Reference-Assisted Microring Weightbank for Signed Multiplication

Minkyu Cho, Yongjin Ji, Minsung Kang, and Woo-Young Choi*

Department of Electrical and Electronic Engineering, Yonsei University, South Korea
**wchoi@yonsei.ac.kr*

We propose and demonstrate a new type of incoherent photonic tensor core for signed multiplication. This structure extends a microring resonator (MR) weightbank by introducing a fixed reference power for differential activation encoding, making both operands differentially defined: the activation is represented by the difference between a fixed reference power and a dynamically modulated power, while the weight is represented by the through/drop response of an MR. This scheme enables signed multiplication with reduced hardware overhead by sharing each weighting MR between the two counter-propagating inputs. The trade-offs are signal attenuation and potential reflection sensitivity. We demonstrate this concept using a fabricated silicon photonic integrated circuit containing two thermally tuned MRs, one for each operand of $x \times w$. Lookup tables (LUTs) spanning -1 to $+1$ are constructed for both operands, and randomly ordered LUT sweeps confirmed signed multiplication with residual error standard deviations of 0.022 for the x-sweep and 0.032 for the w-sweep, both below the nominal 31-level spacing of 0.067.

Artificial neural networks have become a powerful machine learning technique and have been widely applied across fields such as natural language processing, image generation, and scientific data analysis. However, the rapid growth of model sizes and data volumes has greatly increased the computational requirements of neural network inference, challenging the bandwidth and energy efficiency of conventional digital processors. Optical neural networks have therefore emerged as an attractive platform for neural network acceleration because they can exploit low-latency optical transmission, high bandwidth, and energy-efficient optical-domain linear operations. In particular, photonic tensor cores (PTCs) have been investigated as accelerators for the linear algebra operations that dominate neural networks.

For these matrix operations, the operands—activations x and weights w—generally include positive and negative values. From the viewpoint of the operand-encoding domain, PTCs can be broadly classified into two categories: coherent structures that encode operands in the optical field domain and incoherent structures that encode operands in the intensity domain. Coherent PTCs can naturally represent signed values using optical phase, but they require accurate phase control across interferometric paths [1–3]. Incoherent PTCs, in contrast, benefit from simple intensity encoding, but they have difficulty representing signed operands because optical intensity is inherently nonnegative. For activation encoding, this issue is partly mitigated after a rectified linear unit (ReLU), because ReLU activations are nonnegative [4–6], but signed weights and general signed matrix operations still require sign handling. To overcome this nonnegativity constraint, time-division multiplexing (TDM) and spatial-division multiplexing (SDM) have been introduced by separating positive and negative components [7–10]. However, these methods impose additional system overhead, such as extra programming steps or duplicated modulators, motivating an architecture that can encode signed operands directly while preserving the simplicity of incoherent operation.

In this paper, we propose a reference-assisted microring resonator (MR) weightbank architecture in which both operands are represented by intensity differences within the multiplication core, without TDM sign scheduling or duplicated activation modulators. Conventional MR weightbanks have defined a signed weight from the difference between the through- and drop-port transmissions of an MR [11, 12]. In the proposed reference-assisted MR weightbank, this differential concept is extended to the activation operand by defining x as the difference between a fixed reference power and a dynamically varying signal power controlled by an MR. Thus, each MR encoder represents one operand for general matrix-vector multiplication (MVM), avoiding the need for time scheduling or duplicated modulators to implement operand signs. The multiplied result is converted into photodetected currents and read out electrically. We experimentally demonstrate the concept using a fabricated on-chip 1×1 unit cell multiplier.

The proposed structure is shown in Figure 1(a) for a 1×1 unit cell schematic implementing $x \times w$. In the unit cell, two counter-propagating continuous-wave (CW) optical inputs pass through the same w-encoding MR, while one input is modulated by the x-encoding MR and the other is not modulated. We denote the dimensionless through- and drop-port power transmissions of the w-encoding MR by T_w and D_w, respectively, and the dimensionless nonnegative intensity coefficients of the fixed and modulated inputs by f and g, respectively. The structure produces four branch signals whose powers are proportional to fT_w, fD_w, gT_w, and gD_w. The four branch signals are independently detected by the PDs, and the output is obtained

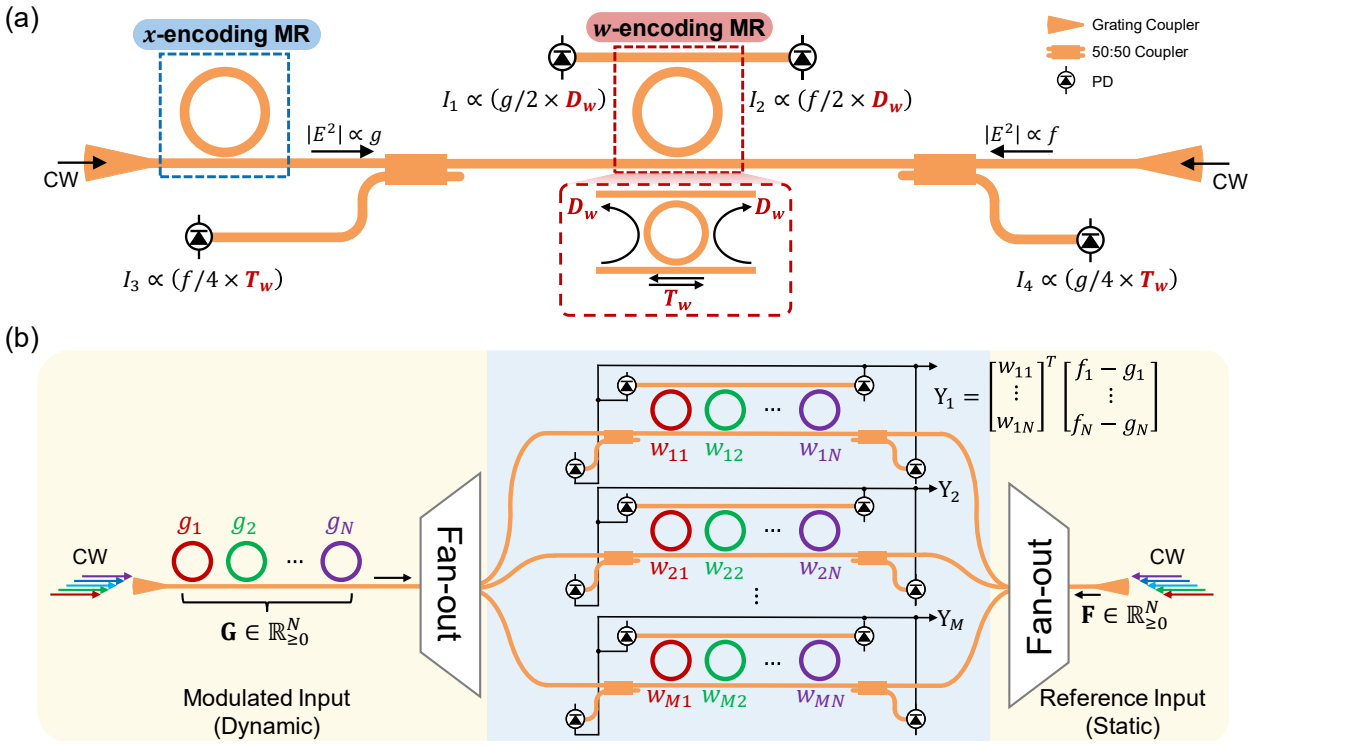


Figure 1. Reference-assisted MR weightbank for signed multiplication. (a) Unit cell containing two MRs, one for each operand of $x \times w$. The x-encoding MR modulates one of the two counter-propagating CW inputs to produce g, while the other CW input is used as the reference, denoted by f; the two counter-propagating inputs share the same w-encoding MR transmissions, T_w and D_w. (b) Scaled-up version in which multiple wavelength channels and spatially separated weightbank floors form signed matrix-vector multiplication. In each weightbank floor, the four photodiodes (PDs) are connected to the current-summing node with appropriate anode/cathode polarities to realize vector-vector multiplication.

as

$$y = (T_w - 2D_w)(f - g) = wx \propto I_1 - I_2 + I_3 - I_4. \quad \textbf{(1)}$$

Here, $x = f - g$ is the signed activation and $w = T_w - 2D_w$ is the signed weight. The sign of x is determined by whether f or g is larger, while the sign of w is determined by the relative through/drop contribution of the w-encoding MR. The factor of 2 in front of D_w accounts for the additional 50:50 coupler encountered by the through-port path, which attenuates the through contribution by half relative to the drop contribution.

The structure can be scaled up as shown in Figure 1(b). To implement multiplication between a weight matrix $\mathbf{W} \in \mathbb{R}^{M \times N}$ and an activation vector $\mathbf{X} \in \mathbb{R}^N$, the scaling strategy can be summarized in two directions. To expand N, wavelength-division multiplexing (WDM) can be used. Across the WDM channels in Figure 1(b), the activation vector is formed as $\mathbf{X} = \mathbf{F} - \mathbf{G}$, where the dynamically modulated vector $\mathbf{G}$ is generated by the x-encoding MRs and the reference vector $\mathbf{F}$ is supplied from the opposite direction. The w-encoding MRs define the signed weights as $w = T_w - 2D_w$, and each weightbank floor, containing the w-encoding MRs corresponding to one matrix row from w_{i1} to w_{iN}, performs a vector-vector inner product over the WDM channels. To expand M, multiple weightbank floors are placed spatially, and the light power of each wavelength channel is split into these floors by optical fan-out. Thus, the MVM result is obtained as

$$\mathbf{Y} = \mathbf{WX} = \begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1N} \\ \vdots & \vdots & \ddots & \vdots \\ w_{M1} & w_{M2} & \cdots & w_{MN} \end{bmatrix} \begin{bmatrix} f_1 - g_1 \\ \vdots \\ f_N - g_N \end{bmatrix}. \quad \textbf{(2)}$$

This structure has several architectural advantages. First, both operands are expressed differentially, with $x = f - g$ and $w = T_w - 2D_w$, so signed operation is native rather than implemented by duplicated encoders or additional time slots. Second, for practical MVM engines, the MR encoders can be implemented with PN-junction tuning, enabling dynamic operand updates. Third, optical phase is not used as an encoding variable, so phase calibration is not required when reflection-induced nonidealities are sufficiently small.

The proposed structure has two trade-offs. The first is signal attenuation. In the scaled MVM architecture, increasing the output dimension M requires optical fan-out to M spatially separated weightbank floors, so the signal intensity delivered to each floor decreases approximately in proportion to $1/M$. In addition, the 50:50 couplers route only half of the propagating optical power to main signal paths. As a result, the output current of the unit cell is given by $I_{\text{out}} = RP_0(f - g)(T_w - 2D_w)/4$, where R is the PD responsivity in A/W and P_0 is the CW input optical power in W, whereas a conventional MR weightbank gives $I_{\text{out}} = RP_0 g(T_w - D_w)$ under the same notation [11, 12]. See Supplement 1 for the noise and resolution-scaling analy-

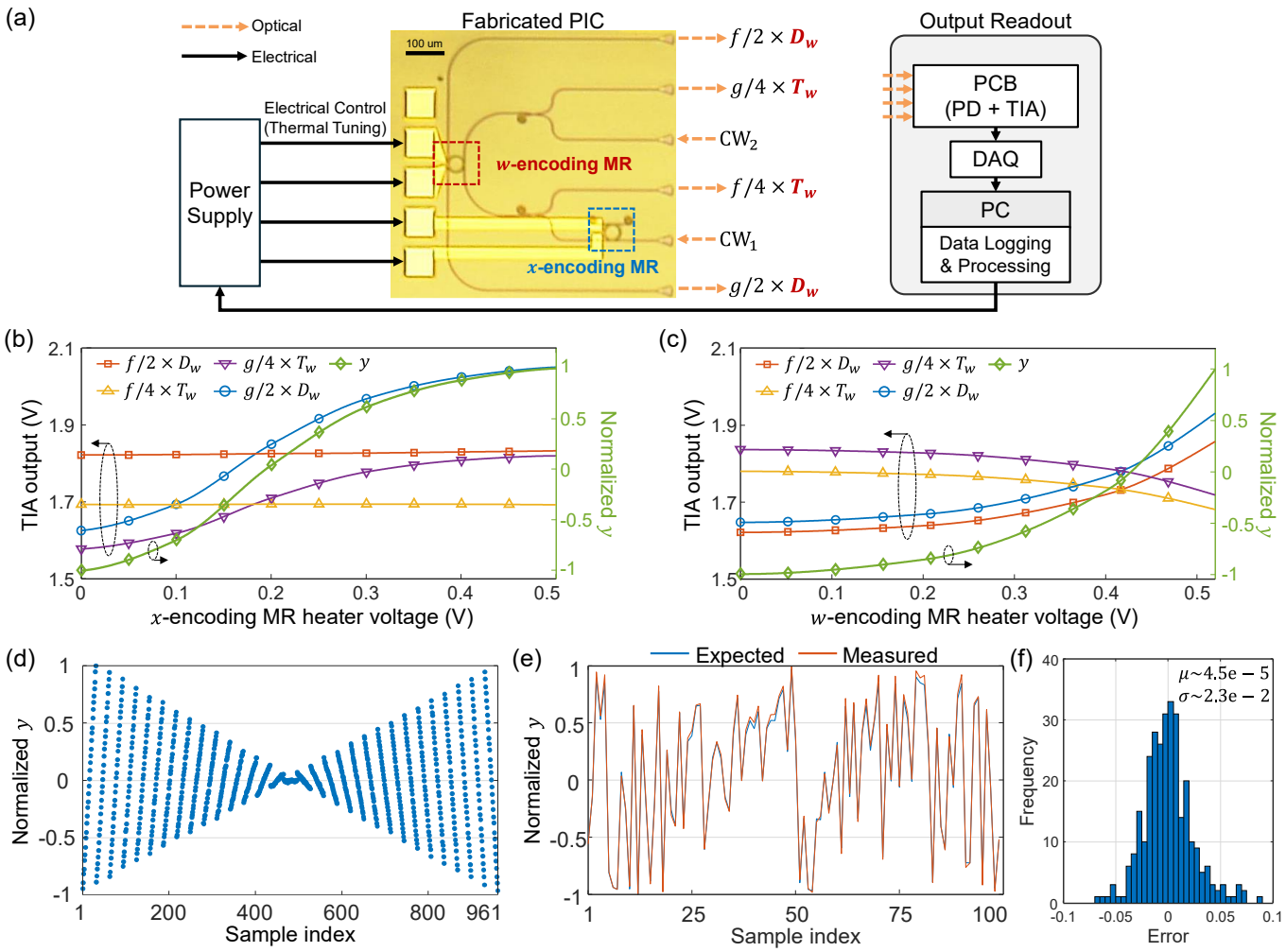


Figure 2. (a) Experimental setup for verifying the 1×1 unit cell. The silicon PIC contains one x-encoding MR and one w-encoding MR, both controlled by integrated heaters. The four optical branch signals are detected by off-chip PDs, converted to voltages by TIAs, and digitally processed to form $y = x \times w$. Microring heater voltage sweep results for (b) the x-encoding MR and (c) the w-encoding MR, used to construct the corresponding operand LUTs. Multiplication results for (d) 961 combinations from two 31-level LUTs and (e) 100 randomly ordered pairs. (f) Error histogram for 300 repeated random LUT sweeps.

sis. The second trade-off is potential sensitivity to reflections from imperfections such as MR sidewall roughness, which may cause parasitic beating in the photocurrent. Such beating can be suppressed by introducing an optical path-length difference exceeding the mutual coherence length of the two input fields [13].

The concept of the proposed unit cell structure was experimentally demonstrated. Figure 2(a) shows the experimental setup. The unit cell structure was fabricated as a photonic integrated circuit (PIC) using NanoSOI fabrication. The PIC contains two microrings, used as the x-encoding and w-encoding MRs, respectively, and both were thermally tuned with integrated metal heaters. The 50:50 couplers were implemented with directional couplers, and optical input/output couplings for the two input lights and four branch output signals were performed using grating couplers. During measurement, the thermal tuning voltages were controlled by a multi-channel power supply. Two CW inputs at the same nominal wavelength were launched into the chip for bidirectional propagation. The two inputs were generated from independently tuned laser channels without active phase locking. The four optical output branch signals were converted into voltages by four off-chip PDs with transimpedance amplifiers (TIAs), captured by a data-acquisition (DAQ). The PD bias voltage V_b was set to 1.5 V during the measurement.

Microring heater voltage sweep results are shown in Figures 2(b) and 2(c). Figure 2(b) shows the x-encoding MR sweep, where, among the four branches, only those containing the dynamically modulated signal power g change significantly, rather than the reference power f. This response confirms that the differential activation definition $x = f - g$ works as intended. Figure 2(c) shows the w-encoding MR sweep, where the branch signals containing T_w and D_w follow the through- and drop-port transmissions, respectively. This through/drop response enables the signed weight definition $w = T_w - 2D_w$.

For the experimental readout, digital post-processing was limited to emulating the signed summation in Eq. (1) and compensating a static channel imbalance caused by unequal grating coupler coupling efficiencies. The latter was corrected by multiplying the corresponding $gT_w/4$ branch signal by a fixed correction factor. With the proposed on-chip PD current summation, Eq. (1) is directly realized by the PD polarities, while the

Table 1. Comparison of signed operation capabilities and hardware overheads in incoherent photonic processors.

Architecture	Signed operand	Modulator duplication; temporal overhead	Weighting devices ($M \times N$ MVM)	Readout devices ($M \times N$ MVM)	**Hardware cost per sign state[a]**	Other drawbacks
PCM core [14]	–	–; –	MN PCM cells	M PDs; M TIAs	1 PCM + 1 PD + 1 TIA	Programming latency
TDM [15]	w	–; 2×	MN MRs	M PDs; M TIAs	1 MR + 1 PD + 1 TIA	–
Duplicated modulators [16]	w	2×; –	$2MN$ PIN attenuators	$2MN$ PDs; M TIAs	1 PIN Mod. + N PDs + 0.5 TIA	Device mismatch
Conventional MR WB [11]	w	–; –	MN MRs	$2M$ PDs; M TIAs	0.5 MR + 1 PD + 0.5 TIA	–
Proposed MR WB	$\boldsymbol{x, w}$	**–; –**	MN **MRs**	$4M$ **PDs;** M **TIAs**	**0.25 MR + 1 PD + 0.25 TIA**	**Signal loss / potential reflection sensitivity**

[a]Excluding input encoders, hardware cost per sign state is $\tau 2^{-n_s}\left[(N_{\mathrm{W}}/MN)\,\mathrm{W} + (N_{\mathrm{PD}}/M)\,\mathrm{PD} + (N_{\mathrm{TIA}}/M)\,\mathrm{TIA}\right]$, where n_s is the number of signed operands and τ is the temporal overhead.

grating coupler calibration is eliminated; hence, neither digital processing nor separate ADC channels are required. Thermal crosstalk was modeled by relating the effective heater drives of the x- and w-encoding MRs, $V^2_{x,\mathrm{eff}}$ and $V^2_{w,\mathrm{eff}}$, respectively, to the applied heater voltages V_x and V_w:

$$\begin{bmatrix} V^2_{x,\mathrm{eff}} \\ V^2_{w,\mathrm{eff}} \end{bmatrix} = \begin{bmatrix} 1 & 0.018 \\ 0.008 & 1 \end{bmatrix} \begin{bmatrix} V^2_x \\ V^2_w \end{bmatrix}, \tag{3}$$

The inverse of this model was used to convert each LUT command into the corresponding applied heater voltages. After calibration, the LUTs and thermal crosstalk coefficients were fixed throughout the evaluation; no result-based fitting or recalibration was applied to the data in Figures 2(d) and 2(e).

Figure 2(d) shows the multiplication result $y = x \times w$, while x and w are sequentially swept over the 31×31 LUT with an inner 31-step sweep of x and an outer 31-step sweep of w. Both x and w are normalized from -1 to $+1$, and the resulting trajectory forms the expected signed multiplication pattern. Figure 2(e) shows the multiplication result for 100 randomly ordered LUT pairs of x and w, where the measured output closely follows the expected product. Figure 2(f) shows the error histogram from 300 repeated randomly ordered LUT sweeps of both x and w, where the error is defined as $e = y_{\mathrm{meas}} - y_{\mathrm{expected}}$. The standard deviation was $\sigma = 2.3 \times 10^{-2}$, reflecting the effects of LUT-mapping error, residual thermal crosstalk, static channel imbalance, and readout noise. For additional single-operand sweeps, the standard deviations were 0.022 for the x-sweep and 0.032 for the w-sweep, corresponding to effective resolutions of 6.51 and 5.97 bits, respectively. The effective resolution was calculated from the ratio between the full-scale range and the RMS noise [17],

$$b_{\mathrm{eff}} = \log_2\left(\frac{\Delta y}{\sigma_e}\right), \tag{4}$$

where $\Delta y = 2$ is the normalized full-scale output range and σ_e is the residual error standard deviation.

Table 1 summarizes the signed operand capabilities and hardware cost of the compared incoherent photonic processors. The main metric, hardware cost per sign state, quantifies the weighting and readout resources required for each concurrently supported sign combination. Among the compared architectures, the proposed design has the lowest normalized device count, attributable to sharing each weighting MR between the counter-propagating inputs.

In conclusion, we have proposed a reference-assisted MR weightbank architecture for signed multiplication and verified its concept using a fabricated 1×1 unit cell multiplier. In the proposed structure, both operands of MVM are defined differentially, enabling the signs of operands to be represented directly in the core. This enables incoherent photonic tensor cores to perform signed MVM without requiring time scheduling or duplicated modulators. Looking forward, this architecture can serve as a compact building block for scalable incoherent signed photonic MVM engines.

Funding. National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2024-00455038).

Disclosures. The authors declare no conflicts of interest.

Data availability. Data underlying the results presented in this paper may be obtained from the authors upon reasonable request.

REFERENCES

1. Y. Shen *et al.*, Nature Photonics **11**, 441–446 (2017).
2. R. Hamerly *et al.*, Physical Review X **9**, 021032 (2019).
3. N. Youngblood, IEEE Journal of Selected Topics in Quantum Electronics **29**, 1–11 (2023).
4. F. Ashtiani, A. J. Geers, and F. Aflatouni, Nature **606**, 501–506 (2022).
5. S. Ohno *et al.*, ACS Photonics **9**, 2614–2622 (2022).
6. M. Becker *et al.*, NeurIPS ML4PS Workshop (2023).
7. S.-Y. Ma *et al.*, Nature Communications **16**, 359 (2025).
8. B. Wu *et al.*, Light: Science & Applications **14**, 328 (2025).
9. J. Li *et al.*, Advanced Photonics **1**, 046001 (2019).
10. S. R. Ahmed *et al.*, Nature **640**, 368–374 (2025).
11. A. N. Tait, T. Ferreira de Lima, E. Zhou, *et al.*, Scientific Reports **7**, 7430 (2017).
12. A. N. Tait *et al.*, IEEE Journal of Selected Topics in Quantum Electronics **22**, 312–325 (2016).
13. R. Xue *et al.*, Opt. Lett. **51**, 2312–2315 (2026).
14. B. Dong, S. Aggarwal, W. Zhou, *et al.*, Nature Photonics **17**, 1080–1088 (2023).
15. S. Xu, J. Wang, S. Yi, and W. Zou, Nature Communications **13**, 7970 (2022).
16. F. Ashtiani, M. H. Idjadi, and K. Kim, Nature **651**, 927–932 (2026).
17. Analog Devices, Technical Article (2012).