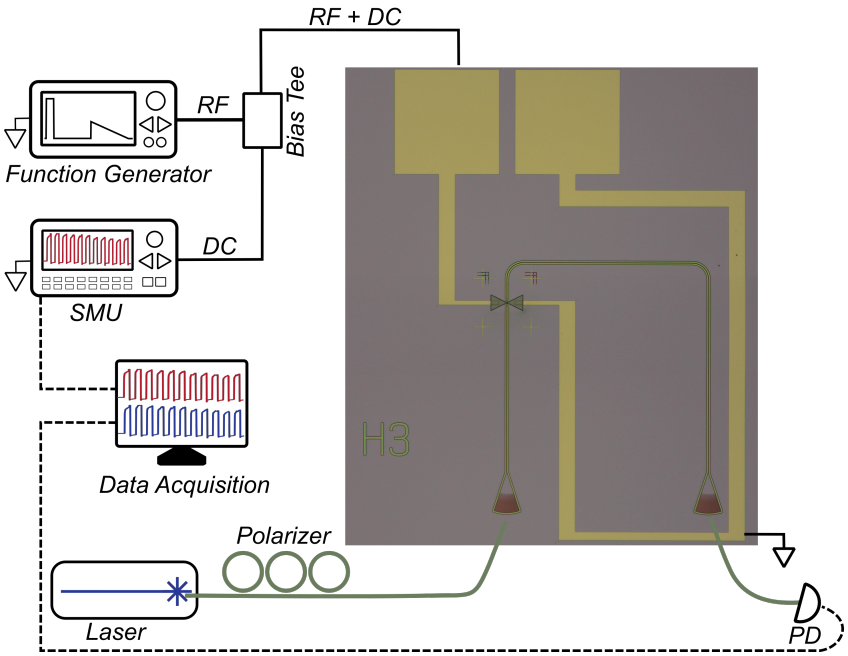

Figure 3.13: Electrical switching setup with both electrical and optical readouts. SMU: source measure unit; PD: photodetector. *Inset*: optical image of a device (green: waveguide layer; gold: electrical pads).

A probe line, with a probe laser (TSL-570, Santec), a polarizer before the device, and a photodetector, is used to continuously monitor device transmission, as introduced in Section 3.3.2.

Electrical programming pulses are programmed by the function generator (AFG3102C, Tektronix) and combined with the SMU-generated DC bias (2614B, Keithley) through a bias tee (ZFBT-4R2GW+, Mini-Circuits). Here, the bias tee provides isolation from the RF+DC port to the DC port, thus enabling the SMU to continuously monitor the status of the device. The pulses have been sent to the device via RF probes (Model 40A, GGB). Both optical and electrical readouts are collected by the DAQ (USB-6009, NI).

4

Plasmonic Mixed-mode Phase-change Devices on the SOI Platform

Contents

4.1 Background

Small footprint and low energy consumption are long sought-after goals for computing applications. As discussed in previous chapters, non-volatile programmable phase-change devices are uniquely suited for low-power computing applications in both electrical and optical domains. Plasmonic nanogap-enhanced phase-change devices have demonstrated tens of picojoule-scale programming energy with phase-change material $Ge_2Sb_2Te_5$ (GST) on a Si_3N_4 platform and meanwhile provided

dual electrical-optical functionality [158]. Yet, due to the low coupling efficiency between the waveguide mode and the plasmonic mode, such devices have achieved a contrast between set and reset states that does not exceed 5%, posing challenges for practical applications and limiting the cascadability for large systems. However, should lower loss and higher contrast be achievable, they become attractive for SOI-based platforms as they can be readily integrated with active photonic computing components such as photodetectors and modulators.

This chapter discusses the design considerations and experimental implementations of plasmonic phase-change devices on a SOI platform. The objective here is to transition to the SOI platform while improving the coupling efficiency and programming contrast.

4.2 A Self-Aligned Fabrication Method

Different types of plasmonic mode converters have been proposed to maximize the mode conversion efficiency [172–178]. However, few of them have been implemented experimentally [173, 175, 178], due to the stringent requirements for the fabrication and alignment process.

In this section, a self-aligned method is exploited to mitigate the challenge of alignment. Using one e-beam lithography mask for both the first silicon etching step and the gold deposition step, the silicon taper and gold pads are automatically aligned and contacted.

A detailed process flow is as follows (Figure 4.1):

(1) Prepare and clean the diced silicon-on-insulator (SOI) substrate.

(2) An e-beam lithography (EBL) process is used to define a photoresist mask for the plasmonic structure. Then RIE is used to fully etch the 220-nm silicon layer down to the SiO_2 substrate.

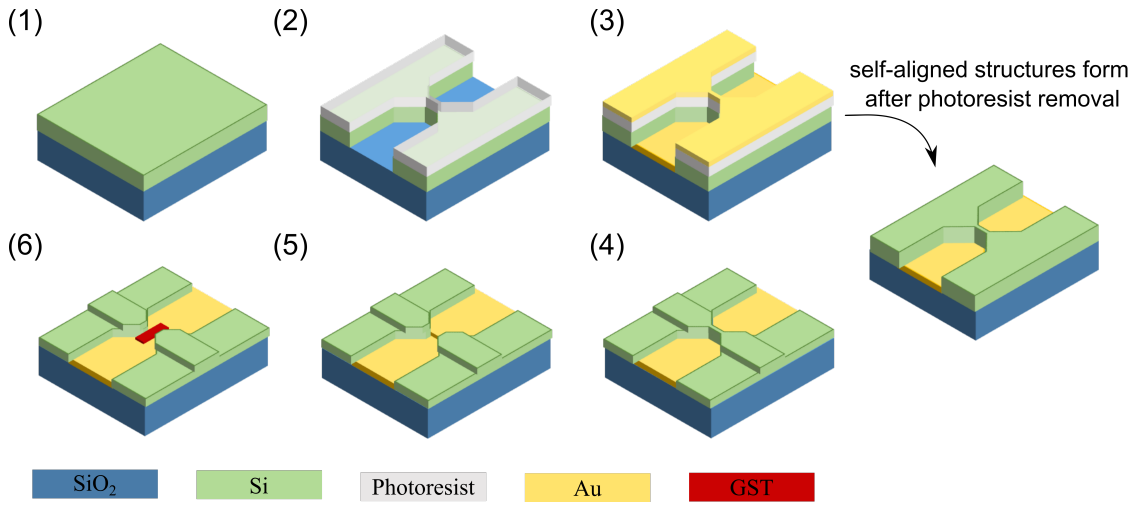

Figure 4.1: Self-aligned fabrication process flow.

(3) Use the same mask for thermal evaporation to deposit Au contacts (75 *nm* after a 5 *nm* Cr adhesion layer) before removing the photoresist, so that the Au contacts and the silicon nanogap are aligned automatically. A second EBL step and another thermal evaporation step are performed to deposit the large Au pads (details in Section 3.2), enabling access to RF probes during measurements.

(4) Third EBL step followed by an RIE step to define half-etched waveguides.

(5) Fourth EBL step followed by an RIE step to fully etch the silicon layer within the nanogap.

(6) Sputter phase change materials (80 nm, GST in specific for this work) with a SiO_2 capping layer (~10 nm).

4.3 Device Design and Simulations

This section introduces detailed design and simulations of a plasmonic phase-change device on a silicon platform. First, the design of the waveguide-plasmonic mode converter structure is discussed. Rather than the conventional taper structure, a grated, or so-called corrugated, taper structure [172] has been exploited to improve

the coupling efficiency between the silicon waveguide mode and the plasmonic mode. Furthermore, propagation loss of the plasmonic nanogap is estimated through numerical simulations. Finally, substrate choices of the plasmonic phase-change device are evaluated using optical-thermal simulations.

4.3.1 The design of the corrugated structure

Figure 4.2 presents a comparison of the electromagnetic field distribution between plasmonic nanogap structures with a conventional taper and a corrugated taper. The field is modeled using a 3D finite-difference time-domain (FDTD) simulation.

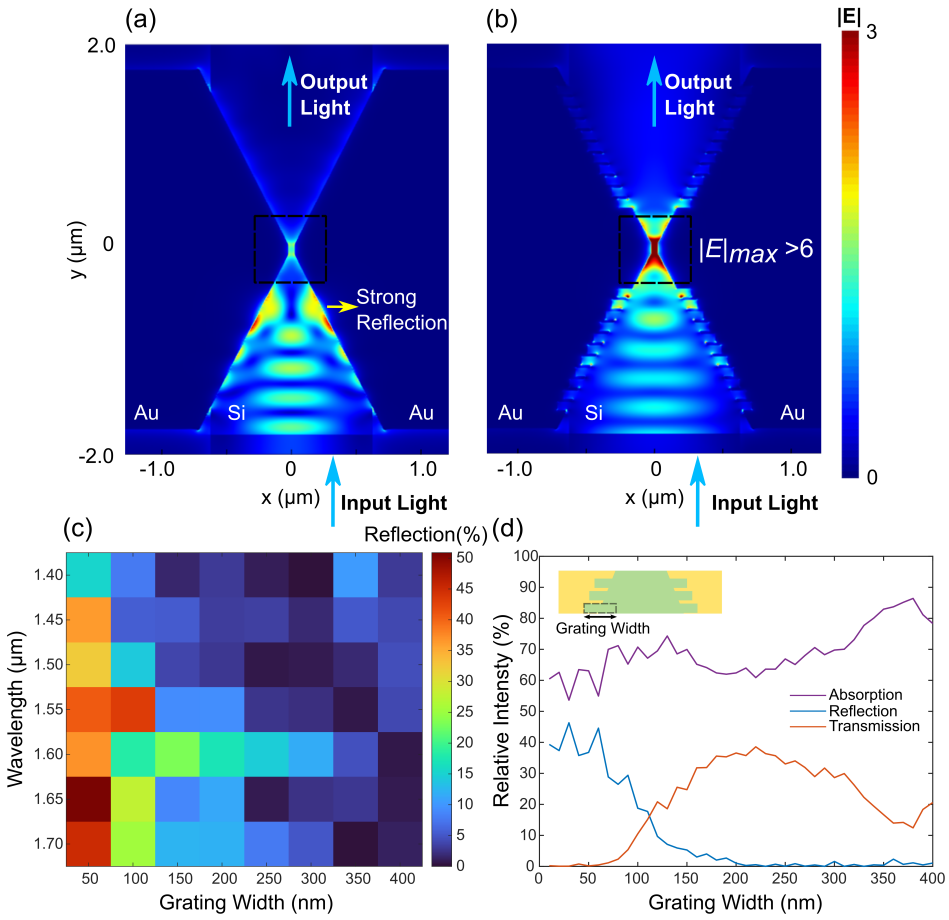

Figure 4.2: Grating structures for improving mode conversion efficiency. (a)-(b) 3D finite-difference time-domain (FDTD) simulated field distribution of structures (a) without and (b) with gratings. Light ($\lambda = 1545\ nm$) is incident from the bottom port. (c) COMSOL simulated reflection profiles with different grating widths. (d) COMSOL simulated transmission and reflection profiles at $\lambda = 1550\ nm$ for structures with different grating widths.

The conventional taper [Figure 4.2(a)] shows a strong reflection of light before the nanogap due to mode mismatching, while stronger field confinement within the nanogap (|E|>6, more than 4× compared to the conventional design) is realized in the corrugated taper [Figure 4.2(b)] with reduced reflection.

The design parameters of the grating structure have been further optimized to maximize the output transmission. As the grating width increases, the transmission first increases while the reflection decreases, reaching a maximum when the grating width is around 200 nm. Beyond this point, transmission decreases as gold absorption begins to dominate [Figure 4.2(c)-(d)]. A further discussion of other design parameters can be found in Section 4.5.2.

4.3.2 Propagation loss

The total transmission loss of the plasmonic devices is dominated by two contributions: propagation loss of the plasmonic mode along the nanogap, and coupling loss between the plasmonic mode and the waveguide mode at the input and output facets of the nanogap. To obtain the coupling/conversion efficiency of the device, the propagation loss of the plasmonic mode is estimated through numerical simulations.

The mode profiles of the plasmonic nanogap structure with different gap sizes have been simulated using a finite difference eigenmode (FDE) solver (Ansys Lumerical). As shown in Figure 4.3 and discussed in Section 3.1.3, the narrower the gap, the stronger the field enhancement. However, the increased field overlap with the metal led to greater absorption loss and, consequently, higher propagation loss. In particular, the total propagation loss of the plasmonic nanogap structure used in the conversion efficiency measurements (Section 4.4.1) is 0.49 dB, with a propagation length of 100 nm and a silicon gap size of 30 nm.

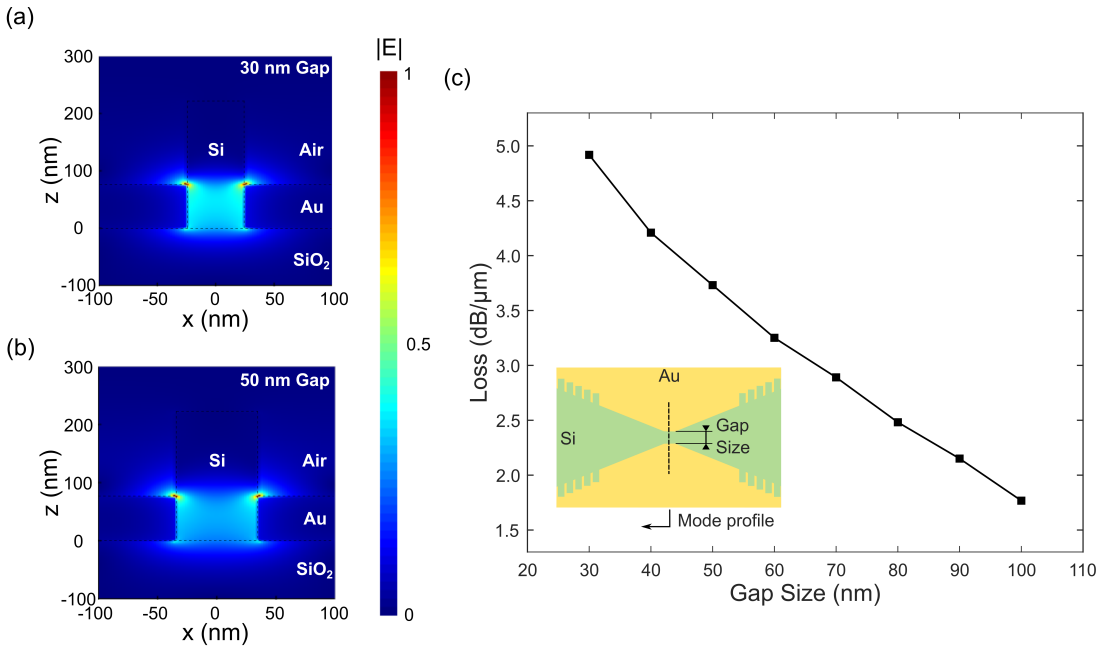

Figure 4.3: Simulated propagation Loss. (a) Cross-sectional mode profile for a device with a 30-nm silicon gap between gold contacts. (b) Cross-sectional mode profile for a device with a 50-nm silicon gap between gold contacts. (c) Simulated propagation loss for devices with different gap size.

4.3.3 Comparison of fully- and half-etched SOI substrates

Given their lower propagation and scattering losses compared to fully-etched or wire waveguides, half-etched or rib waveguides are widely used for waveguide and grating coupler designs in this thesis. However, considering the high thermal conductivity of silicon (148 $Wm^{-1}K^{-1}$ [179], compared to 43 [179] and 1.38 [180] $Wm^{-1}K^{-1}$ for silicon nitride and silicon dioxide), i.e. silicon waveguides exhibit faster heat dissipation than silicon nitride [179], thermal effects should also be taken into account when designing plasmonic phase-change devices on a SOI platform.

To investigate the switching behavior of plasmonic nanogap devices with fully-etched and half-etched SOI substrates, optical-thermal coupling simulations have been exploited via COMSOL Multiphysics. Here, fully-etched refers to the process of etching the 220-nm silicon layer completely down to the SiO_2 substrate in steps (2) and (5) of the self-aligned fabrication process, while half-etched substrates

refer to etching only halfway (110 nm) through the 220-nm silicon layer [Figure 4.4(a)].

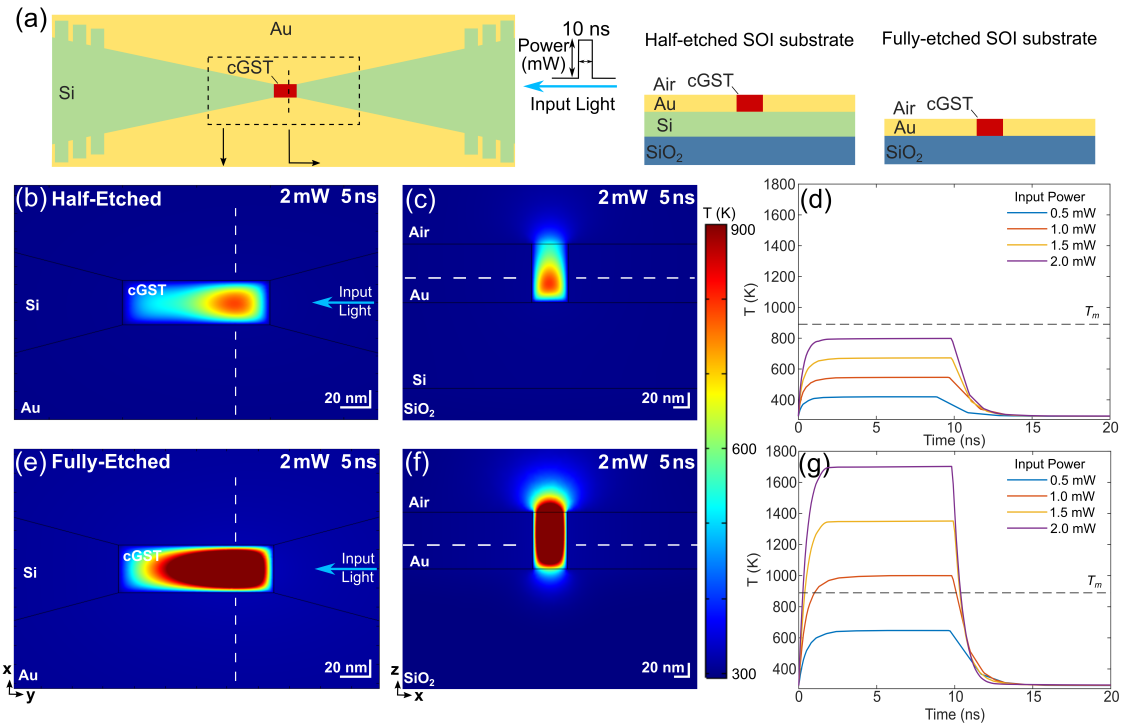

Figure 4.4: Simulated temperature distribution (via COMSOL Multiphysics) for fully-etched and half-etched nanogap structures. The input optical stimuli is fixed as a 10-ns rectangular pulse with different power amplitudes, while the temperature distribution of the cross-sections are captured at $t = 5\ ns$. (a) Schematics for the half- and fully-etched SOI substrates. (b) Temperature distribution for the Z-cut cross-section at $z = 30\ nm$ of a half-etched device (white dashed line in (c)). The reference plane $z = 0$ is defined as the bottom surface of Au. (c) Temperature distribution for the Y-cut cross-section at $y = 25\ nm$ of a half-etched device (white dash line in (b)). The reference plane $y = 0$ is defined as the center of the device. (e)-(f) are their fully-etched comparisons. (d) and (g) show simulated temporal peak temperature response of half-etched and fully-etched devices, respectively. cGST: crystalline $Ge_2Sb_2Te_5$; T_m: melting temperature for crystalline GST [181].

The input for the simulation is set as 10-ns rectangular amorphization pulses with different amplitudes, while the temperature distribution of the nanogap region is recorded during the pulses. With the same pulse amplitude, the phase change material within the fully-etched structure [Figure 4.4 (e)-(g)] reaches a much higher temperature compared to the half-etched structure [Figure 4.4(b)-(d)]. Specifically, an input pulse with 1-mW amplitude is sufficient to heat the phase-change material

above the melting temperature (for amorphization) in a fully-etched structure, while 2-mW is still below the switching threshold for a half-etched structure.

The preference for half-etched waveguides over fully-etched waveguides is primarily attributed to the higher propagation loss of the latter caused by the strong light scattering on the side-wall roughness [182]. However, propagation loss of a plasmonic structure is dominated by metal absorption as explained in Section 4.3.2; thus, the effect of the substrate is less critical. To enable energy-efficient optical switching, fully-etched designs are utilized for the corrugated taper and nanogap structures in this work.

4.4 Experimental Results

Following the design considerations discussed above, plasmonic devices with grating structures have been fabricated using the self-aligned fabrication process. This section presents the experimental results for waveguide-plasmonic mode conversion efficiency, microscopic characterizations, and switching performance.

4.4.1 Conversion efficiency

Devices with varied fill factors and periods of the gratings have been fabricated within the same chip. Figure 4.5 exhibits the experimental transmission of the devices before GST deposition (after fabrication step (4) as described in Section 4.2). Their transmission is compared to plain waveguides on the same chip, showing a maximum efficiency exceeding 43%, corresponding to a total insertion loss of 3.57 dB and 1.54 dB per facet for the converter (taking the propagation loss as simulated in

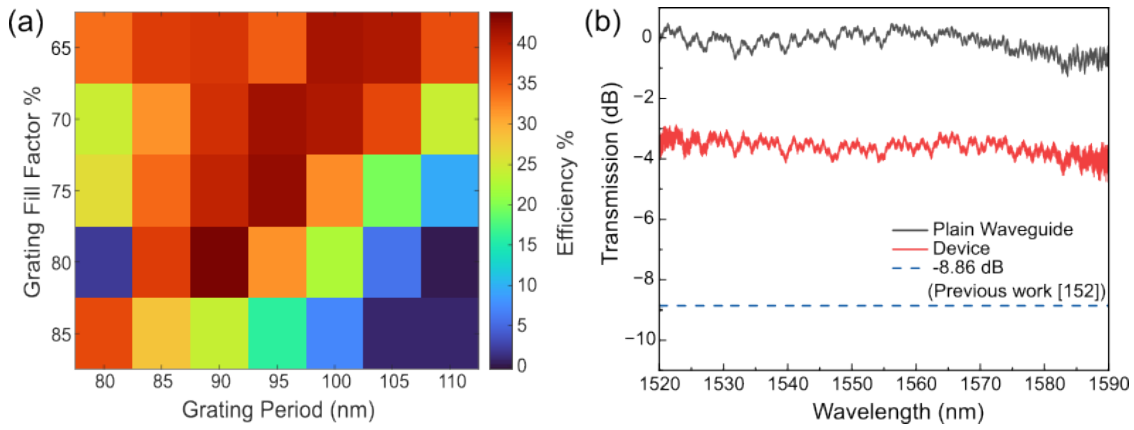

Figure 4.5: Experimental transmission efficiency. (a) Experimental transmission map for devices with different grating parameters, compared to 100% transmission for a plain waveguides on the same chip. (b) Normalized broadband transmission response compared to plain waveguides on the same chip and the previous plasmonic implementation [159].

Section 4.3.2), which outperforms the state-of-the-art experimental implementation [178]. A discussion of fabrication variations compared to simulations can be found in Section 4.5.2.

The broadband feature of silicon devices ($\sim 70\ nm$ bandwidth, limited by the bandwidth of the grating couplers) is maintained with the grating structure, making it promising for future wavelength-multiplexing applications.

4.4.2 Microscopic characterization

After characterizing the coupling efficiency, GST is deposited on the devices, following steps (5)-(6) of the self-aligned fabrication process. Figure 4.6(a)-(c) present optical microscopy and SEM images of the fabricated devices, where the grated silicon taper and the gold pads are well aligned, defining a nanogap of $\sim 30\ nm$.

To confirm the fabrication of the nanogap structure, focused ion beam (FIB) was performed to prepare the sample for cross-sectional imaging. Figure 4.6(d) exhibits the cross-sectional image, showing that GST is deposited into the nanogap as expected. A platinum layer was deposited on the device at the beginning of the FIB process to prevent damage during milling.

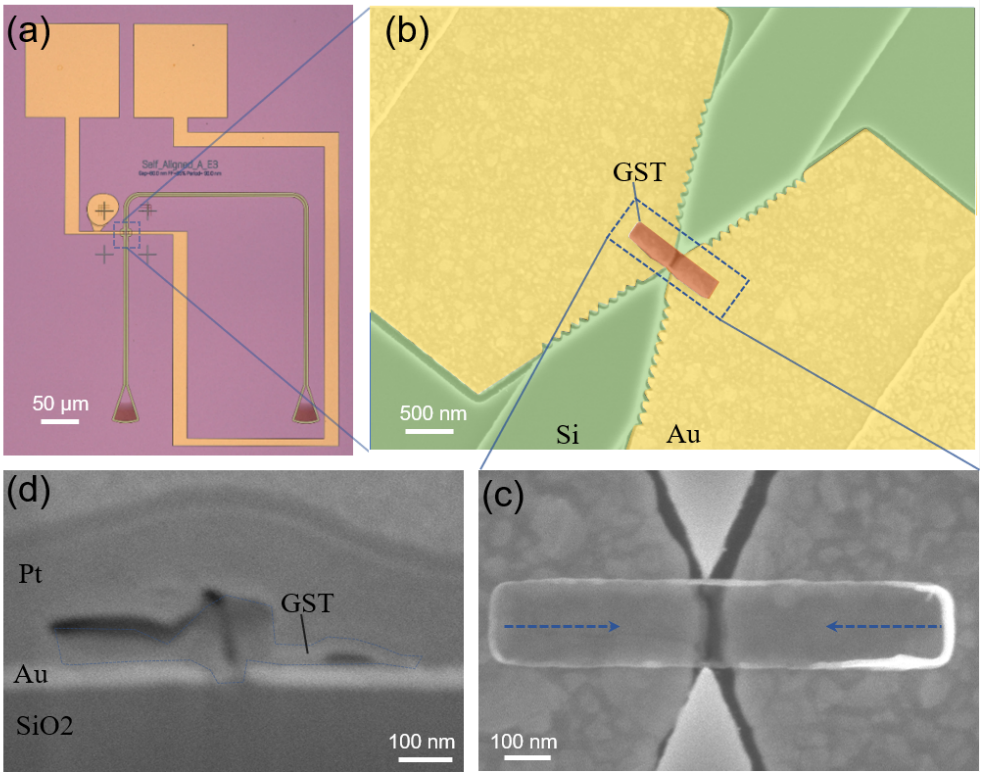

Figure 4.6: Device characterization. (a) Optical microscope image of a full device. The size of the contact pads is 130 μm with a spacing of 150 μm. Scale bar: 50 μm. (b) False-color scanning electron microscopy (SEM) image of the central region of a device. Scale bar: 500 nm. (c) Zoom-in SEM image of the device nanogap ($\sim$ 30 nm). Scale bar: 100 nm. (d) Cross-sectional image of the device after FIB. Scale bar: 100 nm.

4.4.3 Optical switching

The devices were thermally annealed on a hot plate (250 °C for 3 min) to transition to the crystalline state prior to optical switching experiments.

Optical measurements have been performed with the setup described in Section 3.3.2. A low-power probe light ($\lambda = 1550\ nm$) is used to monitor the transmission change through the device while a pump light ($\lambda = 1570\ nm$) is modulated by an electrical-optic modulator (EOM) then amplified by an erbium-doped fiber amplifier (EDFA) to provide programming and erasing pulses.

Rectangular pulses are used to partially switch GST from the crystalline state to the amorphous state (lower transmission to higher transmission), and double step pulses [20] or triangular decay pulses are used to reset the material back to the

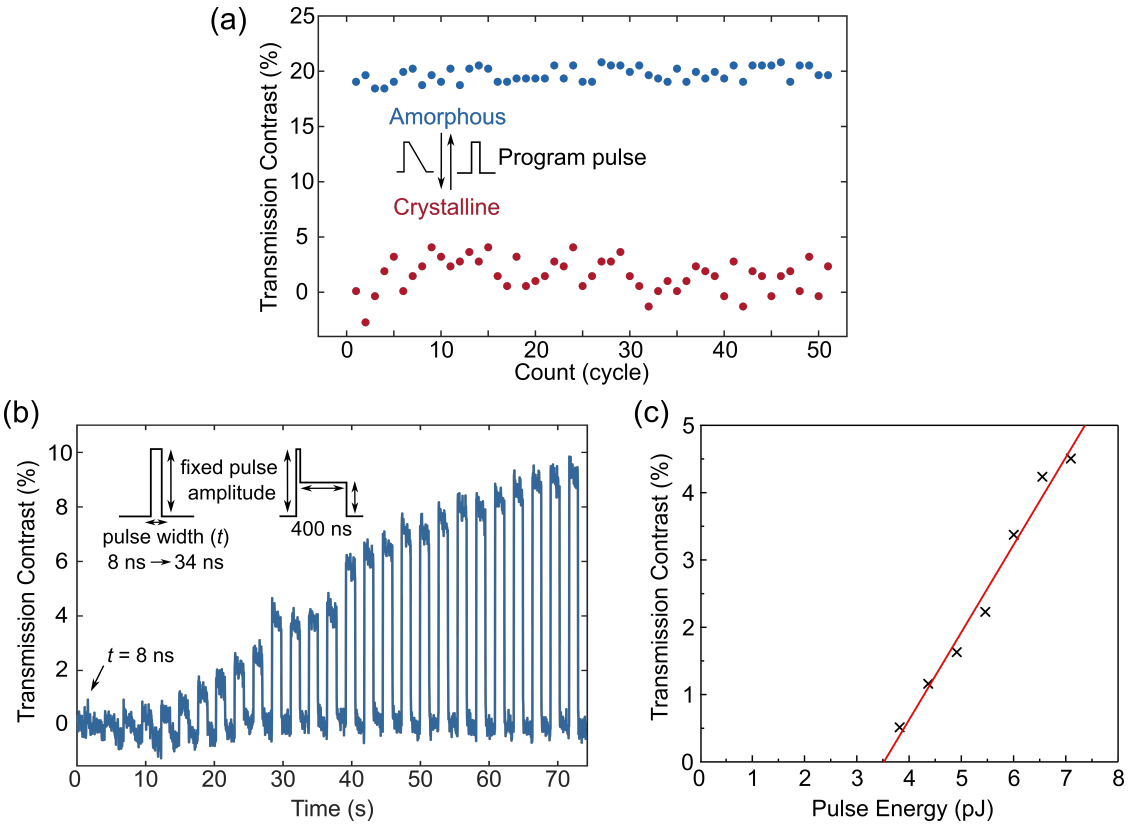

Figure 4.7: Optical Switching Performance. (a) 51 cycles of high contrast binary switching. The amorphization pulse is fixed as a 2.10 mW, 10 ns rectangular pulse, and the crystallization pulse is fixed as a 2.10 mW, 10 ns pulse followed by a 200 ns triangular decay. (b)Recorded transmission with a fixed programming pulse amplitude of 0.41 mW, and monotonically increasing pulse widths from 8 ns to 34 ns in 1 ns increments. The erasing pulses are fixed as a 0.41 mW, 1 ns rectangular pulse with a 0.25 mW, 400 ns rectangular tail. (c) Linear fitting for contrast and amorphization energy relation, with x-intercept at 3.52 pJ. The contrast and pulse energy were obtained for multilevel switching measurements. The amorphization pulse is fixed as a 0.55 mW rectangular pulse with pulse widths varying from 7 ns to 13 ns in 1 ns increments. The crystallization pulse is fixed as a 10 ns, 0.55 mW rectangular pulse followed by a 0.30 mW, 100 ns rectangular tail.

crystalline state. Amorphization (2.10 mW, 10 ns) and crystallization pulses (2.10 mW, 10 ns pulse followed by a 200 ns triangular decay) are sent in a sequence (i.e., one cycle) repeatedly. Around 20% contrast has been maintained without obvious degradation after 50 cycles [Figure 4.7(a)].

Multilevel transmission responses have also been demonstrated, as shown in Figure 4.7(b). Here, the pulse amplitude of the input program is fixed at $0.41mW$, with pulse widths increasing from 8 to 34 ns in 1-ns increments. To estimate

the minimum switching energy, a linear fit is performed for the contrast-energy relationship in the multilevel switching tests, with the x-intercept showing a switching energy threshold of 3.52 pJ [Figure 4.7(c)]. 0.5% contrast has been experimentally demonstrated with an amorphization energy of 3.82 pJ (pulse energy for a 0.55 mW, 7 ns amorphization pulse).

Table 4.1: Optical Switching performance of integrated photonic devices with $Ge_2Sb_2Te_5$

Devices type	Active area (μm^2)	Mini. switching energy (pJ)	Max. programmable contrast	Insertion loss (dB)	Ref.
Plasmonic nanogap devices on Si_3N_4	0.05×0.10	16 ± 2	5%	-10.45 -8.86	[172] [159]
Si_3N_4 waveguide	5.0× 1.3 1.0×1.3	42 ± 0.3 13.4 ± 0.6	$58.2 \pm 2.0\%$ $16.2 \pm 0.46\%$	-0.395 -0.079	[18] [179]
SOI waveguide	4.0×0.45	388.4	15%	-0.236	[179]
Plasmonic nanogap devices on SOI	**0.03×0.10**	**3.82**	**20%**	**-3.57**	**This work**

Table 4.1 provides a comparison of the optical switching performance of different integrated photonic devices. The minimal optical switching energy (amorphization) of this work is reduced to one-fourth of the previous designs Si_3N_4 and two orders lower than previous silicon-based phase-change devices.

4.4.4 Electrical switching

To verify the electrical properties of the plasmonic nanogap devices, electrical switching measurements were carried out. Nanogap devices with different gap sizes were fabricated on SOI substrates according to the self-aligned fabrication procedure outlined in Section 4.2, omitting the waveguide layer (step (4)) for simplicity.

Figure 4.8 exhibits current-voltage (IV) characteristics of the device before and after hot plate annealing, following the RF sputtering step. The resistance of the device drops from 3.3 $M\Omega$ (with GST in the as-deposited state) to 437 $k\Omega$

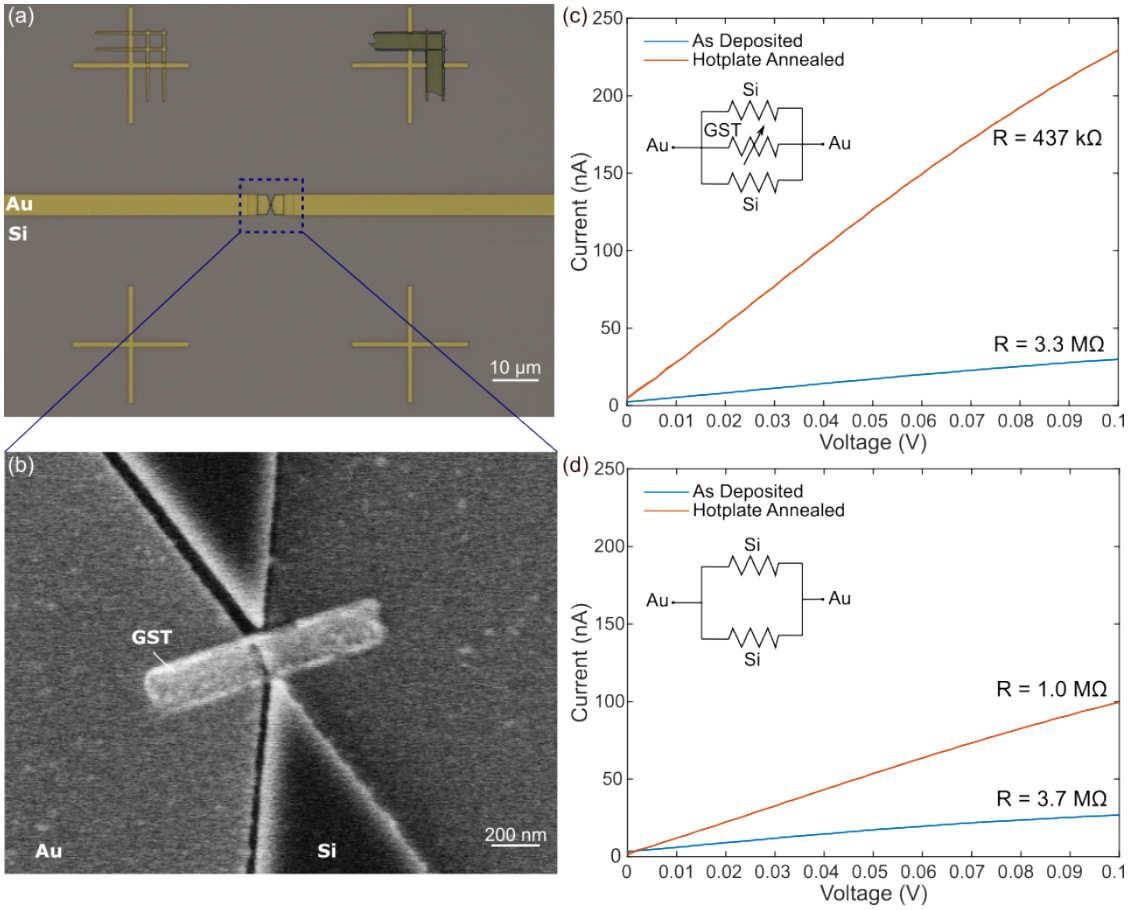

Figure 4.8: Current-voltage measurements of the device resistance with a voltage range below the switching threshold. (a) Optical image of a typical device. (b) Zoom-in SEM image of a typical device (nanogap size: $\sim 30\ nm$). (c)-(d) are current-voltage curves for nanogap devices with and without deposited GST.

(with GST in the crystalline state) after annealing. Here, the resistance of the as-deposited (amorphous) device is less than previous plasmonic nanogap devices on the Si_3N_4 platform ($10 - 300\ M\Omega$ [158, 159]), given the high conductivity of the silicon substrate ($\sim 3.7\ M\Omega$ for a reference device after the fabrication steps (1)-(3) and before GST sputtering, as shown in Figure 4.8(d)).

Meanwhile, some of the devices have shown low resistance ($40 - 50\ \Omega$) close to the resistance of closed-gap devices after annealing, which can be attributed to thermally activated diffusion [183] of gold into silicon waveguides, thus closing the gap. To mitigate the diffusion-related uncertainty of device performance, electrical switching measurements have been carried out for as-deposited devices, replacing

the commonly used hot-plate annealing step.

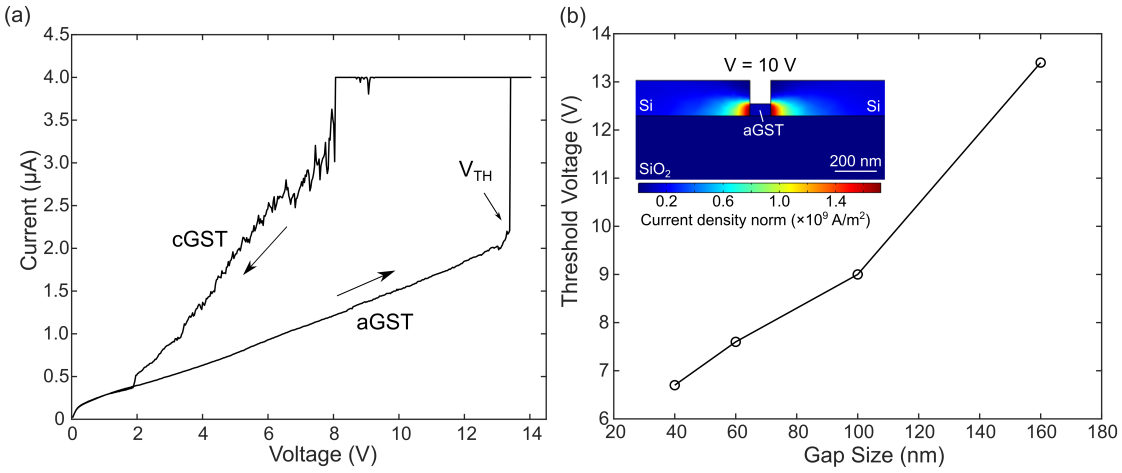

Figure 4.9: Electrical threshold switching. (a) A typical current-voltage curve for electrical switching. A current compliance of 4 μA is imposed to prevent device damage caused by the abrupt current change. (b) Experimental threshold voltages required to induce crystallization of the GST for different nanogap widths. Inset: current density distribution across the device cross-section when apply a voltage bias of 10 V on the gold electrodes.

Voltage sweeps were performed for the devices to switch from the high-resistance amorphous state to the low-resistance crystalline state (Figure 4.9). Threshold switching has been observed, and the switching threshold increases with increased gap sizes.

Nevertheless, the threshold is much higher than previously achieved for plasmonic Si_3N_4 devices ($\sim$ 1 V). This effect can be attributed to the relatively high conductivity of the silicon substrate. The SOI substrates used in this thesis have a p-doped device layer of resistivity $\sim 10 - 20\ \Omega \cdot cm$, which is much lower than amorphous GST ($> 10^4\ \Omega \cdot cm$) and close to the resistivity of crystalline GST ($\sim 1\ \Omega \cdot cm$) [184]. A simulated current density profile across the cross section of an amorphous device is presented in the inset of Figure 4.9(b). Higher current density is concentrated in the low-resistance silicon region, slowing down the heating process for GST thus increasing the switching threshold.

Similarly, electrical switching measurements have also been conducted for plasmonic nanogap devices with waveguides. Transmission drops have been observed

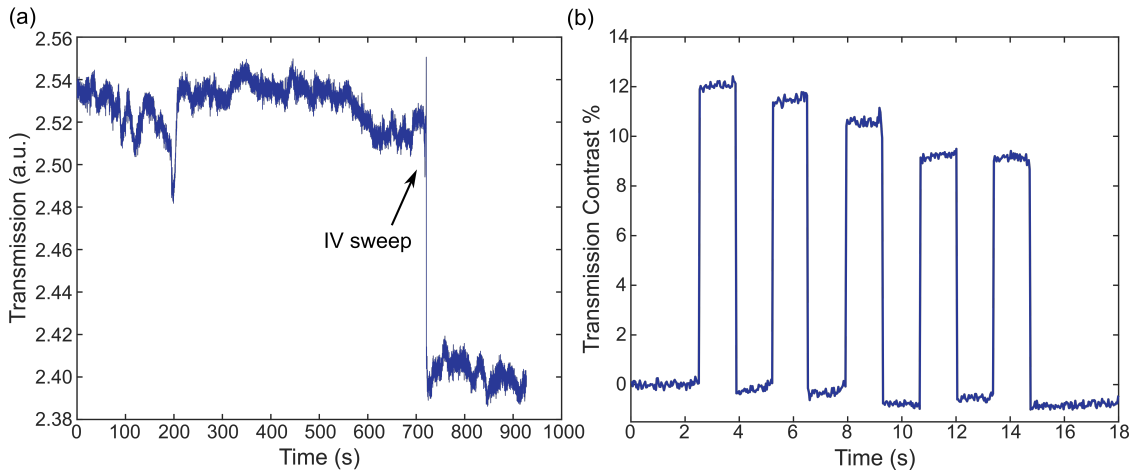

Figure 4.10: Optical switching performance after electrical annealing. (a) Transmission drops after electrical IV annealing. (b) Recorded transmission with a fixed programming pulse amplitude of 0.6 mW, and 10 ns. The erasing pulses are fixed as a 0.6 mW, 5 ns rectangular pulse with a 0.28 mW, 250 ns rectangular tail.

after electrical IV annealing [Figure 4.10(a)], indicating partial crystallization. Optical switching measurements were performed subsequently [Figure 4.10(b)], and low-energy optical programming has been demonstrated similar to the results in Section 4.4.3.

However, mixed-mode read-out has not yet been achieved for the above switching measurements. On the basis of previous discussions, possible solutions might be adding a thin isolation layer between gold and silicon, or changing the SOI substrate to the one with a lightly doped, high-resistance device layer so that the self-aligned process and the corrugated design can be maintained.

4.5 Discussion

In this section, the choice of plasmonic materials and the design parameters of the corrugated tapers are discussed in detail, followed by an outlook on future applications of the corrugated plasmonic mode converters to MZIs.

4.5.1 Analysis for material choices

This section provides a brief discussion on the choice of electrode materials for the plasmonic structure.

Noble metals such as silver (Ag) and gold (Au) are the most commonly used plasmonic materials. Silver provides a lower loss (smaller ε'') and has shown a theoretically higher coupling efficiency than gold [172]. However, given that silver is chemically unstable and prone to oxidation, it is not suitable for programmable phase-change devices, where stability is crucial. Likewise, the requirement of thermal stability up to the melting temperature of phase-change materials excludes plasmonic materials with low melting points, such as aluminum (Al).

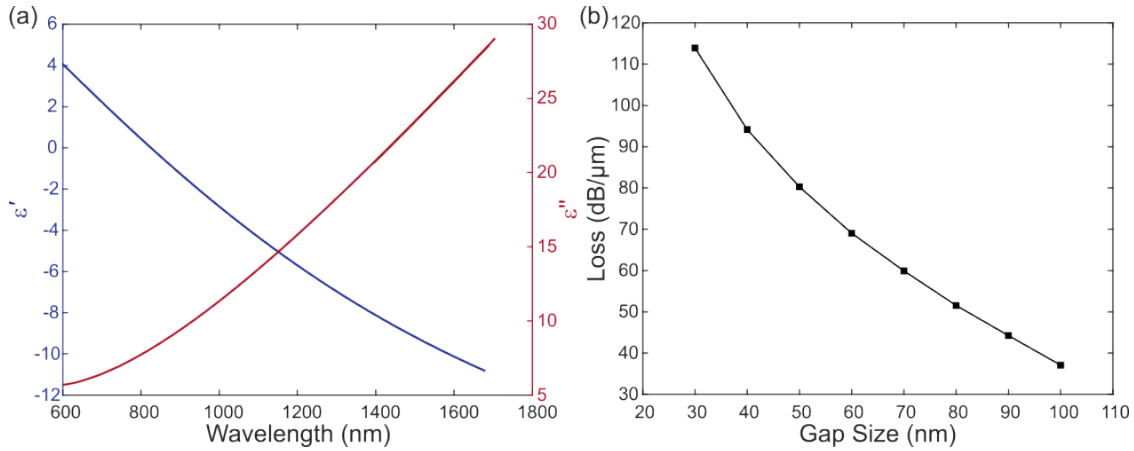

Figure 4.11: Simulated propagation loss when exploiting TiN as the electrodes material. (a) Ellipsometry measured permittivity of TiN. (b) Simulated propagation loss for plasmonic TiN nanogap devices with different gap sizes.

In addition to metals, CMOS-compatible materials, such as titanium nitride (TiN), also support plasmonic enhancements at the near-infrared wavelengths, offering the possibility of CMOS-compatible implementations in the future. To compare the performance of Au and TiN in the proposed plasmonic devices, thin-film TiN was sputtered on a silicon substrate by RF sputtering and its permittivity data were extracted after ellipsometry measurements [Figure 4.11(a)], exhibiting negative real permittivity for wavelength >830 nm. Mode simulations of the plasmonic

nanogap structure formed with TiN were carried out. The simulated propagation loss is shown in Figure 4.11, which is more than one order higher than the results for Au in Section 4.3.2. To implement CMOS-compatible plasmonic phase-change devices, future optimization of the material index (increase the ratio of $|\varepsilon'|/\varepsilon''$ or δ_z/δ_d as discussed in Section 3.1.3) is required to reduce the propagation loss.

4.5.2 Design parameters and fabrication variations

In this section, to investigate how the design parameters of the corrugated plasmonic mode converter affect the coupling efficiency, simulated transmission and dispersion results are discussed and compared to experimental results. The design parameters considered in this section include the taper angle, the waveguide width at which the grating begins, and the period, fill factor, and width of the gratings.

Taper angles

Firstly, simulated transmission and reflection profiles of plasmonic nanogap devices with various taper angles have been obtained via COMSOL as shown in Figure 4.12. Within the considered half-angle range of 20°- 45°, the taper length ranging from 1.72 μm to 625 nm accordingly, devices without grating structures constantly exhibit a high reflection ratio while grating structures significantly

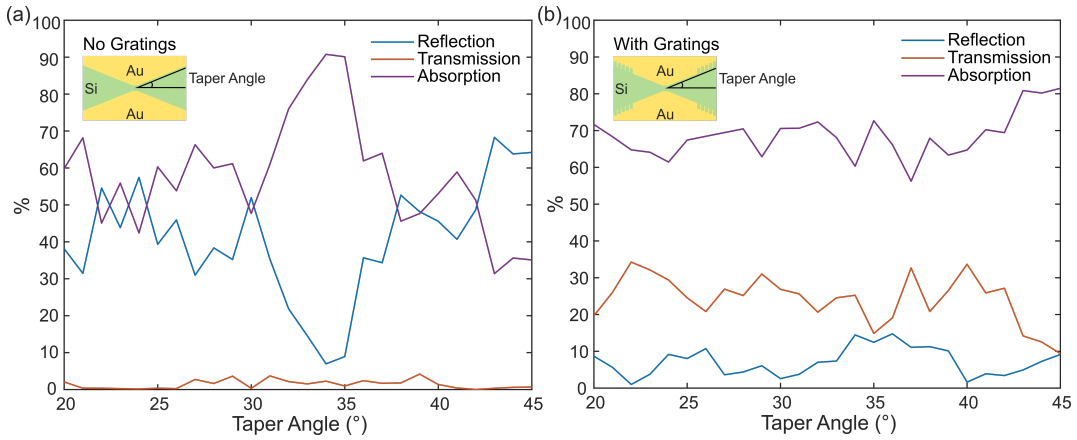

Figure 4.12: Simulated reflection and transmission profiles for the plasmonic nanogap devices with various taper angles. (a) Simulated profiles for devices without grating structures. (b) Simulated profiles for devices with grating structures.

suppress reflection and increase transmission. For experimental implementations, the taper angle is fixed at 22°, forming a taper length of $\sim 1.55\ \mu m$.

Grating parameters

Next, to investigate how different grating parameters (width, fill factor, and period) affect light propagation, dispersion analysis has been performed for a simplified periodic grating structure (schematics in Figure 4.13(a), where β indicates the propagation constant along the metal-dielectric interface). The detailed theory underlying the design is provided in [172]. Strong mode coupling occurs when the dispersion relation of the corrugated waveguide intersects with the plasmonic

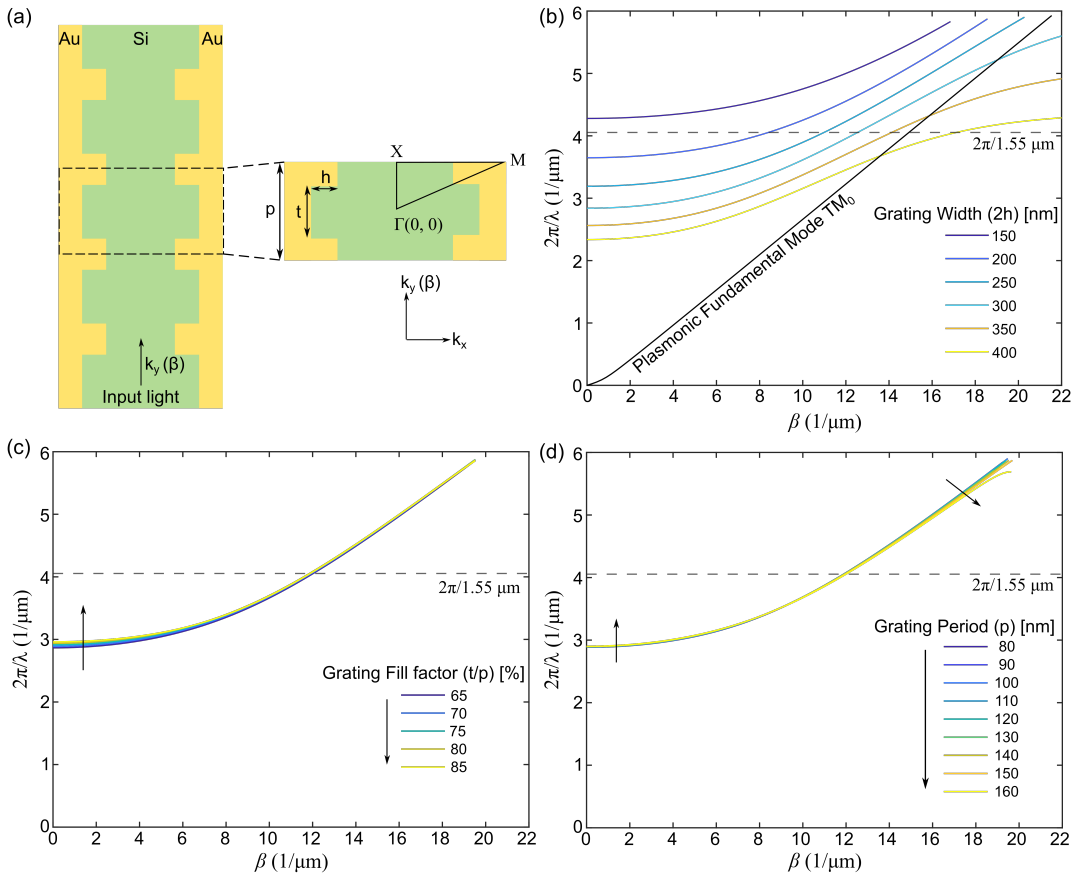

Figure 4.13: Simulated dispersion diagrams with different grating parameters. (a) Schematics for the simulated periodic grating structure. (b) Dispersion relations of grating structures with different grating widths (2h). The dispersion curve of the plasmonic fundamental mode refers to a Au-Si-Au (MIM) structure with a 400-nm Si width. (c) Dispersion relations of grating structures with different grating fill factors (t/p). (d) Dispersion relations of grating structures with different grating periods (p).

fundamental mode within the considered wavelength range. As discussed in Section 4.3.1, the width of the grating (2h) is the key parameter that modulates the reflection of the converter. This point is further confirmed in Figure 4.13(b).

In summary, increased grating widths shift the dispersion curve of the corrugated waveguide closer to that of the plasmonic fundamental mode, indicating stronger mode interaction and coupling. On the other hand, period (p) and fill factor (t/p) of the grating structure contribute insignificantly to the dispersion curve [Figure 4.13(c)-(d)].

Fabrication variations

Lastly, to further explore how fabrication variations contribute to conversion efficiency, transmission simulations have been performed via Lumerical FDTD solutions to compare the differences between experimental and simulation results. As shown in Figure 4.14(a), the grating fill factor contributes insignificantly to the mode conversion efficiency, which is in agreement with previous dispersion analysis. The different conversion efficiency obtained from the experiments in relation to the grating fill factor [Figure 4.5(a)] can be attributed to fabrication variations, where a larger fill factor leads to a poorer fabrication resolution of the grating widths, decreasing conversion efficiency.

Moreover, to identify how the grating period affects both the simulated and experimental conversion efficiency, the effect of the waveguide width at which the grating starts should also be taken into consideration [Figure 4.14 (b)]. When fixing the starting widths, Figure 4.14(c) indicates that the grating period has a minor impact on the conversion efficiency, which aligns with the dispersion analysis.

To further explain the effect of the starting width, effective index profiles for both plasmonic modes and the slab mode are simulated as shown in Figure 4.14(d). When the metallic waveguide width drops to around 400 nm, the higher-order plasmonic mode is cutoff, and adding grating structure to the narrower waveguide region could increase the absorption loss, as the fundamental plasmonic mode

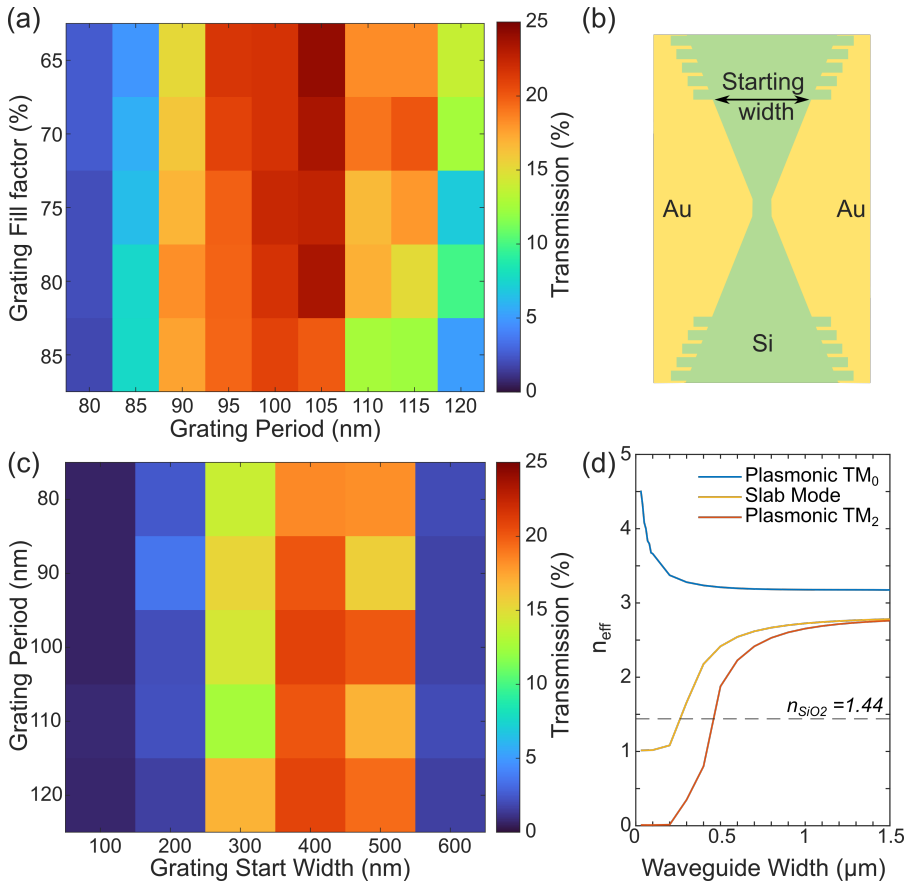

Figure 4.14: Simulated transmission profiles with different grating parameters. (a) Simulated transmission of the plasmonic nanogap structures with different grating period and fill factors. (b) Schematic indicates the waveguide width at which the grating begins. (c) Simulated transmission of the plasmonic nanogap structures with different grating period when fixing the grating start point at different waveguide widths. (d) Effective index profiles for both plasmonic modes and the slab mode. Dashed line: refractive index of SiO_2 (substrate material) at $\lambda = 1550\ nm$. The slab mode and the higher-order plasmonic mode do not exist when the effective index is smaller than this value.

concentrates more on the corrugated surfaces. As a result, fixing the starting width around 400 *nm* provides a higher conversion efficiency.

4.5.3 Plasmonic MZI

To explore potential applications of the proposed device design, the same structure is applied to plasmonic MZI as shown in Figure 4.15. The device exhibits a measured free spectral range (FSR) of 8 *nm*, as designed, and provides excellent extinction ratio >500 (27 dB). After further optimization of the device, these structures can

be utilized effectively in low-power applications, such as modulators [105] and bio-sensing applications[185].

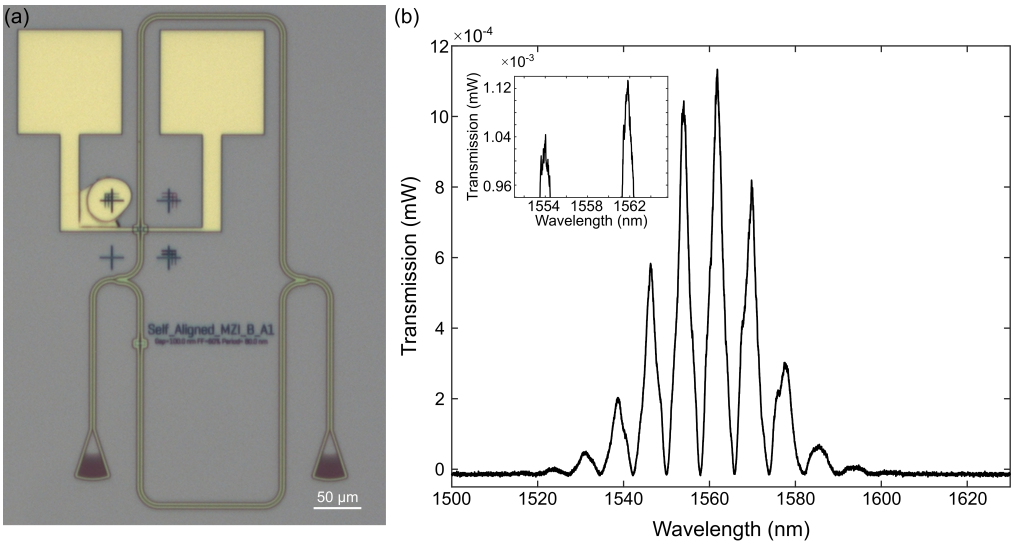

Figure 4.15: Transmission profile of the plasmonic MZI device. (a) Optical image of the plasmonic MZI device. (b) Transmission profile of the device.

4.6 Chapter Summary

In this chapter, by designing corrugated taper structures between the waveguides and the plasmonic waveguides, an active SOI-based plasmonic cell with ultra-high coupling efficiency (-3.57 dB) and high switching contrast has been experimentally demonstrated. A novel self-aligned fabrication process is proposed, which increases yield and reduces fabrication difficulty. The minimum switching energy has been shown to be less than 4pJ, which is a staggering two-order-of-magnitude reduction compared to previous silicon-based phase-change devices [179] and a 75% reduction compared to comparable devices in the literature [158]. The devices also show high programmable contrast (20 % compared to 5% in the previous work [158]) and stable multilevel switching performance, which makes them promising for future high-efficiency hybrid electro-optical computing systems.

All steps for device fabrication and measurements have been performed by the author, with the exception of the ellipsometry measurements conducted by June

Sang Lee, and focused ion beam (FIB) performed with the assistance of Gareth Hughes at Oxford David Cockayne Center for Electron Microscopy (DCCEM). All simulations presented were performed by the author, with the contribution of Nikolaos Farmakidis for Lumerical FDTD simulations, and guidance from Samarth Aggarwal and Angel Ortega in COMSOL simulations and dispersion simulations, respectively. All analyses were performed by the author. The work presented in this chapter has been delivered as an oral presentation in the following conference proceeding with a manuscript in preparation:

- Y. He, N. Farmakidis, S. Aggarwal, J. S. Lee, M. Wang, W. Zhou, and H. Bhaskaran, "Ultra-Efficient Plasmonic Phase-Change Devices on SOI Platform," in CLEO 2023, Technical Digest Series (Optica Publishing Group, 2023), paper STh1O.7.
- Y. He, N. Farmakidis, et al. "Self-aligned corrugated plasmonic mode converters for energy-efficient phase-change devices," (in preparation).

5

Mixed-mode Phase-change Devices with Constriction Structures

Contents

5.1 Background

To address the challenge of size mismatch between integrated electrical and optical phase change devices, two promising pathways have been explored [Figure 5.1]. The first is to enhance light–matter interactions at the nanometer scale by exploiting

plasmonic structures, thus reducing the dimensions of photonics. This is a highly promising approach that enables low switching energy and a compact device footprint. However, as discussed in the previous chapter, such an approach requires a uniform high-resolution fabrication and alignment process, posing challenges in scaling up. Moreover, plasmonics impose stringent constraints on material choices, limiting CMOS compatibility.

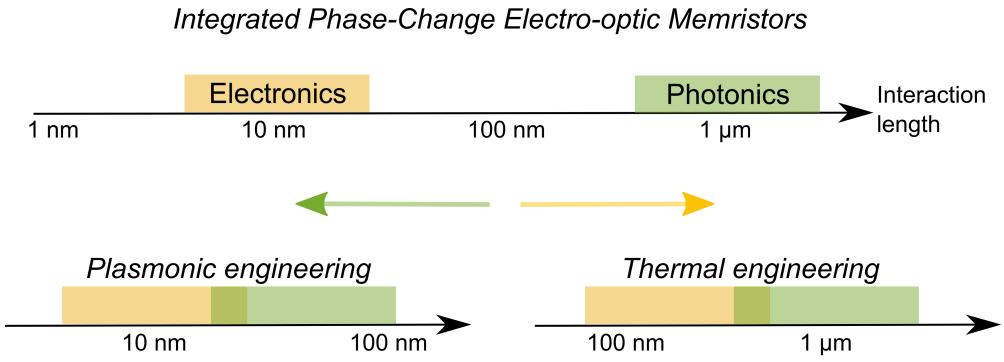

Figure 5.1: Two directions to tackle the size mismatch between integrated electrical and optical phase change devices for implementing integrated electro-optic memristors.

Alternatively, engineering the thermal distribution in the electrical domain, i.e. engineering the device structure such that the thermal distribution is optimized when an electrical field is applied, offers the possibility of aligning the size of programmable electronics with photonics. For example, combining external heaters with silicon waveguides provides uniform Joule heating for electrically switching large-area phase-change materials [103, 104, 160–162]. However, dual electrical–optical functionality has not been achieved in these devices because their electrical readout is independent of the phase-change material state.

In the electronic domain, there has been huge progress in the realization of low-energy electrical switching using thermal engineering techniques such as self-aligned carbon nanotube electrodes [25], superlattices [26], and self-confined cells [28]. In this chapter, these concepts are applied to optoelectronics, realizing electrical heat confinement for scalable, energy-efficient electro-optic memristors by tailoring the geometry of the phase-change material as a nanoscale constriction.

5.2 Device Design and Simulations

This section covers the design and simulations for the constriction-structure-based mixed-mode device.

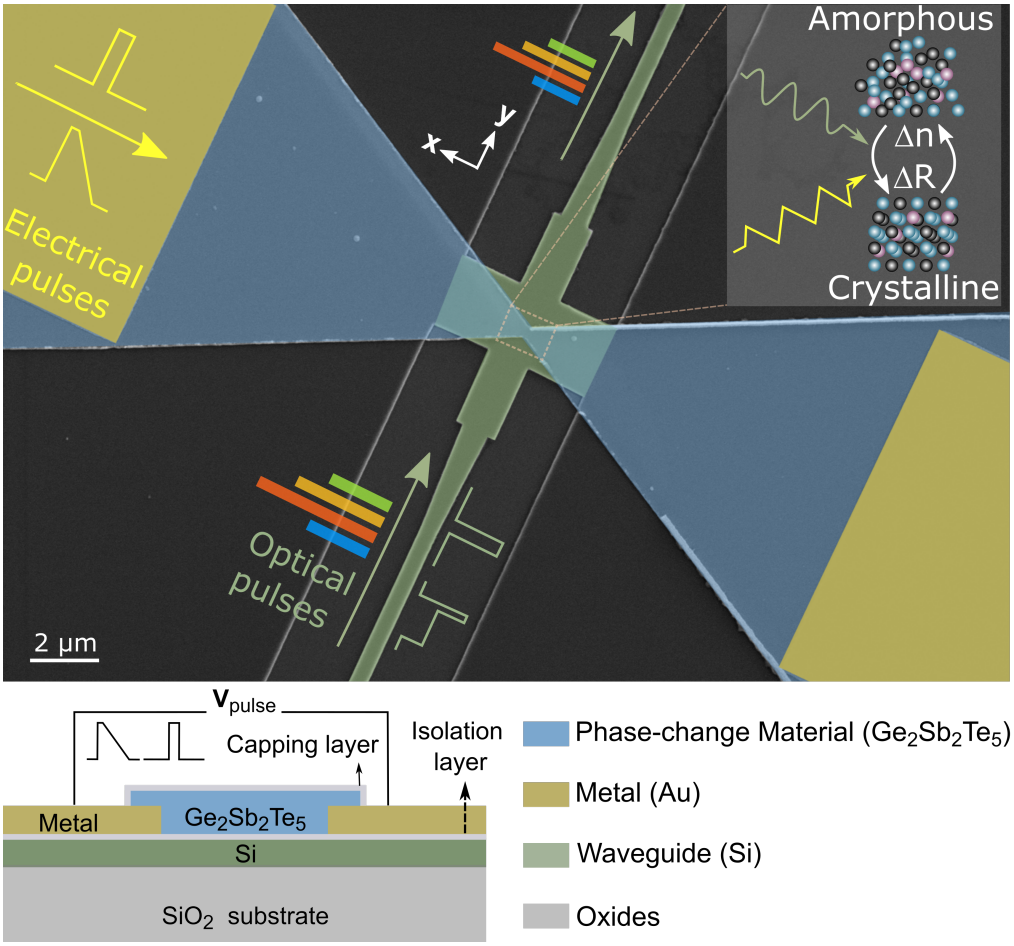

Figure 5.2: Device schematic. False color SEM image for a device with a 450-nm constriction (green: the waveguide with crossing structure; blue-gray: GST; dark-gold: gold pads). Scale bar: 2 μm.

5.2.1 Device schematics

A false color SEM image and the cross-sectional schematic of the proposed device are shown in Figure 5.2. The device combines a waveguide crossing structure with one-step lithography-defined phase-change material. The phase change material (blue-gray) is designed as a bow-tie constriction structure with the narrowest feature size ranging from 100 to 450 nm (details in Section 5.2.3). Electrical signals are

supplied and read out via metal contact pads (dark-gold) away from waveguides, and optical signals via the waveguide underneath (green), coupled by an optimized crossing structure.

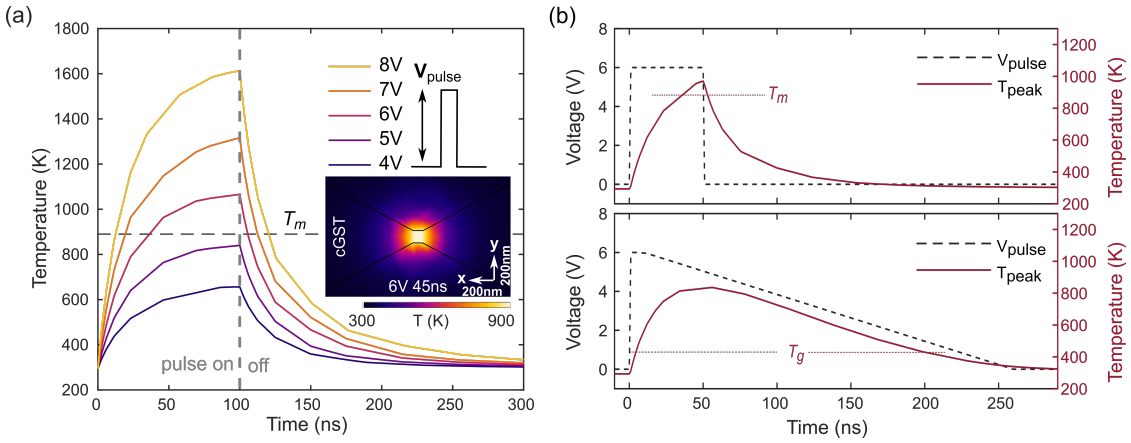

Figure 5.3: Simulated (via COMOSL Multiphysics) temporal peak temperature profiles. (a) Temporal peak temperature curves for GST region when applying square pulses on the pads with different voltage amplitudes and a fixed 100 ns pulse width. A pulse of 6 V with a 40-ns duration suffices to heat the GST above its melting temperature. T_m: melting temperature for GST. *Inset*: simulated temperature profile for the constriction region after a 6-V, 45-ns amorphization square pulse. *cGST*: crystalline GST. Scale bar: 200 *nm*. (b) Temporal peak temperature profile (red line) during an 6-V, 50-ns amorphization pulse (dashed black line) and a crystallization pulse (a 6-V, 10-ns pulse followed by a 250-ns triangular decay) for a device with a 100-nm constriction T_g: Glass transition temperature for GST.

When a voltage pulse is applied to the metal pads, the current-induced Joule heating will be confined to the narrow high resistance region of the device, as validated through the FEM simulation profile in Figure 5.3(a). For a crystalline-state device, a voltage pulse of sufficient pulse amplitude and pulse width will melt-quench the phase-change material, leading to a phase transition to amorphous phase, i.e. switching of the device. The induced transmission and current change are read out simultaneously, and likewise the phase transition induced by optical stimuli (high-energy optical pulses) can also be read out both electrically and optically.

Figure 5.3(b) provides simulated temporal peak temperature profiles for phase change material during amorphization and crystallization pulses. The short amor-

phization pulse rapidly heats the device above the melting temperature ($T_m \approx 890\ K$ [181]) of GST with a fast quenching, enabling the phase transition from the crystalline to amorphous state. Similarly, the long triangular decay pulse heats the device above the glass transition temperature ($T_g \approx 415\ K$ [181]) with slow cooling, leading to crystallization.

5.2.2 Design of the crossing structure

In the waveguide crossing design, a multimode interferometer (MMI) crossing structure is exploited [186] to create interactions between the light and phase-change material.

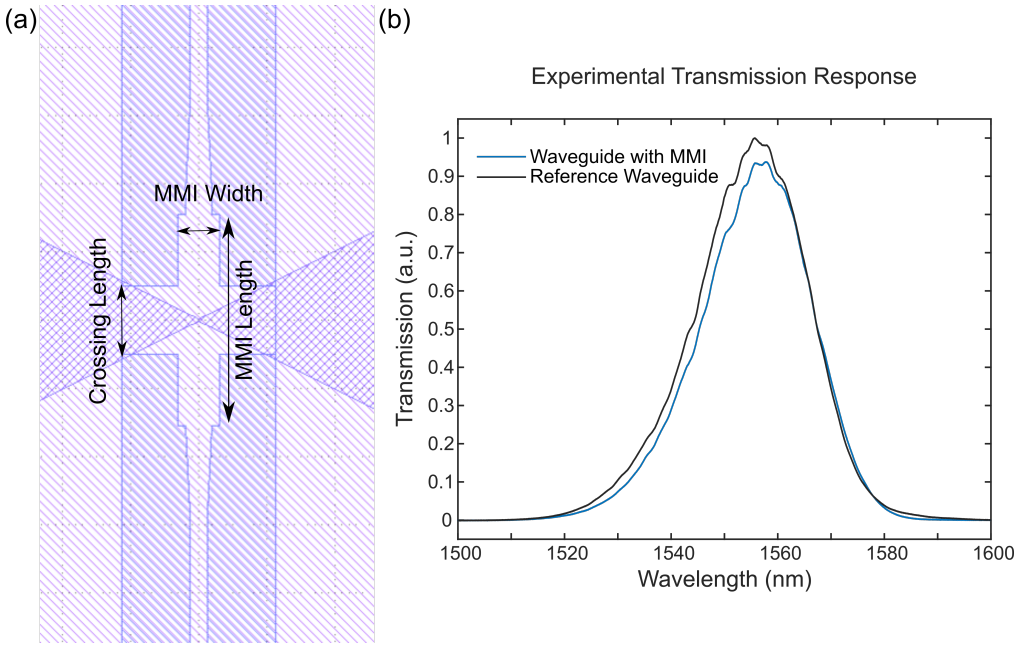

Figure 5.4: Details for the multi-mode interferometer (MMI) crossing design. (a) Design parameters for the MMI structure, with a 3-μm taper that transitions from a 500 nm to an 800 nm waveguide width at both ends. (b) Experimental transmission response for the optimized MMI design with crossing length = 2 μm, MMI length = 6.2 μm and MMI width = 1.2 μm. The bandwidth is limited by the grating couplers.

The MMI is extended to the GST region to create planar devices, ensuring that the GST film is deposited on the same height and maintaining good conductivity. Based on this requirement, the crossing length is fixed at 2 μm, while the other MMI parameters as indicated in Figure 5.4 (a) are optimized with Lumerical FDTD

solution to minimize the insertion loss. 94% ($-0.27\ dB$) experimental transmission is obtained before phase change material deposition, compared to 100% transmission for a plain waveguide on the same chip [Figure 5.4(b)].

5.2.3 Design of the constriction structure

The constriction structure used in this work is designed with the following parameters [Figure 5.5(a)-(c)]: constriction length, constriction width and thickness.

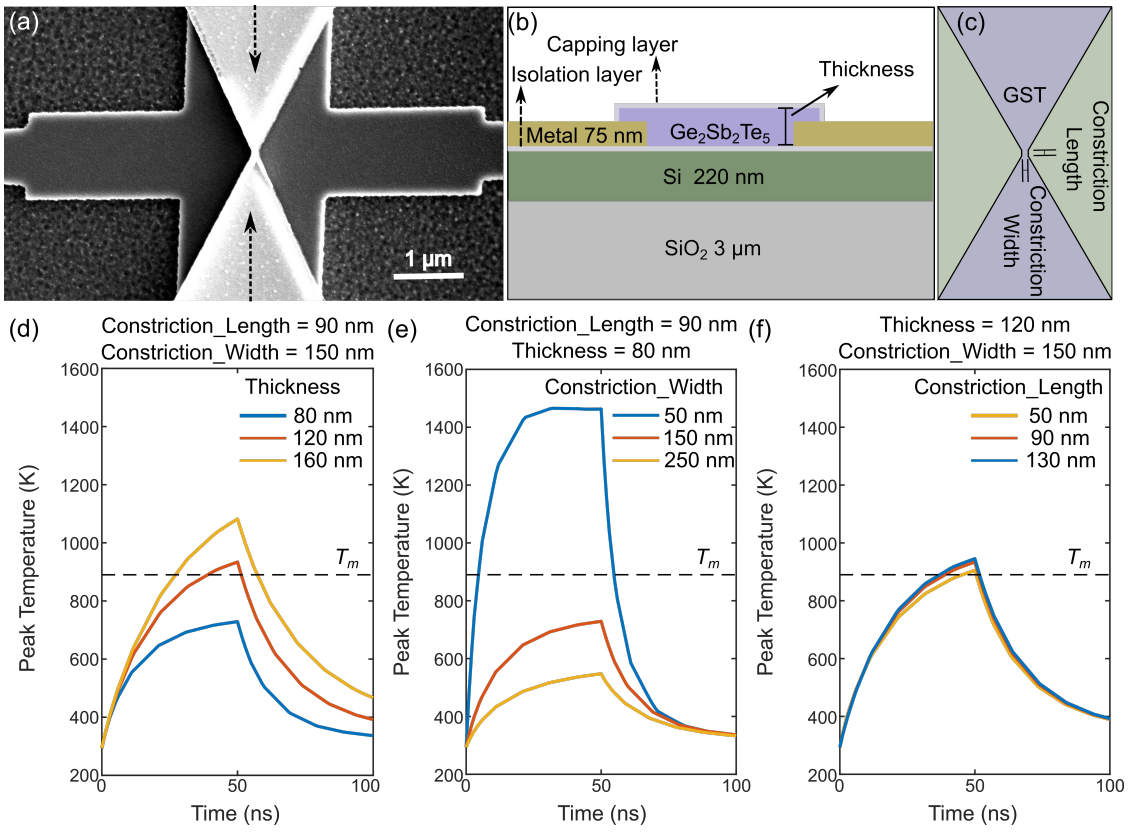

Figure 5.5: Design parameters for the phase-change material constriction. (a) SEM image for the central region of the device. (b) Cross-section of the device at the dashed line in (a). (c) Top view of the device with indications of constriction width and constriction length. (d)-(f) COMSOL simulated temporal peak temperature distributions for different material thickness, constriction width, and length with an 8-V, 50-ns pulse applied at the device metal contacts. For better comparison, the curve colors is mapped from higher device resistance (blue) to lower resistance (yellow).

COMSOL simulations have been carried out to explore how the above design parameters of the phase-change geometry affect the electrical switching performance of the device. Figures 5.5(d)-(f) plot the simulated temporal peak temperature

profile of the phase-change materials region when an 8 V, 50 ns square pulse is applied to the metal electrodes.

Devices with higher thicknesses [Figure 5.5(d)] exhibit lower resistance and provide a larger current. Thus, the temperature rises faster to the melting temperature and supports faster switching speed. Conversely, devices with smaller constriction widths [Figure 5.5(e)] exhibit higher resistance and lower current, but better heat confinement for the constriction region. The heat confinement accelerates the Joule-heating process for the constriction, enabling fast and low-energy electrical switching as well. Lastly, the effect of constriction length is not as significant as the other two parameters [Figure 5.5(f)].

In summary, for fast and low-energy electrical switching, higher thickness and smaller constriction widths are preferred. However, there is an upper thickness limit (around 150 nm) for phase-change materials to achieve reversible switching, constrained by the required critical cooling rate [187]. Narrower constriction widths also limit the maximum transmission contrast (shorter evanescent-coupling length). Thus, the design of the constriction parameters must balance complex trade-offs among the parameters, with careful consideration of the specific application context.

In this work, 100-450 nm constriction width has been used to balance the trade-off between switching energy and transmission contrast. The thickness and length of the constriction have been fixed at 150 nm and 90 nm, respectively.

5.2.4 Optical readout of the electrical switching

Thermal and optical simulations are further conducted to investigate the electrical switching performance and related optical responses of the device. Figure 5.6(a) presents the simulated light propagation profiles with GST in fully crystalline (cGST) and fully amorphous (aGST) states, showing clear field contrast at the

output ports (64% and 75 %, respectively normalized to the field at the input port). The following discussion focuses on the amorphization process.

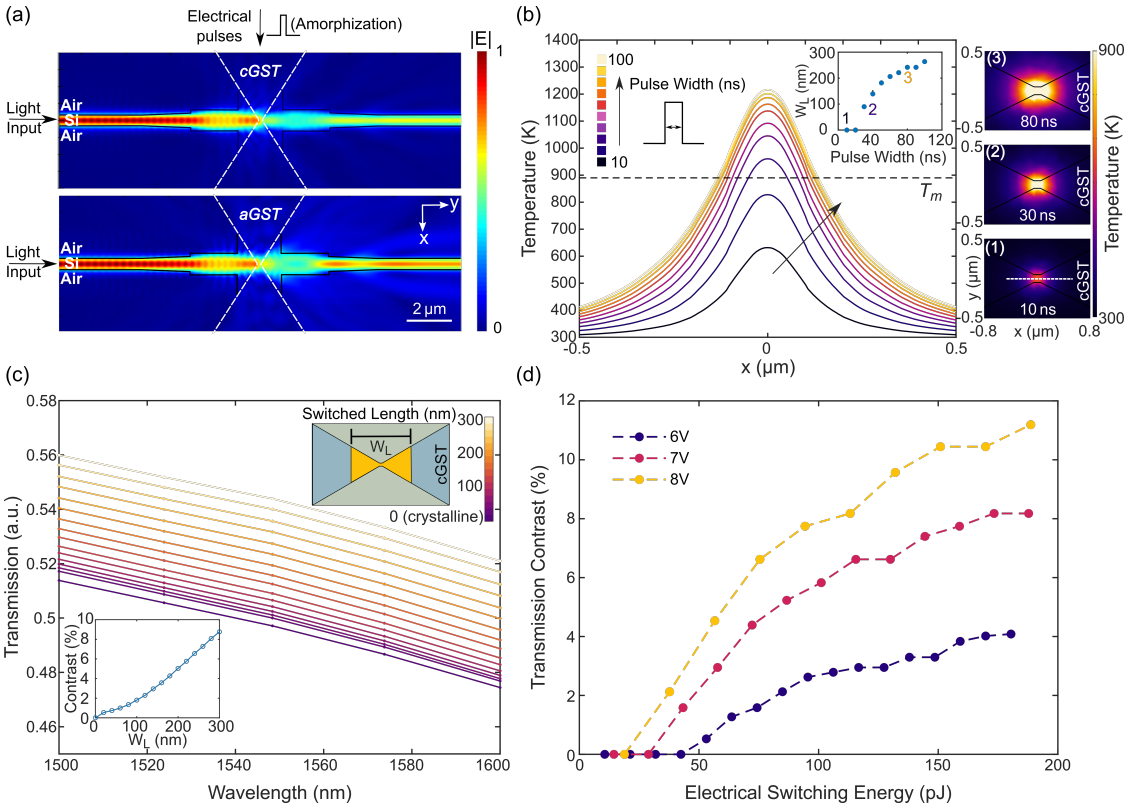

Figure 5.6: Simulated electrical switching performance. (a) Simulated (via Lumerical FDTD solutions) light propagation profiles of the devices with amorphous GST (aGST) and crystalline GST (cGST). The constriction width is 150 nm for the simulated device. (b) Simulated temperature distribution (via COMSOL Multiphysics) at a device y-cutline (white dashed line in inset (1)) when applying voltage pulses with different pulse widths (10 ns-100 ns with a 10-ns increment, fixed amplitude at 7 V). The switched lengths (W_L, device region with temperature over melting temperature) are further obtained based on these cut-line temperature profiles. *Right column inset*: corresponding simulated 2D temperature distribution for the device center slice with varied pulse widths (80 ns, 30 ns, 10 ns, from top to bottom). *Middle inset*: the relationship between voltage pulse width and device switched length(W_L). The constriction width is 100 nm for the simulated device. (c) The simulated (via Lumerical FDTD Solutions) transmission change with different switched lengths. *Inset*: the relationship between switched length and transmission contrast at $\lambda = 1574\ nm$. (d) Relationship between calculated electrical switching energy and transmission contrast based on (b) and (c). Pulses parameters are 6 V 10-170 ns, 7 V 10-130 ns, 8 V 10-100 ns with a 10-ns increment. Programming current and energy are calculated based on initial resistance (crystalline state) of the device.

With increasing amplitude [Figure 5.3(a)] or width [Figure 5.6(b)] of the electrical pulse, a larger portion of the material is heated above the melting

temperature, thereby forming a larger amorphous region. Different volume ratios of amorphous and crystalline material provide intermediate electrical and optical readout levels between fully crystalline and fully amorphous states, enabling multilevel accessibility.

To explore the relationship between device transmission and electrical switching energy, a model is built to calculate the broadband transmission response with different switched lengths [Figure 5.6(c)]. Here, the switched length (W_L) is defined as the length of the region with temperature over the melting temperature after an amorphization pulse. The transmission contrast increases with an increase in electrical switching energy, and a higher pulse amplitude provides a lower minimum switching energy due to the shorter pulse width required [Figure 5.6(d)]. The simulated minimum switching energy is less than 37 pJ (8 V, 20 ns).

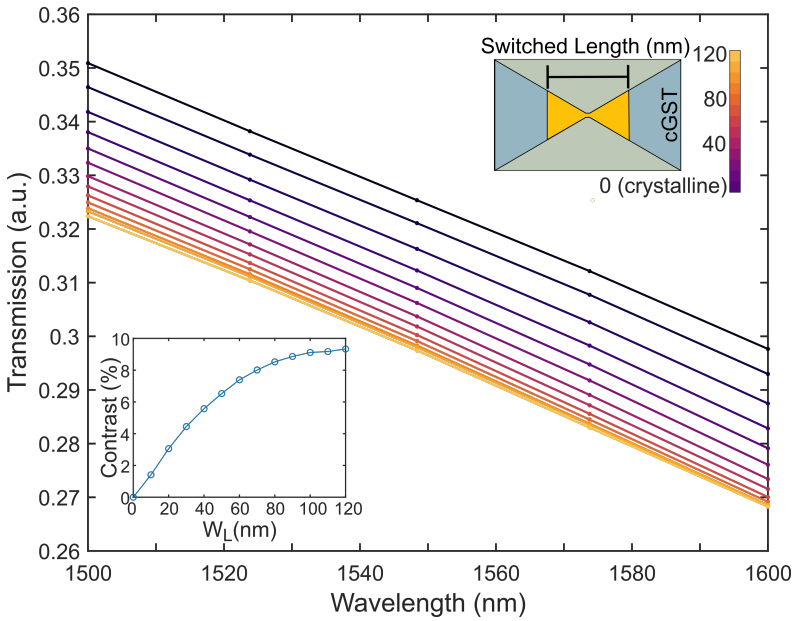

Figure 5.7: The simulated (via Lumerical FDTD Solutions) transmission change with different switched lengths for a 300-nm constriction device. *Inset*: the relationship between switched length and transmission contrast at $\lambda = 1574\ nm$.

Here, the transmission contrast is defined as $\Delta T/T_{cry}$, where ΔT and T_{cry} represent the transmission change and the transmission for the crystalline state, respectively.

For small constriction widths [Figure 5.6(c)], phase transitions from crystalline to amorphous increase transmission, so $\Delta T = T_{amo} - T_{cry}$ (T_{amo} represents the transmission readout of the partially amorphous state). For larger constriction widths, the phase transition from crystalline to amorphous reduces the transmission, so ΔT is defined as $T_{cry} - T_{amo}$. The variation in the direction of the pulse-switched transmission change can be ascribed to the increased scattering loss at the boundary between the amorphous and crystalline regions as the constriction width increases. The simulated transmission change for a representative device with a 300-nm constriction is shown in Figure 5.7, where the transmission decreases with increased switched lengths. The experimental results in Figure 5.12 and Figure 5.16 also match the simulation.

The current contrast is defined as $(I_{cry} - I_{amo})/I_{cry}$, where the crystalline state always provides higher current than the partially amorphous state.

5.3 Device Fabrication and Characterization

This section begins with proof-of-concept experiments using electrical phase-change devices on SiO_2 substrates. Then, the complete fabrication process flow is introduced, followed by basic morphological, electrical, and optical characterizations of the devices.

5.3.1 Electrical constriction devices

As a proof of concept, the first set of experiments have been carried out for devices without waveguides on SiO_2 substrates to confirm that constriction structure-based GST devices are electrically switchable.

150 *nm* GST plus a 10 *nm* ZnS−SiO_2 capping layer has been deposited on SiO_2 substrates, which were pre-patterned with 5 *nm* Cr/75 *nm* Au electrodes [Figure 5.8(a)]. Fabricated devices are thermally annealed (250 °C for 5 min) to the crystalline state before electrical switching experiments.

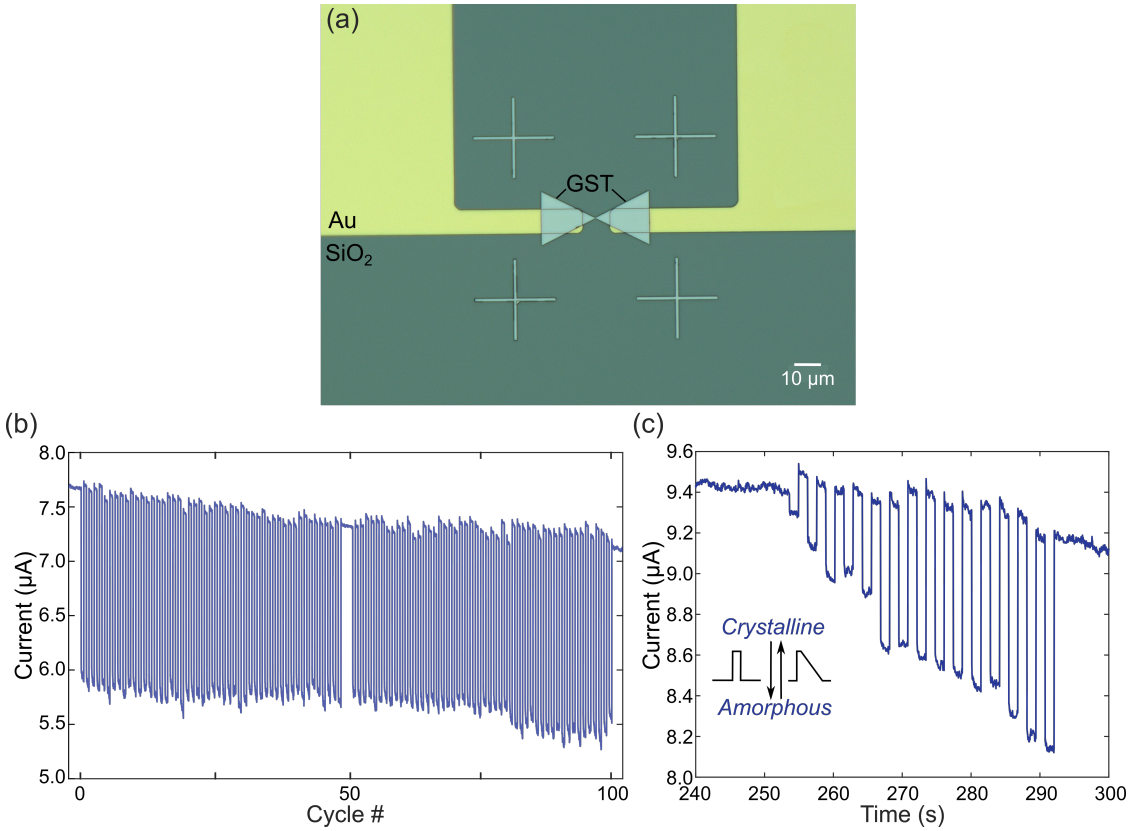

Figure 5.8: Electrical switching performance on a SiO_2 substrate. (a) Optical microscope image for an electrical constriction device on a SiO_2 substrate. (b) Reversible switching for 100 cycles. The initial state of the device is crystalline (high conductivity). One switching cycle includes switching the device to the amorphous state (low conductivity) with an amorphization pulse then back to the crystalline with a crystallization pulse. The amorphization pulse amplitude is fixed as a 1.8-V, 100-ns pulse (25 pJ), and the crystallization pulse as a 1.6-V, 100-ns pulse with a 500-ns triangular decay tail. The interval between amorphization and crystallization pulses is ~ 1 s. (c) Multilevel switching results. The amorphization pulse amplitude is fixed at 1.3 V, with the pulse width increasing from 10 ns (1.6 pJ) to 150 ns (24 pJ) in 10-ns increments. The crystallization pulse is fixed at a 1.1-V, 100-ns pulse with a 500-ns triangular decay tail.

Both binary and multilevel switching measurements have been carried out with the switching pulse parameters obtained experimentally. 100 cycles of binary switching between the crystalline and amorphous states have been demonstrated with stable contrast $> 20\%$ [Figure 5.8(b)]. Fixing the rectangular amorphous pulse amplitude as 1.3 V and varying the pulse width from 10 ns to 150 ns in 10-ns increments, multilevel switching has been achieved [Figure 5.8(c)]. 1% current contrast has been attained for the shortest pulse, providing an ultralow switching

energy of 1.6 pJ.

5.3.2 Device fabrication process flow

The full device is fabricated on diced silicon-on-insulator (SOI) substrates.

Waveguides, MMIs, and grating couplers are patterned using EBL with positive photoresist CSAR 62, followed by RIE with an etching depth of 110 nm. Then 5 nm AlO_x is deposited via atomic layer deposition (ALD, Savannah S200) for electrical isolation prior to EBL patterning and thermal evaporation of 5 nm Cr/75 nm Au for the metal pads.

Finally, 150 nm GST plus a 10 nm ZnS$-SiO_2$ capping layer (protecting GST from oxidation) is patterned with a third EBL step and deposited with an RF sputtering system (AJA). Fabrication details for each step have been introduced in Section 3.2.

5.3.3 Basic characterization

After device fabrication, atomic force microscope (AFM) images [Figure 5.9 (a)] have first been collected to confirm the thickness of the device. The height profiles in Figure 5.9(b)-(c) exhibit a height transition from around 150 nm [Figure 5.9(b)] to 70 nm [Figure 5.9(c)], from the side to the center, attributed to the shadowing effect during the RF sputtering process. Such gradual height transitions provide additional heat confinement for the constriction structure.

The basic transmission and resistance characterizations have been conducted thereafter. The optical and electrical measurement setups are introduced in Sections 3.3.2 and 3.3.3, respectively. The loss of a typical device in the amorphous state (as deposited) with a 100-nm constriction width is around 2 dB [Figure 5.10(a)], increasing to 5 dB for a device with a 450-nm constriction width. Figure 5.10(b) presents the current-voltage characteristics of a typical device in its two states, the

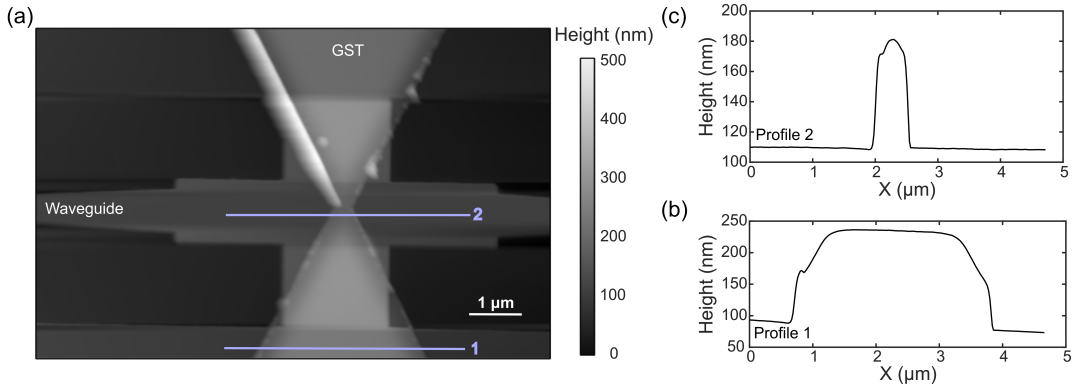

Figure 5.9: Atomic force microscope (AFM) Characterization. (a) AFM collected height profile for the constriction region of the device. (b)-(c) Height profiles along cut-lines 1-2 in (a), respectively.

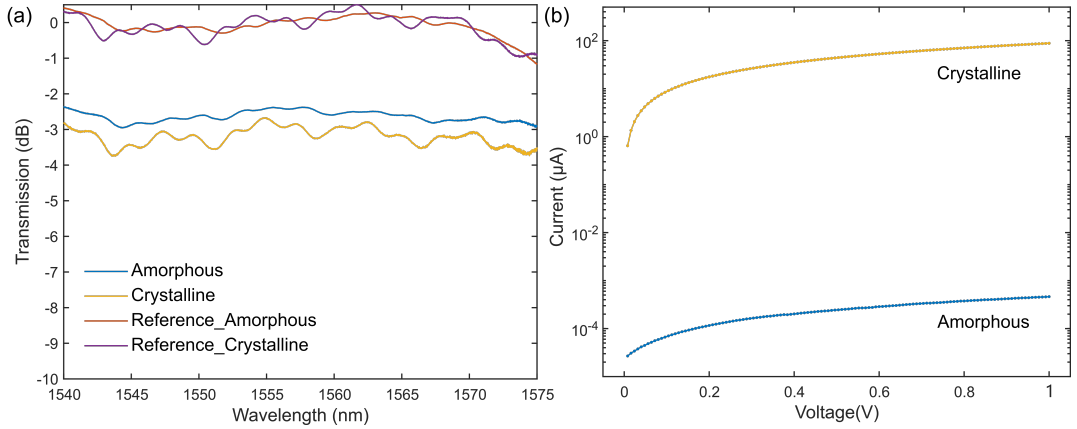

Figure 5.10: Basic Device Characterization. (a) Transmission for a typical device in its two states with a 100-nm constriction width. Transmission is normalized to the transmission of plain waveguides with MMIs on the same chip. (b) Current-voltage measurements for a typical device in its two states with a 130-nm constriction width.

resistance of the device dropping from 2.2 $G\Omega$ in its amorphous state to 11 $k\Omega$ in the crystalline state.

5.4 Switching Performance Characterization

Experimental measurements for both optical and electrical switching are carried out using a custom electro-optic setup (Section 3.3). Fabricated devices are thermally annealed (250 °C for 5 min) to the crystalline state before switching experiments.

5.4.1 Optical switching with mixed-mode read-out

Optical switching measurements are first performed to demonstrate dual electrical-optical functionality. Amplified pump pulses (λ=1571 nm) are used to program the devices, while low-power CW probe light (λ=1570 nm) is used to read the device states.

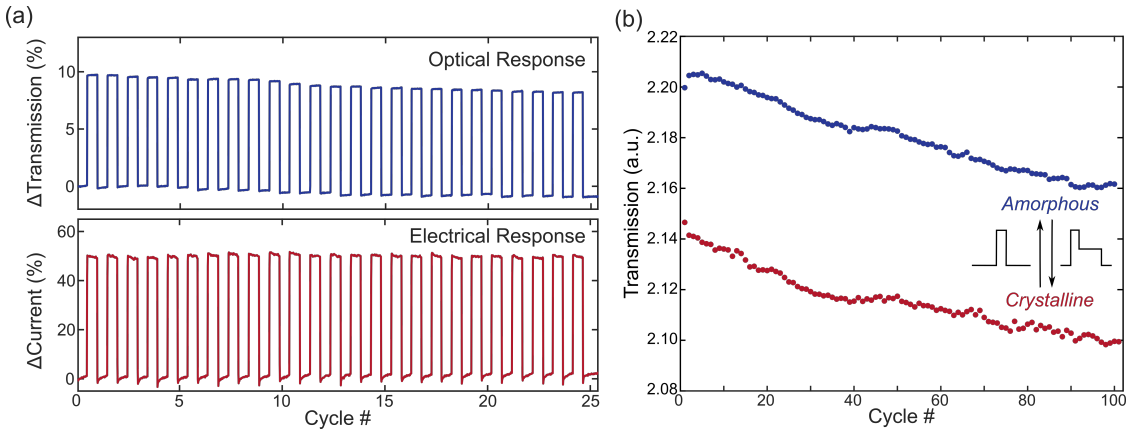

Figure 5.11: Binary optical switching performance. (a) Experimental optical switching with both optical and electrical readouts for a device with a 310 nm constriction. The amorphization pulse is fixed at 5.04 mW with a 25 ns pulse width, and the crystalline pulse is a 5.04 mW, 10 ns pulse with a 1.51 mW, 250 ns rectangular decay tail. (b) 100 cycles of reversible optical switching for a device with a 110 nm constriction. The amorphization pulse is fixed at 14.39 mW with a 20 ns pulse width, and the crystalline pulse is a 14.39 mW, 5 ns pulse with a 6.48 mW, 250 ns rectangular decay tail.

The programming pulse parameters are obtained experimentally. For binary switching, the optical pulse amplitude is fixed at 5.04 mW, with a 25 ns pulse for amorphization and a 10 ns pulse followed by a 250 ns, 1.51 mW rectangular tail for crystallization. Successively reversible switching events are obtained for around 10% transmission contrast and more than than 40% current contrast [Figure 5.11(a)]. The cyclability test in Figure 5.11(b) demonstrates $\sim$ 3% reversible optical switching contrast with no obvious degradation after 100 cycles. The constant drift in optical transmission can be attributed to the mechanical motion of the chip induced by the electrical probe assembly and to the formation of larger crystalline domains [162]. After initial pulsing and amorphization of the phase-change material,

ordered structures form which act as seeds for further crystal growth and are not fully amorphized with the same pulse [188]. The optical transmission thus exhibits a trend toward a more crystalline state.

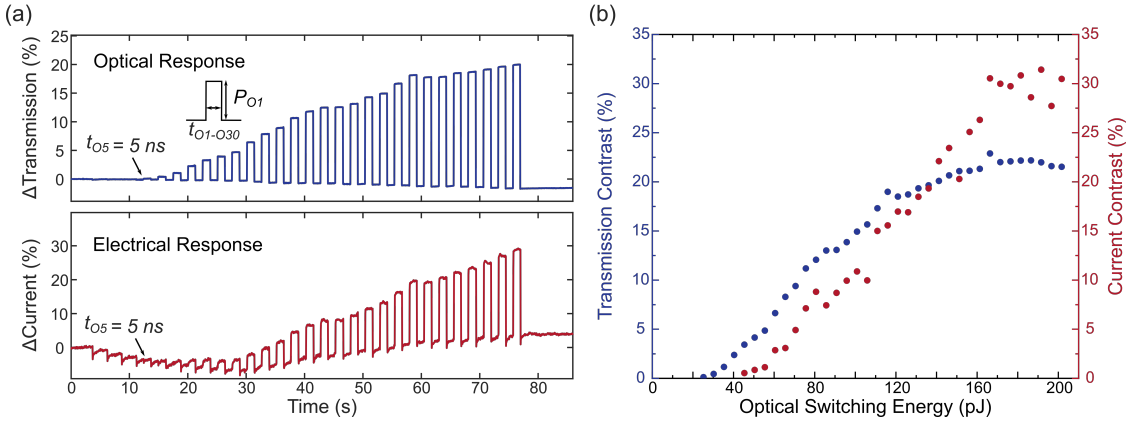

Figure 5.12: Multilevel optical switching performance. (a) Experimental multilevel optical switching with both optical and electrical readout for a device with a 310 nm constriction. Amorphization pulse amplitude is fixed at 5.04 mW (P_{O1}) with pulse widths increasing from 1 to 30 ns (t_{O1}-t_{O30}) in 1 ns increments. The crystallization pulse is fixed as a 5.04 mW, 10 ns pulse followed by a 1.51 mW, 250 ns rectangular tail. (b) Relationship between optical switching energy and switching contrast for both transmission and current. Amorphization pulse amplitude is fixed at 5.04 mW with pulse widths increasing from 1 to 40 ns in 1 ns increments, and the crystallization pulses are kept the same as in (a). The optical contrast saturates at around 160 pJ (32 ns).

In addition to binary switching, multilevel switching is also achieved with optical programming. The crystallization pulse is fixed as before and the amorphous pulse widths vary from 1 to 30 ns, maintaining the amplitude at 5.04 mW [Figure 5.12(a)]. For short pulses (1-4 ns), the pulse power is not enough for amorphization, whereas the crystallization pulse further anneals the materials, as indicated in the current trend. No optical transmission contrast is detected as a result of the limited signal-to-noise ratio. Increasing the amorphous pulse width to 5 ns induces over 0.05% transmission contrast using 25.2 pJ switching energy, and a 30 ns pulse achieves more than 20% contrast in both the electrical and optical domains. Figure 5.12(b) further illustrates the relationship between optical switching energy and switching contrast. The switching contrast increases first with increasing switching

energy, then saturates at around 160 pJ with over 20% transmission contrast and around 30% current contrast.

5.4.2 Electrical switching with mixed-mode readout

Electrical switching measurements were further carried out. Electrical programming voltage pulses are applied to the metal pads of the devices. A low-power DC bias signal (100 mV unless otherwise specified) and probe light ($\lambda = 1570\ nm, 10\ \mu W$) are used to monitor the current and transmission, respectively.

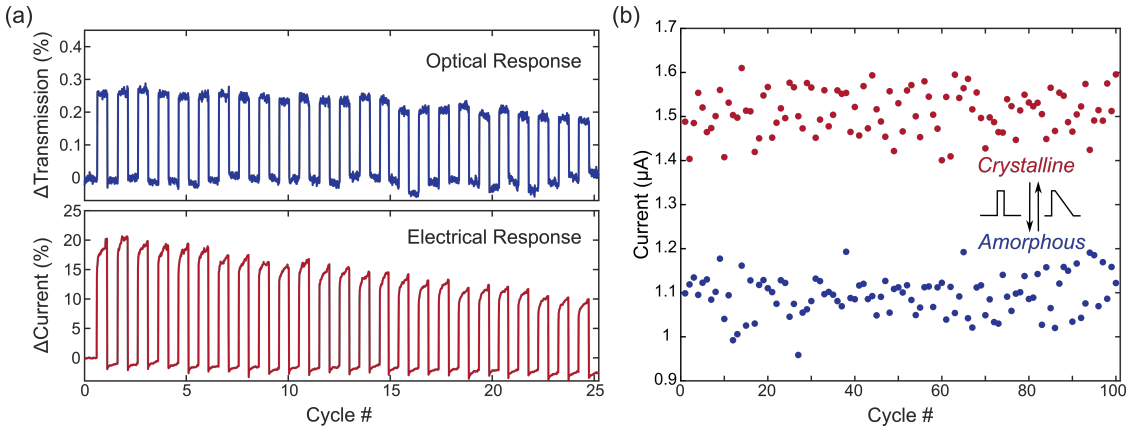

Figure 5.13: Binary electrical switching performance. (a) Experimental electrical switching with both optical and electrical readout for a device with a 120 nm constriction. The amorphization pulse is fixed at 4.5 V and 30 ns with a crystallization pulse as a 30 ns, 2.7 V pulse followed by a 250 ns triangular decay. The interval between amorphization and crystallization pulses is $\sim 1\ s$. (b) 100 cycles of reversible electrical switching for a device with a 265 nm constriction. The amorphization pulse is fixed at 3.5 V with a 10 ns pulse width, and the crystalline pulse is a 2 V, 10 ns pulse with a 250 ns triangular decay tail.

The device switching parameters for both amorphization and crystallization were experimentally determined. In binary switching measurements, the amorphization pulse is fixed at 4.5 V and 30 ns (24 pJ including bias power) with a crystallization pulse as a 30 ns, 2.7 V pulse followed by a 250 ns triangular decay. Using these switching parameters, 25 cycles of sequential switching events have been demonstrated with both electrical and optical readouts, as shown in Figure 5.13(a). The maximum current contrast is over 15%, and a transmission contrast of over 0.2%

is obtained. The dynamic response of electrical switching is discussed in Section 5.4.4. To investigate device reliability, the cyclability test in Figure 5.13(b) demonstrates 100 cycles of reversible electrical switching with $\sim$ 20% current contrast. The confinement of heat to the constriction area results in a low switching energy. Here, the amorphization pulse (3.5 V, 10 ns) for electrical switching consumes only 2.1 pJ energy (switching current $\sim$ 58 μA).

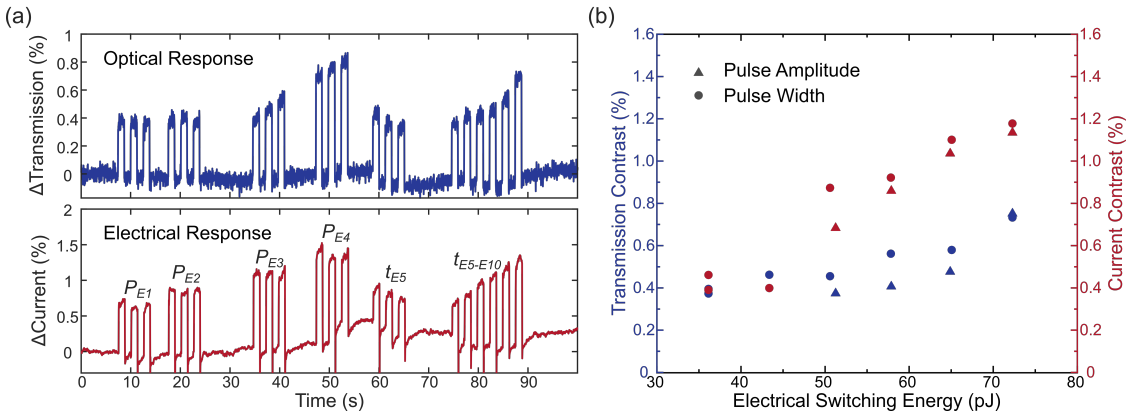

Figure 5.14: Multilevel electrical switching performance. (a) Experimental electrical switching with both optical and electrical readouts for a device with a 130 nm constriction. P_{E1}-P_{E4}: amorphization pulse amplitude 8–9.5 V with a 0.5 V increment and fixed pulse width at 10 ns (each parameter repeated 3 times); t_{E5}-t_{E10}: 5–10 ns with a 1 ns increment and fixed pulse amplitude at 9.5 V. The crystallization pulse is fixed as an 8 V, 10 ns pulse followed by a 250 ns triangular decay tail. (b) Relationship between electrical switching energy and switching contrast for both transmission and current. The contrast is taken as the average of different switching cycles with the same amorphization pulse.

Next, to demonstrate multilevel switching, the crystallization pulse is fixed as an 8 V, 10 ns pulse followed by a 250 ns triangular decay tail, with the amplitude and pulse widths of the amorphous pulses varying from 8 V to 9.5 V (P_{E1}-P_{E4} in Figure 5.14(a)) and from 5 to 10 ns (t_{E5}-t_{E10} in Figure 5.14(a)), respectively. Distinguishable contrasts are obtained with increasing pulse widths or amplitudes, leading to higher contrast in both optical and electrical domains. Figure 5.14(b) illustrates the relationship between contrast and pulse energy. The experimental switching energy is similar to the simulated results in Figure 5.6(d) but with a lower transmission contrast. The deviation can be attributed to the simplified simulation

model, which only considers the length of the switched region, but the whole region is not fully switched in the experiments, leading to a smaller optical contrast.

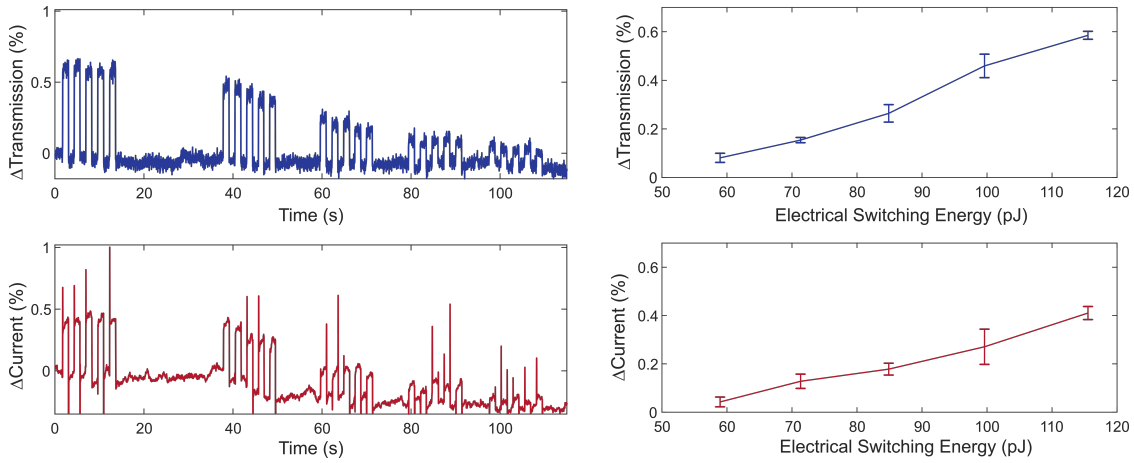

Figure 5.15: Multilevel electrical switching and multilevel stability for a device with a 130 *nm* constriction. Amorphization pulse amplitude is varied from 7 V to 5 V with a 0.5-V decrement and the pulse width is fixed at 30 ns. The crystallization pulse is fixed as a 5-V, 10-ns pulse followed by a 250-ns triangular decay tail. The standard deviations for 5 repeated write cycles are calculated to be 47%, 23%, 14%, 27%, 6% for current contrast and 23%, 7%, 14%, 10%, 3% for transmission contrast.

The stability of electrically induced multilevel switching has been further investigated [Figure 5.15]. Five unique levels are resolved with reasonable stability. The drift in current levels / resistance can be attributed to the randomness of the atomic rearrangement and structural relaxation of the amorphous phase [8].

5.4.3 Discussion on electrical switching conditions

During experiments, it was also observed that the amplitude of read-out DC bias has an impact on the electrical switching contrast.

Increasing the DC voltage from 100 mV to 500 mV (probe power as 10 μW) improves the electrically switched transmission contrast. 50 cycles of electrical switching with around 3% (0.13 dB) transmission contrast have been demonstrated [Figure 5.16(a)]. The electrical switching energy is calculated to be 19.5 pJ (30 ns, 9 V plus 0.5-V bias voltage), providing ultra-low energy consumption at 0.15 nJ/dB. The gradual drift to the amorphous state can be attributed to elemental segregation,

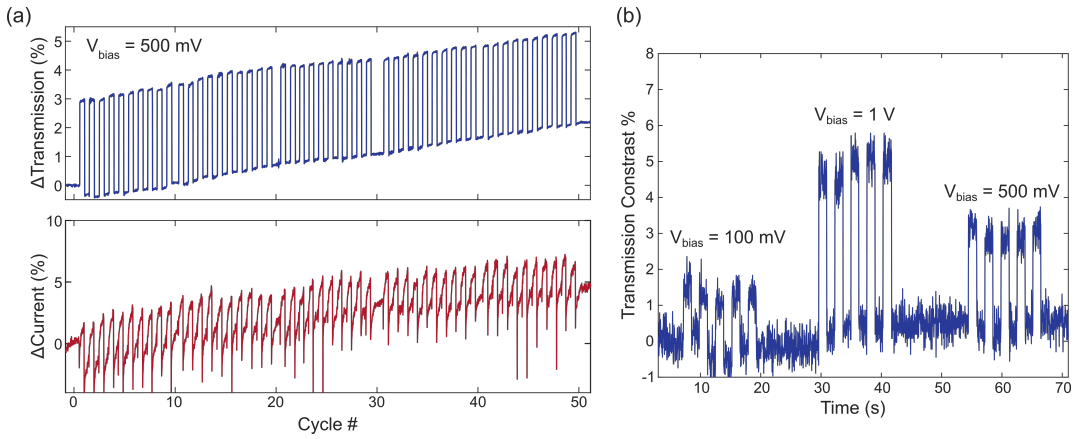

Figure 5.16: Electrical switching under various DC bias voltages. (a) 50 cycles reversible switching with both optical and electrical readout under 500 mV DC bias. The amorphization pulse is 9 V, 30 ns and the crystallization pulse is fixed at 5 V, 10 ns with a 250 ns triangular tail, with 500 mV DC bias and 10 μW probe light power to readout current and transmission change. The interval between amorphization and crystallization pulses is $\sim 1\ s$. The device resistance is around 143 kΩ, with a 310-nm constriction width. (b) Electrically switched transmission contrast under different bias conditions. 5 cycles of reversible switching have been collected under 100 mV, 1V and 500 mV bias voltages, respectively. The amorphization pulse is fixed at 6 V, 50 ns and the crystallization pulse is fixed as 5 V, 10 ns with a 250 ns triangular tail. The device is around 187 kΩ with a 450-nm constriction width.

i.e. elements Te and Sb move towards different directions along the electrical field or thermal gradient after electrical switching cycles, leading to smaller crystalline domains [189].

The impact of the read-out DC bias voltage is further presented in Figure 5.16(b). Varying the bias voltage from 100 mV to 1 V boosts the switched transmission contrast from $\sim 2\%$ to $\sim 5\%$ ($\sim$1-order improvement compared to the first plasmonic phase-change electro-optic memristor [158]). Increasing the bias voltage improves the maximum transmission contrast by boosting the absorption of the amorphization pulse through both thermo-optic and electronic effects in GST [190–193].

5.4.4 Dynamic response

The experimentally measured thermo-optic response of the constriction device is presented in Figure 5.17. The setup is customized based on the mixed-mode electrical switching setup introduced in Section 3.3.3. The output probe light signal

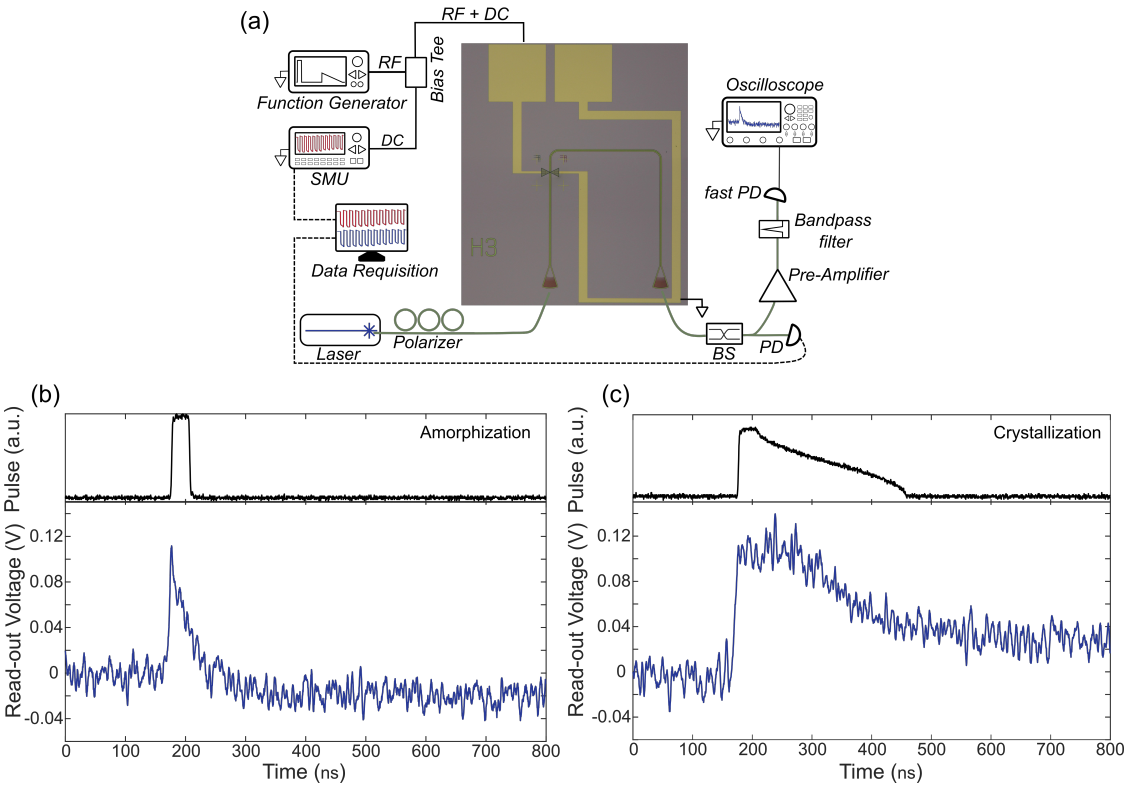

Figure 5.17: Experimental dynamic response for the device with a 450-nm constriction. (a) Measurement setup. SMU: source measure unit; PD: photodetector; BS: 99:1 beam splitter, directing 99% of the output signal to the pre-amplifier and 1% to the PD. (b)-(c) Read-out voltage profiles of the high-speed photodetector, indicating device transmittance when applying a 9.5-V, 30-ns amorphization pulse and a 7.5-V crystallization pulse (a 30-ns square pulse followed by a 250-ns triangular tail). The read-out voltage curve is an average over 3 cycles.

is amplified (AEDFA-PA-35-B-FA, Amonics) after a 99:1 splitter (TF1550R1A1, Thorlabs), and after the band-pass filter (OTF-320, Santec) collected by the high-speed photoreceiver (model 1811, New Focus) and oscilloscope (TDS7404, Tektronix).

The post-excitation 1/e decay time [18] for a 30-ns amorphization pulse is calculated to be 30 ns, and the total settling time for a crystallization pulse is 290 ns, providing operational speeds of 60 ns and 290 ns for amorphization and crystallization, respectively. The reset speed (amorphization) of the constriction device is faster than that of heater-based integrated phase-change electro-optical devices (>100 ns [103, 104, 160–162]), and the set speed (crystallization) is

comparable ($\sim$ 200 ns and 556 ns for P^{++} doped-Si heaters [160, 161], $>$ 80 μs for PIN heaters [103, 104] and $>$ 220 μs for graphene heater-based devices [162]).

5.4.5 Switching energy analysis

Figure 5.18 provides a summary of the energy performance of the constriction devices. The tens of picojoule switching energy obtained for mixed-mode readouts, i.e. simultaneous readouts in both electrical and optical domains, is similar to those for plasmonic nanogap-based implementations and is more than one order smaller than those for heater-based state-of-the-art integrated electro-optical phase-change devices.

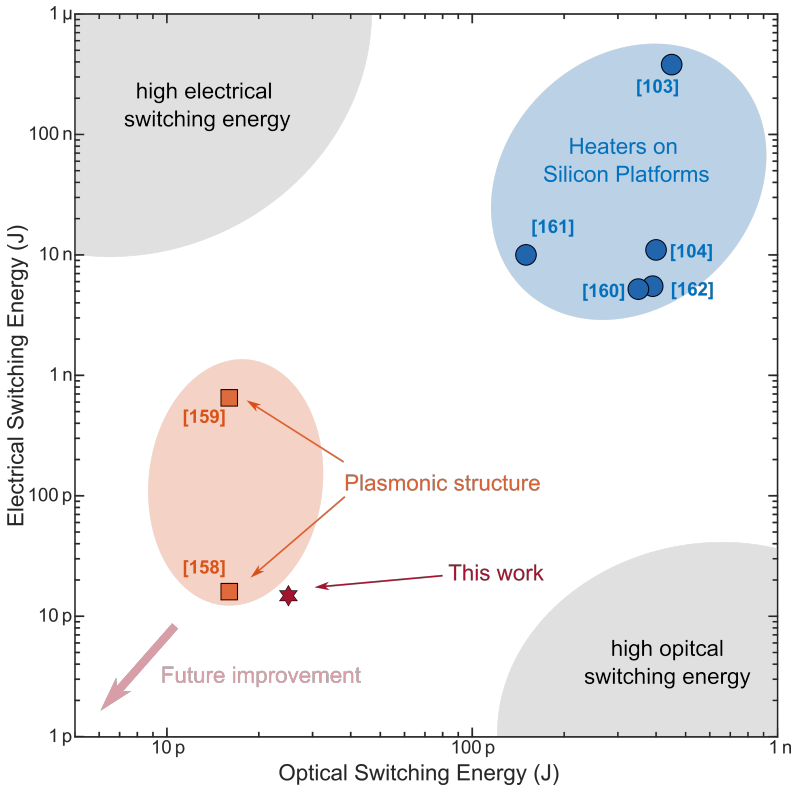

Figure 5.18: Switching energy map for different phase-change electro-optic memristor implementations. Heaters on a silicon platform ([103, 104, 160–162]): The optical switching energy is estimated based on [179]. Plasmonic structures: based on [158, 159].

Table 5.1 highlights the significance of this work by comparing its metrics with those of other nonvolatile electro-optical devices based on GST. This work reduces

Table 5.1: Performance comparison of electro-optic memristor implementations based on $Ge_2Sb_2Te_5$

Electro-optical Device with GST	Optical Switching Energy (pJ)	Electrical Switching energy** (pJ)	Active PCM Volume $W \times L \times H$ (μm^3)	Insertion loss for amorphous devices (dB)	Energy per unit modulation depth (nJ/dB)	Mixed-Mode Read	Ref
Silicon PIN diode heater		$11,000$ $(650,000)$	$0.5 \times 4 \times 0.02$	-1.6	4.0	No	[104]
Graphene heaters	~ 388.4*	$5,550$ $(860,000)$	$0.4 \times 4.73 \times 0.23$	-1	1.85	No	[162]
P^{++} doped Si heater		$5,200$ $(6,900)$	$0.5 \times 3.5 \times 0.03$	-1.78	1.7	No	[160]
Si_3N_4 plasmonic nanogap	16 ± 2	16.0 (312)	$0.05 \times 0.05 \times 0.075$	-10.45	0.62	Yes	[158]
Si_3N_4 plasmonic heaters		650 (4000)	$0.07 \times 0.2 \times 0.075$	-8.86	9.8	Yes	[159]
Constriction Structure on SOI	25.2	19.5 (26.5)	$0.3 \times 0.3 \times 0.07$ $0.3 \times 0.1 \times 0.07$	-5 -2	0.15	Yes	This work

* Using the experimental data for 4 µm $Ge_2Sb_2Te_5$ on Si from [179] as a reference.

** Amorphization (crystallization).

the high insertion loss relative to previous plasmonic implementations (from ~ 10 dB to 2–5 dB) and demonstrates a very low electrical switching energy per unit modulation depth at 0.15 nJ/dB.

5.5 Photo-detection Behavior

Besides switching performance, to fully exploit the dual electrical-optical functionality, the photoresponse behavior of the device has been further investigated.

5.5.1 Responsivity

The current-voltage (I-V) characteristics of a constriction device in the amorphous and crystalline states have been measured at different input optical power levels [Figures 5.19 (a)-(b)]. The input light is incident through the waveguides beneath

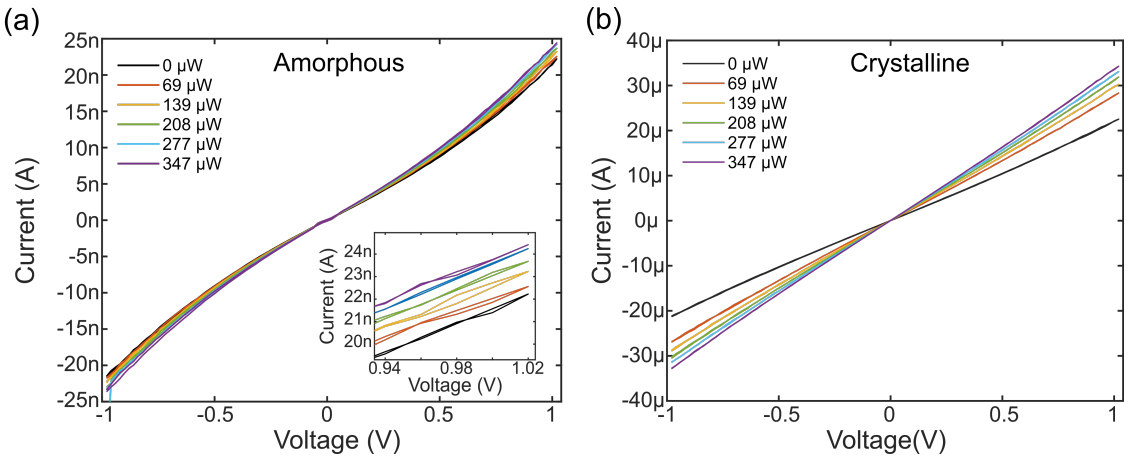

Figure 5.19: Phase-dependent photo-detection behavior. IV curves under different input optical power levels ($\lambda = 1550\ nm$) for the same device in (a) the amorphous state (as-deposited, 61 $M\Omega$) and (b) the crystalline (thermally annealed, 48 $k\Omega$) state, respectively.

the GST constriction structure. The measured photoresponse is symmetric in the positive and negative bias regions with negligible response at zero bias, indicating that the device operates in photoconductive mode [151, 152, 154].

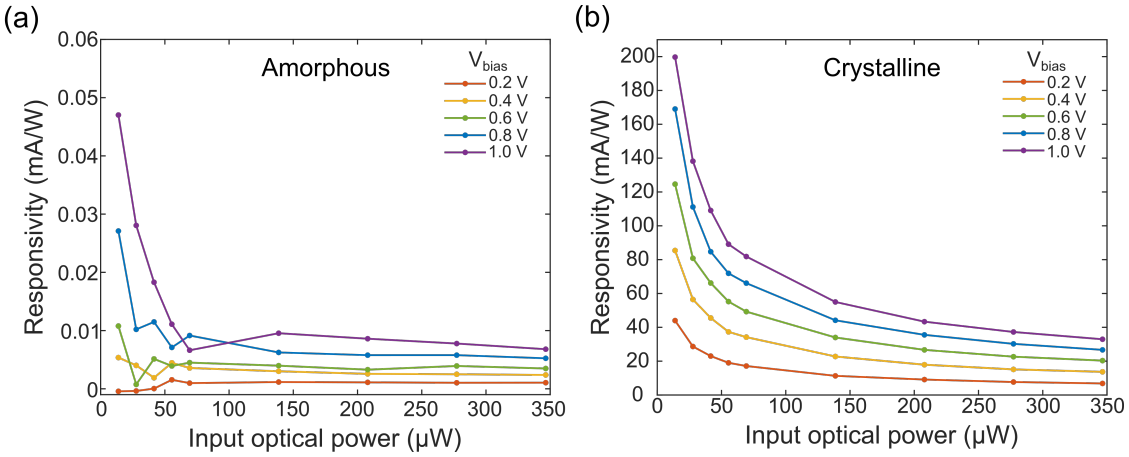

Figure 5.20: Phase-dependent responsivity ($\lambda = 1550\ nm$) of the constriction device under different bias voltages. (a) Calculated responsivity for a constriction device in the amorphous state. (b) Calculated responsivity for the same devices in the crystalline state. The device is with a 450-nm constriction.

Figure 5.20 further presents the relationship between the input optical power and the calculated responsivities of the device in its two states under different bias voltages. The photocurrent increases with higher input optical power and bias

voltages. However, the responsivity and incident optical power exhibit a negative correlation, i.e. the photocurrent saturates at high optical input power. The saturation can be attributed to saturation absorption and refractive-index-related absorption coefficient change. Similar behavior has been discussed in previous work [154].

The calculated responsivity for the same device in the crystalline state is more than 3 orders of magnitude higher than that of the amorphous state given the large resistance change between the two states. With bias voltage at 1 V, the highest responsivity of 0.2 A/W has been obtained at an incident optical power of 13.9 μW, which is close to the unity photoresponsivity of commercialized integrated photodetectors and comparable with emerging photodetectors based on two-dimensional materials [151, 152]. Notably, the active length of the device used in this work ($\sim$ 450 nm) is much shorter than that of the aforementioned integrated photodetectors (several to tens of μm). Therefore, with a sufficient active length for light absorption, the responsivity of the constriction device can be further boosted.

5.5.2 Random access and potential applications

The dynamic response of the photoconductivity has been further characterized. Figures 5.21(a)-(b) exhibit the dynamic on-off response of a constriction device in its two states, where incident light was switched on and off at 10 s intervals, with 9.3 μW increments in optical power for each cycle. The photocurrent increases sublinearly with the incident light power, showing a one-to-one mapping relationship.

Stable access to different photocurrent levels in response to different input optical power has been demonstrated for an example device in the crystalline state [Figure 5.22(a)]. The one-to-one mapping relationship between the input optical power and the photocurrent adds a new avenue for data transfer from the optical

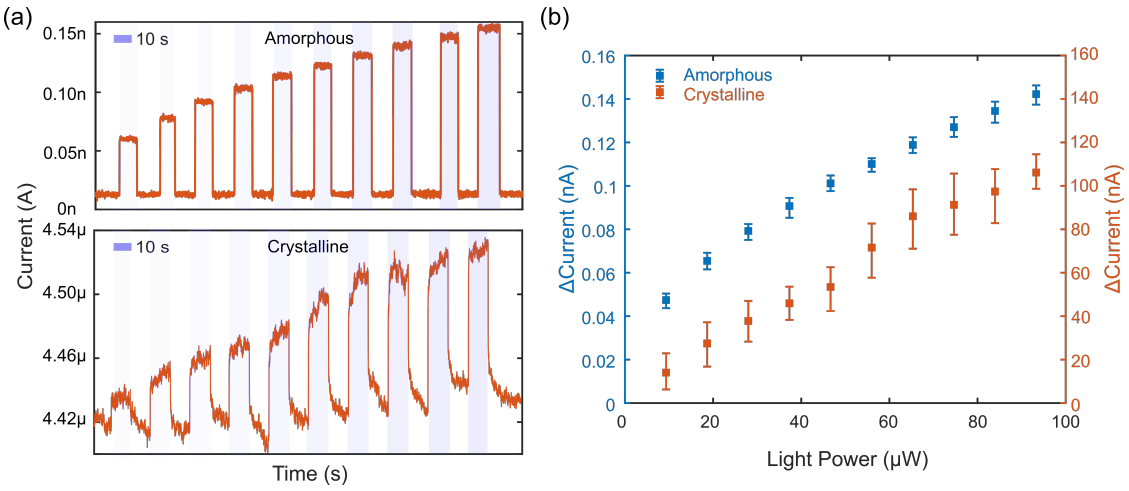

Figure 5.21: Dynamic readout of the input optical power at $\lambda = 1550\ nm$. (a) Current-time response of a device in its amorphous and crystalline states under pulsed illumination (on and off at 10 s intervals) with a 100-mV bias. The incident power varies from 9.3 μW to 93 μW with a 9.3 μW increment. (b) The relationship between photocurrent and input light power based on (a).

to the electrical domain within a computing core (example schematic in Figure 5.22(b)).

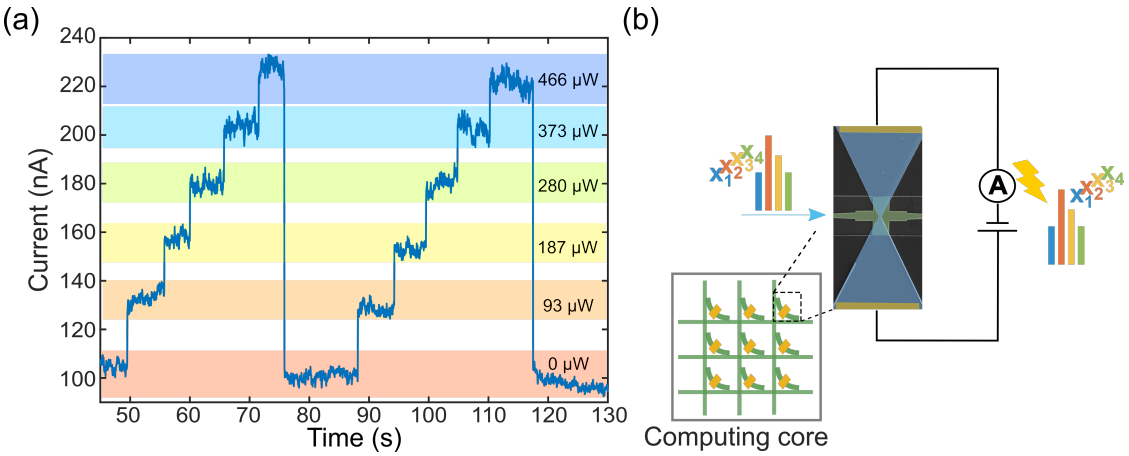

Figure 5.22: Potential applications. (a) Stable photocurrent response with different incident optical power at $\lambda = 1550\ nm$. (b) Schematic for a constriction device in an in-memory computing core. The optical power from the input port could be readout electrically from the device without adding extra optical ports or photodetectors, which provides in-situ feedback of the signal intensity.

According to the available literature, this is the first characterization of the in-plane photosensitivity for an integrated nonvolatile reconfigurable photonic device. Combining photodetection behavior and the nonvolatile reconfigurable feature in

one cell, this design opens up new opportunities to simplify the large-footprint electro-optical conversion for a photonic computing core, such as providing in situ feedback of the optical signal intensity without extra area for optical monitor ports [74] or implementing tuneable threshold detection functions.

5.6 Future Optimization

In this section, existing limitations and potential future optimizations of the constriction-based electro-optic phase-change memristor are discussed, focusing on three key aspects: the optical coupling structure, electrical switching performance, and applications.

The optical coupling structure

As explained in Section 5.2.2, the MMI crossing coupling structure is designed to create a planar device while minimizing the insertion loss. This optical coupling structure can be further optimized to enhance the optical interaction with the phase-change material constriction, utilizing methods such as inverse design [194].

Electrical switching performance

This constriction-based electro-optic memristor design has demonstrated excellent switching energy efficiency; however, other metrics of the electrical switching performance require further improvements, such as cyclibility and electrical switched transmission contrast, especially when compared to more developed heater-based technologies, which have achieved an endurance of over 1,000 cycles [162] and 4-bit programming resolution for GST [160].

The electrically switched transmission contrast of this work could be further enhanced with an optimized isolation layer or a better coupling structure, as discussed in the previous aspect. The drift and noise of the device could also be reduced with engineering efforts, such as exploiting a carefully chosen material as

the projected layer to decouple the resistance storage and the information retrieval process [195].

Applications

Although different implementations of integrated electro-optic memristors have been implemented, there are not yet practical applications to make full use of the mixed-mode readout property.

The function of electrical programming with optical readout has been extensively investigated to implement electrically programmed photonic computing core, whereas the optical programming and electrical readout properties remain underdeveloped. Free-space implementations of electro-optic memristors have explored applications such as multi-factor learning [155] and combinatorial optimization problems [196], which are still underdeveloped for integrated implementations.

5.7 Chapter Summary

In this chapter, the concept of thermal engineering is exploited to propose a novel design for implementing integrated phase-change electro-optic memristors. As far as is known, this is the first demonstration of integrated electro-optic memristors to achieve dual electro-optical functionality, low-energy switching, and potential for CMOS-compatible scaling at the same time. The electrical heat profile of the device is confined and combined with photonics by carefully designing the phase-change material as a self-confined nanoscale constriction. This design achieves heat enhancement within the GST without placing constraints on the metallic layer, making the entire process foundry compatible and scalable for future applications. Benefiting from this thermal confinement, electrical switching has been demonstrated with tens to hundreds of µA programming current corresponding to ultralow electrical switching energy (sub-10 pJ), which is two orders lower compared to heater-based demonstrations. The devices show low energy for both electrical

(19.5 pJ) and optical switching (25.2 pJ) with readouts in both electrical and optical domains, a strong modulation depth of 0.15 nJ/dB, and, importantly, multilevel operation. With further efforts to improve the device performance, this work will enable versatile electro-optical memory cells, which offer the potential for fully integrated, energy-efficient electro-optical computing systems combining in-memory programming and sensing abilities.

Device fabrication, characterizations and measurements were performed by the author, as well as the simulations. The work presented in this chapter has been published in the following article:

- Y. He, N. Farmakidis, S. Aggarwal, B. Dong, J. S. Lee, M. Wang, Y. Zhang, F. Parmigiani, and H. Bhaskaran, "Energy-Efficient Integrated Electro-Optic Memristors," Nano Lett. 24, 16325–16332 (2024).

6
System Integration

Contents

In addition to cleanroom device fabrication, well-optimized foundry manufacturing through a multi-project wafer (MPW) offers improved uniformity and reliability, making it the preferred choice for scaling up photonic integrated circuits (PICs). Besides its availability for integration with nonvolatile phase-change material devices, foundry manufacturing provides access to high-speed volatile optoelectronic devices. This chapter discusses the design and basic characterizations of modulators and photodetectors with IMEC iSiPP50G technology, compared with non-volatile GST-based optoelectronic devices, and then provides a brief introduction to the design of the peripheral data transceiver system utilizing RFSoC FPGA.

6.1 Foundry Run Design and Characterizations

Python-based software Luceda IPKISS has been used to design the PIC layout, exploiting the component libraries from IMEC iSiPP50G technology. Upon successful Design Rule Check (DRC) validation, the PIC was fabricated by IMEC foundry and later characterized by the author in the lab.

Figure 6.1 presents the complete schematics of the PIC layout design, with black dashed boxes that highlight the contributions of the author. Specifically, this section discusses the design and basic characterizations of electro-absorption modulators (EAMs) and integrated germanium photodiodes.

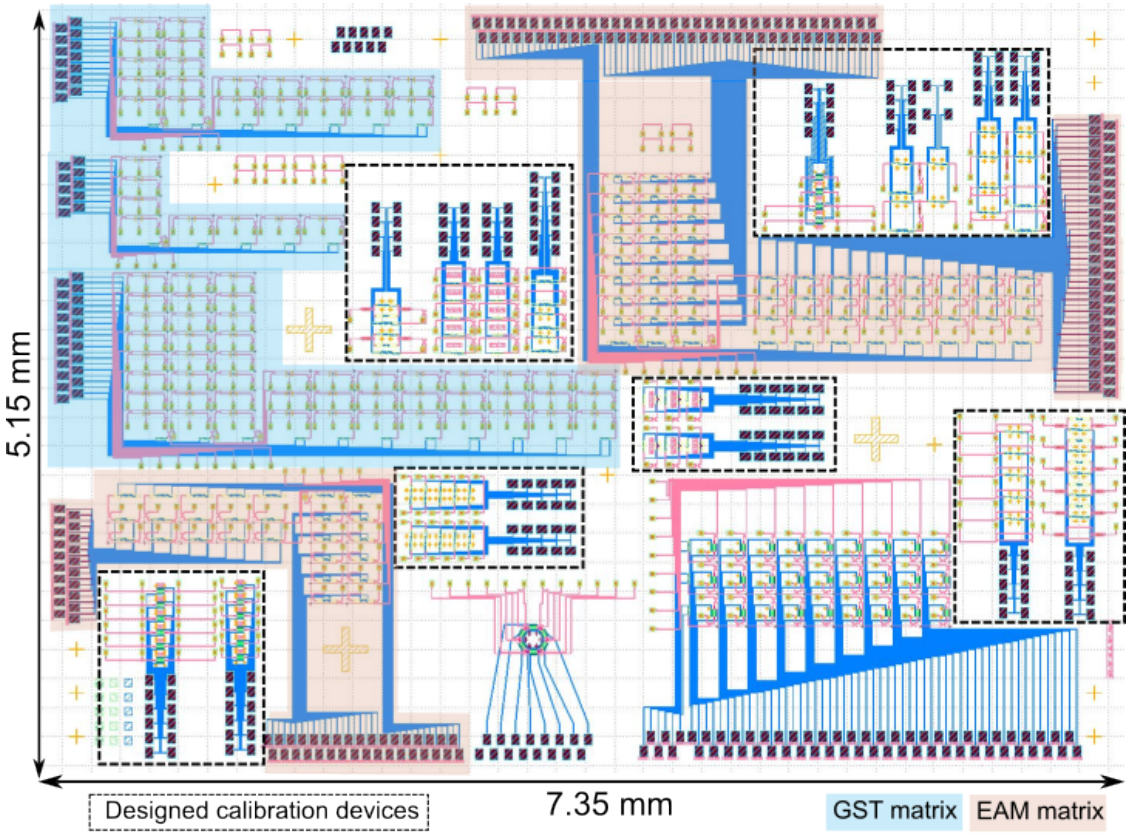

Figure 6.1: PIC layout with IMEC iSiPP50G technology, including both GST- and EAM-based matrices. Layouts designed by the author is highlighted with dashed boxes.

6.1.1 SiGe EAM characterizations

EAM components from IMEC iSiPP50G technology are germanium (Ge) devices grown on 200 mm SOI wafers with a 220 nm-thick top Si layer. With dedicated doping implantation for the germanium layer to create lateral p-i-n diodes, a strong electrical field contrast ($> 10\ kV/cm$) is realized to tilt the energy bands, thus changing the absorption coefficient of Ge (Franz-Keldysh effect [147, 197]).

Figures 6.2(a)-(b) present the layout and optical images of EAM devices,

designed with waveguides and AlCu contact pads. The footprint length of the EAM component is measured as $\sim$ 140 μm, with a $\sim$ 40 μm long germanium modulation region. The transmission of the device has been measured under different bias conditions to calculate the insertion loss. When increasing the reverse bias from 0 V to 2 V, the transmission decreases (insertion loss increases) from -2.7 dB (53.6%) to -5.1 dB (31.2%) at 1550 nm [Figures 6.2(c)].

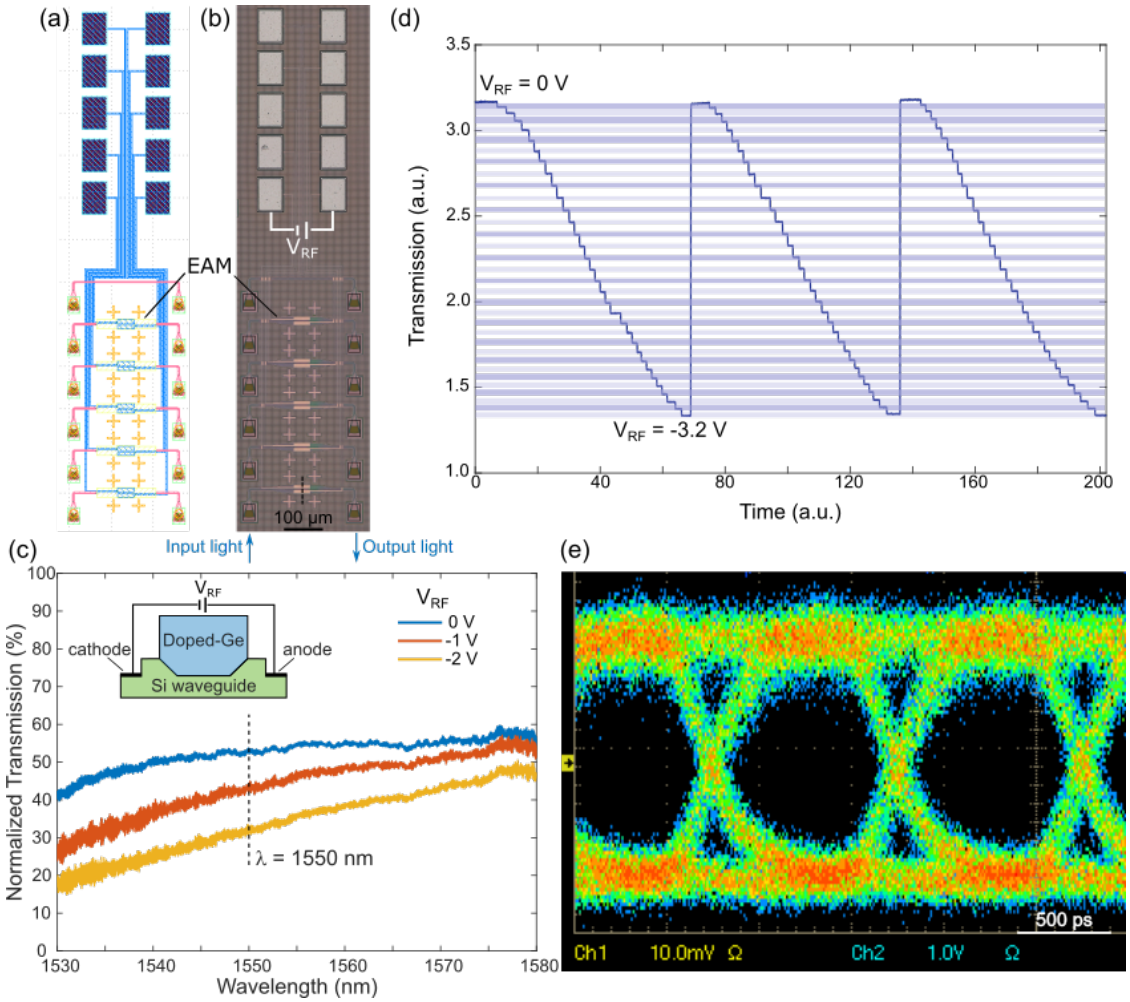

Figure 6.2: Characterizations for SiGe EAMs. (a) Layout of 5 EAMs with waveguides and AlCu contact pads. (b) Optical image of the EAM device. (c) Transmission of the EAM under different bias voltage, normalized to the transmission of a plain waveguide on the same chip. *Inset*: cross-sectional schematic of the EAM. (d) Stable 32 transmission levels under bias varying from 0 V to -3.2 V in 0.1 V decrements. (e) Eye diagram for the EAM device under 3-V, 1 GHz modulation at $\lambda = 1550$ nm.

Similarly, by tuning the reverse bias (V_{RF}), the device is programmable to differ-

ent transmission levels. Figure 6.2(d) shows stable accessibility of 32 transmission levels when modulating the bias from 0 V to -3.2 V, similar to the 5-bit precision of phase-change photonic devices [20]. Moreover, speed measurements have been carried out. Figure 6.2(e) exhibits a clear and wide eye opening of transmission signals when sending 1 GHz pseudo-random binary sequence (PRBS) at a voltage swing of 3 V. The measured bandwidth is limited by the electrical bandwidth of the photoreceiver in the lab, and the maximum bandwidth designed for EAM is 50 GHz [147].

6.1.2 Integrated SiGe photodetector characterizations

Integrated photodetectors of the IMEC iSiPP50G technology are also based on SiGe laterial PIN diodes, but with a longer footprint length for effective light absorption (measured as ~ 180 μm, with a ~ 80 μm-long germanium modulation region). The calibration devices are designed with a directional coupler to split

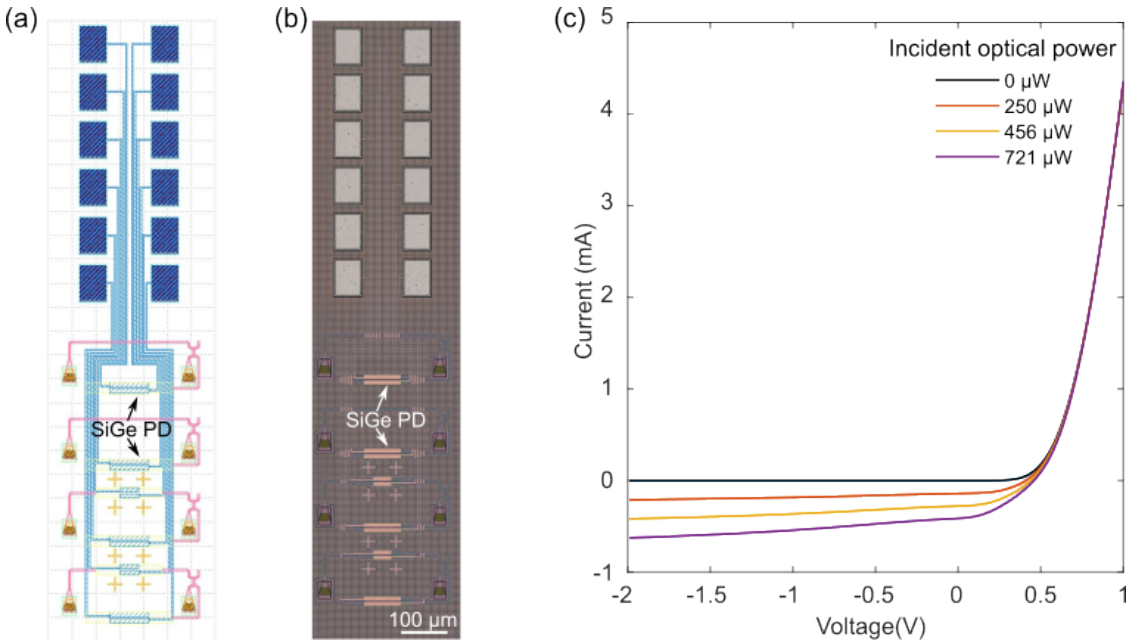

Figure 6.3: Characterization for SiGe photodetectors. (a) Layout of the SiGe photodetectors with pads and waveguides. (b) Optical image of the SiGe photodetector. (c) Current-voltage relationship of the SiGe photodetector under different incident optical power at $\lambda = 1550$ nm.

the output optical signal equally between the integrated photodetectors and the output grating couplers [Figures 6.3 (a)-(b)].

The asymmetric current-voltage relationship of the photodetector and the nonzero photocurrent at zero bias indicate that the device operates in photovoltaic mode [Figures 6.3 (c)]. The photocurrent and responsivity of the device under different voltage biases is calculated as shown in [Figures 6.4 (a)-(b)]. The generated photocurrent increases linearly with the power of the incident light, and no saturation trend is observed till an incident power of 721 μW. The responsivity of the devices increases from ~ 0.55 A/W with zero bias to ~ 0.83 A/W with -2 V bias.

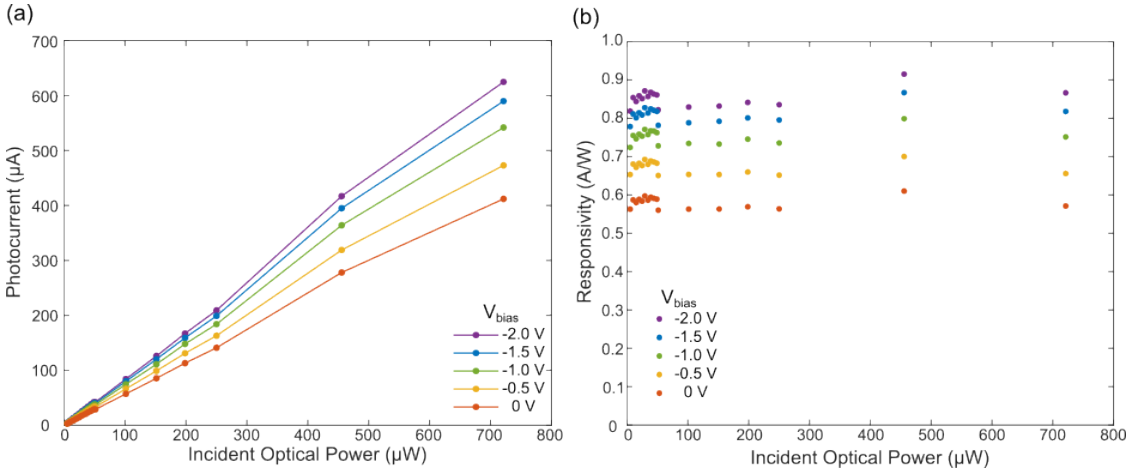

Figure 6.4: Responsivity of SiGe photodetectors at $\lambda = 1550$ nm. (a) Photocurrent of the SiGe photodetector under different bias. (b) Responsivity of the SiGe photodetector under different bias.

6.2 Comparison between GST and SiGe devices

The phase-change devices proposed in previous chapters can be considered non-volatile counterparts to the volatile EAMs, as they both modulate the amplitude of the incident light. Table 6.1 provides a comparison between SiGe-based EAMs and GST-based optical memristors.

Given its high bandwidth, 50 GHz or 0.02 ns refreshing time (which operates over two orders of magnitude faster than GST-based devices), EAM outperforms GST-

based nonvolatile devices in terms of programming energy and modulation efficiency. Therefore, EAMs are better suited to the training of computing applications where frequent weight updating is imperative.

Table 6.1: Performance comparison of SiGe EAM and $Ge_2Sb_2Te_5$-based optical memristors

Devices type	Active device Area $L \times W$ (μm^2)	Insertion loss (dB)	Programming energy (pJ)	Precision (bit)	Energy per modulation depth (pJ/dB)
SOI waveguide [179]	4.0×0.45	**-0.236**	388.4	**5**	780
Doped-Si heater[160]	3.5×0.45	-1.78	$5,200$	4	1700
Plasmonic device	**0.1×0.03**	-3.57	3.82	>4	**26**
Constriction device	0.1×0.30 0.3×0.30	-2 -5	25.2 (Optical) 19.5 (Electrical)	>4 (Optical) >2 (Electrical)	150 190
SiGe EAM	$\sim 40 \times 0.6$	-2.7	**0.037**	**5**	**0.010**

On the other hand, nonvolatile GST-based devices are well-positioned for the inference of computing applications, where the transmission level is required to be held constantly and thus the static voltage configuration of the volatile EAM devices will dominate the energy consumption. From this perspective, the two types of devices proposed in previous chapters demonstrate superior energy efficiency with a more compact footprint.

Table 6.2: Performance comparison of SiGe and $Ge_2Sb_2Te_5$-based photodetectors

Photodetector type	Active length (μm)	Responsivity* at $\pm 1V$ (mA/W)	Responsivity* per unit length (mA/W/μm)
Crystalline GST constriction Amorphous GST constriction	**0.45**	33 (200) 0.007 (0.05)	73 (444) 0.02 (0.11)
SiGe	~ 80	730 (800)	9 (10)

* Stable (peak).

Similarly, Table 6.2 provides a comparison between SiGe-based and GST-constriction-based photodetectors. SiGe devices surpass GST-based devices in terms of bandwidth (50 GHz compared to ~ 2 MHz [154]) and device responsivity.

However, given its elevated responsivity per unit length, GST-based constriction devices demonstrate potential for high-responsivity applications.

6.3 FPGA Transceiver System

For integrated circuits, the requirement of simultaneously generating and receiving signals for multiple inputs and outputs is difficult to be fulfilled by discrete instruments as introduced in Section 3.3. Instead, a multi-port, programmable tool such as an FPGA board is preferred to support the versatile functionalities required for computing cores.

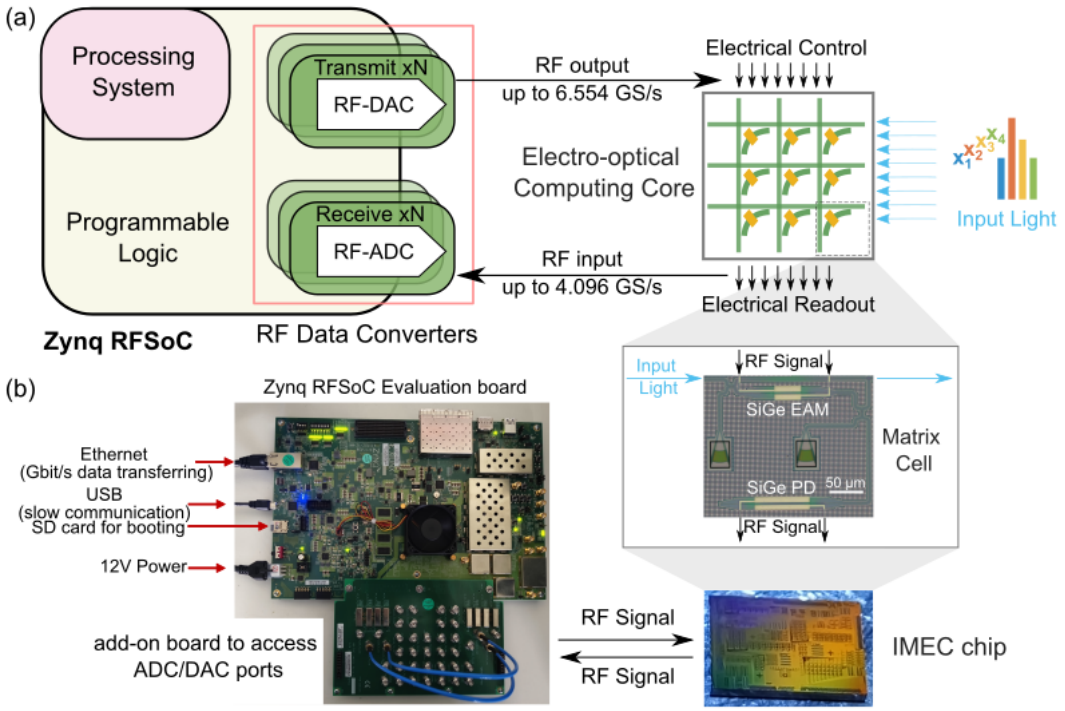

Figure 6.5: Schematics for the RFSoC FPGA. (a) High level illustration of Zynq RFSoC FPGA (Modified from [198]) as a high-speed transceiver system for the electro-optical computing core. *Middle inset*: a zoom-in optical image of a EAM-based matrix cell. Scale bar: 50 μm. *Bottom inset*: an optical image of the IMEC chip. The chip will be wire-bonded onto a customized printed circuit board (PCB), which facilitates RF signal communication with the FPGA. (b) Optical image of the RFSoC FPGA evaluation board, with indications of the required connections.

Specifically, this section focuses on the usage of RFSoC FPGA (ZCU111 Evaulation Board from AMD), for which the analogue and digital conversion

units are integrated with the digital programming and processing units on a single board. The working principle of the RFSoC FPGA is briefly introduced in this section, followed by primary function validations.

Schematics for RFSoC FPGA

Figure 6.5 presents the high-level illustrations of the RFSoC FPGA, including three main modules: RF data converters for high-speed digital and analog conversions, a processing system to support software execution (e.g., in C/Python for user interface) and digital programmable logics (PL, the core of FPGA) for low-latency routing. By carefully programming and coordinating the three modules, the multi-channel RFSoC FPGA is capable of transmitting electrical control signals up to GHz for the inputs and weights of the electro-optical computing core, and meanwhile receiving the electrical readouts [97]. An optical image of an RFSoC FPGA board in operation is presented in Figure 6.5(b).

Function Validation

As a proof of concept, the task of transmitting and receiving sine waves is discussed, with an oscilloscope and a function generator to simulate the data transceiving of an electro-optical computing core.

Figures 6.6(a)-(b) illustrate the working principle of the RF data converters. Customized waveform data (sine wave in this task) are programmed in the PL or PS module, upconverted to the required transmission frequency exploiting interpolation and a frequency mixer, and then converted to analog for transmitting. Similarly, readout signals can be converted into the digital domain and then downconverted to the operation frequency of the PL with appropriate mixer frequency and decimation factors. Figure 6.6(c) presents the oscilloscope-recorded transmit signal, with a 3 dB bandwidth of 1 GHz. Sine waves, with a frequency of 150 MHz limited by the function generator, are also properly received by the RF-ADC as shown in Figure 6.6(d).

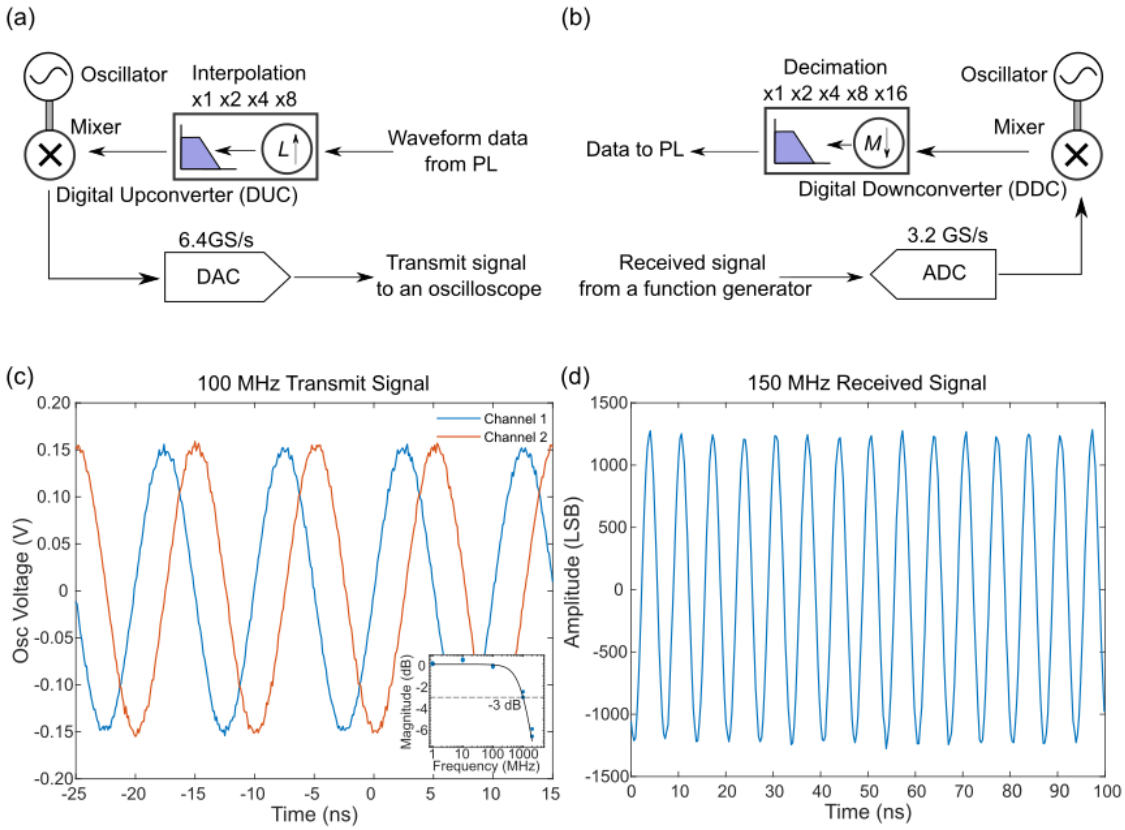

Figure 6.6: Sine wave transmission and reception via RFSoC FPGA. (a)-(b) The block diagrams of RF-DAC and RF-ADC. PL: programmable logic. Based on [198]. (c) RFSoC FPGA transmitted 100 MHz sine wave monitored by oscilloscope. *Inset*: frequency response of the RFSoC FPGA measured with sine waves of different frequencies. (d) RFSoC FPGA received 150 MHz sine wave from the function generator. LSB: least significant bit.

With further efforts to program the RFSoC FPGA for implementing versatile functions [199], the performance of nonvolatile GST-based and volatile EAM-based computing systems can be experimentally compared from a system perspective.

6.4 Chapter Summary

In this chapter, by exploiting a foundry-fabricated integrated photonic chip, the performance of integrated volatile modulators and photodetectors is experimentally compared with those of the aforementioned GST-based nonvolatile devices. Further, the development of a data transceiver system based on RFSoC FPGA is introduced

with validation of primary functions.

SiGe devices were fabricated by the IMEC foundry. The layout design, measurements, and analysis have been performed by the author, with the guidance of Bowei Dong on the eye diagram measurements.

7 Conclusion and Outlook

7.1 Conclusions

In this thesis, two types of electro-optic devices have been developed to address the challenges of building energy-efficient electro-optical computing systems on SOI platforms.

First, plasmonic phase-change devices have been achieved on a SOI platform with a compact footprint. Exploiting an optimized corrugated grating design and a self-aligned fabrication process, the devices have demonstrated high mode-coupling efficiency with an ultra-low optical switching energy of less than 4 pJ, which is lower than both previous silicon-based phase-change devices and comparable devices on Si_3N_4 platforms.

Next, a novel non-plasmonic electro-optic memristor has been experimentally demonstrated by structuring the phase-change material as a constriction structure. As far as is known, it is the first experimental demonstration of integrated electro-optic memristors with dual electro-optical functionality that provides the potential for CMOS-compatible scaling. This design places no constraints on the metallic layers and achieves a high electro-optical modulation efficiency of 0.15 nJ/dB. Moreover, the in-plane photosensitivity of the device complements its non-volatile reconfigurability, unlocking new possibilities in energy-efficient electro-optical systems that combine both in-memory programming and sensing abilities.

Finally, system integration of electro-optical computing units has been explored, exploiting a foundry-fabricated integrated photonic chip. The performance of

integrated volatile SiGe-based optoelectronic devices has been compared to GST-based nonvolatile optoelectronic devices, highlighting their respective advantages in different applications. The development of a data transceiver system based on RFSoC FPGA is further discussed, paving the way for the integration of compact electro-optical computing systems.

7.2 Outlook

The field of photonic computing has witnessed rapid advancement during the course of this thesis, with novel insights from both academia and industry. Following current achievements discussed in this thesis, potential future development directions are highlighted in this section from the perspective of materials and applications.

Emerging materials

As reviewed in Chapter 2, four types of materials, i.e. metal-oxide, magnetic, phase-change or ferroelectric materials, have been commonly used as electrical non-volatile devices, however, only phase-change materials have been widely used in photonic platforms.

Recently, emerging materials, such as $BaTiO_3$ thin films[200], Al-doped HfO_2 (HAO) [201, 202] and cerium-substituted yttrium iron garnet (Ce:YIG) [203], have been explored to implement photonic components. These ferroelectric or magneto-optic materials unlock additional benefits for photonic computing systems, including excellent endurance (2.4 billion optical programming cycles [203]) and ultra-low programming energy (143 fJ/bit or 275 fJ/dB [203]). Moreover, conductive-filament-based resistive memristors have been recently combined with ring modulators to implement non-volatile photonic memory, offering low switching energy (0.15 pJ) and fast switching speed (300 ps) [204]. Meanwhile, novel phase-change materials such as antimony (Sb) [205–207] and antimony selenide (Sb_2Se_3) [105, 208] stand out as promising candidates for integrated non-volatile optoelectronic devices, supporting

fast programmability (subpicoseconds) and low absorption loss at telecommunication wavelengths, respectively.

The metrics and analytical methods discussed in this thesis for devices based on $Ge_2Sb_2Te_5$ also apply to devices based on the above emerging materials. With careful material engineering, fast, low-loss, high-endurance, and energy-efficient integrated non-volatile optoelectronic devices can serve as critical building blocks for future computing systems.

Novel applications

The computing applications mentioned in this thesis are primarily related to neural networks, incorporating two equally critical units of distinct functions: linear matrix-vector multiplications (MVMs) and nonlinear activation functions. While passive, high-bandwidth in-memory photonic computing cores are well-positioned solutions for implementing linear MVMs in hardware, implementation of nonlinear activation functions are still dominated by power-intensive electronics, primarily in silico.

Efforts have been made to explore integrated programmable electro-optical [75, 209, 210] or all-optical [211, 212] nonlinear units; however, challenges remain in managing the trade-off between speed and energy overhead from optoelectronic nonlinearities, as well as the limited cascadability of weak all-optical nonlinearities. As a result, practical advantages of photonic computing cores are still limited within combinatorial optimizations [213, 214], where the requirements for nonlinear activation functions are relatively less critical compared to applications related to deep neural networks.

Meanwhile, endeavors have been devoted to interconnect applications [215]. High-bandwidth and low-latency optical interconnects have been widely employed for data centers, while the current wave of interest has extended the applications of optical interconnects to chip-to-chip or even on-chip communication[216, 217].

With energy-efficient non-volatile properties, low-loss phase-change photonic devices [105, 208] also emerge as appealing candidates.

List of Publications

Journal Articles

1. **Y. He**†, N. Farmakidis†, S. Aggarwal, B. Dong, J. S. Lee, M. Wang, Y. Zhang, F. Parmigiani, and H. Bhaskaran. "Energy-efficient integrated electro-optic memristors," Nano Lett. 24, 16325–16332 (2024).

2. B. Dong†, F. Brückerhoff-Plückelmann†, L. Meyer, J. Dijkstra, I. Bente, D. Wendland, A. Varri, S. Aggarwal, N. Farmakidis, M. Wang, G. Yang, J. S. Lee, **Y. He**, et al. "Partial coherence enhances parallelized photonic computing," Nature 632, 55–62 (2024).

3. B. Dong, S. Aggarwal, W. Zhou, U. E. Ali, N. Farmakidis, J. S. Lee, **Y. He**, et al. "Higher-dimensional processing using a photonic tensor core with continuous-time data," Nat. Photonics 17, 1080–1088 (2023).

4. W. Zhou†, B. Dong†, N. Farmakidis, X. Li, N. Youngblood, K. Huang, **Y. He**, et al. "In-memory photonic dot-product engine with electrically programmable weight banks," Nat. Commun. 14, 2887 (2023).

5. N. Farmakidis, J. S. Lee, J. Feldmann, M. Wang, **Y. He**, S. Aggarwal, B. Dong, W. H. P. Pernice, and H. Bhaskaran. "Scalable high-precision trimming of photonic resonances by polymer exposure to energetic beams." Nano Lett. 23, 4800–4806 (2023).

6. M. Wang, J.S. Lee, S. Aggarwal, N. Farmakidis, **Y. He**, et al. "Varifocal metalens using tunable and ultralow-loss dielectrics," Adv. Sci. 10, 2204899 (2023).

7. W. Zhou, N. Farmakidis, J. Feldmann, X. Li, J. Tan, **Y. He**, et al. "Phase-change materials for energy-efficient photonic memory and computing," MRS Bull. 47, 502–510 (2022).

Conference Presentations and Proceedings

1. **Y. He**, et. al. "Electro-optical phase-change devices approaching single pico-joule switching", SPIE Photonic West, San Francisco, USA, 27 January – 1 February 2024. (Oral presentation)

2. **Y. He**, et. al. "Mixed-Mode Phase-Change Devices with Picojoule Switching Energy", E/PCOS 2023, Rome, Italy, 17 - 20 September 2023. **(Invited oral presentation)**

3. **Y. He**, et. al. "Ultra-Efficient Plasmonic Phase-Change Devices on SOI Platform", Conference on Lasers and Electro-Optics (CLEO), San Jose, USA, 7-12 May 2023. (Oral presentation)

4. Y. Zhang, N. Farmakidis, J. S. Lee, B. Dong, S. Aggarwal, **Y. He**, and H. Bhaskaran. "Bosonic control in integrated photonics", Proc. SPIE PC12889, Integrated Optics: Devices, Materials, and Technologies XXVIII, PC128890R (13 March 2024)

5. S. Aggarwal, **Y. He**, I.Z. Esmaeil, and H. Bhaskaran. "Tunable Silicon Carbide Photonics using Phase Change Materials." Proc. SPIE PC12010, Photonic and Phononic Properties of Engineered Nanostructures XII, PC120100C (5 March 2022)

Supervised Part II Thesis

1. Lee, Y (2022). Novel Electro-Optic Approaches to Analog/Digital Conversions in AI Applications [Unpublished MEng thesis]. University of Oxford.

References

1. Weik, M. H. The ENIAC Story. *Ordnance* **45,** 571–575 (1961).
2. Von Neumann, J. First draft of a report on the EDVAC. *IEEE Annals of the History of Computing* **15,** 27–75 (1993).
3. Datta, S., Chakraborty, W. & Radosavljevic, M. Toward attojoule switching energy in logic transistors. *Science* **378,** 733–740 (Nov. 2022).
4. Mehonic, A. & Kenyon, A. J. Brain-inspired computing needs a master plan. *Nature* **604,** 255–260 (Apr. 2022).
5. Zhang, W. *et al.* Neuro-inspired computing chips. *Nature Electronics* **3,** 371–382 (July 2020).
6. Huang, Y. *et al.* Memristor-based hardware accelerators for artificial intelligence. *Nature Reviews Electrical Engineering* **1,** 286–299 (Apr. 2024).
7. Song, M.-K. *et al.* Recent Advances and Future Prospects for Memristive Materials, Devices, and Systems. *ACS Nano* **17,** 11994–12039 (July 2023).
8. Lanza, M. *et al.* Memristive technologies for data storage, computation, encryption, and radio-frequency communication. *Science* **376,** eabj9979 (June 2022).
9. Liu, Q. *et al. 33.2 A Fully Integrated Analog ReRAM Based 78.4TOPS/W Compute-In-Memory Chip with Fully Parallel MAC Computing* in *2020 IEEE International Solid- State Circuits Conference - (ISSCC)* (IEEE, Feb. 2020), 500–502.
10. Le Gallo, M. *et al.* A 64-core mixed-signal in-memory compute chip based on phase-change memory for deep neural network inference. *Nature Electronics* **6,** 680–693 (Aug. 2023).
11. McMahon, P. L. The physics of optical computing. *Nature Reviews Physics* **5,** 717–734 (Oct. 2023).
12. Shastri, B. J. *et al.* Photonics for artificial intelligence and neuromorphic computing. *Nature Photonics* **15,** 102–114 (Feb. 2021).
13. Farmakidis, N., Dong, B. & Bhaskaran, H. Integrated photonic neuromorphic computing: opportunities and challenges. *Nature Reviews Electrical Engineering* **1,** 358–373 (June 2024).
14. Feldmann, J. *et al.* Parallel convolutional processing using an integrated photonic tensor core. *Nature* **589,** 52–58 (Jan. 2021).
15. Xu, X. *et al.* 11 TOPS photonic convolutional accelerator for optical neural networks. *Nature* **589,** 44–51 (Jan. 2021).
16. Lee, J. S., Farmakidis, N., Wright, C. D. & Bhaskaran, H. Polarization-selective reconfigurability in hybridized-active-dielectric nanowires. *Science Advances* **8,** eabn9459 (June 2022).
17. Youngblood, N., Ríos Ocampo, C. A., Pernice, W. H. P. & Bhaskaran, H. Integrated optical memristors. *Nature Photonics* **17,** 561–572 (July 2023).
18. Rios, C. *et al.* Integrated all-photonic non-volatile multi-level memory. *Nature Photonics* **9,** 725–732 (Oct. 2015).

19. Ríos, C. *et al.* In-memory computing on a photonic platform. *Science Advances* **5,** eaau5759 (Feb. 2019).
20. Li, X. *et al.* Fast and reliable storage using a 5 bit, nonvolatile photonic memory cell. *Optica* **6,** 1 (Jan. 2019).
21. Feldmann, J., Youngblood, N., Wright, C. D., Bhaskaran, H. & Pernice, W. H. P. All-optical spiking neurosynaptic networks with self-learning capabilities. *Nature* **569,** 208–214 (May 2019).
22. Ghazi Sarwat, S. *et al.* An integrated photonics engine for unsupervised correlation detection. *Science Advances* **8,** eabn3243 (June 2022).
23. Wu, C. *et al.* Programmable phase-change metasurfaces on waveguides for multimode photonic convolutional neural network. *Nature Communications* **12,** 96 (Jan. 2021).
24. Wuttig, M. & Yamada, N. Phase-change materials for rewriteable data storage. *Nature Materials* **6,** 824–832 (Nov. 2007).
25. Xiong, F., Liao, A. D., Estrada, D. & Pop, E. Low-Power Switching of Phase-Change Materials with Carbon Nanotube Electrodes. *Science* **332,** 568–570 (Apr. 2011).
26. Khan, A. I. *et al.* Ultralow–switching current density multilevel phase-change memory on a flexible substrate. *Science* **373,** 1243–1247 (Sept. 2021).
27. He, Y. *et al.* Energy-Efficient Integrated Electro-Optic Memristors. *Nano Letters* **24,** 16325–16332 (Dec. 2024).
28. Park, S.-O. *et al.* Phase-change memory via a phase-changeable self-confined nano-filament. *Nature* **628,** 293–298 (Apr. 2024).
29. Lanza, M., Molas, G. & Naveh, I. The gap between academia and industry in resistive switching research. *Nature Electronics* **6,** 260–263 (Apr. 2023).
30. Strukov, D. B., Snider, G. S., Stewart, D. R. & Williams, R. S. The missing memristor found. *Nature* **453,** 80–83 (May 2008).
31. Yang, F. *et al.* Carbon-based memristors for resistive random access memory and neuromorphic applications. *Chip* **3,** 100086 (June 2024).
32. Si, J. *et al.* A carbon-nanotube-based tensor processing unit. *Nature Electronics* **7,** 684–693 (July 2024).
33. Nirmal, K. A., Kumbhar, D. D., Kesavan, A. V., Dongale, T. D. & Kim, T. G. Advancements in 2D layered material memristors: unleashing their potential beyond memory. *npj 2D Materials and Applications* **8,** 83 (Dec. 2024).
34. Li, J. *et al.* Polymeric Memristor Based Artificial Synapses with Ultra-Wide Operating Temperature. *Advanced Materials* **35,** 2209728 (June 2023).
35. Zhang, B. *et al.* 90% yield production of polymer nano-memristor for in-memory computing. *Nature Communications* **12,** 1984 (Mar. 2021).
36. Chua, L. Memristor-The missing circuit element. *IEEE Transactions on Circuit Theory* **18,** 507–519 (1971).
37. Marković, D., Mizrahi, A., Querlioz, D. & Grollier, J. Physics for neuromorphic computing. *Nature Reviews Physics* **2,** 499–510 (July 2020).

38. Shao, Q., Wang, Z. & Yang, J. J. Efficient AI with MRAM. *Nature Electronics* **5,** 67–68 (Feb. 2022).
39. Rao, M. *et al.* Thousands of conductance levels in memristors integrated on CMOS. *Nature* **615,** 823–829 (Mar. 2023).
40. DerChang Kau *et al. A stackable cross point Phase Change Memory* in *2009 IEEE International Electron Devices Meeting (IEDM)* (IEEE, Dec. 2009), 1–4.
41. *Intel and Micron Produce Breakthrough Memory Technology* 2015.
42. Fong, S. W., Neumann, C. M. & Wong, H.-S. P. Phase-Change Memory—Towards a Storage-Class Memory. *IEEE Transactions on Electron Devices* **64,** 4374–4385 (Nov. 2017).
43. Le Gallo, M. & Sebastian, A. An overview of phase-change memory device physics. *Journal of Physics D: Applied Physics* **53,** 213002 (May 2020).
44. Raoux, S. Phase Change Materials. *Annual Review of Materials Research* **39,** 25–48 (Aug. 2009).
45. Kim, S., Burr, G. W., Kim, W. & Nam, S.-W. Phase-change memory cycling endurance. *MRS Bulletin* **44,** 710–714 (Sept. 2019).
46. Molinari, A. *et al.* Configurable Resistive Response in BaTiO3 Ferroelectric Memristors via Electron Beam Radiation. *Advanced Materials* **32,** 1907541 (Mar. 2020).
47. Yan, X. *et al.* A Robust Memristor Based on Epitaxial Vertically Aligned Nanostructured BaTiO3-CeO2 Films on Silicon. *Advanced Materials* **34,** 2110343 (June 2022).
48. Garcia, V. & Bibes, M. Ferroelectric tunnel junctions for information storage and processing. *Nature Communications* **5,** 4289 (July 2014).
49. Song, W. *et al.* Programming memristor arrays with arbitrarily high precision for analog computing. *Science* **383,** 903–910 (Feb. 2024).
50. Ambrogio, S. *et al.* An analog-AI chip for energy-efficient speech recognition and transcription. *Nature* **620,** 768–775 (Aug. 2023).
51. Vaswani, A. *et al. Attention is All you Need* in *Advances in Neural Information Processing Systems* (eds Guyon, I. *et al.*) (Curran Associates, Inc., 2017), 6000–6010.
52. Graves, A. Sequence Transduction with Recurrent Neural Networks (Nov. 2012).
53. He, K., Zhang, X., Ren, S. & Sun, J. *Deep Residual Learning for Image Recognition* in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, June 2016), 770–778.
54. Yao, P. *et al.* Fully hardware-implemented memristor convolutional neural network. *Nature* **577,** 641–646 (2020).
55. LeCun, Y., Cortes, C. & Burges, C. *Mnist Handwritten Digit Database* 2010.
56. Huang, W.-H. *et al. A Nonvolatile AI-Edge Processor with 4MB SLC-MLC Hybrid-Mode ReRAM Compute-in-Memory Macro and 51.4-251TOPS/W* in *2023 IEEE International Solid- State Circuits Conference (ISSCC)* (IEEE, Feb. 2023), 15–17.

57. Hung, J.-M. *et al.* 8-b Precision 8-Mb ReRAM Compute-in-Memory Macro Using Direct-Current-Free Time-Domain Readout Scheme for AI Edge Devices. *IEEE Journal of Solid-State Circuits* **58,** 303–315 (Jan. 2023).

58. Krizhevsky, A. *Learning Multiple Layers of Features from Tiny Images* PhD thesis (Univ. Toronto, 2009).

59. Deaville, P., Zhang, B., Chen, L.-Y. & Verma, N. *A Maximally Row-Parallel MRAM In-Memory-Computing Macro Addressing Readout Circuit Sensitivity and Area* in *ESSCIRC 2021 - IEEE 47th European Solid State Circuits Conference (ESSCIRC)* (IEEE, Sept. 2021), 75–78.

60. Jung, S. *et al.* A crossbar array of magnetoresistive memory devices for in-memory computing. *Nature* **601,** 211–216 (Jan. 2022).

61. Xu, Y. *et al.* Ferroelectric FET-based context-switching FPGA enabling dynamic reconfiguration for adaptive deep learning machines. *Science Advances* **10,** eadk1525 (Jan. 2024).

62. Lu, T. *et al.* Two-dimensional fully ferroelectric-gated hybrid computing-in-memory hardware for high-precision and energy-efficient dynamic tracking. *Science Advances* **10,** eadp0174 (Sept. 2024).

63. Jia, H. *et al. 15.1 A Programmable Neural-Network Inference Accelerator Based on Scalable In-Memory Computing* in *2021 IEEE International Solid- State Circuits Conference (ISSCC)* (IEEE, Feb. 2021), 236–238.

64. Mori, H. *et al. A 4nm 6163-TOPS/W/b 4790-TOPS/mm2/b SRAM Based Digital-Computing-in-Memory Macro Supporting Bit-Width Flexibility and Simultaneous MAC and Weight Update* in *2023 IEEE International Solid- State Circuits Conference (ISSCC)* (IEEE, Feb. 2023), 132–134.

65. Wu, P.-C. *et al. A 28nm 1Mb Time-Domain Computing-in-Memory 6T-SRAM Macro with a 6.6ns Latency, 1241GOPS and 37.01TOPS/W for 8b-MAC Operations for Edge-AI Devices* in *2022 IEEE International Solid- State Circuits Conference (ISSCC)* (IEEE, Feb. 2022), 1–3.

66. Miller, D. A. B. Are optical transistors the logical next step? *Nature Photonics* **4,** 3–5 (Jan. 2010).

67. Wetzstein, G. *et al.* Inference in artificial intelligence with deep optics and photonics. *Nature* **588,** 39–47 (Dec. 2020).

68. Wherrett, B. S. & Goodman, J. W. in *Optical Computing* (eds Tooley, F. A. P. & Wherrett, B. S.) 1–21 (Taylor & Francis, United Kingdom, 1989).

69. Ambs, P. Optical Computing: A 60-Year Adventure. *Advances in Optical Technologies* **2010,** 1–15 (May 2010).

70. Minzioni, P. *et al.* Roadmap on all-optical processing. *Journal of Optics* **21,** 063001 (June 2019).

71. De Marinis, L., Cococcioni, M., Castoldi, P. & Andriolli, N. Photonic Neural Networks: A Survey. *IEEE Access* **7,** 175827–175841 (2019).

72. Fu, T. *et al.* Optical neural networks: progress and challenges. *Light: Science & Applications* **13,** 263 (Sept. 2024).

73. Shen, Y. *et al.* Deep learning with coherent nanophotonic circuits. *Nature Photonics* **11,** 441–446 (July 2017).

74. Pai, S. *et al.* Experimentally realized in situ backpropagation for deep learning in photonic neural networks. *Science* **380,** 398–404 (Apr. 2023).

75. Bandyopadhyay, S. *et al.* Single-chip photonic deep neural network with forward-only training. *Nature Photonics* **18,** 1335–1343 (Dec. 2024).

76. Lawson, C. L. & Hanson, R. J. *Solving Least Squares Problems* 18–22 (SIAM, 1995).

77. Carolan, J. *et al.* Universal linear optics. *Science* **349,** 711–716 (Aug. 2015).

78. Ramey, C. *Silicon Photonics for Artificial Intelligence Acceleration (Lightmatter)* in *IEEE Hot Chips 32 Symposium (HCS)* (IEEE, Aug. 2020), 1–26.

79. Hughes, T. W., Minkov, M., Shi, Y. & Fan, S. Training of photonic neural networks through in situ backpropagation and gradient measurement. *Optica* **5,** 864 (July 2018).

80. Bogaerts, W. *et al.* Programmable photonic circuits. *Nature* **586,** 207–216 (2020).

81. Zhuang, L., Roeloffzen, C. G. H., Hoekman, M., Boller, K.-J. & Lowery, A. J. Programmable photonic signal processor chip for radiofrequency applications. *Optica* **2,** 854 (Oct. 2015).

82. Pérez-López, D. & Torrijos-Morán, L. Large-scale photonic processors and their applications. *npj Nanophotonics* **2,** 32 (Aug. 2025).

83. Miller, D. A. B. Self-configuring universal linear optical component [Invited]. *Photonics Research* **1,** 1 (June 2013).

84. Xu, X. *et al.* Self-calibrating programmable photonic integrated circuits. *Nature Photonics* **16,** 595–602 (Aug. 2022).

85. Roques-Carmes, C., Fan, S. & Miller, D. A. B. Measuring, processing, and generating partially coherent light with self-configuring optics. *Light: Science & Applications* **13,** 260 (Sept. 2024).

86. Tait, A. N., Nahmias, M. A., Shastri, B. J. & Prucnal, P. R. Broadcast and weight: An integrated network for scalable photonic spike processing. *Journal of Lightwave Technology* **32,** 3427–3439 (2014).

87. Tait, A. N. *et al.* Microring Weight Banks. *IEEE Journal of Selected Topics in Quantum Electronics* **22,** 312–325 (Nov. 2016).

88. Prucnal, P. R., Shastri, B. J. & Teich, M. C. *Neuromorphic Photonics* (eds Prucnal, P. R. & Shastri, B. J.) (CRC Press, May 2017).

89. Ferreira De Lima, T., Shastri, B. J., Tait, A. N., Nahmias, M. A. & Prucnal, P. R. Progress in neuromorphic photonics. *Nanophotonics* **6,** 577–599 (2017).

90. Peng, H.-T., Nahmias, M. A., de Lima, T. F., Tait, A. N. & Shastri, B. J. Neuromorphic Photonic Integrated Circuits. *IEEE Journal of Selected Topics in Quantum Electronics* **24,** 1–15 (Nov. 2018).

91. Tait, A. N. *et al.* Neuromorphic photonic networks using silicon photonic weight banks. *Scientific Reports* **7,** 1–10 (2017).

92. Tait, A. N. *et al.* Silicon Photonic Modulator Neuron. *Physical Review Applied* **11,** 064043 (June 2019).

93. Bai, B. *et al.* Microcomb-based integrated photonic processing unit. *Nature Communications* **14,** 66 (Jan. 2023).

94. Tait, A. N., de Lima, T. F., Nahmias, M. A., Shastri, B. J. & Prucnal, P. R. Multi-channel control for microring weight banks. *Optics Express* **24,** 8895 (Apr. 2016).

95. Li, X. *et al. On-chip Phase Change Optical Matrix Multiplication Core* in *2020 IEEE International Electron Devices Meeting (IEDM)* (IEEE, Dec. 2020), 1–7.

96. Dong, B. *et al.* Higher-dimensional processing using a photonic tensor core with continuous-time data. *Nature Photonics* **17,** 1080–1088 (Dec. 2023).

97. Dong, B. *et al.* Partial coherence enhances parallelized photonic computing. *Nature* **632,** 55–62 (Aug. 2024).

98. Von Keitz, J. *et al.* Reconfigurable Nanophotonic Cavities with Nonvolatile Response. *ACS Photonics* **5,** 4644–4649 (Nov. 2018).

99. Stern, B., Ji, X., Okawachi, Y., Gaeta, A. L. & Lipson, M. Battery-operated integrated frequency comb generator. *Nature* **562,** 401–405 (Oct. 2018).

100. Raja, A. S. *et al.* Electrically pumped photonic integrated soliton microcomb. *Nature Communications* **10,** 680 (Feb. 2019).

101. Gehring, H., Eich, A., Schuck, C. & Pernice, W. H. P. Broadband out-of-plane coupling at visible wavelengths. *Optics Letters* **44,** 5089 (Oct. 2019).

102. Youngblood, N. Coherent Photonic Crossbar Arrays for Large-Scale Matrix-Matrix Multiplication. *IEEE Journal of Selected Topics in Quantum Electronics* **29,** 1–11 (Mar. 2023).

103. Chen, R. *et al.* Broadband Nonvolatile Electrically Controlled Programmable Units in Silicon Photonics. *ACS Photonics* **9,** 2142–2150 (June 2022).

104. Zheng, J. *et al.* Nonvolatile Electrically Reconfigurable Integrated Photonic Switch Enabled by a Silicon PIN Diode Heater. *Advanced Materials* **32,** 2001218 (Aug. 2020).

105. Ríos, C. *et al.* Ultra-compact nonvolatile phase shifter based on electrically reprogrammable transparent phase change materials. *PhotoniX* **3,** 26 (Oct. 2022).

106. Lin, X. *et al.* All-optical machine learning using diffractive deep neural networks. *Science* **361,** 1004–1008 (Sept. 2018).

107. Yan, T. *et al.* Fourier-space Diffractive Deep Neural Network. *Physical Review Letters* **123,** 23901 (2019).

108. Zhou, T. *et al.* Large-scale neuromorphic optoelectronic computing with a reconfigurable diffractive processing unit. *Nature Photonics* **15,** 367–373 (May 2021).

109. Psaltis, D., Brady, D., Gu, X. G. & Lin, S. Holography in artificial neural networks. *Nature* **343,** 325–330 (1990).

110. Rumelhart, D. E., Hinton, G. E. & Williams, R. J. Learning representations by back-propagating errors. *Nature* **323,** 533–536 (Oct. 1986).

111. Xu, Z. *et al.* Large-scale photonic chiplet Taichi empowers 160-TOPS/W artificial general intelligence. *Science* **384,** 202–209 (Apr. 2024).

112. Chen, Y. *et al.* All-analog photoelectronic chip for high-speed vision tasks. *Nature* **623,** 48–57 (Nov. 2023).

113. Yan, T. *et al.* All-optical graph representation learning using integrated diffractive photonic computing units. *Science Advances* **8,** eabn7630 (June 2022).

114. Wang, Z. *et al.* On-chip wavefront shaping with dielectric metasurface. *Nature Communications* **10,** 3547 (Aug. 2019).

115. Zarei, S., Marzban, M.-r. & Khavasi, A. Integrated photonic neural network based on silicon metalines. *Optics Express* **28,** 36668 (Nov. 2020).

116. Chen, Z. *et al.* Deep learning with coherent VCSEL neural networks. *Nature Photonics* **17** (eds Kitayama, K.-i. & Jalali, B.) 723–730 (Aug. 2023).

117. Rahimi Kari, S., Nobile, N. A., Pantin, D., Shah, V. & Youngblood, N. Realization of an integrated coherent photonic platform for scalable matrix operations. *Optica* **11,** 542 (Apr. 2024).

118. Hamerly, R., Bernstein, L., Sludds, A., Soljačić, M. & Englund, D. Large-Scale Optical Neural Networks Based on Photoelectric Multiplication. *Physical Review X* **9,** 021032 (May 2019).

119. Sludds, A. *et al.* Delocalized photonic deep learning on the internet's edge. *Science* **378,** 270–276 (Oct. 2022).

120. Zhao, H., Li, B., Li, H. & Li, M. Enabling scalable optical computing in synthetic frequency dimension using integrated cavity acousto-optics. *Nature Communications* **13,** 5426 (Sept. 2022).

121. Fan, L., Wang, K., Wang, H., Dutt, A. & Fan, S. Experimental realization of convolution processing in photonic synthetic frequency dimensions. *Science Advances* **9,** eadi4956 (Aug. 2023).

122. Meng, X. *et al.* Compact optical convolution processing unit based on multimode interference. *Nature Communications* **14,** 3000 (May 2023).

123. Zhang, Y., Zhang, R., Zhu, Q., Yuan, Y. & Su, Y. Architecture and Devices for Silicon Photonic Switching in Wavelength, Polarization and Mode. *Journal of Lightwave Technology* **38,** 215–225 (Jan. 2020).

124. He, Y., Zhang, Y., Wang, H., Sun, L. & Su, Y. Design and experimental demonstration of a silicon multi-dimensional (de)multiplexer for wavelength-, mode- and polarization-division (de)multiplexing. *Optics Letters* **45,** 2846 (May 2020).

125. Marandi, A., Wang, Z., Takata, K., Byer, R. L. & Yamamoto, Y. Network of time-multiplexed optical parametric oscillators as a coherent Ising machine. *Nature Photonics* **8,** 937–942 (Dec. 2014).

126. Ramanujam, J. & Sadayappan, P. Mapping combinatorial optimization problems onto neural networks. *Information Sciences* **82,** 239–255 (1995).

127. McMahon, P. L. *et al.* A fully programmable 100-spin coherent Ising machine with all-to-all connections. *Science* **354,** 614–617 (Nov. 2016).

128. Inagaki, T. *et al.* A coherent Ising machine for 2000-node optimization problems. *Science* **354,** 603–606 (2016).

129. Roques-Carmes, C. *et al.* Heuristic recurrent algorithms for photonic Ising machines. *Nature Communications* **11,** 249 (Dec. 2020).

130. Tezak, N. *et al.* Integrated Coherent Ising Machines Based on Self-Phase Modulation in Microring Resonators. *IEEE Journal of Selected Topics in Quantum Electronics* **26,** 1–15 (Jan. 2020).

131. Böhm, F., Alonso-Urquijo, D., Verschaffelt, G. & Van der Sande, G. Noise-injected analog Ising machines enable ultrafast statistical sampling and machine learning. *Nature Communications* **13,** 5847 (Oct. 2022).

132. Pierangeli, D., Marcucci, G. & Conti, C. Photonic extreme learning machine by free-space optical propagation. *Photonics Research* **9,** 1446 (Aug. 2021).

133. Pierangeli, D., Marcucci, G. & Conti, C. Large-Scale Photonic Ising Machine by Spatial Light Modulation. *Physical Review Letters* **122,** 213902 (May 2019).

134. Zangeneh-Nejad, F., Sounas, D. L., Alù, A. & Fleury, R. Analogue computing with metamaterials. *Nature Reviews Materials* **6,** 207–225 (Mar. 2021).

135. Hughes, T. W., Williamson, I. A. D., Minkov, M. & Fan, S. Wave physics as an analog recurrent neural network. *Science Advances* **5,** eaay6946 (Dec. 2019).

136. Estakhri, N. M., Edwards, B. & Engheta, N. Inverse-designed metastructures that solve equations. *Science* **363,** 1333–1338 (2019).

137. Van der Sande, G., Brunner, D. & Soriano, M. C. Advances in photonic reservoir computing. *Nanophotonics* **6,** 561–576 (May 2017).

138. Dong, J., Rafayelyan, M., Krzakala, F. & Gigan, S. Optical Reservoir Computing Using Multiple Light Scattering for Chaotic Systems Prediction. *IEEE Journal of Selected Topics in Quantum Electronics* **26,** 1–12 (Jan. 2020).

139. Nahmias, M. A. *et al.* Photonic Multiply-Accumulate Operations for Neural Networks. *IEEE Journal of Selected Topics in Quantum Electronics* **26,** 1–18 (Jan. 2020).

140. Ahmed, S. R. *et al.* Universal photonic artificial intelligence acceleration. *Nature* **640,** 368–374 (Apr. 2025).

141. Jouppi, N. P. *et al. In-Datacenter Performance Analysis of a Tensor Processing Unit* in *Proceedings of the 44th Annual International Symposium on Computer Architecture* (ACM, New York, NY, USA, June 2017), 1–12.

142. Jouppi, N. P. *et al. Ten Lessons From Three Generations Shaped Google's TPUv4i : Industrial Product* in *2021 ACM/IEEE 48th Annual International Symposium on Computer Architecture (ISCA)* (IEEE, June 2021), 1–14.

143. Reed, G. T., Mashanovich, G., Gardes, F. Y. & Thomson, D. J. Silicon optical modulators. *Nature Photonics* **4,** 518–526 (Aug. 2010).

144. Leuthold, J., Koos, C. & Freude, W. Nonlinear silicon photonics. *Nature Photonics* **4,** 535–544 (2010).

145. Harris, N. C. *et al.* Efficient, compact and low loss thermo-optic phase shifter in silicon. *Optics Express* **22,** 10487 (May 2014).

146. Liu, A. *et al.* High-speed optical modulation based on carrier depletion in a silicon waveguide. *Optics Express* **15,** 660 (Jan. 2007).

147. Gupta, S. *et al. 50GHz Ge Waveguide Electro-Absorption Modulator Integrated in a 220nm SOI Photonics Platform* in *Optical Fiber Communication Conference* **1** (OSA, Washington, D.C., 2015), Tu2A.4.

148. Wang, C. *et al.* Integrated lithium niobate electro-optic modulators operating at CMOS-compatible voltages. *Nature* **562,** 101–104 (2018).

149. He, M. *et al.* High-performance hybrid silicon and lithium niobate Mach–Zehnder modulators for 100 Gbit s -1 and beyond. *Nature Photonics* **13,** 359–364 (2019).

150. Michel, J., Liu, J. & Kimerling, L. C. High-performance Ge-on-Si photodetectors. *Nature Photonics* **4,** 527–534 (Aug. 2010).

151. Liu, C. *et al.* Silicon/2D-material photodetectors: from near-infrared to mid-infrared. *Light: Science & Applications* **10,** 123 (June 2021).

152. Ahn, G. H. *et al.* Platform-agnostic waveguide integration of high-speed photodetectors with evaporated tellurium thin films. *Optica* **10,** 349 (Mar. 2023).

153. Hosseini, P., Wright, C. D. & Bhaskaran, H. An optoelectronic framework enabled by low-dimensional phase-change films. *Nature* **511,** 206–211 (July 2014).

154. Sarwat, S. G. *et al.* Engineering Interface-Dependent Photoconductivity in Ge 2 Sb 2 Te 5 Nanoscale Devices. *ACS Applied Materials and Interfaces* **10,** 44906–44914 (Dec. 2018).

155. Sarwat, S. G., Moraitis, T., Wright, C. D. & Bhaskaran, H. Chalcogenide optomemristors for multi-factor neuromorphic computation. *Nature Communications* **13,** 2247 (Dec. 2022).

156. Youngblood, N. *et al.* Reconfigurable Low-Emissivity Optical Coating Using Ultrathin Phase Change Materials. *ACS Photonics* **9,** 90–100 (Jan. 2022).

157. Popescu, C. C. *et al.* Electrically Reconfigurable Phase-Change Transmissive Metasurface. *Advanced Materials* **36,** 2400627 (July 2024).

158. Farmakidis, N. *et al.* Plasmonic nanogap enhanced phase-change devices with dual electrical-optical functionality. *Science Advances* **5,** eaaw2687 (Nov. 2019).

159. Farmakidis, N. *et al.* Electronically Reconfigurable Photonic Switches Incorporating Plasmonic Structures and Phase Change Materials. *Advanced Science* **9,** 2200383 (Apr. 2022).

160. Zhou, W. *et al.* In-memory photonic dot-product engine with electrically programmable weight banks. *Nature Communications* **14,** 2887 (May 2023).

161. Zhang, H. *et al.* Miniature Multilevel Optical Memristive Switch Using Phase Change Material. *ACS Photonics* **6,** 2205–2212 (Sept. 2019).

162. Fang, Z. *et al.* Ultra-low-energy programmable non-volatile silicon photonics based on phase-change materials with graphene heaters. *Nature Nanotechnology* **17,** 842–848 (Aug. 2022).

163. Chen, R. *et al.* Opportunities and Challenges for Large-Scale Phase-Change Material Integrated Electro-Photonics. *ACS Photonics* **9,** 3181–3195 (Oct. 2022).

164. Chen, R. *et al.* Non-volatile electrically programmable integrated photonics with a 5-bit operation. *Nature Communications* **14,** 3465 (June 2023).

165. Okamoto, K. in *Fundamentals of Optical Waveguides* 13–55 (Elsevier, 2006).

166. Ríos-Ocampo, C. *Phase-Change Materials for Photonic Memories and Optoelectronic Applications* PhD thesis (University of Oxford, 2016).

167. Marchetti, R., Lacava, C., Carroll, L., Gradkowski, K. & Minzioni, P. Coupling strategies for silicon photonics integrated chips [Invited]. *Photonics Research* **7,** 201 (Feb. 2019).

168. Cheng, L., Mao, S., Li, Z., Han, Y. & Fu, H. Grating Couplers on Silicon Photonics: Design Principles, Emerging Trends and Practical Issues. *Micromachines* **11,** 666 (July 2020).

169. Bozzola, A., Carroll, L., Gerace, D., Cristiani, I. & Andreani, L. C. Optimising apodized grating couplers in a pure SOI platform to 05 dB coupling efficiency. *Optics Express* **23,** 16289 (June 2015).

170. Wang, H. *et al.* Ultralow-loss optical interconnect enabled by topological unidirectional guided resonance. *Science Advances* **10,** eadn4372 (Mar. 2024).

171. Maier, S. A. *Plasmonics: Fundamentals and Applications* (Springer US, New York, NY, 2007).

172. Liu, Y., Lai, Y. & Chang, K. Plasmonic coupler for silicon-based micro-slabs to plasmoinc nano-gap waveguide mode coversion enhancement. *Journal of Lightwave Technology* **31,** 1708–1712 (2013).

173. Chen, L., Shakya, J. & Lipson, M. Subwavelength confinement in an integrated metal slot waveguide on silicon. *Optics Letters* **31,** 2133 (2006).

174. Veronis, G. & Fan, S. Theoretical investigation of compact couplers between dielectric slab waveguides and two-dimensional metal-dielectric-metal plasmonic waveguides. *Optics Express* **15,** 1211 (2007).

175. Tian, J., Yu, S., Yan, W. & Qiu, M. Broadband high-efficiency surface-plasmon-polariton coupler with silicon-metal interface. *Applied Physics Letters* **95,** 7–10 (2009).

176. Thomas, R., Ikonic, Z. & Kelsall, R. Silicon based plasmonic coupler. *Optics Express* **20,** 21520 (2012).

177. Chen, C.-T., Xu, X., Hosseini, A., Pan, Z. & Chen, R. T. High efficiency silicon strip waveguide to plasmonic slot waveguide mode converter. *Optical Interconnects XV* **9368,** 936809 (2015).

178. Ono, M. *et al.* Deep-subwavelength plasmonic mode converter with large size reduction for Si-wire waveguide. *Optica* **3,** 999 (2016).

179. Li, X. *et al.* Experimental investigation of silicon and silicon nitride platforms for phase-change photonic in-memory computing. *Optica* **7,** 218 (2020).

180. Rios, C. *et al.* Controlled switching of phase-change materials by evanescent-field coupling in integrated photonics [Invited]. *Optical Materials Express* **8,** 2455 (Sept. 2018).

181. Yamada, N., Ohno, E., Nishiuchi, K., Akahira, N. & Takao, M. Rapid-phase transitions of GeTe-Sb2Te3 pseudobinary amorphous thin films for an optical disk memory. *Journal of Applied Physics* **69,** 2849–2856 (Mar. 1991).

182. Vivien, L. *et al.* Comparison between strip and rib SOI microwaveguides for intra-chip light distribution. *Optical Materials* **27,** 756–762 (Feb. 2005).

183. Shields, J., Galarreta, C. R. d., Bertolotti, J. & Wright, C. D. Enhanced Performance and Diffusion Robustness of Phase-Change Metasurfaces via a Hybrid Dielectric/Plasmonic Approach. *Nanomaterials* **11,** 525 (Feb. 2021).

184. Lazarenko, P. I. *et al. Electrical properties of the Ge2Sb2Te5 thin films for phase change memory application* in *AIP Conf. Proc* (2016), 020013.

185. Gao, Y., Gan, Q., Xin, Z., Cheng, X. & Bartoli, F. J. Plasmonic Mach–Zehnder Interferometer for Ultrasensitive On-Chip Biosensing. *ACS Nano* **5,** 9836–9844 (Dec. 2011).

186. Chen, H. & Poon, A. W. Low-Loss Multimode-Interference-Based Crossings for Silicon Wire Waveguides. *IEEE Photonics Technology Letters* **18,** 2260–2262 (Nov. 2006).

187. Zhang, Y. *et al.* Myths and truths about optical phase change materials: A perspective. *Applied Physics Letters* **118,** 210501 (May 2021).

188. Loke, D. K. *et al.* Ultrafast Nanoscale Phase-Change Memory Enabled By Single-Pulse Conditioning. *ACS Applied Materials & Interfaces* **10,** 41855–41860 (Dec. 2018).

189. Martin-Monier, L. *et al.* Endurance of chalcogenide optical phase change materials: a review. *Optical Materials Express* **12,** 2145 (June 2022).

190. Youngblood, N. *et al.* Tunable Volatility of Ge2Sb2Te5 in Integrated Photonics. *Advanced Functional Materials* **29,** 1807571 (Mar. 2019).

191. Stegmaier, M., Ríos, C., Bhaskaran, H. & Pernice, W. H. P. Thermo-optical Effect in Phase-Change Nanophotonics. *ACS Photonics* **3,** 828–835 (May 2016).

192. Waldecker, L. *et al.* Time-domain separation of optical properties from structural transitions in resonantly bonded materials. *Nature Materials* **14,** 991–995 (Oct. 2015).

193. Loke, D. *et al.* Breaking the Speed Limits of Phase-Change Memory. *Science* **336,** 1566–1569 (June 2012).

194. Albrechtsen, M. *et al.* Nanometer-scale photon confinement in topology-optimized dielectric cavities. *Nature Communications* **13,** 6281 (Oct. 2022).

195. Koelmans, W. W. *et al.* Projected phase-change memory devices. *Nature Communications* **6,** 8181 (Sept. 2015).

196. Syed, G. S., Zhou, Y., Warner, J. & Bhaskaran, H. Atomically thin optomemristive feedback neurons. *Nature Nanotechnology* **18,** 1036–1043 (Sept. 2023).

197. Miller, D. A. B., Chemla, D. S. & Schmitt-Rink, S. Relation between electroabsorption in bulk semiconductors and in quantum wells: The quantum-confined Franz-Keldysh effect. *Physical Review B* **33,** 6976–6982 (May 1986).

198. Crockett, L. H., Northcote, D. & Stewart, R. W. (eds.) Software Defined Radio with Zynq UltraScale+ RFSoC (First Edition). *Strathclyde Academic Media,* https://www.RFSoCbook.com. (2023).

199. Zhong, Z. *et al. Lightning: A Reconfigurable Photonic-Electronic SmartNIC for Fast and Energy-Efficient Inference* in *SIGCOMM 2023 - Proceedings of the ACM SIGCOMM 2023 Conference* (2023), 452–472.

200. Geler-Kremer, J. *et al.* A ferroelectric multilevel non-volatile photonic phase shifter. *Nature Photonics* **16,** 491–497 (July 2022).

201. Chen, Y. *et al. First Demonstration of Fully CMOS-compatible Non-volatile Programmable Photonic Switch Enabled by Ferroelectric-SOI Waveguide for Next Generation Photonic Integrated Circuit* in *2022 IEEE Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits)* (IEEE, June 2022), 405–406.

202. Zhang, G. *et al.* Thin film ferroelectric photonic-electronic memory. *Light: Science & Applications* **13,** 206 (Aug. 2024).

203. Pintus, P. *et al.* Integrated non-reciprocal magneto-optics with ultra-high endurance for photonic in-memory computing. *Nature Photonics* **19,** 54–62 (Jan. 2025).

204. Tossoun, B. *et al.* High-speed and energy-efficient non-volatile silicon photonic memory based on heterogeneously integrated memresonator. *Nature Communications* **15,** 551 (Jan. 2024).

205. Salinga, M. *et al.* Monatomic phase change memory. *Nature Materials* **17,** 681–685 (Aug. 2018).

206. Cheng, Z. *et al.* Antimony thin films demonstrate programmable optical nonlinearity. *Science Advances* **7,** eabd7097 (Jan. 2021).

207. Aggarwal, S. *et al.* Antimony as a Programmable Element in Integrated Nanophotonics. *Nano Letters* **22,** 3532–3538 (May 2022).

208. Delaney, M., Zeimpekis, I., Lawson, D., Hewak, D. W. & Muskens, O. L. A New Family of Ultralow Loss Reversible Phase-Change Materials for Photonic Integrated Circuits: Sb2S3 and Sb2Se3. *Advanced Functional Materials* **30,** 2002447 (Sept. 2020).

209. Williamson, I. A. D. *et al.* Reprogrammable Electro-Optic Nonlinear Activation Functions for Optical Neural Networks. *IEEE Journal of Selected Topics in Quantum Electronics* **26,** 1–12 (Jan. 2020).

210. Pour Fard, M. M. *et al.* Experimental realization of arbitrary activation functions for optical neural networks. *Optics Express* **28,** 12138 (Apr. 2020).

211. Huang, C. *et al.* On-Chip Programmable Nonlinear Optical Signal Processor and Its Applications. *IEEE Journal of Selected Topics in Quantum Electronics* **27,** 1–11 (Mar. 2021).

212. Jha, A., Huang, C. & Prucnal, P. R. Reconfigurable all-optical nonlinear activation functions for neuromorphic photonics. *Optics Letters* **45,** 4819 (2020).

213. Kalinin, K. P. *et al.* Analog Iterative Machine (AIM): using light to solve quadratic optimization problems with mixed variables. *arXiv preprint arXiv:2304.12594* (Apr. 2023).

214. Hua, S. *et al.* An integrated large-scale photonic accelerator with ultralow latency. *Nature* **640,** 361–367 (Apr. 2025).

215. Daudlin, S. *et al.* Three-dimensional photonic integration for ultra-low-energy, high-bandwidth interchip data links. *Nature Photonics* **19,** 502–509 (May 2025).

216. Fathololoumi, S. *et al. Highly Integrated 4 Tbps Silicon Photonic IC for Compute Fabric Connectivity* in *2022 IEEE Symposium on High-Performance Interconnects (HOTI)* (IEEE, Aug. 2022), 1–4.

217. Steinman, M. *Hummingbird™ Low-Latency Computing Engine* in *2023 IEEE Hot Chips 35 Symposium (HCS)* (IEEE, Aug. 2023), 1–20.