Microring-based Photonic Tensor Cores and their Calibration Methods

Minkyu Cho

Department of Electrical and Electronic Engineering
Graduate School
Yonsei University

Microring-based Photonic Tensor Cores and their Calibration Methods

A Master's Thesis Submitted
to the Department of Electrical and Electronic Engineering
and the Committee on Graduate School
of Yonsei University in Partial Fulfillment of the
Requirements for the Degree of
Master of Science

Minkyu Cho

August 2026

Microring-based Photonic Tensor Cores and their Calibration Methods

This Certifies that the Master's Thesis
of Minkyu Cho is Approved

Committee Chair	**Woo-Young Choi**
Committee Member	**Kyoungsik Yu**
Committee Member	**Myung-Jae Lee**

Department of Electrical and Electronic Engineering
Graduate School
Yonsei University
August 2026

TABLE OF CONTENTS

List of Figures

List of Tables

Abstract

Microring-based Photonic Tensor Cores and their Calibration Methods

This thesis investigates microring-based photonic tensor cores from the viewpoints of architecture and calibration. It first establishes the conventional silicon microring weightbank as a baseline for wavelength-multiplexed matrix–vector multiplication by reviewing its transfer characteristics, weight definition, scaling behavior, and experimental operation with a two-channel prototype and a small neural inference task. This baseline confirms the compactness and WDM compatibility of microring weightbanks, while also revealing two major limitations for realistic AI workloads: inefficient support for signed inputs and a dataflow mismatch between MVM-oriented hardware and GEMM-dominant computation.

To address the first limitation, this thesis proposes a bidirectional microring weightbank for signed multiplications. The proposed structure embeds signed representation in a differential bidirectional intensity-modulation framework, avoiding phase-based sign encoding while reducing the optical duplication typically required in conventional signed schemes. One-channel experiments, heater-sweep characterization, and 2D LUT calibration are used to validate the signed encoding principle and to show that the architecture can improve effective density and calibration efficiency, although multi-channel calibration and reflection sensitivity remain open problems.

To address the second limitation, this thesis proposes a RAMZM-based photonic tensor-core architecture that is more naturally aligned with matrix–matrix multiplication. The architecture uses independently encoded row and column operands, complex-plane trajectory analysis, and LUT-based stable operand encoding, and it is validated through direct RAMZM-output measurements using sequential and random LUT sweeps together with repeated-trial error statistics. Based on this operating principle, a reusable 2×2 GEMM tile layout is designed as a building block for larger photonic matrix–matrix multiplication engines. Together, these results show that microring-based photonic tensor cores can be extended beyond conventional non-negative MVM weightbanks toward signed-operation support and GEMM-native organization, with calibration emerging as a central requirement for practical photonic computing systems.

Key words: Photonic Tensor Core, Photonic Neural Network, Photonic Computing, Silicon Photonics

1. Introduction

1.1. AI Scaling and Computational Pressure

Artificial intelligence has become a central driver of advanced computing demand because state-of-the-art models are now trained and deployed at scales that were unusual only a few years ago. This growth is not limited to the absolute number of operations executed during training; it also appears in model size, dataset size, and the amount of infrastructure required to serve large models after training. Figure 1.1 illustrates the rapid rise of training compute reported for notable AI models. As this trend continues, the hardware question is no longer simply how to build a faster processor, but how to provide the required throughput within practical limits of power, area, and cost. For accelerator research, this context is important because the pressure on hardware originates from the workload itself rather than from a temporary inefficiency in software implementation [1].

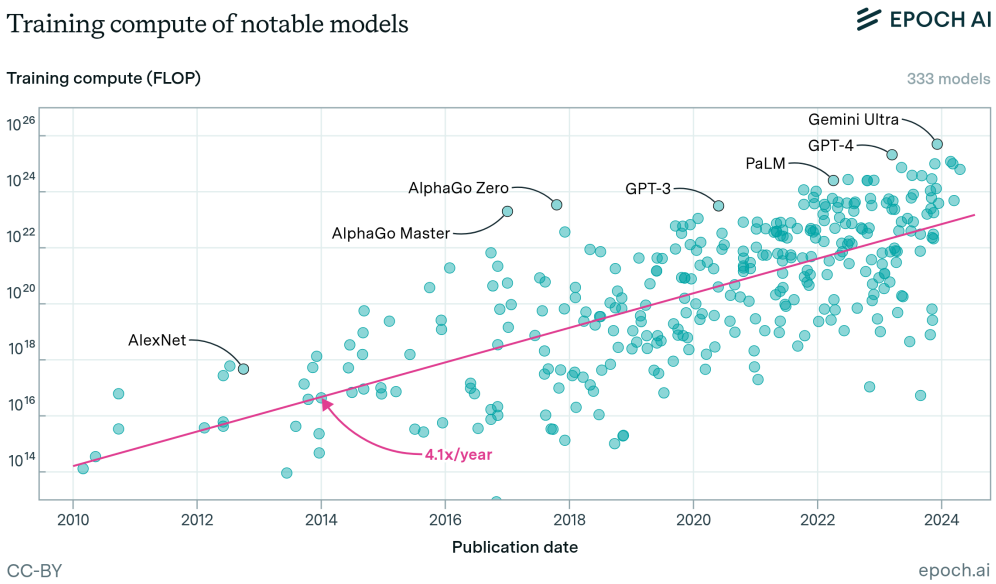


Figure 1.1: Rapid growth of AI training compute for notable modern models, adapted from Epoch AI [1].

The significance of this growth becomes clearer when it is compared with the rate of improvement in conventional machine-learning hardware. Figure 1.2 shows that the energy-efficiency gains of leading ML hardware are much slower than the rise in AI compute demand. Digital electronics

continue to improve, and those improvements remain indispensable. Nevertheless, when workload growth outpaces hardware efficiency gains, the system-level response is usually to add more chips, enlarge memory capacity, expand interconnect bandwidth, and accept a higher power budget. That response can sustain progress, but it also increases cost and design complexity and places greater emphasis on data movement and communication. In this sense, the contrast between Figures 1.1 and 1.2 does not imply that conventional hardware is ineffective; rather, it indicates that conventional scaling alone may not be sufficient to support the long-term growth of AI workloads [2, 3]. This widening pressure on throughput and communication also helps motivate interest in alternative interconnect and accelerator substrates, including optical approaches [4].

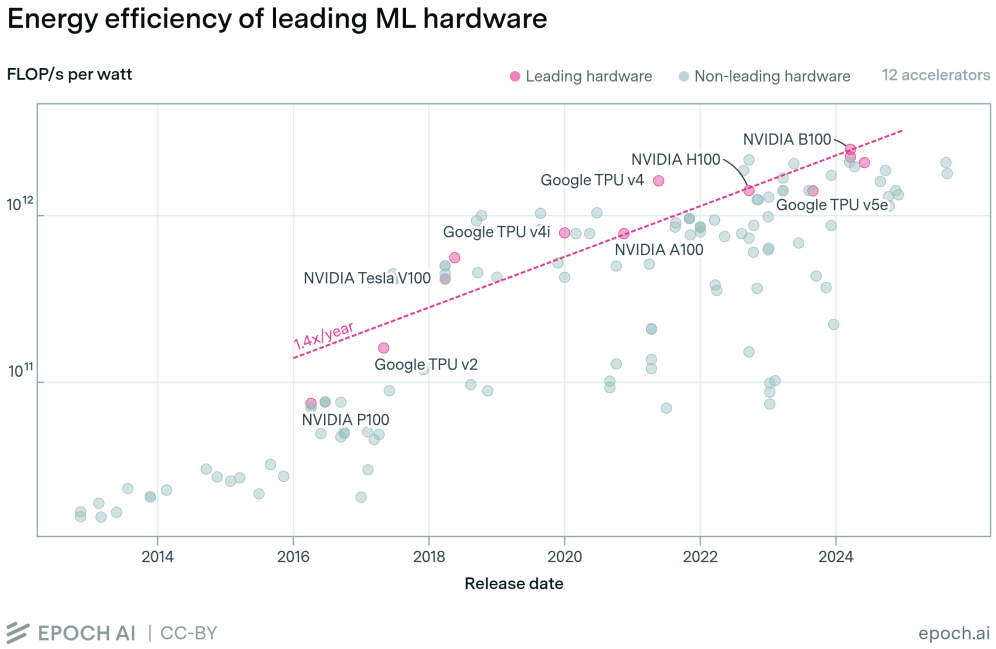


Figure 1.2: Improvement trend of leading machine-learning hardware energy efficiency, which progresses more slowly than AI compute demand, adapted from Epoch AI [3].

This pressure is unlikely to disappear soon because it is reinforced by empirical scaling laws. As summarized in Figure 1.3, modern AI models continue to exhibit systematic performance improvements when compute, dataset size, and parameter count are increased in a coordinated manner. The precise rate of improvement depends on model class, data quality, and training strategy, and practical limits will certainly arise from cost and resource constraints. Even so, the general implication is important for hardware architecture: larger models and larger training runs are not pursued only for scale itself, but because additional scale still tends to yield measurable gains. Consequently, the incentive to develop more capable computing substrates is likely to persist. The architectural problem

is therefore not only one of peak speed, but also one of sustaining useful computation at scales that remain economically and energetically reasonable [5].

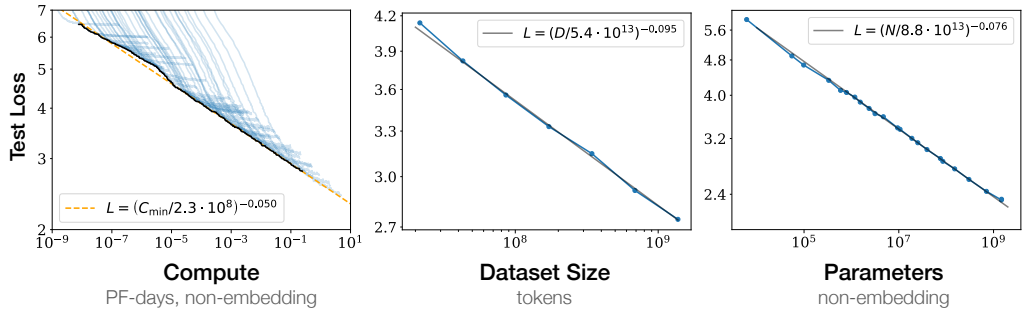


Figure 1.3: Representative empirical scaling laws of modern AI models, showing continued performance improvement with increasing model size, data, and compute, adapted from [5].

1.2. Linear Arithmetic and the Case for Photonic Computing

The persistence of scaling pressure leads directly to a second question: which parts of modern AI workloads most urgently deserve architectural acceleration? In Transformer-based models, a large share of the arithmetic burden comes from linear operations, especially matrix multiplications and linear projections. Query, key, and value formation, attention output projections, and the expansion and compression stages of the feed-forward network are all dominated by dense linear algebra. Figure 1.4 highlights this point for a representative Transformer block. In terms of repeated arithmetic kernels, linear transforms remain the main computational workload. This makes alternative accelerator substrates for massively parallel linear computation particularly attractive.

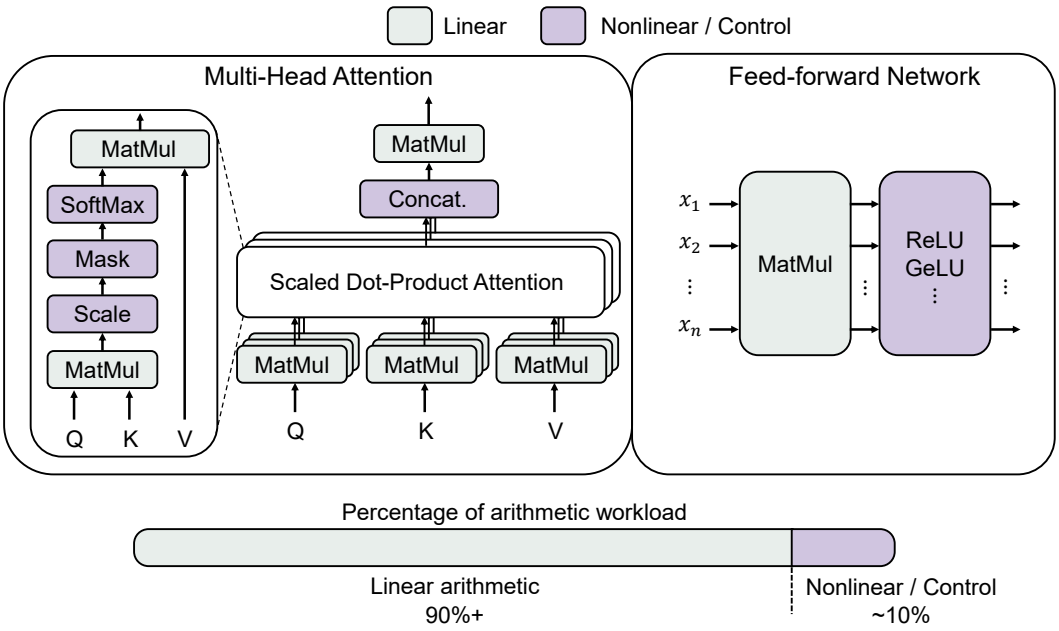


Figure 1.4: Dominance of linear arithmetic, including matrix multiplications and projections, in a representative Transformer block.

One such substrate is photonic computing. The appeal of photonics does not lie in replacing all digital computation with optics, but in exploiting the physical properties of light where they align naturally with the structure of the workload. Optical systems offer high bandwidth, native superposition, and straightforward wavelength-division multiplexing, all of which can be advantageous for high-throughput linear transforms. As a result, a broad family of photonic computing architectures has emerged. Figure 1.5 provides a schematic taxonomy of representative photonic architectures, grouped by parallelism scheme and dataflow organization rather than as a strict one-to-one catalog of six published systems. On the parallelism-scheme side, the figure contrasts decomposed coherent transforms, free-space diffractive propagation, resonant fan-out/WDM weighting, and mesh- or array-style interferometric processing. On the dataflow side, it highlights organizations that rely on backend demultiplexing as well as direct crossbar dot-product arrays aimed at matrix–matrix execution. Representative examples across these categories include coherent mesh-based processors [6], diffractive optical networks [7], microring weightbanks [8], phase-change-material photonic tensor-core devices [9], Mach–Zehnder-based matrix–matrix processors [10], and integrated coherent crossbar platforms [11]. Despite their differences, all of them pursue the same architectural objective: to accelerate high-throughput linear computation in a form that is useful for modern AI workloads.

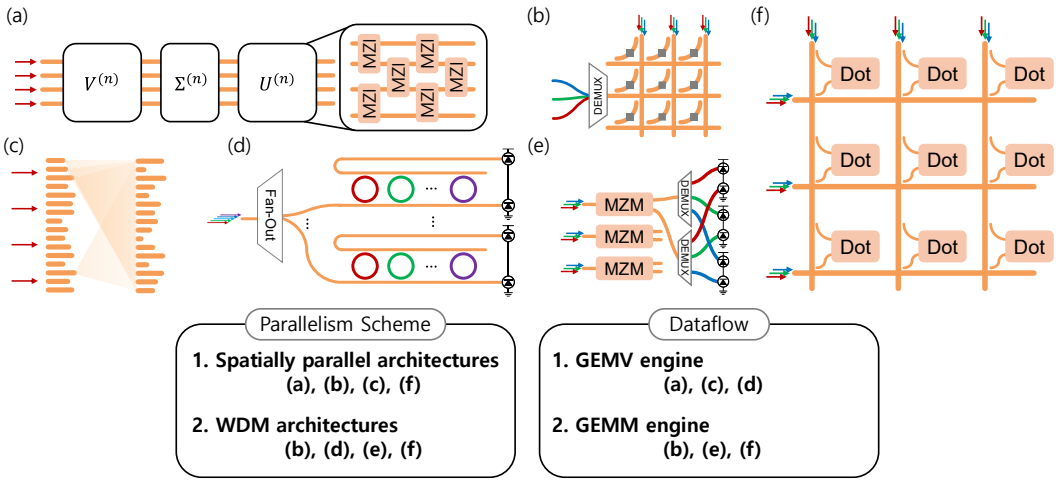


Figure 1.5: Representative photonic computing architecture families, schematically grouped by parallelism scheme and dataflow organization.

The architectural landscape shown in Figure 1.5 motivates a broader system-level question concerning where photonic accelerators may be most competitive. In this regard, neuromorphic photonics has emerged as a promising regime in the efficiency–speed space. Figure 1.6 situates this regime relative to representative computing approaches. The central idea is to combine photonic parallelism and bandwidth with electronic control, storage, and nonlinear processing, thereby targeting workloads in which very high linear throughput is more valuable than exact digital sequential logic. In this context, the efficiency–speed regime associated with neuromorphic photonics remains compelling because it suggests a practical middle ground between purely digital acceleration and all-optical computing [12, 13, 14].

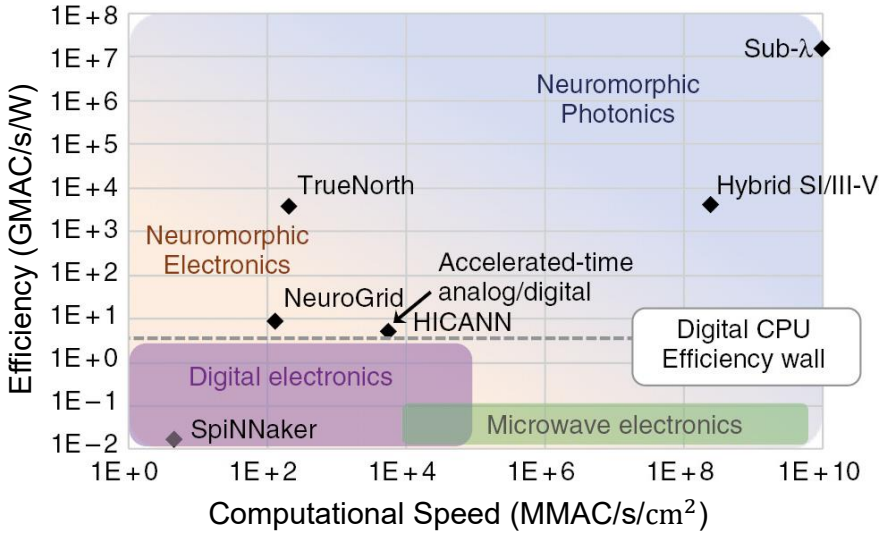


Figure 1.6: Efficiency–speed regime of neuromorphic photonics relative to representative computing approaches, adapted from [15].

1.3. Microring-Based Photonic Tensor Cores and Their Calibration

Among the many photonic approaches represented in Figures 1.5 and 1.6, microring-based architectures are particularly attractive for scalable tensor-core design. A microring resonator provides a compact and wavelength-selective optical element whose transmission can be programmed through resonance tuning. This gives it several properties that are valuable for photonic linear algebra: a small footprint, natural compatibility with wavelength-division multiplexing, the ability to densely integrate many tunable weight elements, and a direct connection between device tuning and weighted optical summation. These advantages have made microring weightbanks an influential platform in neuromorphic photonics [8, 16, 17, 18]. In their conventional form, however, microring weightbanks are structurally awkward for signed multiplication and are more naturally aligned with matrix–vector multiplication than with GEMM-style matrix–matrix execution [8, 9, 14, 16].

This thesis is motivated by that combination of opportunity and constraint. It investigates microring-based photonic tensor cores as accelerator structures for high-throughput linear algebra and studies the calibration methods required to make such structures practically usable. The discussion begins from the conventional microring weightbank as a baseline, then moves to a bidirectional microring architecture for signed multiplications, and finally considers a tensor-core organization aimed at matrix–matrix multiplication. Throughout the thesis, the discussion identifies the operating regime in which microring-based photonic tensor cores may provide a meaningful advantage and clarifies the calibration strategies needed to realize that advantage in practice.

2. Conventional Si Microring Weightbank

2.1. Overview

Before proposing new architectures, it is necessary to establish a clear baseline. This section describes the conventional MR weightbank used for photonic weighted summation, explains how weights are encoded through through-transmission and drop-transmission responses, and reviews a representative two-channel experimental implementation [8, 14, 16].

In this thesis, a "weightbank" refers to an array of tunable optical transfer elements that apply scalar coefficients to optical input channels. When multiple weighted signals are combined in the optical domain without additional overhead, the resulting photocurrent returns a weighted sum. This optical weighted-summation mechanism, which maps physical optical signals to mathematical dot-product terms, is central to photonic realizations of neural layers.

2.2. MR Transfer Characteristics and Weight Definition

The MR provides two complementary transmission responses: through transmission and drop transmission. As resonance is tuned, the two transmission responses vary in opposite directions around the resonance region. This complementary behavior is useful for weight encoding because differential forms enable transfer ranges centered around zero.

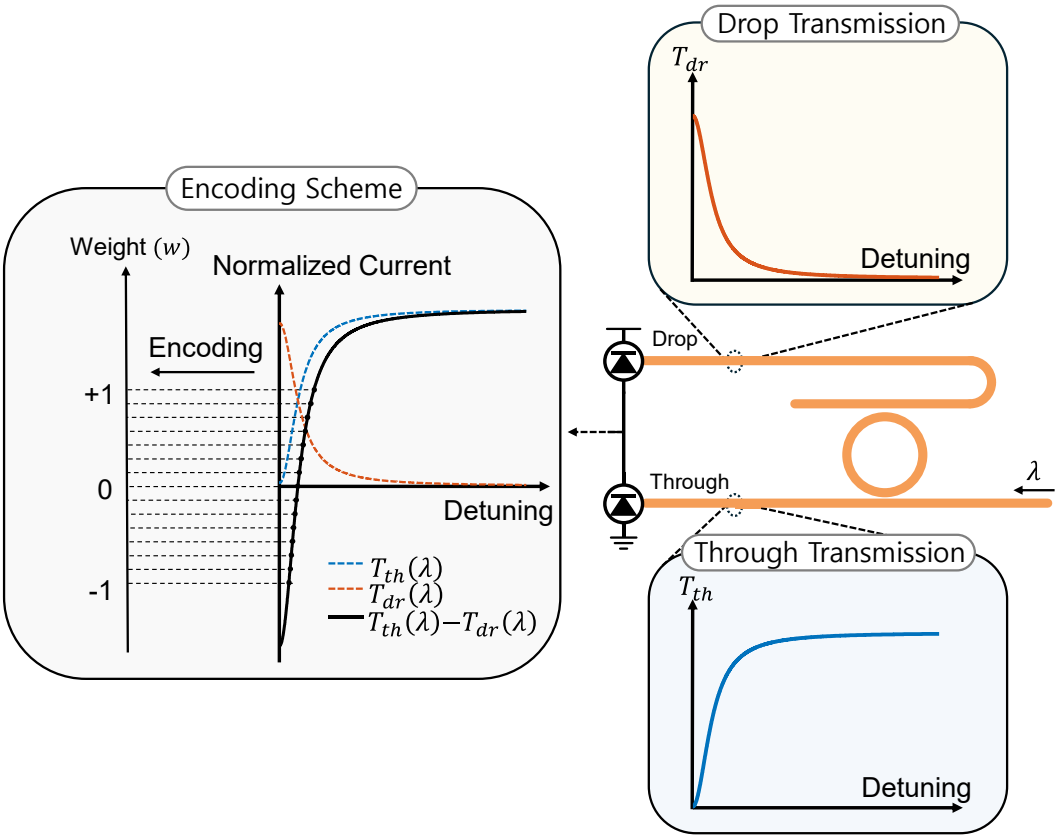


Figure 2.1: Fundamental MR operation and transfer curves for through and drop transmissions. The effective weight can be defined from differential transmission and then linearized over a usable operating range.

Let $T_{\mathrm{th}}(\lambda)$ and $T_{\mathrm{dr}}(\lambda)$ denote normalized through and drop transmissions. A practical differential weight variable can be written as

$$w = \alpha \left(T_{\mathrm{th}} - T_{\mathrm{dr}}\right), \tag{2.1}$$

where α is a scaling factor chosen to match the target neural-network weight range, such as $[-5, 5]$ or $[-3, 3]$, to the physically usable differential-transmission range of the MR. This sign convention follows the differential curve plotted in Figure 2.1; at system level the sign can be flipped in calibration if an alternate convention is preferred. Because raw transfer is nonlinear with respect to the electrical control variable (e.g., heater tuning command or PN-junction bias voltage), calibration maps target numerical weights to linearized electrical drive levels suitable for integer datatype encoding.

The conventional workflow is therefore:

1. characterize transfer response per ring,

2. select monotonic and near-linear operating window,

3. generate LUT from desired weight to electrical control codes.

2.3. Scaling to Matrix–Vector Multiplication

A key strength of MR weightbanks is their native compatibility with WDM, which enables straightforward scale-up from single-channel weighting to wavelength-parallel matrix–vector multiplication. As shown in Figure 2.2, the input encoders are composed of intensity modulators such as MR modulators. Several WDM continuous-wave (CW) wavelengths are intensity-modulated to form the input vector $\mathbf{X}$, and this WDM signal is distributed to multiple parallel weightbank branches. Here, "intensity-modulated input mapping" means that each numerical input x_i is encoded as the optical power of wavelength channel λ_i, i.e., $x_i \mapsto P_i(\lambda_i)$, after normalization to the modulator operating range. In Figure 2.2, each horizontal branch represents one output path of the weightbank. Along one branch, all wavelength channels are weighted by the corresponding MR elements and then combined at a photodetector, so that the detected value from that branch becomes one component of the final output vector.

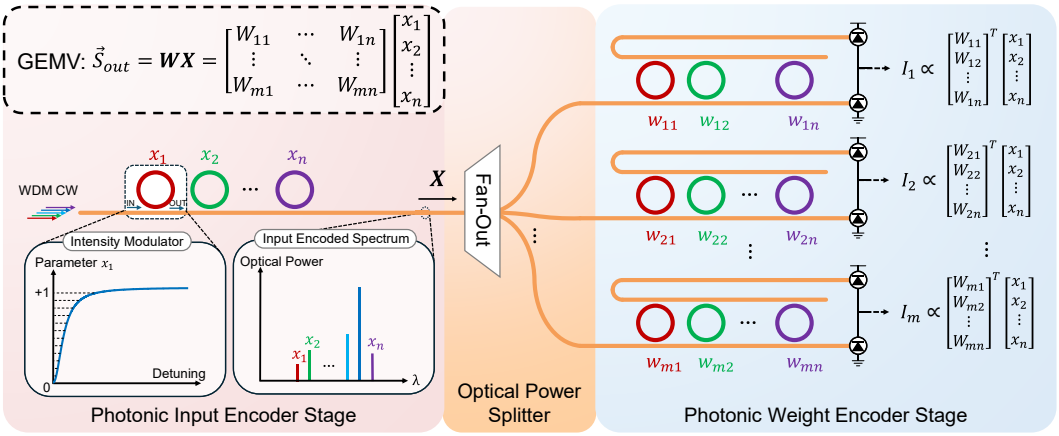


Figure 2.2: Scaled conventional MR weightbank for matrix–vector multiplication. Input channels are WDM-encoded, distributed across parallel output branches, weighted by MR encoders, and summed at photodetectors.

Assume an input vector $\mathbf{x} \in \mathbb{R}^N$ mapped to N wavelength channels. For the jth output branch, the detected signal is written as

$$y_j = \sum_{i=1}^{N} w_{j,i} x_i. \tag{2.2}$$

That is, branch j applies the weights $w_{j,1}, \ldots, w_{j,N}$ to the shared input channels and produces the jth element of the output vector.

This architecture has several attractive properties:

- natural channel parallelism through WDM,
- compact per-weight footprint using resonators,
- direct physical accumulation at detection stage.

This structure is especially useful because the same WDM input can be broadcast to many parallel branches, allowing one optical input vector to be reused while different weight settings generate multiple outputs simultaneously. As a result, the conventional MR weightbank provides an intuitive and physically direct implementation of matrix–vector multiplication that serves as a useful baseline for the later architectures in this thesis.

2.4. Two-Channel Experimental Prototype

To validate baseline feasibility before large-scale integration, a two-channel conventional weight-bank prototype is used as an experimental test vehicle. In the setup captured in Figure 2.3, input values are encoded using variable optical attenuators (VOAs), then injected into an on-chip two-channel MR weightbank. The goal is to test whether commanded weighted sums match measured outputs over the chosen operating range.

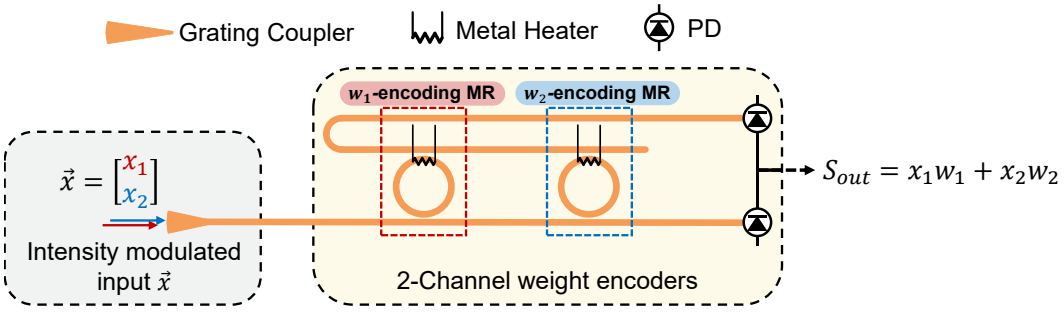


Figure 2.3: Experimental architecture of a two-channel conventional MR weightbank. Input encoding is performed by variable optical attenuators before on-chip weighting and readout.

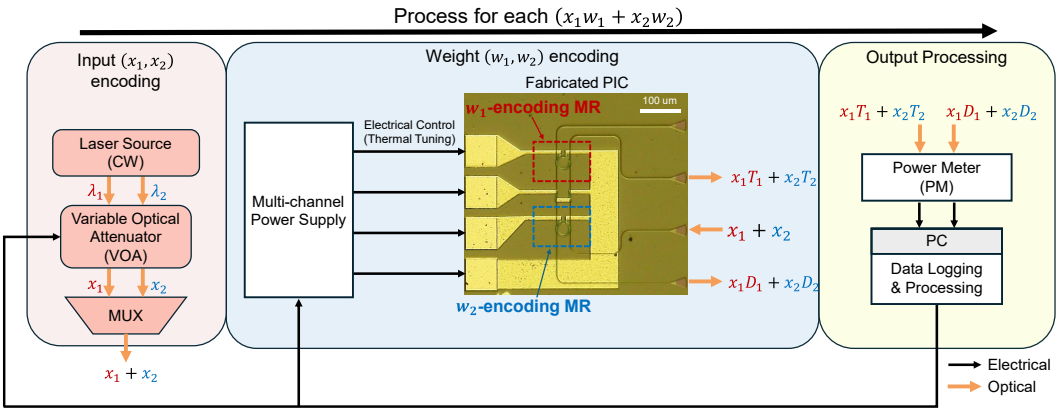


Figure 2.4: Overall measurement flow for the two-channel conventional MR weightbank. The figure summarizes the input-encoding stage, the fabricated PIC used for w_1/w_2 encoding, and the output-processing path used to read out the weighted optical sums.

Figure 2.4 summarizes the full measurement flow of the baseline experiment, from input encoding through the fabricated PIC to output processing. The setup includes laser sources, wavelength control elements, a VOA-based input-encoding stage, on-chip electrical heater drivers, and photodetector readout electronics.

To verify the prototype behavior before task-level demonstration, Figure 2.5 summarizes the key device-level results: (a) voltage-to-weight transfer curves for two channels and (b) commanded versus measured optical weighted-sum behavior.

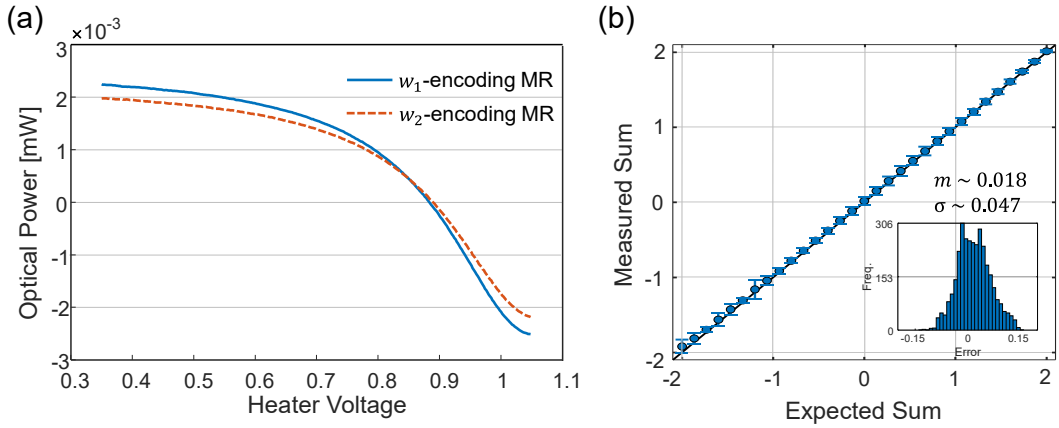


Figure 2.5: Representative results of the two-channel conventional weightbank: (a) voltage-to-weight curves and (b) commanded vs. measured optical weighted sum.

These results provide a direct baseline for this prototype. The transfer curves are usable but

channel dependent, so per-channel LUT mapping remains necessary. The commanded weighted sums and the measured outputs agree closely within the selected operating window.

An additional observation from Figure 2.5(a) is that the two channels exhibit different absolute optical-power scales. Therefore, channel-to-channel normalization is required. In this experiment, separate LUTs were constructed so that the same encoding value produces the same emitted optical power in both channels. Although direct fitting of input light intensity can also normalize the transfer curves, LUT-level adjustment is more general for subsequent realization because it calibrates input-power mismatch and channel-dependent transfer differences within a unified encoding framework.

2.5. Demonstration in a Small Neural Inference Task

A valuable way to evaluate the baseline weightbank is to embed it in a minimal neural inference pipeline. Figure 2.6 illustrates an Iris dataset classification experiment implemented with two-channel optical partial sums. The inner product term $x_1 w_1 + x_2 w_2$ is computed in the optical domain, while accumulation across layers and nonlinear activation are executed in the digital domain.

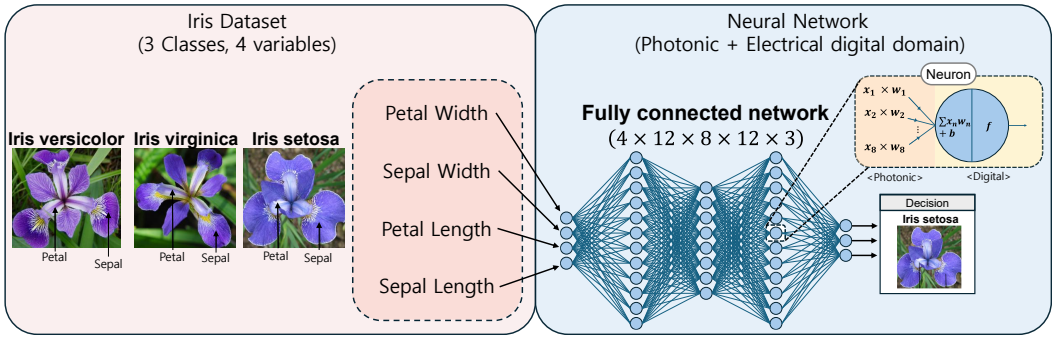


Figure 2.6: Iris dataset demonstration with two-channel optical partial sums. Optical hardware computes weighted partial terms, while higher-level accumulation and nonlinear functions are performed digitally.

This hybrid partitioning is representative of many practical photonic AI systems. Optics is used where parallel weighted summation is dominant; electronics remains responsible for nonlinear operations, global control, and memory orchestration. The partition avoids forcing optics into tasks where digital logic is inherently better (e.g., exact branching, dynamic control flow).

The corresponding task-level result is shown in Figure 2.7, which presents the confusion matrix for the Iris classification demonstration. The values in the matrix indicate the ratio between the ground-truth label and the predicted label for each class, so diagonal entries represent correct classification and off-diagonal entries represent misclassification.

Using the photonic weightbank, the experiment implements MAC operations from a pre-trained

fully connected network for Iris classification. The network uses the layer configuration $4 \times 12 \times 8 \times 12 \times 3$ and, for simplicity, includes weights and activation functions without bias terms. For a test set of 30 samples composed of 10 Setosa, 9 Versicolor, and 11 Virginica examples, the measured classification accuracy is 88.67%, whereas the corresponding pre-trained model achieves 100% accuracy.

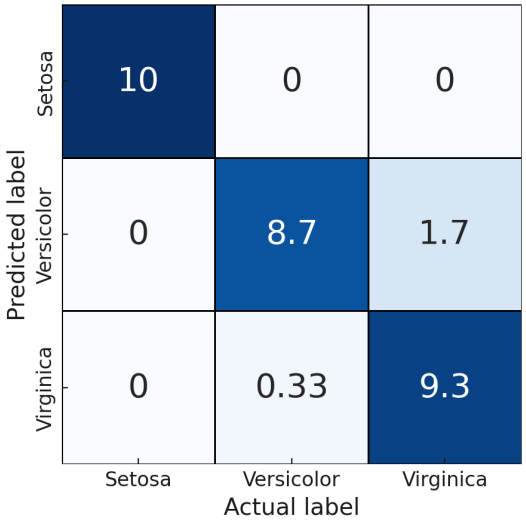


Figure 2.7: Confusion matrix for the Iris classification demonstration using the two-channel conventional MR weightbank.

2.6. Strengths of the Conventional Architecture

The conventional MR weightbank remains a strong baseline for several reasons.

Compactness and integration density. MR devices provide high functional density per footprint, enabling many weights in limited chip area.

Natural WDM compatibility. The architecture scales channel count through wavelength multiplexing, allowing vector parallelism without proportional routing duplication.

Direct optical accumulation. Weighted channels combine physically before readout, reducing explicit digital adder depth for front-end linear computation.

These strengths justify continued use of conventional weightbanks in workloads where inputs are mostly non-negative, such as ReLU-dominant inference paths. However, the conventional structure is not optimized for signed-input processing or GEMM-native matrix–matrix multiplication, so it serves as a baseline for the improved architectures introduced later.

2.7. Structural Limitations

Despite its practicality, the conventional architecture reveals critical limitations when targeting realistic AI workloads.

- **Signed-input overhead.** Most conventional mappings naturally realize non-negative transmission factors, so handling signed-input values with the conventional MR weightbank structure introduces additional overhead. Detailed methods are discussed in Section 3.
- **Dataflow mismatch for GEMM.** The conventional design excels in MVM, but large neural workloads often require GEMM with high operand reuse. Executing GEMM through repeated MVM passes can produce control and data-movement bottlenecks that reduce effective utilization. Section 4 addresses this limitation by introducing a new architecture for a GEMM-oriented tensor core.

3. Proposed Bidirectional Microring Weightbank for Signed Multiplications

3.1. Motivation and Problem Statement

Prior work has established MR weightbanks as compact, WDM-compatible primitives for photonic weighted summation [8, 16]. However, in intensity-based mappings, optical signals are non-negative by construction, so the main representational difficulty lies in signed-input representation. In conventional MR-based weightbanks, signed-weight values can already be realized comparatively easily through differential through/drop responses, whereas signed-input values usually require auxiliary mechanisms such as complex time scheduling (time-division multiplexing, TDM) or spatial-division multiplexing (SDM) with duplicated optical paths [19, 20, 21]. These workarounds are functional, but they introduce either scheduling overhead or structural duplication that must later be calibrated and stabilized.

Figure 3.1 summarizes this limitation. Two common workarounds are typically used:

- **Method 1 – Complex time scheduling (TDM):** place positive and negative contributions in separate time slots, then recombine them at readout.
- **Method 2 – Spatial path duplication (SDM):** map positive and negative contributions to separate optical paths using extra modulators, then subtract at readout.

Both methods are functional, but their penalties differ: TDM mainly reduces effective throughput and increases timing/control complexity, while SDM mainly increases area, routing complexity, and calibration load.

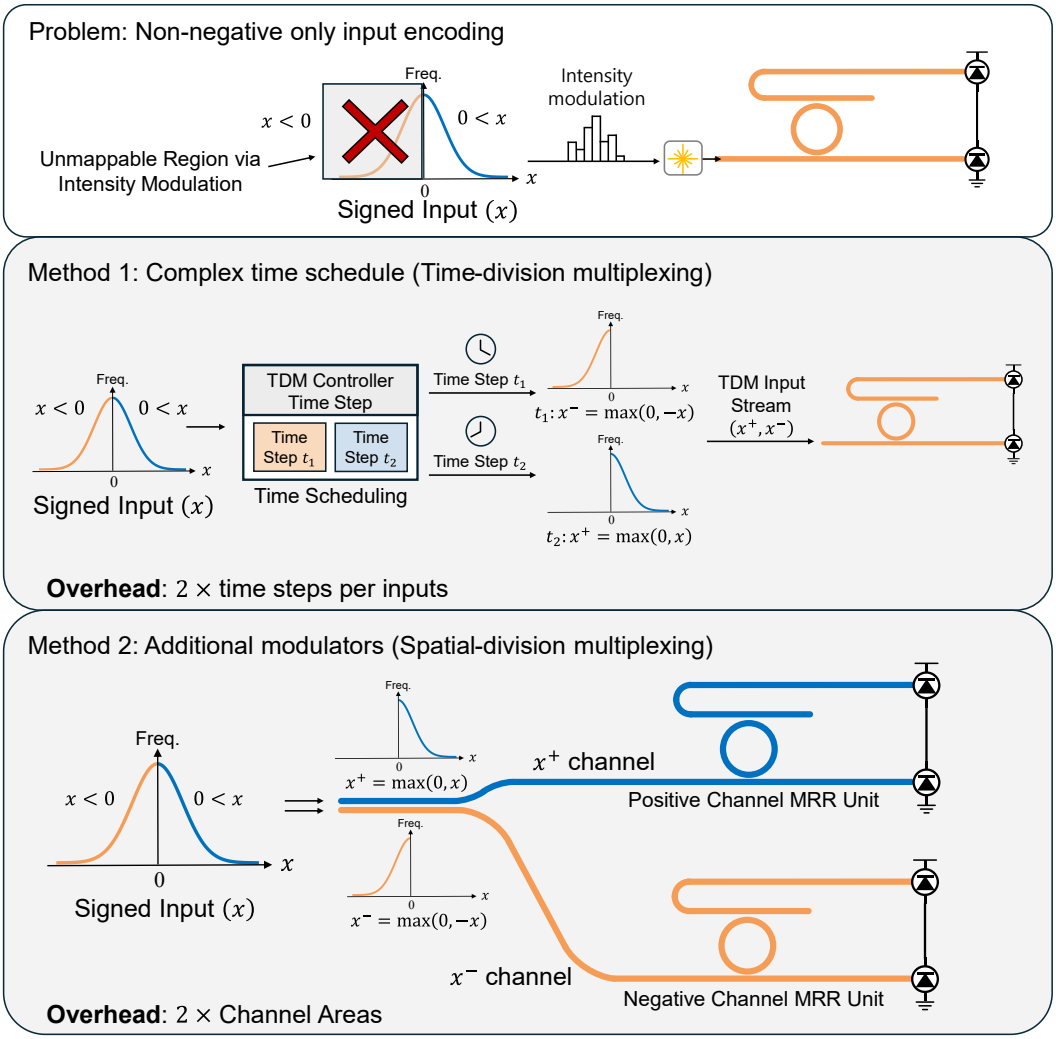


Figure 3.1: Signed-input workarounds for conventional MR weightbanks. TDM splits positive and negative contributions across time steps, whereas SDM duplicates the optical path into separate positive and negative channels, leading to throughput or area overhead.

The proposed architecture addresses this gap by extending the differential encoding principle that is already useful for weight encoding to the input path as well. The design objective is to realize signed computation without relying on SDM-style full optical-path duplication or complex TDM scheduling.

3.2. Bidirectional Weightbank Concept

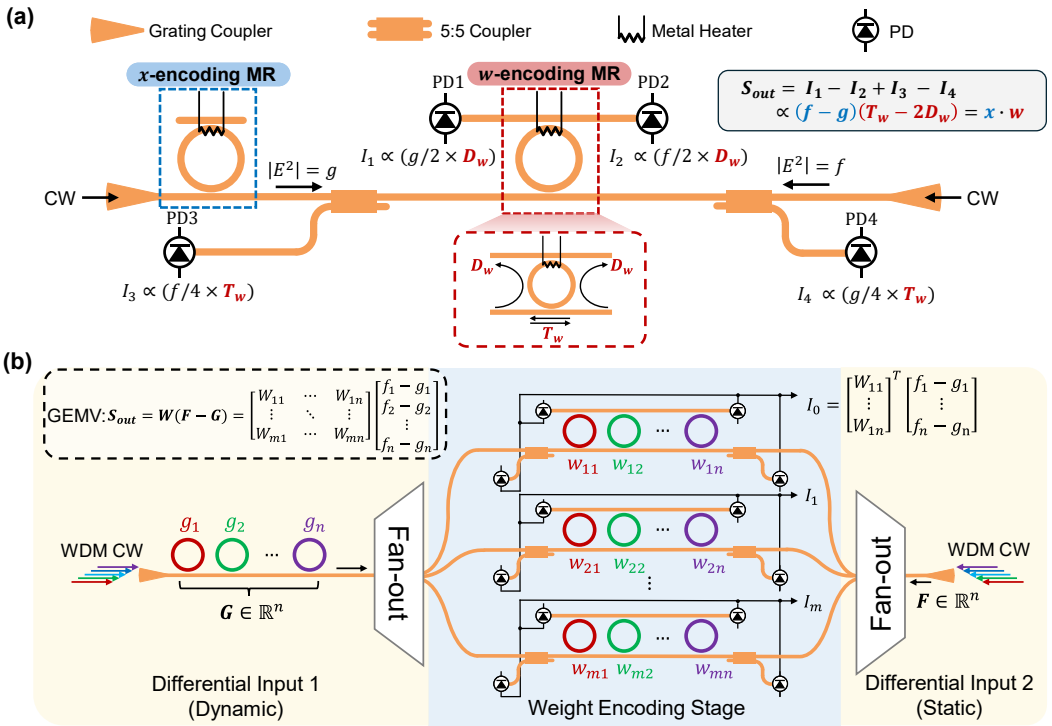


Figure 3.2: Proposed bidirectional MR weightbank. (a) Unit-cell structure for signed multiplication using the input term $(f-g)$ and the w response $(T_w - 2D_w)$. (b) Scaled architecture for multi-channel tensor core operation.

The core idea is to encode the input x by the imbalance between a fixed CW reference intensity and a counter-propagating intensity tuned by the x-encoding MR, and to encode the weight w by the state of the w-encoding MR. Let f denote the constant CW intensity injected from one direction, and let g denote the opposite-direction intensity dynamically tuned by the x-encoding MR. Let T_w and D_w denote the through and drop transmissions of the w-encoding MR. The numerical input and weight are then represented as

$$x = \beta_x \, (f - g) \,, \qquad w = \beta_w \, (T_w - 2D_w) \,, \tag{3.1}$$

where β_x and β_w are scaling factors. If one keeps the fixed readout normalization explicit, the same structure may also be interpreted as $x = \frac{1}{2}(f-g)$ with the factor $\frac{1}{2}$ absorbed into β_x or the readout gain.

In the unit structure, the four photodiodes produce branch signals proportional to

$$I_1 = \frac{g}{2} D_w, \qquad I_2 = \frac{f}{2} D_w, \qquad I_3 = \frac{f}{4} T_w, \qquad I_4 = \frac{g}{4} T_w. \tag{3.2}$$

After appropriate electrical recombination of these four branches, the output signal becomes

$$S_{\text{out}} = \alpha \, (f - g) \, (T_w - 2D_w) = \frac{\alpha}{\beta_x \beta_w} xw, \tag{3.3}$$

where $\alpha = \frac{1}{2}$ under the explicit unit-normalization convention and can be absorbed into the overall gain. Therefore, signed multiplication is realized by combining the input term $(f - g)$ with the w response $(T_w - 2D_w)$.

Figure 3.2 shows the proposed unit cell and scale-up architecture. At the unit level, a CW laser provides the fixed reference intensity f, the x-encoding MR dynamically tunes g, and the w-encoding MR sets T_w and D_w. The four PD branches are then merged electrically so that the cell directly produces the signed product term. At scale, many unit cells are tiled into wavelength/channel-parallel arrays while preserving the same signed arithmetic semantics.

Compared with an SDM-style additional-modulator implementation, the bidirectional approach reduces repeated photonic paths. More importantly, it reduces the number of independently tuned elements required to represent the same signed operation, which directly improves calibration scalability.

3.3. Experimental Setup for 1-Channel Validation

Figure 3.3 presents the 1-channel unit-cell experiment used to validate the proposed principle. Heater tuning is applied to the x-encoding MR that tunes g and to the w-encoding MR, while the counter-propagating optical signals and the through/drop responses of the w-encoding MR are monitored. In this section, f, g, T_w, and D_w follow the same definitions introduced above.

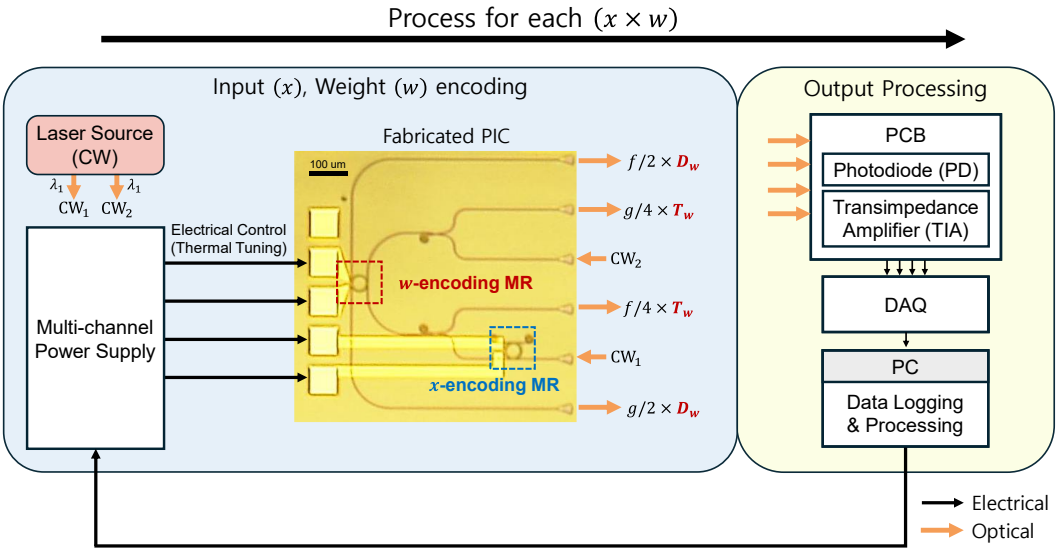


Figure 3.3: One-channel experimental setup for the proposed unit cell. Heater tuning controls the x-encoding MR that tunes g and the state of the w-encoding MR, with the fixed reference f, tuned opposite-direction intensity g, and through/drop responses used for calibration.

To verify the proposed operation, we measured the heater-dependent responses of the x-encoding MR and w-encoding MR, as well as their combined two-dimensional behavior, so that the signed-input and signed-weight mappings could be experimentally confirmed.

3.4. Heater Sweep Characterization

Figure 3.4 reports heater sweep responses of the x-encoding MR and the w-encoding MR under heater tuning. The central objective is to identify operating windows where the input term $f - g$ and the w term $T_w - 2D_w$ are smooth, monotonic, and sufficiently linear after LUT correction.

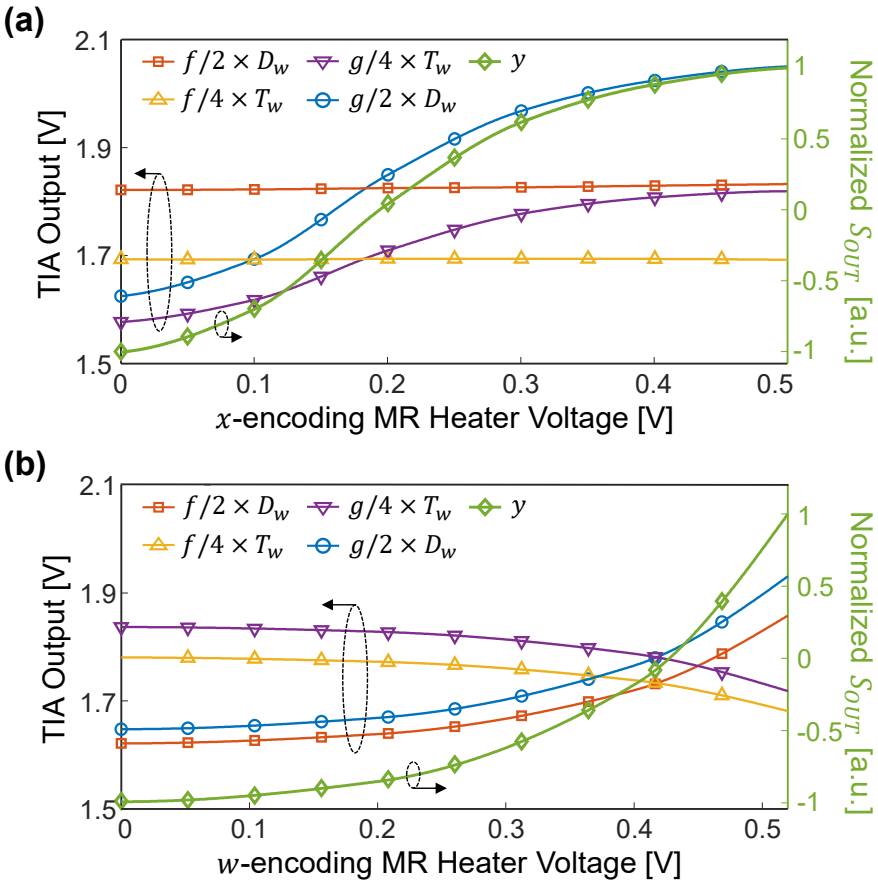


Figure 3.4: Measured transfer behavior from sweeping the heater controls of the proposed unit cell. (a) x-encoding MR heater sweep. (b) w-encoding MR heater sweep.

The heater sweeps also reveal several practical non-idealities that must be identified and calibrated before reliable operation:

- **Non-uniform slope:** sensitivity changes across heater range, implying resolution is not uniform in numerical space. In the x-encoding MR sweep, the steep transition appears in the lower-voltage region, whereas in the w-encoding MR sweep it appears in the higher-voltage region. This asymmetry means that, under finite heater-control resolution, the input x can be distributed more evenly over its usable range than the weight w, whose effective resolution is more compressed near the high-slope high-voltage region.

- **Readout asymmetry from grating-coupler imbalance:** ideally, the heater setting at which $\frac{g}{2}D_w = \frac{f}{2}D_w$ should also satisfy $\frac{g}{4}T_w = \frac{f}{4}T_w$, because both conditions correspond to the same optical balance point. The measured mismatch indicates unequal coupling strengths in the grating-coupler-monitored branches, which is a practical readout imperfection that must be compensated before final LUT construction.

- **Thermal crosstalk:** the effective input term and the w response can shift with the bias point of the other control because heating one MR perturbs the thermal state of the neighboring MR. This shows that one-dimensional calibration is insufficient and directly motivates the later thermal-domain correction and two-dimensional LUT validation.

Therefore, calibration cannot rely only on the nominal equations $x = \beta_x(f-g)$ and $w = \beta_w(T_w - 2D_w)$, because those relations assume balanced branches and negligible thermal coupling. The heater sweeps already reveal the main non-idealities—slope asymmetry, grating-coupler imbalance, and thermal crosstalk—so compensation and two-dimensional LUT validation are required, as shown later in Figures 3.5 and 3.6. In practice, the architecture is calibrated from measured transfer surfaces with branch-balancing correction and interpolation.

Based on these primitive measurements, the full 1-channel experiment in this section proceeds in five stages:

1. **Heater-sweep characterization:** first, the heaters of the x-encoding MR and w-encoding MR are swept to obtain the voltage-domain responses used for the initial x and w mappings, as shown in Figure 3.4.

2. **Grating-coupler imbalance compensation:** the experiment is measured through grating couplers rather than on-chip photodiodes, so the four sub-signals that form S_{out} do not experience identical coupling efficiency. In this compensation step, the monitoring curve is obtained by sweeping the x-encoding MR heater voltage and tracking the branch responses. Ideally, the heater voltage at which $\frac{g}{2}D_w = \frac{f}{2}D_w$ should coincide with the voltage at which $\frac{g}{4}T_w = \frac{f}{4}T_w$, but a residual coupling mismatch shifts the balance point of one monitored branch relative to the other. To compensate this readout asymmetry, the mismatched monitored branch is digitally rescaled by a constant factor of 1.1645 before forming the signed output.

3. **Voltage-domain LUT mapping:** after coupler compensation, the measured sweeps are inverted to construct initial LUTs between heater-voltage commands (V_x, V_w) and the target signed operands (x, w).

4. **Thermal-crosstalk modeling and thermal-domain LUT conversion:** a subtle but repeatable thermal coupling between the x-encoding MR and w-encoding MR is modeled before final biasing. Because the heater-generated thermal power is proportional to voltage squared and

each MR experiences a linear combination of both heaters, the thermal state is written as

$$\begin{bmatrix} T_x \\ T_w \end{bmatrix} = \begin{bmatrix} 1 & M_{12} \\ M_{21} & 1 \end{bmatrix} \begin{bmatrix} V_x^2 \\ V_w^2 \end{bmatrix}. \tag{3.4}$$

The coefficients are identified by sweeping M_{12} and M_{21} to minimize symmetry-error objectives over the four corner states $(x, w) \in \{-1, +1\}^2$, yielding $M_{12} = 0.018$ and $M_{21} = 0.008$. The resulting voltage-domain LUTs are then converted into one-to-one thermal-domain LUTs $T_x \leftrightarrow x$ and $T_w \leftrightarrow w$, which are more convenient for final runtime lookup under crosstalk.

5. **LUT-sweep validation:** at runtime, a target pair (x, w) is first mapped to (T_x, T_w) through the thermal-domain LUTs, then converted back to heater commands through

$$\begin{bmatrix} V_x^2 \\ V_w^2 \end{bmatrix} = \begin{bmatrix} 1 & M_{12} \\ M_{21} & 1 \end{bmatrix}^{-1} \begin{bmatrix} T_x \\ T_w \end{bmatrix}, \tag{3.5}$$

after which (V_x, V_w) is applied to the chip and the compensated output is evaluated through the final LUT sweep reported later in this section.

This flow clarifies why the uncompensated two-dimensional sweep shows a small asymmetry and a non-zero response near the nominal $w \times 0 = 0$ operating point: those residual offsets originate from thermal crosstalk and motivate the thermal-domain correction described above.

3.5. 2D LUT Calibration and Validation

After heater sweeps and grating-coupler compensation, Figure 3.5 shows the two-dimensional LUT sweeps measured before applying the thermal compensation. The two sweep orders reveal a slight asymmetry caused by uncompensated thermal crosstalk. The mismatch is especially visible near the zero region, where thermal crosstalk causes a small residual response even when the target condition is $w \times 0 = 0$. Thermal-domain correction was introduced to achieve more accurate calibration.

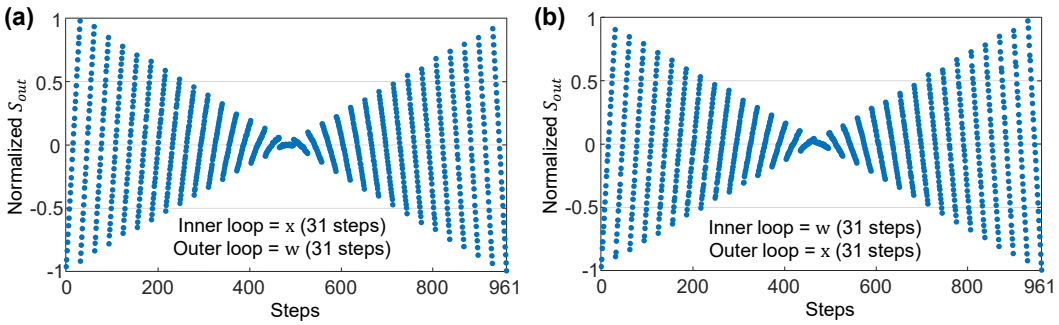


Figure 3.5: Two-dimensional LUT sweeps before thermal compensation for the proposed bidirectional weightbank. The two ordered scans use different inner/outer loop orders and show the asymmetry caused by uncompensated thermal crosstalk.

After applying the thermal compensation and using the final thermal-domain LUTs, Figure 3.6 presents the final validation results. They include compensated two-dimensional sweeps, randomized LUT queries, and the corresponding error histograms.

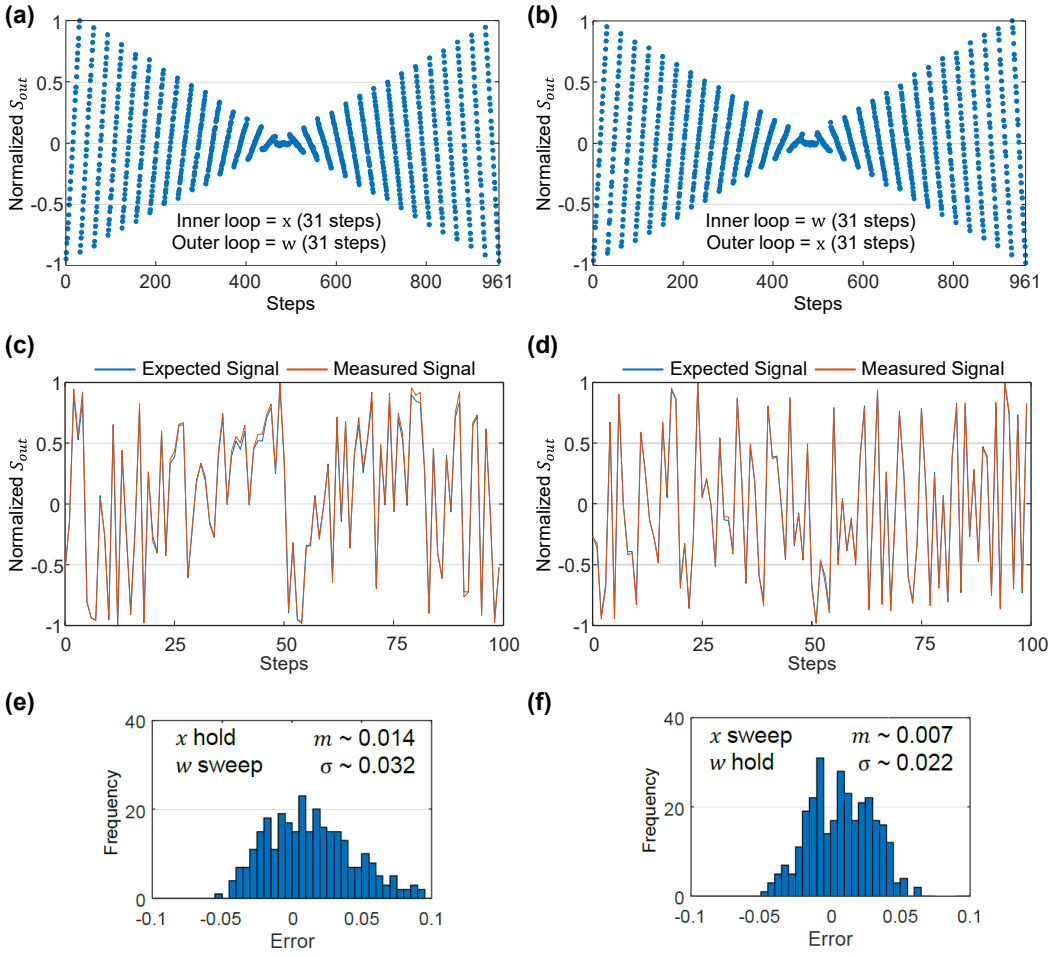


Figure 3.6: Final LUT-validation results after thermal compensation for the proposed bidirectional weightbank. (a) and (b) show compensated two-dimensional sequential LUT sweeps over x and w. (c) and (d) show random-sweep validation results for w and x, respectively. (e) and (f) show histograms of the output error for the random sweeps of w and x, respectively. In (e) and (f), the reported standard deviations were obtained from 300 repeated random trials.

The sweep-order comparison is useful because thermal systems may show path dependence. In the compensated results, close agreement between the ordered sweeps indicates that the calibrated bidirectional mapping remains stable under different scan orders. Randomized sweeps then test whether the LUT remains accurate for non-sequential commands, and the error histograms summarize the residual deviation from the target output.

A practical error metric is

$$e = S_{\text{out,meas}} - S_{\text{out,target}}, \tag{3.6}$$

which is evaluated over the measured random-sweep trials. The histogram of e reflects two main error categories: systematic errors and random residuals. Systematic errors arise from repeatable effects such as residual thermal crosstalk, finite resolution of the heater voltage, grating-coupler imbalance, and photodiode responsivity mismatch among the four branch signals. Random residuals mainly arise from input optical power fluctuation, heater voltage noise, and photodiode readout noise.

The measured distributions are not perfectly Gaussian, and their residual skew or tail components imply that small systematic errors remain after compensation. One example is the difference between the x-sweep and w-sweep histograms in Figures 3.6(e) and 3.6(f), which can be understood from the heater sweep characteristics in Figure 3.4. For the x-encoding MR, the low voltage region lies near resonance, where the transfer slope is steep. For the w-encoding MR, however, the steep slope region appears at higher heater voltage. Because thermally induced detuning is approximately proportional to heater power, and therefore to V^2, a fixed voltage step produces a smaller change in detuning at low voltage than at high voltage. It is therefore preferable to place the steep transfer slope in a lower heater-voltage region, as in the x-sweep case. This operating point difference explains why the x-sweep validation shows a narrower and more accurate error distribution than the w-sweep validation.

The key architectural point is that the proposed signed scheme based on $S_{\text{out}} \propto (f-g)(T_w-2D_w)$ is calibratable. A signed architecture is not useful if control overhead makes it impractical. These results support the claim that the proposed bidirectional structure can be stabilized with realistic measurement and firmware logic.

3.6. Scaling to Multi-Channel Tensor Core Layout

After validating the unit cell, the next step is physical scaling. Figure 3.7 shows the scaled four-channel bidirectional MR tensor core layout. This layout represents the transition from device-level feasibility to array-level architecture.

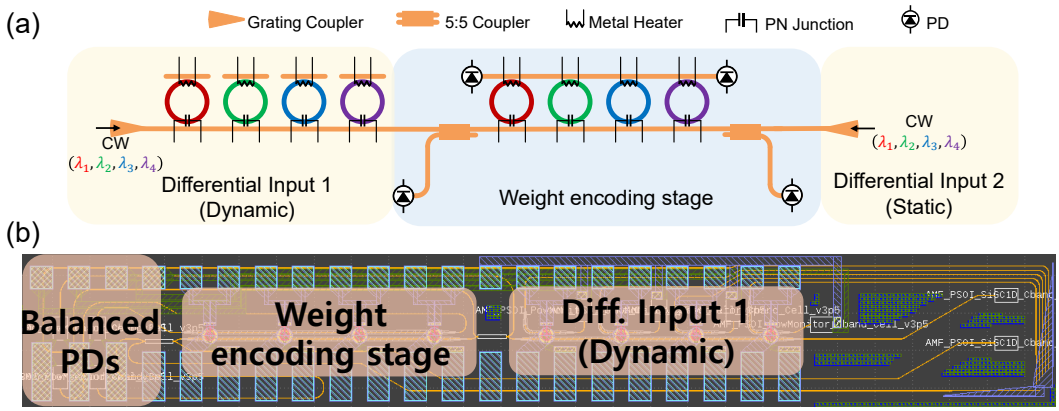


Figure 3.7: Scaled four-channel bidirectional MR tensor core layout. (a) Schematic of the layout. (b) Designed layout, where the four photodiodes for balanced readout are placed in one section and merged into one electrical pad.

Scaling introduces three practical layout issues:

- skew between the four branch signals,
- reflections at terminations,
- placement of the balanced PDs for merging the four branch signals into one electrical pad.

These issues are addressed directly in the layout. First, the skew problem is mitigated by balancing the waveguide lengths to the four PDs. Second, reflections are reduced by using long, gently tapered Ge absorbers at the terminations. Third, the four PDs are placed in one section and merged as shown in Figure 3.7(b).

At the device level, the core was designed around a target MR operating speed of 1 GHz. The readout used 15 GHz photodiodes provided by the AMF PDK. Within this speed target, the input and weight stages follow different optimization principles. For the w-encoding MRs, the design objective is to maximize the variation in the normalized w response, $w \propto T_w - 2D_w$. For the x-encoding MRs, the design objective is to maximize the x response, $x \propto f - g$, where f is a constant reference intensity and g follows the MR through transfer curve. In this case, maximizing the x response is equivalent to maximizing the optical modulation amplitude (OMA) of the through response at the target speed.

3.7. Complexity and Efficiency Discussion

To compare conventional and bidirectional architectures, consider a simplified metric set. The key difference is that the proposed signed MAC structure uses only one x-encoding MR and one w-

encoding MR to perform a signed multiplication. In many conventional signed architectures, positive and negative contributions must be separated by duplicated optical paths or extra encoding elements, so the same signed MAC operation effectively requires more tunable photonic devices. Let N be channel count and let k_{dup} denote path duplication factor for signed support. In a duplicated conventional signed scheme, effective tunable elements may scale as $O(2N)$. In the proposed bidirectional scheme, signed MAC operation is realized with one MR for x and one MR for w, so tunable optical elements can remain closer to $O(N)$ plus modest differential readout logic.

While constants matter in real hardware, this scaling intuition is important. Reduced optical duplication affects:

- required source power,
- number of control channels,
- calibration complexity,
- probability of drift-induced misalignment events.

3.8. Architectural Significance for Neuromorphic Workloads

The proposed bidirectional weightbank is significant because it addresses the main input-side sign bottleneck in otherwise non-negative MR datapaths. Many neural workloads rely on signed activations and signed partial sums after normalization and residual operations, so even systems that can encode signed-weight values still need efficient signed-input handling. If signed-input handling is inefficient, overall accelerator advantage weakens.

By embedding signed representation directly in a compact bidirectional differential structure, the architecture improves effective density for practical workloads. This can increase usable MAC throughput per unit area and per unit calibration effort, especially when scaling from proof-of-concept channels to larger tensor core arrays.

In addition, this architectural direction should be understood in the broader context of signed-input photonic MAC design. Prior signed-input MAC implementations have mostly relied on coherent optical schemes that use the phase of light to represent sign. Although effective, coherent schemes must explicitly control phase, which typically requires additional heaters or other phase-tuning elements and makes the system vulnerable to phase noise as well as intensity noise. By contrast, the proposed bidirectional design resolves the signed-input problem within an intensity-modulation framework, avoiding the need to encode sign through optical phase while preserving the benefits of differential computation.

3.9. Limitations and Open Questions

Although the experimental results are encouraging, the present calibration procedure still leaves several open questions.

First, the present experimental verification was performed only for the single-product case $x_1 w_1$. How the calibration should be extended to multi-channel monitored accumulation, i.e., $\sum_i x_i w_i$, remains unknown. Second, because the architecture is bidirectional, it is vulnerable to optical reflections. For cases beyond low-level reflection, a reliable calibration methodology has not yet been established.

These limitations should be addressed in future work before the architecture is extended to larger and more practical systems.

4. Proposed GEMM Tensor Core Architecture with RAMZMs

4.1. Choosing a GEMM-Friendly Multiplication Strategy

Before discussing array-level organization, it is useful to compare the two optical multiplication strategies shown in Figure 4.1. Method 1 cascades composite functions along a single optical path. In Method 1, light propagates in a single direction, and the implementation of the operator and operand is performed simultaneously along the same path. This makes it straightforward to realize multiple scalar encodings and form vector–vector multiplication, so the structure is naturally GEMV-friendly. To implement GEMM with this scheme, however, all elements of the encoded vector must ultimately be demultiplexed and redistributed as shown in Figure 4.2 (d) and (e). Method 2 instead performs multiplication between two separately encoded optical signals at the photodetectors. Because the two operands remain on separate waveguide channels until readout, this structure is much more convenient for arranging row–column interactions in a two-dimensional multiplication array. For that reason, Method 2 is more naturally GEMM-friendly, and it is the configuration targeted in this work.

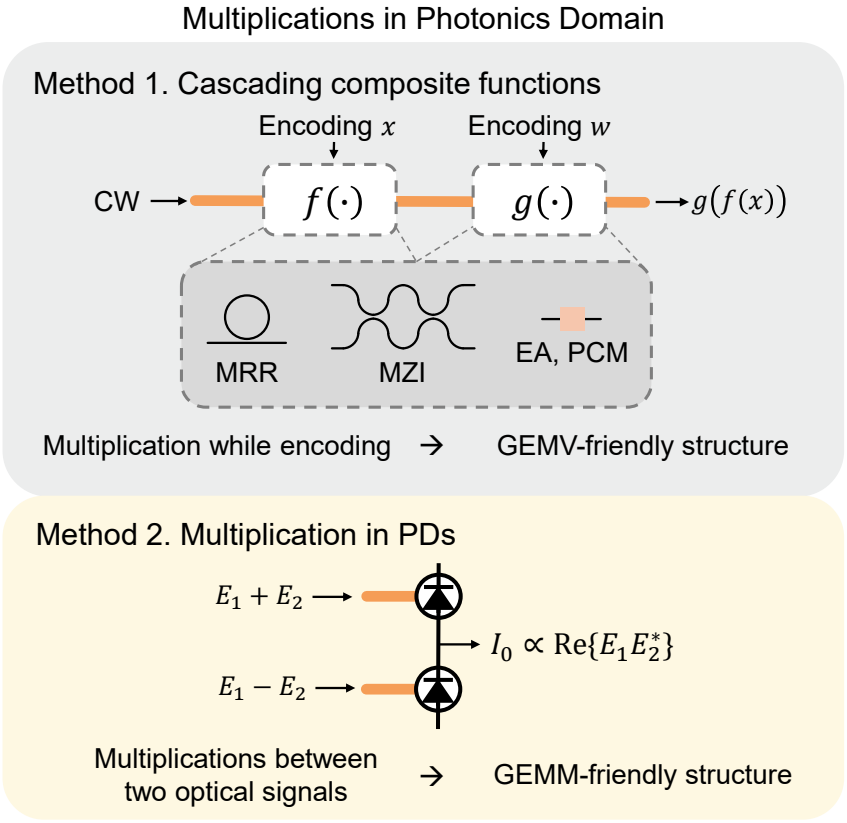


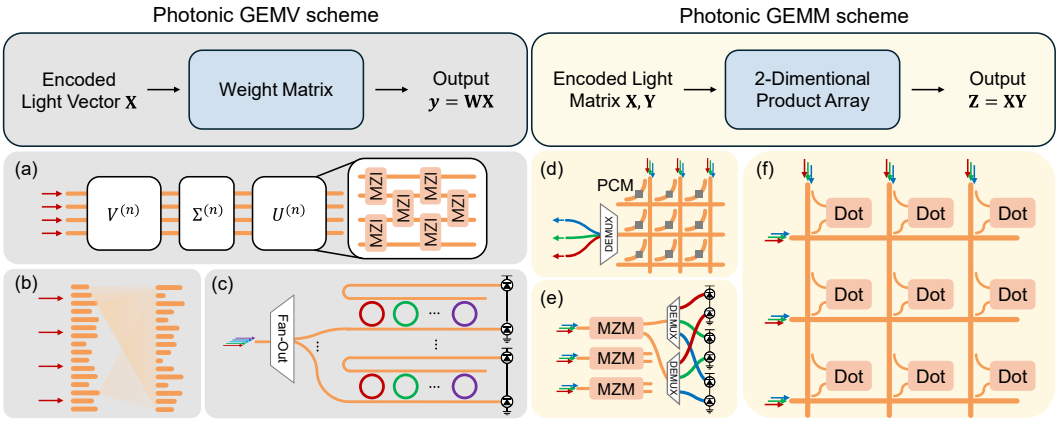
Figure 4.1: Comparison of two optical multiplication strategies in the linear photonics domain. Method 1 performs multiplication while encoding along a single optical path and is therefore GEMV-friendly, whereas Method 2 performs multiplication between two separately encoded optical signals at balanced photodetectors and is therefore more suitable for two-dimensional GEMM arrays.

For Method 2, let the two encoded optical fields be E_1 and E_2. One detector reads the constructive-interference output $|E_1 + E_2|^2$, and the other reads the destructive-interference output $|E_1 - E_2|^2$. Subtracting these two detector signals removes the individual intensity terms and leaves only the interaction term between the two fields:

$$I_{\text{out}} \propto |E_1 + E_2|^2 - |E_1 - E_2|^2 = 4\,\mathrm{Re}(E_1 E_2^*). \tag{4.1}$$

Thus, after proper biasing and differential readout, the photodetector pair directly produces a signal proportional to the multiplication of the two encoded optical fields. This comparison also explains why the remainder of this section focuses on GEMM-native organization rather than repeated GEMV decomposition. Figure 4.2 then places this choice in the broader landscape of photonic linear-algebra

architectures. On the GEMV side, Figure 4.2(a) shows an SVD-based structure, Figure 4.2(b) shows a diffractive neural-network structure, and Figure 4.2(c) shows an MR-based weightbank. All three are naturally suited to vector-to-matrix style processing. On the GEMM side, Figure 4.2(d) shows a PCM-based crossbar that encodes multiple wavelengths at once and then demultiplexes them, and Figure 4.2(e) shows a similar organization in which multiple wavelengths are modulated first and then demultiplexed. By contrast, Figure 4.2(f) shows the target crossbar organization of this work, where two optical paths are injected directly into one multiplication-product unit. Because this structure does not rely on a large demultiplexing backend for every product location, it is more convenient for constructing a two-dimensional crossbar array. The key message is that architecture-level dataflow must match workload-level arithmetic patterns: repeatedly decomposing GEMM into GEMV passes forces extra reconfiguration and partial-sum movement, whereas the structure in Figure 4.2(f) is more natural for direct photonic GEMM.



	Photonic GEMV (a, b, c)	Photonic GEMM (d, e, f)
Throughput	O(N^2)	O(N^3)
Complexity	Medium	High
Data Reuse	Limited (All-to-one interaction)	High (All-to-all interaction)

Figure 4.2: Representative photonic GEMV and GEMM organizations. (a)–(c) show SVD-based, diffractive-neural-network, and MR-weightbank GEMV structures, while (d) and (e) show multi-wavelength GEMM organizations that require backend demultiplexing. (f) shows the targeted crossbar dot-product array, where two optical paths enter one multiplication-product unit directly for scalable two-dimensional GEMM.

This section therefore moves from the underlying optical multiplication scheme to the proposed RAMZM crossbar array, then to the operating principle of the proposed architecture, stable operand

encoding, single-path experimental validation, and finally to core layout.

4.2. Proposed RAMZM Crossbar Array Structure

Having selected Method 2 as the multiplication strategy, the next step is to organize the multiplication units into a GEMM-native array. Figure 4.3 shows the proposed 4×4 RAMZM-based GEMM core. The four horizontal channels carry row operands of one matrix and the four vertical channels carry column operands of the other matrix, so each crosspoint acts as a multiplication unit with a two-heater local photonic dot-product cell. For a 4×4 matrix multiplication, the architecture uses time-division multiplexing (TDM) over the four terms of each inner product. At cycle $t \in \{1, 2, 3, 4\}$, the row inputs encode $X_{m,t}$ and the column inputs encode $Y_{t,n}$, so the (m, n) crosspoint produces the partial product for that cycle. After four cycles, the detected outputs are accumulated to obtain

$$\mathbf{C} = \mathbf{XY}, \qquad C_{m,n} = \sum_{t=1}^{4} X_{m,t} Y_{t,n}. \tag{4.2}$$

Thus, the full 4×4 GEMM is completed after four TDM cycles while the physical core remains a regular two-dimensional crossbar.

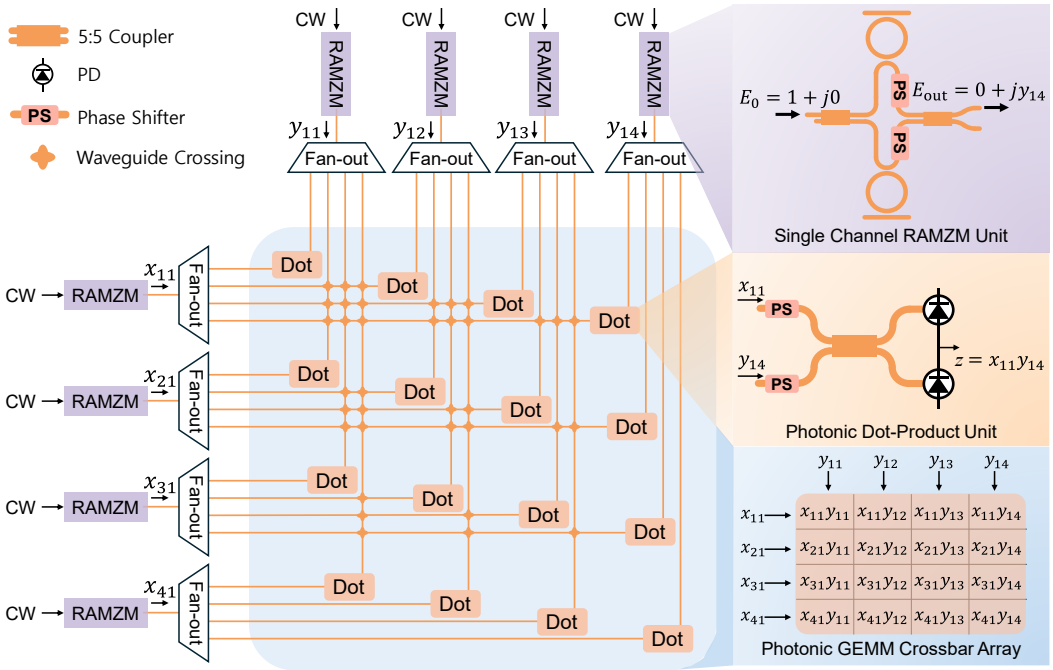


Figure 4.3: Proposed 4×4 RAMZM-based GEMM core operated with single-wavelength time-division multiplexing. Each crosspoint contains a local photonic dot-product unit with two heaters. At cycle t, row inputs encode $X_{m,t}$ and column inputs encode $Y_{t,n}$; after four cycles and accumulation, the full output matrix is obtained.

The intent of this structure is to place the independently encoded elements of the two matrices to be multiplied—row-wise operands from one matrix and column-wise operands from the other—on a regular crossbar so that each crosspoint acts as a reusable multiplication unit, now drawn with two local heaters in the dot-product cell. In this way, the array directly follows the row–column organization of GEMM and provides a clear architectural template for tiled photonic matrix–matrix multiplication.

4.2.1. Operating Principle of the Proposed Architecture

The next step is to explain how the proposed RAMZM architecture uses complex-plane trajectories for complex-domain encoding and photonic multiplication. Figures 4.4 and 4.5 summarize this principle through the MR trajectory, the RAMZM trajectory, and the resulting dot-product operation.

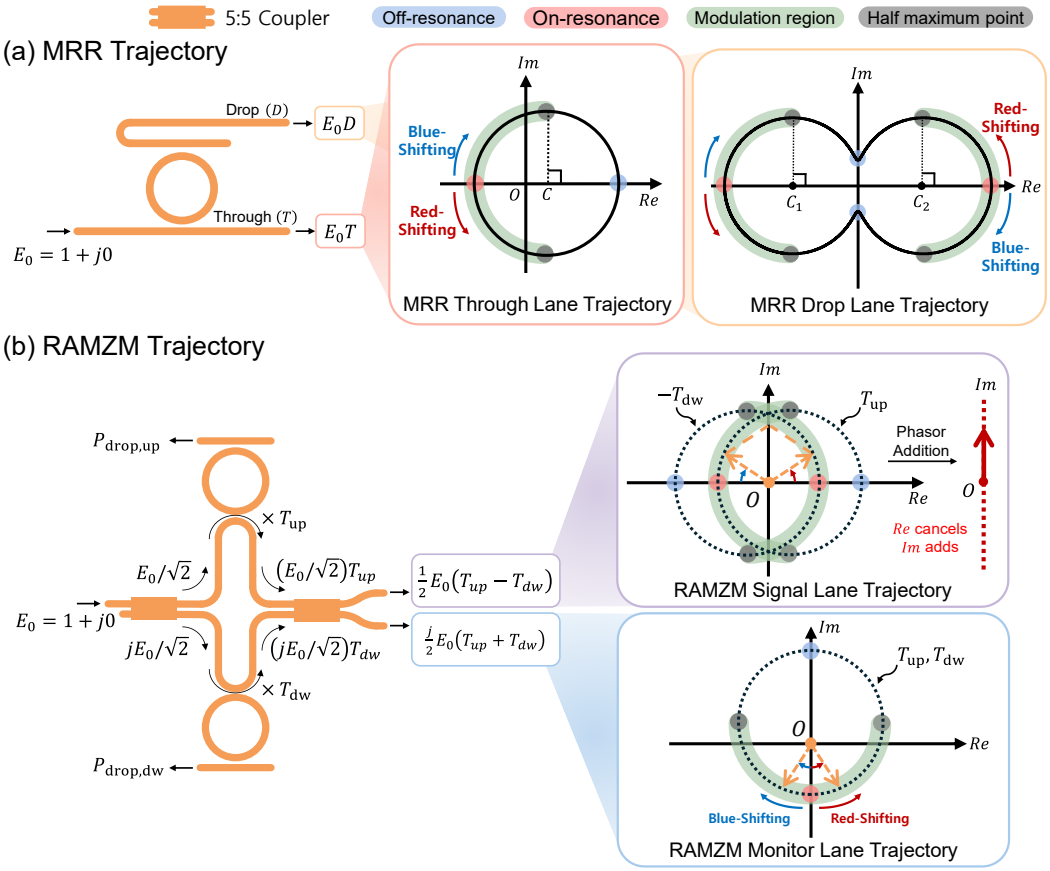


Figure 4.4: Complex-plane trajectories used in the proposed RAMZM-based GEMM architecture. In (a), the through-lane trajectory $T(\phi)$ and drop-lane trajectory $D(\phi)$ are drawn from the MR transfer equations in Eq. (4.3). Here, on-resonance denotes the state where the input wavelength matches the ring resonance, off-resonance denotes the detuned state, the half-maximum points denote the two states where the drop-lane power is one half of its on-resonance maximum, and the modulation region denotes the selected trajectory segment used for encoding. In (b), the RAMZM trajectories are shown for two output lanes: the signal lane, which is routed to the following dot-product unit, and the monitor lane, which is used for monitoring.

Figure 4.4(a) shows the complex-plane trajectories of a conventional add-drop MR. As the round-trip phase detuning ϕ is swept, the complex through-lane and drop-lane responses can be written as [17, 18]

$$T(\phi) = \frac{\gamma\left(1 - \alpha e^{-j\phi}\right)}{1 - \alpha\gamma^2 e^{-j\phi}}, \qquad D(\phi) = \frac{-\kappa^2\sqrt{\alpha}\, e^{-j\phi/2}}{1 - \alpha\gamma^2 e^{-j\phi}}, \tag{4.3}$$

where α is the round-trip field amplitude, γ is the through coefficient, κ is the coupling coefficient, and $\gamma^2 + \kappa^2 = 1$. Accordingly, Figure 4.4(a) shows the trajectories traced by $T(\phi)$ and $D(\phi)$ on the complex plane when α, γ, and κ are fixed. The through-lane trajectory $T(\phi)$ is ideally a circle whose center lies on the real axis. In the lossless case $\alpha = 1$, this trajectory passes through the point 1 on the real axis, and its center and radius are determined by α, γ, and κ. As shown in Figure 4.4(a), when the MR is tuned to the resonance state for the input wavelength, the through-lane response lies on the real axis, and the corresponding through-lane power, proportional to $|T(\phi)|^2$ or equivalently to the square of the distance from the origin on the complex plane, becomes minimum. With the trajectory convention used in the figure, red shifting the MR moves the response clockwise along the trajectory, whereas blue shifting moves it in the opposite direction. The half-maximum points are defined as the two points on the trajectory where the drop-lane power is one half of its on-resonance maximum. In Figure 4.4(a), these are the two labeled points on the trajectory whose connecting line passes through the center of the trajectory. At the half-maximum points, the absolute value of the imaginary component is maximized; therefore, the upper half-maximum state and lower half-maximum state naturally define the $+1$ and -1 encoding limits, respectively. The modulation region is the selected segment of the trajectory between the half-maximum points that is used for operand encoding, so that it spans all usable imaginary values on the trajectory.

Figure 4.4(b) shows the corresponding RAMZM behavior on the same complex plane. In the proposed structure, the RAMZM contains two MRs placed in the two arms of the MZI and provides two output ports, denoted as the signal lane and the monitor lane. The arm phase is tuned so that the two optical fields arrive with opposite phase at the signal lane, which means that the MR trajectories embedded in the two arms are oppositely aligned there, as shown in Figure 4.4(b). The two MRs are then driven in different directions, with one red shifting and the other blue shifting. Because the two arm phasors move symmetrically in opposite directions, the real components cancel at the signal lane and only the imaginary components remain. As a result, the output at the signal lane follows an approximately straight trajectory on the imaginary axis, as shown in Figure 4.4(b). In this operating mode, the output phase is restricted near $+90^\circ$ or $+270^\circ$, while the magnitude can be varied largely independently of the phase; this is the low-frequency chirp method used in the proposed structure. By contrast, at the monitor lane the two arm trajectories are overlapped rather than canceled.

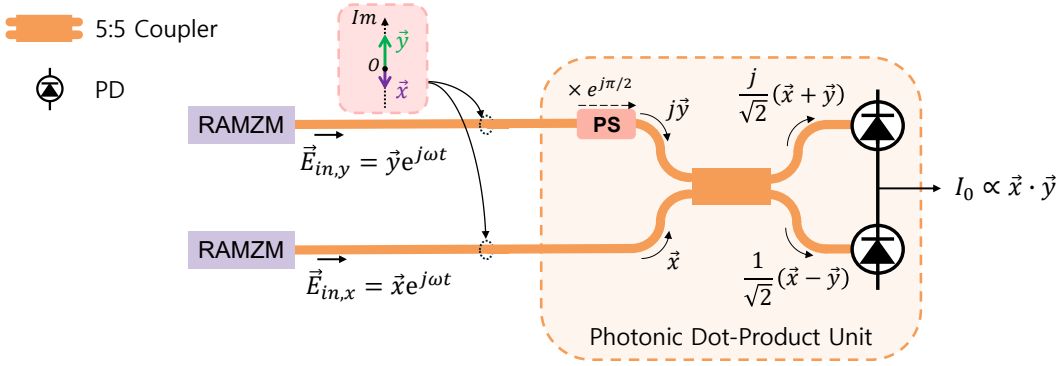


Figure 4.5: Dot-product extraction in the proposed RAMZM-based GEMM architecture. The local photonic dot-product unit includes two heaters, and two operands encoded as complex-domain phasors along the imaginary axis are combined through optical interference and balanced detection.

Figure 4.5 shows the dot-product process between two phasors encoded by the RAMZM in the complex plane. For a general derivation, let the two input phasors be

$$E_x = a + jx, \qquad E_y = b + jy,$$

where a and b are the real components and x and y are the imaginary components. In the ideal operating condition of the proposed signal lane, the phasors are aligned on the imaginary axis, so $a = b = 0$. After passing through the 5:5 coupler, the two detector-side fields become

$$E_+ = \frac{E_x + E_y}{\sqrt{2}}, \qquad E_- = \frac{E_x - E_y}{\sqrt{2}}.$$

The balanced photodetector (BPD) then reads the differential signal

$$I_{\mathrm{BPD}} \propto |E_+|^2 - |E_-|^2 = 2\,\mathrm{Re}(E_x E_y^*) = 2\,\mathrm{Re}[(a + jx)(b - jy)] = 2ab + 2xy.$$

Thus, in the general case the BPD output contains contributions from both the real-axis and imaginary-axis components of the two phasors. For the intended operating point used in this work, $a = b = 0$, so the differential output reduces to $I_{\mathrm{BPD}} \propto 2xy$. This shows how the 5:5 coupler and BPD produce an output proportional to the multiplication result of the two encoded phasors. Each crosspoint in the proposed crossbar therefore directly generates the multiplication result needed for matrix–matrix multiplication.

4.3. Stable Operand Encoding and LUT Validation

Once the crossbar topology in Figure 4.3 and the operating principle illustrated in Figures 4.4 and 4.5 are fixed, the next requirement is stable operand encoding within the usable RAMZM operating region. Figure 4.6 summarizes the push–pull microring behavior and the per-ring LUT construction used for RAMZM encoding. Figure 4.6(a) illustrates push–pull driving of a microring, which is typically employed in conventional RAMZM calibration with two microrings [22, 23]. As shown by the through-port field trajectory in Figure 4.6(b), the push and pull states can produce unbalanced phase shifts, so the through-port trajectory alone does not provide a sufficiently direct calibration coordinate. Instead, the drop-port field trajectory is used to extract an angular coordinate θ, and the resulting drop-port power relation is used to build a LUT for each microring. The upper and lower microrings in the RAMZM are then calibrated independently so that the combined RAMZM output follows the intended imaginary-axis field trajectory.

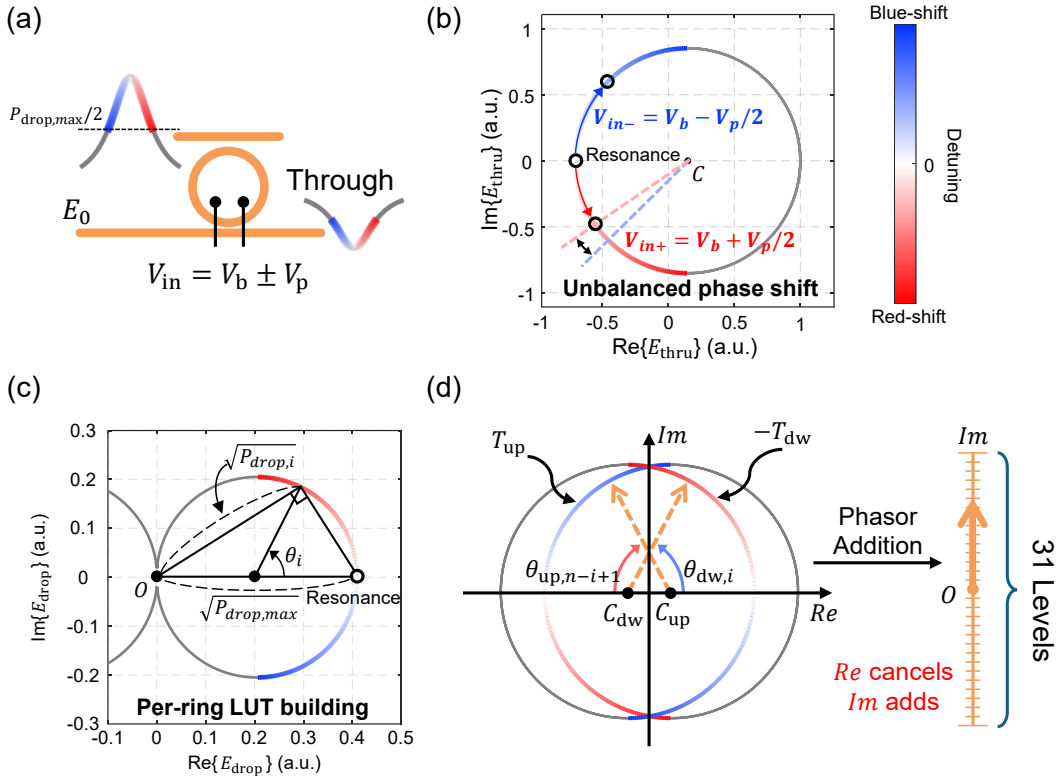


Figure 4.6: Stable operand encoding and per-ring LUT construction for RAMZM operation. (a) Schematic of a microring driven in push–pull mode. (b) Through-port field trajectory of the microring in push–pull mode, showing unbalanced phase shifts between the push and pull states. (c) Drop-port field trajectory of the microring used to extract the angular coordinate θ and build a per-ring LUT. (d) RAMZM schematic and output field trajectory with independently calibrated upper and lower microrings.

Let $u \in [-1, 1]$ denote the target normalized imaginary value on the through-port field trajectory, so that

$$u = \sin\theta. \tag{4.4}$$

On the drop-port field trajectory, let r be the circle radius. Because the drop-port power is the squared distance from the origin on the complex plane, the maximum drop-port power is

$$P_{\text{drop,max}} = (2r)^2, \qquad r = \frac{\sqrt{P_{\text{drop,max}}}}{2}. \tag{4.5}$$

From the chord geometry in Figure 4.6(c), the distance from the origin to the selected drop-port

phasor can be written directly as

$$\sqrt{P_{\text{drop}}} = 2r\cos\frac{\theta}{2}, \tag{4.6}$$

so that

$$\cos\frac{\theta}{2} = \frac{\sqrt{P_{\text{drop}}}}{2r} = \sqrt{\frac{P_{\text{drop}}}{P_{\text{drop,max}}}}. \tag{4.7}$$

Eliminating θ then gives the signed normalized through-port imaginary component as

$$u = \sin\theta = \pm 2\sqrt{\frac{P_{\text{drop}}}{P_{\text{drop,max}}}\left(1 - \frac{P_{\text{drop}}}{P_{\text{drop,max}}}\right)}. \tag{4.8}$$

Here, the + or − sign is selected according to whether the operating point is in the blue-shifted or red-shifted region from the resonance point, while the magnitude is determined by the drop-port power. Therefore, the LUT can be constructed by choosing P_{drop} so that u is linearly sampled over the target encoding range. For the resonance-connected solution used here, the inverse relation becomes

$$P_{\text{drop}}(u) = \frac{P_{\text{drop,max}}}{2}\left(1 + \sqrt{1-u^2}\right), \qquad u \in [-1, 1], \tag{4.9}$$

with the blue-shifted or red-shifted region from the resonance point providing the sign of u. This construction gives $P_{\text{drop}} = P_{\text{drop,max}}$ at $u = 0$ and $P_{\text{drop}} = P_{\text{drop,max}}/2$ at $u = \pm 1$, corresponding to the resonance point and the half-maximum points used as the encoding limits.

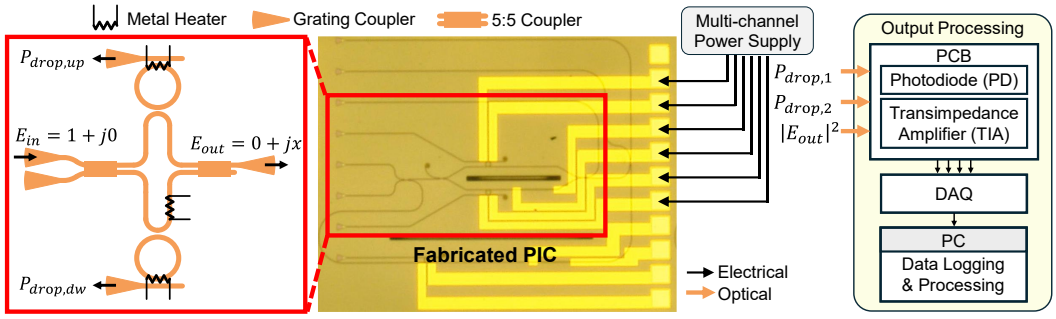


Figure 4.7: Experimental setup for RAMZM-only LUT validation.

The LUT is validated by directly measuring the output of a standalone RAMZM. In this verification, the RAMZM alone is programmed through the independently constructed per-ring LUTs, and its output response is observed directly without introducing a second optical operand or a differential multiplication readout. This experiment focuses on whether LUT-programmed RAMZM operation alone can generate the intended output response reproducibly. Figure 4.7 shows the experimental setup used for this RAMZM-only LUT validation. During LUT validation, the RAMZM output is

observed while the programmed LUT values are swept. This RAMZM-only measurement is used to verify the direct output behavior of the device before integrating it into the full multiplication architecture.

With this setup, the direct RAMZM-output validation results are summarized in Figure 4.8. Figure 4.8(a) first shows how the measured drop-port power of a single microring is converted into extracted imaginary-axis field values using Eq. (4.8). This result illustrates the LUT-mapping step itself: a drop-port power level is mapped to the corresponding normalized field value used for RAMZM programming. Figure 4.8(b) then compares the expected and measured RAMZM output during a sequential LUT sweep. The measured output follows the programmed order with an approximately monotonic response, confirming that the LUT obtained from the drop-port mapping can be applied reproducibly in hardware. Figure 4.8(c) compares the expected and measured RAMZM output under random LUT access. The output still tracks the commanded values under non-sequential programming, indicating that the RAMZM-only encoding remains meaningful even when the LUT entries are visited in arbitrary order. Figure 4.8(d) summarizes the RAMZM-output error distribution obtained from 300 random trials.

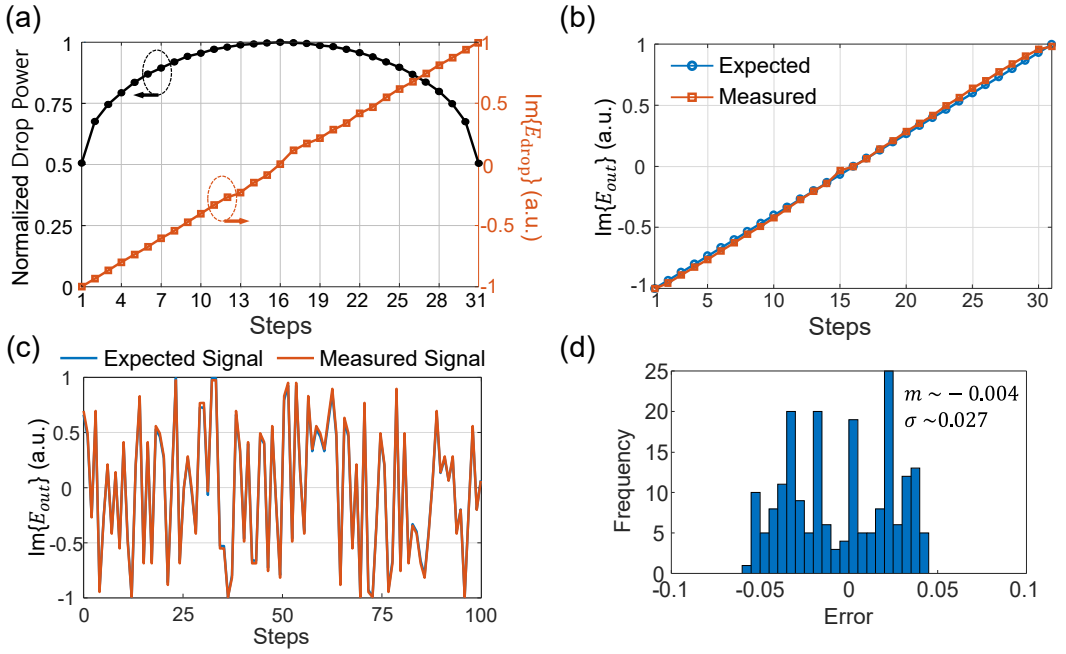


Figure 4.8: Integrated direct RAMZM-output validation results. (a) Drop-port power and extracted imaginary-axis field values of a microring used to construct a LUT. (b) Expected and measured RAMZM output during a sequential LUT sweep. (c) Expected and measured RAMZM output during random LUT access. (d) Error distribution of RAMZM output from 300 random trials.

The histogram contains both measurement noise and residual calibration error. If additive measurement noise were the dominant contribution, the error distribution would be expected to appear closer to a Gaussian shape. The observed distribution is not clearly Gaussian, which suggests that calibration-related residual errors are a significant part of the remaining output error. Possible sources include thermal crosstalk between heaters, LUT entries constructed from drop-port power values that already include measurement noise, and mismatch between the per-ring LUT calibration condition and the final RAMZM-output measurement condition. Taken together, these results show that the proposed LUT flow can control the RAMZM output, while also indicating that calibration accuracy and thermal stability remain important for reducing residual error.

4.4. 2×2 Matrix–Matrix Multiplication Core Layout

After validating the LUT-controlled RAMZM output flow, the next step is to arrange the multiplication cells into a small GEMM core. Figure 4.9 shows the designed layout of a 2×2 matrix–matrix multiplication core based on the proposed RAMZM strategy.

For the minimal case,

$$\mathbf{C} = \mathbf{XY}, \qquad C_{m,n} = \sum_{k=1}^{2} X_{m,k} Y_{k,n}. \tag{4.10}$$

Each output node corresponds to the partial contributions that are accumulated over two cycles to form one final element of $\mathbf{C}$.

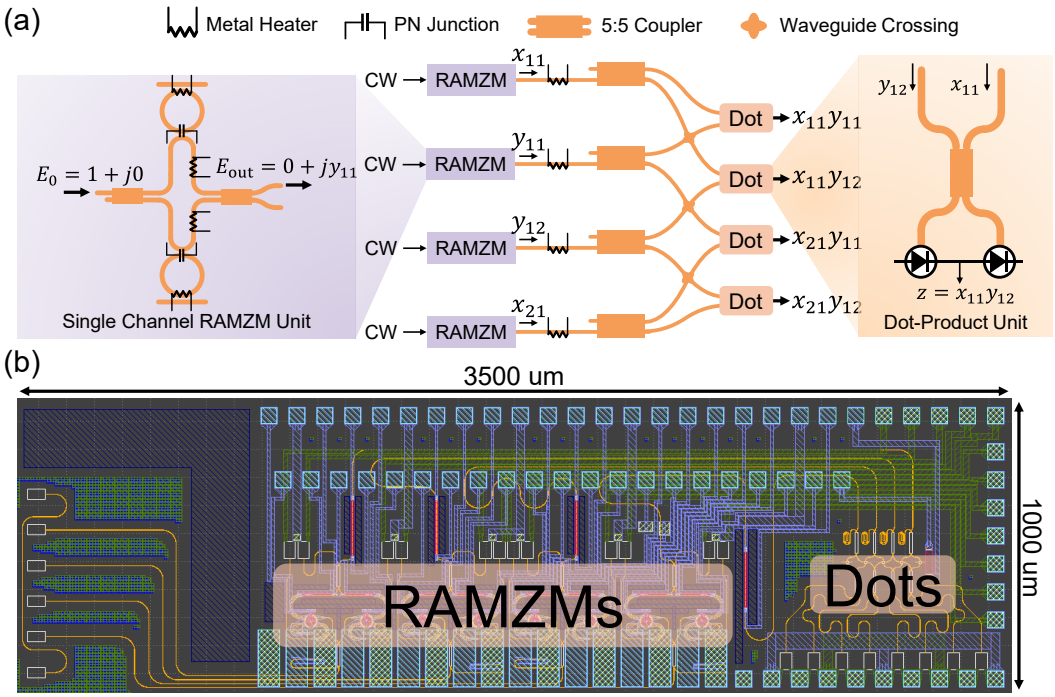


Figure 4.9: Designed layout of the proposed 2×2 matrix–matrix multiplication core. The figure shows the physical organization of the RAMZM-based multiplication cells, routing paths, and readout paths for tile-level GEMM operation.

The layout design considerations include the RAMZM coupling condition and optical skew matching across the tile. The RAMZM coupling coefficients κ_1 and κ_2 must be selected by considering speed, modulation efficiency, and LUT readability together. Here, κ_1 is the field coupling coefficient between the through bus and the MR, and κ_2 is the field coupling coefficient between the drop bus and the MR. The target operating speed of this design was set to 1 GHz, so each MR in the RAMZM was designed for 1 GHz operation. In an MR, stronger over-coupling enlarges the through-lane trajectory on the complex plane and can help satisfy the modulation-speed requirement. However, stronger coupling does not always translate into enhanced modulation efficiency; for the same wavelength detuning, the response can span a smaller portion of the trajectory, so simply increasing κ_1 and κ_2 is not optimal. Another practical constraint comes from the LUT construction. As shown in Figure 4.8(a), when the LUT is determined from the drop-port response, the drop-port power near resonance can become too similar between adjacent LUT indices. Therefore, the selected κ_1 and κ_2 must also provide enough drop-port power contrast to distinguish the LUT entries reliably. In this design, because the target speed was fixed at 1 GHz and the required detuning wavelength was known, κ_1 and κ_2 were selected to satisfy the speed requirement while maintaining usable modulation

efficiency and LUT distinguishability. At the layout level, the waveguide lengths from the RAMZMs to the dot-product units were also adjusted so that the optical-path skew among multiplication cells is matched. This path-length balancing is important because the proposed GEMM operation relies on synchronized partial-product generation and accumulation across the tile.

4.5. Comparison with Existing Photonic GEMM Architectures

The proposed RAMZM-based tensor core can now be positioned against other photonic GEMM-oriented architectures. The comparison in Table 4.1 focuses on architectures that share the same high-level objective of matrix–matrix acceleration, but implement operand encoding, multiplication, and readout in different ways. MZM/MZI-based homodyne crossbars provide a coherent-crossbar baseline because both operands can remain dynamically programmable and balanced photodetection can directly extract the interference term [11, 24]. MZM-modulated WDM–DEMUX tensor processors are also important comparison points because they exploit wavelength parallelism, but their wavelength channels require demultiplexing and readout resources after optical modulation [25]. PCM-based WDM tensor cores further illustrate the benefit of low-power nonvolatile optical weights, while also showing the trade-off introduced by static or slowly updated material states and backend DEMUX/readout complexity [9].

Table 4.1: Architecture-level comparison of representative photonic GEMM tensor-core approaches.

Comparison item	**RAMZM-based homodyne 2D crossbar (Proposed)**	**MZM/MZI-based homodyne 2D crossbar** [11, 24]	**MZM-modulated WDM–DEMUX tensor processor** [25]	**PCM-based WDM tensor core** [9]
Computing density	High	Low	Low	Medium–High
Reconfigurability (Operand1 / Operand2)	Dynamic / Dynamic	Dynamic / Dynamic	Dynamic / Dynamic	Dynamic / Static
Operand update cost	Low	Low	Low	High
Power	Medium	Medium	Medium	Low
Operand range	Full-range	Full-range	Full-range	One-signed
Readout	Balanced PD	Balanced PD	DEMUX + PD / balanced PD	DEMUX + PD
Main bottleneck	Dot-product array calibration	MZM footprint; Dot-product array calibration	MZM footprint; DEMUX/readout area;	Static operand; PCM fidelity; DEMUX/readout area

As summarized in the table, the proposed structure is primarily distinguished by compact dynamic operand encoding. Compared with an MZM/MZI-based homodyne 2D crossbar, the proposed RAMZM-based crossbar preserves dynamic operand programmability and balanced-photodetector readout, but replaces large MZM/MZI modulation cells with microring-assisted cells that are more favorable for dense area scaling. Compared with an MZM-modulated WDM–DEMUX tensor processor, the proposed structure forms the two-dimensional dot product directly in the homodyne crossbar and therefore does not require a DEMUX stage. This reduces backend area and complexity: a WDM–DEMUX tensor processor must multiplex wavelengths at the input and demultiplex them again at the readout stage, whereas the proposed structure would only require wavelength multiplexing at the input if it is extended to WDM operation. Compared with PCM-based WDM tensor cores, the proposed approach does not rely on a static material-state operand; instead, both operands can be dynamically reprogrammed. This dynamic operation is more generally suitable for large and rapidly changing matrix workloads, such as transformer models. From the power perspective, one RAMZM cell includes four heater-controlled elements: two heaters for the MZI arms and two heaters for the microrings, in addition to the driving power required for the two microrings. Therefore, the RAMZM power cost should be considered together with heater efficiency and microring driving energy; if heater implementation becomes more efficient through fabrication and process improvements, the total RAMZM driving power can also be reduced. Overall, the proposed RAMZM tensor core is positioned as a GEMM-native microring-based architecture that prioritizes dense dynamic operand encoding without requiring a backend demuxing stage.

5. Conclusion

This thesis examined microring-based photonic tensor cores from the combined perspectives of architecture and calibration. The central conclusion is that practical photonic tensor-core design requires not only compact optical devices, but also arithmetic dataflow that matches target workloads and calibration methods that remain stable under device nonidealities.

First, this thesis reviewed the conventional silicon microring weightbank as a baseline for wavelength-multiplexed matrix–vector multiplication. That baseline confirmed the compactness and WDM compatibility of microring-based computation, but it also clarified two structural limitations for modern AI workloads: inefficient support for signed operands and a mismatch between MVM-centered hardware and GEMM-dominant computation.

Second, this thesis proposed a bidirectional microring weightbank for signed multiplications. By embedding signed representation in a differential bidirectional intensity-modulation framework, the proposed scheme avoided phase-based sign encoding and reduced the optical duplication often required in conventional signed architectures. Single-channel experiments, heater-sweep characterization, and 2D LUT calibration showed that the signed mapping can be stabilized in hardware, while also identifying multi-channel calibration and reflection sensitivity as the main remaining challenges.

Third, this thesis proposed a RAMZM-based photonic tensor-core architecture that is more naturally aligned with matrix–matrix multiplication. Through complex-plane trajectory analysis, LUT-based operand encoding, direct RAMZM-output validation, and a reusable 2×2 GEMM tile layout, the work established a concrete path from device-level behavior to GEMM-oriented architectural organization.

Taken together, these contributions show that microring-based photonic tensor cores can be extended beyond conventional non-negative MVM weightbanks toward signed-operation support and GEMM-native computation. The final message of this thesis is that calibration is not an auxiliary step but a central design requirement, and future progress depends on scalable multi-channel calibration, reflection-robust integration, and larger system-level demonstrations that connect optical device behavior to workload-level acceleration benefits.

References

[1] Epoch AI, "AI Models," data page reporting that the training compute of notable AI models has grown by approximately 4.4× per year since 2010, accessed Apr. 11, 2026.

[2] Epoch AI, "Machine Learning Hardware," data page reporting that 16-bit performance of leading machine-learning hardware improved by about 36% per year over 2006–2024, accessed Apr. 11, 2026.

[3] Epoch AI, "Leading machine learning hardware is getting 40% more energy efficient every year," data insight, accessed Apr. 11, 2026.

[4] D. A. B. Miller, "Rationale and Challenges for Optical Interconnects to Electronic Chips," *Proceedings of the IEEE*, vol. 88, no. 6, pp. 728–749, 2000, doi:10.1109/5.867687.

[5] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei, "Scaling Laws for Neural Language Models," *arXiv preprint arXiv:2001.08361*, 2020, doi:10.48550/arXiv.2001.08361.

[6] Y. Shen, N. C. Harris, S. Skirlo, M. Prabhu, T. Baehr-Jones, M. Hochberg, X. Sun, S. Zhao, H. Larochelle, D. Englund, and M. Soljačić, "Deep learning with coherent nanophotonic circuits," *Nature Photonics*, vol. 11, pp. 441–446, 2017, doi:10.1038/nphoton.2017.93.

[7] X. Lin, Y. Rivenson, N. T. Yardimci, M. Veli, Y. Luo, M. Jarrahi, and A. Ozcan, "All-optical machine learning using diffractive deep neural networks," *Science*, vol. 361, no. 6406, pp. 1004–1008, 2018, doi:10.1126/science.aat8084.

[8] A. N. Tait, A. X. Wu, T. Ferreira de Lima, E. Zhou, B. J. Shastri, M. A. Nahmias, and P. R. Prucnal, "Microring Weight Banks," *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 22, no. 6, pp. 312–325, 2016, doi:10.1109/JSTQE.2016.2573583.

[9] J. Feldmann, N. Youngblood, M. Karpov, H. Gehring, X. Li, M. Stappers, M. Le Gallo, X. Fu, A. Lukashchuk, A. S. Raja, J. Liu, C. D. Wright, A. Sebastian, T. J. Kippenberg, W. H. P. Pernice, and H. Bhaskaran, "Parallel convolutional processing using an integrated photonic tensor core," *Nature*, vol. 589, no. 7840, pp. 52–58, 2021, doi:10.1038/s41586-020-03070-1.

[10] Y. Huang, H. Yue, W. Ma, Y. Zhang, Y. Xiao, Y. Tang, H. Tang, and T. Chu, "Parallel photonic acceleration processor for matrix–matrix multiplication," *Optics Letters*, vol. 48, no. 12, pp. 3231–3234, 2023, doi:10.1364/OL.488464.

[11] S. Rahimi Kari, N. A. Nobile, D. Pantin, V. Shah, and N. Youngblood, "Realization of an integrated coherent photonic platform for scalable matrix operations," *Optica*, vol. 11, no. 4, pp. 542–551, 2024, doi:10.1364/OPTICA.507525.

[12] T. Ferreira de Lima, H.-T. Peng, A. N. Tait, M. A. Nahmias, H. B. Miller, B. J. Shastri, and P. R. Prucnal, "Machine Learning with Neuromorphic Photonics," *Journal of Lightwave Technology*, vol. 37, no. 5, pp. 1515–1534, 2019, doi:10.1109/JLT.2019.2903474.

[13] B. J. Shastri, A. N. Tait, T. Ferreira de Lima, W. H. P. Pernice, H. Bhaskaran, C. D. Wright, and P. R. Prucnal, "Photonics for artificial intelligence and neuromorphic computing," *Nature Photonics*, vol. 15, pp. 102–114, 2021, doi:10.1038/s41566-020-00754-y.

[14] A. N. Tait, T. Ferreira de Lima, M. A. Nahmias, H.-T. Peng, B. J. Shastri, and P. R. Prucnal, "Multi-channel control for microring weight banks," *Optics Express*, vol. 24, no. 8, pp. 8895–8906, 2016, doi:10.1364/OE.24.008895.

[15] T. Ferreira de Lima, A. N. Tait, H.-T. Peng, B. J. Shastri, and P. R. Prucnal, "Progress in Neuromorphic Photonics," *Nanophotonics*, vol. 6, no. 3, pp. 577–599, 2017, doi:10.1515/nanoph-2016-0139.

[16] A. N. Tait, T. Ferreira de Lima, E. Zhou, M. A. Nahmias, H.-T. Peng, B. J. Shastri, and P. R. Prucnal, "Neuromorphic photonic networks using silicon photonic weight banks," *Scientific Reports*, vol. 7, article 7430, 2017, doi:10.1038/s41598-017-07754-z.

[17] B. E. Little, S. T. Chu, H. A. Haus, J. Foresi, and J.-P. Laine, "Microring resonator channel dropping filters," *Journal of Lightwave Technology*, vol. 15, no. 6, pp. 998–1005, 1997, doi:10.1109/50.588673.

[18] J. Wei, Y. He, Y. Yang, J. Ding, L. Lu, and L. Zhu, "Lightweight optical neural network based on micro-ring resonator," *Optics Express*, vol. 33, no. 5, pp. 10674–10686, 2025, doi:10.1364/OE.543941.

[19] M. A. Nahmias, T. Ferreira de Lima, A. N. Tait, H.-T. Peng, B. J. Shastri, and P. R. Prucnal, "Photonic Multiply-Accumulate Operations for Neural Networks," *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 26, no. 1, pp. 1–18, 2020, doi:10.1109/JSTQE.2019.2941485.

[20] X. Xu, M. Tan, B. Corcoran, J. Wu, A. Boes, T. G. Nguyen, S. T. Chu, B. E. Little, D. G. Hicks, R. Morandotti, A. Mitchell, and D. J. Moss, "11 TOPS photonic convolutional accelerator for optical neural networks," *Nature*, vol. 589, pp. 44–51, 2021, doi:10.1038/s41586-020-03063-0.

[21] F. Ashtiani, A. J. Geers, and F. Aflatouni, "An on-chip photonic deep neural network for image classification," *Nature*, vol. 606, pp. 501–506, 2022, doi:10.1038/s41586-022-04714-0.

[22] A. Geravand, Z. Zheng, F. Shateri, S. Levasseur, L. A. Rusch, and W. Shi, "Ultrafast coherent dynamics of microring modulators," *Nature Photonics*, vol. 19, pp. 740–750, 2025, doi:10.1038/s41566-025-01686-1.

[23] F. G. Vanani, A. Fardoost, Z. Zhu, C. Doerr, S. Pang, and G. Li, "A Scalable Silicon Photonic Tensor Accelerator," in *CLEO 2025, Technical Digest Series*, Optica Publishing Group, 2025, paper SS162_4, doi:10.1364/CLEO_SI.2025.SS162_4.

[24] N. Youngblood, "Coherent Photonic Crossbar Arrays for Large-Scale Matrix–Matrix Multiplication," *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 29, no. 2, pp. 1–11, 2023, doi:10.1109/JSTQE.2022.3171167.

[25] S. Ou, K. Xue, L. Zhou, C. H. Lee, A. Sludds, R. Hamerly, K. Zhang, H. Feng, Y. Yu, R. Kopparapu, E. Zhong, C. Wang, D. R. Englund, M. Yu, and Z. Chen, "Hypermultiplexed integrated photonics–based optical tensor processor," *Science Advances*, vol. 11, no. 23, article eadu0228, 2025, doi:10.1126/sciadv.adu0228.

ABSTRACT IN KOREAN

마이크로링 기반 포토닉 텐서 코어와 그 보정 방법

본 학위논문은 마이크로링 기반 포토닉 텐서 코어를 아키텍처와 보정 방법의 관점에서 연구한다. 먼저 기존 실리콘 마이크로링 웨이트뱅크를 파장분할다중 기반 행렬–벡터 곱셈의 기준 구조로 설정하고, 전달 특성, 가중치 정의, MVM 확장 방식, 2채널 실험 프로토타입, 그리고 소규모 신경망 추론 실증을 통해 그 동작을 정리한다. 이 과정은 마이크로링 웨이트뱅크의 높은 집적도와 WDM 적합성을 확인하는 동시에, 실제 AI workload를 위해서는 부호 있는 입력 처리의 비효율성과 MVM 중심 데이터플로우의 GEMM 부적합성이라는 두 가지 구조적 한계가 있음을 보여준다.

첫 번째 한계를 해결하기 위해, 본 논문은 부호 있는 곱셈을 위한 양방향 마이크로링 웨이트뱅크를 제안한다. 제안 구조는 광 위상이 아니라 차동 양방향 강도 변조 구조 안에 부호 표현을 직접 포함시켜, 기존 signed scheme에서 필요한 광 경로 중복을 줄이도록 설계되었다. 단일 채널 실험, heater sweep 특성 측정, 그리고 2차원 LUT 보정을 통해 이 signed encoding 원리를 검증하였으며, 이를 통해 더 높은 유효 집적도와 더 낮은 보정 복잡도의 가능성을 확인하였다. 다만 다채널 보정과 반사 민감도는 향후 과제로 남는다.

두 번째 한계를 해결하기 위해, 본 논문은 행렬–행렬 곱셈에 보다 자연스럽게 대응하는 RAMZM 기반 포토닉 텐서 코어 아키텍처를 제안한다. 이 구조는 독립적으로 인코딩된 행/열 피연산자, 복소평면 궤적 기반 동작 해석, LUT 기반의 안정적 피연산자 인코딩을 사용하며, 순차 및 랜덤 LUT sweep과 반복 실험 오차 통계를 통한 RAMZM 직접 출력 검증으로 그 타당성을 확인한다. 또한 이 동작 원리를 바탕으로 확장 가능한 포토닉 GEMM 엔진의 기본 블록이 되는 2×2 GEMM 타일 레이아웃을 설계하였다. 종합하면, 본 연구는 기존의 비음수 MVM 중심 마이크로링 웨이트뱅크를 넘어, 부호 연산 지원과 GEMM 친화적 데이터플로우를 갖는 포토닉 텐서 코어로의 확장을 제시하며, 실용적인 포토닉 컴퓨팅 시스템에서 보정이 핵심 설계 요소임을 보여준다.

핵심되는 말: 포토닉 텐서 코어, 포토닉 신경망, 포토닉 컴퓨팅, 실리콘 포토닉스