

# FSR-GeMM: A Scalable FSR-Parallel Photonic Accelerator for Real-Valued GeMM Computing

Peiyu Chen<sup>1</sup>, Yinyi Liu<sup>2</sup>, Minhong Xu<sup>1</sup>, Chongyi Yang<sup>1</sup>, Xiaohan Jiang<sup>2</sup>, Wei Zhang<sup>2</sup> and Jiang Xu<sup>1†</sup>

<sup>1</sup>Microelectronics Thrust, The Hong Kong University of Science and Technology (Guangzhou), China

<sup>2</sup>Dept. of Electronic and Computer Engineering (ECE), Hong Kong University of Science and Technology, Hong Kong

**Abstract**—Photonic computing is poised to revolutionize artificial intelligence (AI) acceleration by offering exceptional speed and energy efficiency for General Matrix Multiplication (GeMM). However, existing works on photonic tensor core architectures face significant challenges in managing real-valued and dynamic operands. Specifically, Mach-Zehnder interferometer (MZI) meshes require computationally intensive singular value decomposition (SVD) for matrix preprocessing, while microring resonator (MRR) weight banks are limited to non-negative operands, complicating operations with dual negative values. Additionally, coherent interference crossbars, although theoretically capable of supporting real-valued multiplication, struggle with fabrication complexities and sensitivity to environmental variations.

To address these limitations, we propose FSR-GeMM schema, a scalable photonic accelerator that leverages free-spectral range (FSR) multiplexing. This architecture eliminates the need for SVD preprocessing, supports direct multiplication of two dynamic real-valued operands, and enhances reliability and scalability. Experimental results from a photonic-electronic prototype demonstrate that FSR-GeMM achieves up to  $57\times$  improvements in area efficiency and  $13.8\times$  gains in energy efficiency compared to existing photonic GeMM accelerators. Furthermore, it reduces energy consumption by 70% relative to MRR-based systems and achieves  $21\times$  speedup against leading photonic GeMM accelerator designs, highlighting its potential to advance practical and scalable AI acceleration.

## I. INTRODUCTION

The slowdown of Moore’s Law has prompted a paradigm shift in computing, necessitating architectures that deliver higher performance and energy efficiency to meet the growing computational demands of AI applications [1]. Photonic computing has emerged as a promising solution, particularly for matrix-heavy operations like GeMM, which accounts for approximately 70% of computations in AI workloads [2]. Photonic accelerators offer significant advantages over traditional electronic systems, including superior speed and energy efficiency in vector operations [3]–[5]. However, the field currently lacks a *scalable, practical, and manufacturable* architecture design that effectively leverages photonic technologies for *large-scale* AI workloads. We reconsider existing designs from the aspects of: algorithm-hardware mapping cost, value range coverage of operands, and fabrication feasibility.

Current architectures primarily rely on three paradigms: (1) MZI-based **unitary meshes** [6], [7], (2) MRR-based **weight**

**banks** [8]–[10], and (3) **crossbar coupling** designs [11]–[14]. However, each approach has inherent drawbacks that restrict its applicability to real-world AI workloads. MZI-based meshes, while suitable for unitary matrix operations, require costly SVD preprocessing for real-valued GeMM, introducing significant overheads. MRR-based systems are limited by their inability to handle negative operands and by inefficient optical bandwidth utilization, processing only a single wavelength per device. Crossbar designs, including those utilizing interference devices or phase-change materials (PCMs), suffer from instability due to either phase noise or the lack of PCM robustness, along with fabrication complexities and large footprints.

This analysis highlights a critical gap in photonic computing technology: **no** existing design fully addresses the **four** essential traits for effective GeMM acceleration: (1) direct algorithm-hardware mapping without heavy preprocessing, (2) efficient optical bandwidth utilization through multicast modulation, (3) support for real-valued range arithmetic, and (4) compact and scalable physical implementation with minimal loss and crosstalk. Bridging this gap is vital for unlocking the potential of photonic computing in AI acceleration.

To tackle these challenges, we propose **FSR-GeMM**, a novel photonic tensor core architecture that utilizes free-spectral range (FSR) multiplexing parallelism in specially designed MRR arrays with symmetric topology. This schema eliminates algorithm-hardware mapping overhead such as SVD preprocessing, supports real-valued matrix multiplication, and also enhances scalability and compactness compared to existing solutions. The FSR-GeMM accelerator integrates optimized optical interconnects and analog-domain resource sharing, significantly improving computational density and energy efficiency while maintaining wavelength-division multiplexing (WDM) compatibility. Our evaluation demonstrates substantial performance gains, establishing FSR-GeMM as a viable solution for next-generation AI systems.

TABLE I: Comparison across Photonic Paradigms

Aspects	Unitary Mesh	Weight Bank	Interference Crossbar	Ours
Mapping Cost	High	Low*	Medium	Low*
Op1 Val Range	Real*	Positive	Real*	Real*
Op2 Val Range	Real*	Real*	Real*	Real*
Compute Type	Coherent	Incoherent	Coherent	Incoherent
Footprint	Large	Small*	Large	Small*
Stability	Medium	High*	Low	High*

\***Note:** asterisk denotes the *best* indicator for that row (aspect).

This work is partially supported by National Natural Science Foundation of China (No. 62474152), Guangzhou Nansha District Key Area S&T Scheme (No. 2024ZD007), Guangdong Science and Technology Department (No. 2021JC02X145 and No. 2025B1212150003).

<sup>†</sup> Corresponding author: jiang.xu@hkust-gz.edu.cn

## II. BACKGROUND

### A. MRR Principles

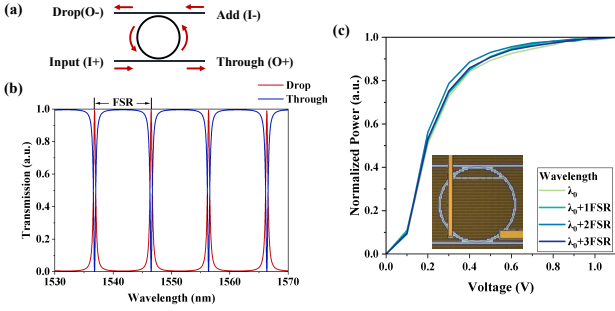


Fig. 1: (a) Dual-bus MRR configuration with its (b) transmission spectra. (c) Multi-FSR spectral response measured on a fabricated MRR and its microscope image (inset).

Micro-ring resonators (MRRs) are wavelength-selective photonic devices that manipulate optical signals through resonance effects. They have attracted significant attention due to their compact size and low power consumption [15]. A dual bus MRR consists of a closed-loop waveguide and two straight bus waveguides, as illustrated in Fig. 1(a). Resonance occurs when the optical path length in the ring satisfies the phase-matching condition. The resonance wavelength is calculated as:

$$\lambda_m = \frac{n_{\text{eff}}L}{m}, \quad m \in \mathbb{Z}^+ \quad (1)$$

where  $\lambda_m$  is the resonant wavelength,  $n_{\text{eff}}$  the effective refractive index,  $L = 2\pi R$  the ring circumference with radius  $R$ , and  $m$  the integer resonance order. When injected from the *Input* port, lights at resonance wavelengths  $\lambda_m$  couple into the ring and transmit to the *Drop* port, while off-resonance wavelengths propagate directly to the *Through* port. The transmission spectra at the *Drop* port and the *Through* port are shown in Fig. 1(b).

### B. Multi-FSR Modulation

The phase-matching condition creates periodic resonant peaks in the transmission spectrum separated by the free spectral range (FSR). The FSR in wavelength domain is derived from the resonance condition:

$$\text{FSR}_\lambda = \frac{\lambda_m^2}{n_g L} \quad (2)$$

where  $n_g$  is the group index. These periodic resonances result in a periodic transfer function, inherently enabling wavelength-division multiplexing (WDM) between adjacent peaks separated by the FSR, thus enhancing the modulation capabilities of the MRR. As shown in Fig. 1(c), a single MRR can process  $N_{\text{FSR}}$ ,  $N_{\text{FSR}}$  denotes the number of FSR channels, independent optical streams at wavelengths  $\lambda_k = \lambda_0 + k\text{FSR}$ . From an algorithmic perspective, the FSR parallelism of each MRR allows for the extension of scalar operations to vector operations without the need for additional hardware overhead, which enhances the efficiency of vector operations significantly.

## III. FSR-PARALLEL TENSOR CORE

This section presents the design of an MRR-based FSR-parallel tensor core (FPTC), which overcomes the limitations of conventional optical tensor cores: low computational throughput and restrictions to nonnegative operands.

TABLE II: Key tensor core design parameters

Notation	Definition
$N_w$	MRRs per array, same as Wavelengths per FSR channel
$N_{\text{FSR}}$	modulation arrays per modulation module, same as FSR channels
$N_h$	computation arrays per computation module

### A. Full Real-valued Multiplication

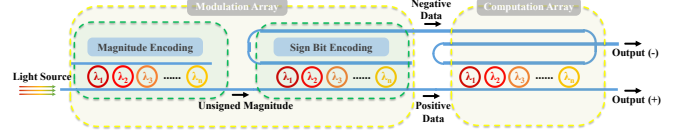


Fig. 2: The proposed real-valued vector multiplication MRR array engine.

Conventional MRR-based multipliers support only nonnegative operands due to sign constraints [5]. Our approach overcomes this by separating signed operands and utilizing all four ports of dual-bus MRRs: *Input*, *Through*, *Drop*, and *Add*.

As shown in Fig. 2, the input vector  $\mathbf{A}_i$  is first **magnitude-encoded** onto a multi-wavelength optical signal via an MRR array. Each MRR modulates an element  $a_{i,k}$  onto a specific wavelength. This unsigned signal then enters a second MRR array for **sign encoding**:

- For positive data  $a_{i,k,+}$ , the MRR remains **off**, directing light to the *Through* port.
- For negative data  $a_{i,k,-}$ , the MRR turns **on**, routing light to the *Drop* port.

Together, these arrays form the *modulation array*, converting  $\mathbf{A}_i$  into signed optical signals separated into the *Through* port (positive data:  $a_{i,k,+}$ ) and the *Drop* port (negative data:  $a_{i,k,-}$ ).

These signed signals next interact with vector  $\mathbf{B}_j$  at the *computation array*. Each element  $b_{k,j}$  is decomposed differentially into  $b_{k,j}^+$  and  $b_{k,j}^-$ , which exit via the *Through* and *Drop* ports, respectively. The spatial separation of signs enables incoherent computation through intensity modulation, avoiding the need for complex coherent maintenance and detection. The multiplication proceeds as follows:

- **Positive Path:** The modulation array's *Through* port ( $a_{i,k,+}$ ) connects to the computation array's *Input*, producing  $a_{i,k,+}b_{k,j}^+$  at *Through* and  $a_{i,k,+}b_{k,j}^-$  at *Drop*.
- **Negative Path:** The modulation array's *Drop* port ( $a_{i,k,-}$ ) connects to the *Add* port. Due to MRR symmetry, this yields  $a_{i,k,-}b_{k,j}^+$  at *Drop* and  $a_{i,k,-}b_{k,j}^-$  at *Through*.

This four-port configuration enables incoherent real-valued multiplication with positive results exit from the *Through* port and negative results from the *Drop* port. By cascading  $N_w$  MRRs per array, vector-vector multiplication is achieved within a single FSR channel:

$$\begin{aligned} \mathbf{O}_{i,j} &= \mathbf{A}_i \cdot \mathbf{B}_j = \sum_{k=1}^{N_w} (a_{i,k} b_{k,j}) \\ &= \underbrace{\sum_{k=1}^{N_w} (a_{i,k,+} b_{k,j}^+ + a_{i,k,-} b_{k,j}^-)}_{\text{Through port (positive results)}} - \underbrace{\sum_{k=1}^{N_w} (a_{i,k,+} b_{k,j}^- + a_{i,k,-} b_{k,j}^+)}_{\text{Drop port (negative results)}} \end{aligned} \quad (3)$$

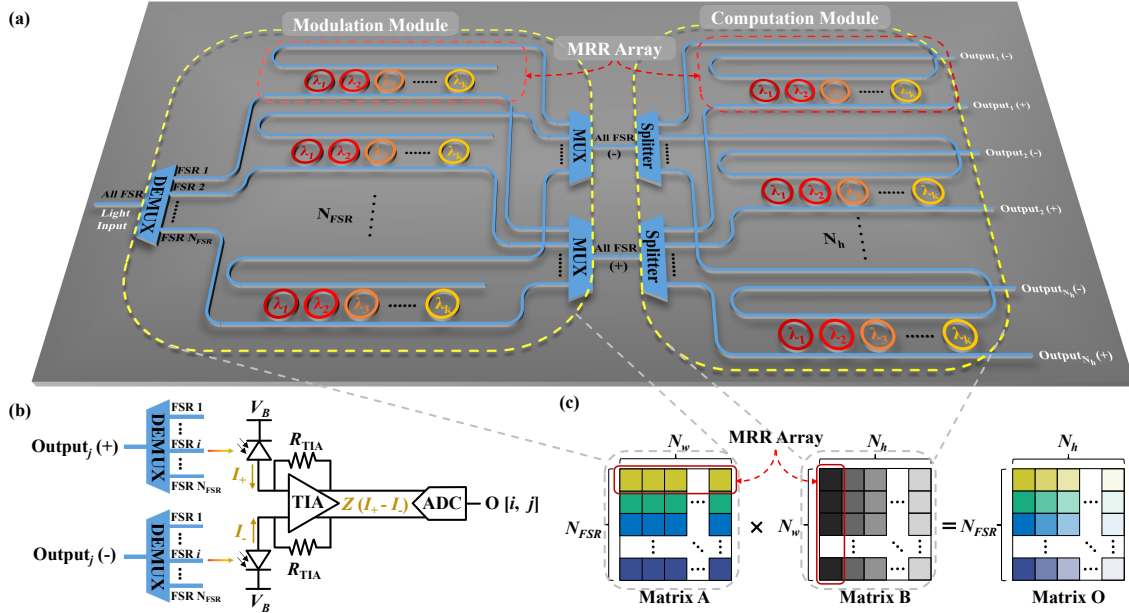


Fig. 3: (a) Overview of the FSR-parallel tensor core, FPTC. (b) Output detection module. (c) Matrix multiplication data flow.

The computation array thus performs  $[1 \times N_w] \times [N_w \times 1]$  multiplication per FSR channel. Simultaneous processing across  $N_{FSR}$  channels via multi-FSR modulation enables  $[N_{FSR} \times N_w] \times [N_w \times 1]$  operations within a single computation array.

### B. FPTC Workflow

The FSR-parallel tensor core (FPTC) performs GeMM by coordinating modulation and computation modules, as illustrated in Fig. 3(a). The modulation module encodes matrix  $\mathbf{A}$  into optical signals using  $N_{FSR}$  modulation arrays described in Section III-A. For clarity, Fig. 3(a) depicts each modulation array as a single MRR array. The computation module, comprising  $N_h$  computation arrays, performs optical-domain multiplication between encoded matrix  $\mathbf{A}$  and matrix  $\mathbf{B}$ .

**a) Optical Source Distribution:** A multi-wavelength optical source (e.g., an optical frequency comb) generates  $N_{FSR} \times N_w$  lights with different wavelengths. A coarse wavelength demultiplexer (DEMUX) separates these into  $N_{FSR}$  FSR groups, each containing  $N_w$  wavelengths. Each group is routed to a distinct modulation array.

**b) Matrix A Modulation:** The modulation module encodes all elements of  $\mathbf{A}$  onto the optical carriers. The  $i$ -th modulation array ( $i \in [1, N_{FSR}]$ ) encodes the row vector  $\mathbf{A}_i$  within the  $i$ -th FSR channel, as shown in Fig. 3(c). The  $k$ -th MRR ( $k \in [1, N_w]$ ) modulates  $\mathbf{A}[i, k]$  at wavelength  $\lambda_{i,k} = \lambda_k + i \cdot \text{FSR}$ . Thus, the entire matrix  $\mathbf{A}$  is represented across all  $N_{FSR}$  channels.

**c) Signal Routing and Preparation:** Optical multiplexers (MUX) combine signed operands from the *Through* (positive) and *Drop* (negative) ports of all modulation arrays. The combined optical signal contains  $N_{FSR}$  channels, each representing a row of  $\mathbf{A}$ . Broadband splitters then broadcast these signals to all computation arrays. Positive signals are routed to the *Input* ports of the computation arrays, while negative signals are directed to the *Add* ports.

**d) Matrix B Modulation and Computation:** In the computation module, the  $j$ -th array ( $j \in [1, N_h]$ ) encodes the column

vector  $\mathbf{B}_j$ , as shown in Fig. 3(c). The  $k$ -th MRR modulates element  $\mathbf{B}[k, j]$ . Each computation array receives optical signals from all FSR channels, enabling concurrent real-valued multiplication of  $\mathbf{B}_j$  with every row of  $\mathbf{A}$  via multi-FSR modulation, as detailed in Section II-B and Section III-A. This allows each MRR to compute  $N_{FSR}$  dot products in parallel through wavelength-division parallelism.

**e) Result Extraction:** Signed results are separated into positive values ( $\text{Output}_j(+)$ ) at the *Through* ports and negative values ( $\text{Output}_j(-)$ ) at the *Drop* ports of each computation array. The output optical signals comprise  $N_{FSR}$  channels, collectively forming a column vector  $\mathbf{O}_j$  of the result matrix. Each FSR channel aggregates contributions from  $N_w$  MRRs according to (3).

The multiplexers (MUXs) and demultiplexers (DEMUXs) in the FPTC can be implemented using various devices, such as cascaded MZI interleavers, arrayed waveguide gratings (AWGs), or ring-assisted MZIs (RAMZIs). Among these, we employ RAMZIs due to their compact footprint [16]. Another key advantage of RAMZIs is that their free spectral range (FSR), determined by the ring length  $L_{\text{ring}}$ , matches the FSR of the MRRs used in the core, as given in (2). This inherent FSR alignment ensures easy implementation of MUXs and DEMUXs with our multi-FSR computing paradigm.

### C. Photonic Output Readout

The photonic computation results undergo domain conversion through the detection module, as illustrated in Fig. 3(b). DEMUXs first separate individual FSR channels from the *Through* and *Drop* ports of each MRR array within the computation module. Subsequently, two photodetectors measure the corresponding photocurrents  $I_+$  and  $I_-$ . A differential transimpedance amplifier (TIA) computes the difference between  $I_+$  and  $I_-$ , yielding an amplified signed result:

$$I_{i,j} = \mathbb{Z}(I_+ - I_-) = \mathbb{Z}\left(|E_{+,j}^{(i)}|^2 - |E_{-,j}^{(i)}|^2\right) \quad (4)$$

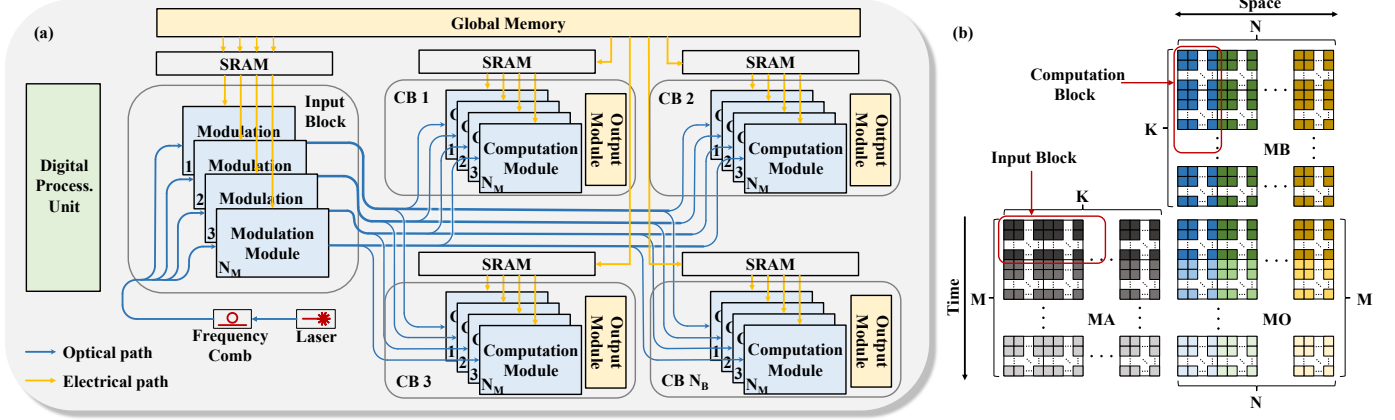


Fig. 4: (a) System architecture of the FSR-GeMM accelerator, with its (b) matrix mapping strategy.

where  $\mathbb{Z}$  denotes the transimpedance gain and  $E$  the optical field amplitude. Analog-to-digital converters (ADCs) then sample these amplified signals, reconstructing the digital result matrix with proper sign resolution.

The FPTC performs  $[N_{FSR} \times N_w] \times [N_w \times N_h]$  matrix multiplication per clock cycle. Each computation array generates  $N_{FSR}$  independent results in parallel, substantially increasing computational throughput. By maintaining full real-valued matrix representations optically throughout the computation, the design effectively eliminates the operand sign constraints inherent in conventional intensity-based optical computing systems. This approach provides a scalable and efficient pathway toward high-performance photonic tensor core implementations.

#### IV. PHOTONIC ACCELERATOR ARCHITECTURE

This section describes the high-level architecture of the proposed FSR-GeMM accelerator. Key design parameters are summarized in Table III.

TABLE III: Key accelerator design parameters

Notation	Definition
$N_B$	Computation blocks in the accelerator
$N_M$	Modules per block

##### A. System Design

The FSR-GeMM accelerator is designed for high-performance large-scale matrix operations through a structured architecture and efficient data mapping. It comprises three core components: the FPTC for real-valued GeMM operations, optical interconnects for data distribution, and electronic circuits for signal conversion, storage, and nonlinear processing. As illustrated in Fig. 4(a), the architecture decouples modulation and computation modules within the FPTC, which has been discussed in Section III-B. Here,  $N_M$  modulation modules form an *input block*, while  $N_M$  computation modules constitute a *computation block* (CB).

##### B. Dataflow and Matrix Mapping

The input block modulates optical signals from a frequency comb source and broadcasts them to  $N_B$  computation blocks via optical splitters. Each computation block contains  $N_M$  computation modules that receive signals from their corresponding modulation modules, along with dedicated output

readout circuits as described in Section III-C. As outlined in Section III-B, each modulation module modulates a matrix of size  $[N_{FSR}, N_w]$ , enabling the input block to modulate  $N_M$  such matrices in parallel. Within each computation block, every module performs a  $[N_{FSR}, N_w] \times [N_w, N_h]$  multiplication, resulting in  $N_M$  parallel GeMM operations per block. This broadcast and sharing scheme reduces resource usage of input block, improving area and power efficiency.

A structured mapping strategy is employed to map large matrices onto the accelerator's modular architecture, as illustrated in Fig. 4(b). The input matrix  $\mathbf{MA}$  is partitioned into sub-matrices of size  $[N_{FSR}, N_w]$ . Within each row,  $N_M$  segments are assigned to the  $N_M$  modulation modules, optically encoded, and broadcast to all computation blocks. Similarly, the weight matrix  $\mathbf{MB}$  is divided into sub-matrices of size  $[N_w, N_h]$ . Segments along the  $K$  dimension are assigned to the  $N_M$  computation modules per block, and  $N_B$  blocks compute segments along the  $N$  dimension via spatial parallelism. The  $k_M$ -th sub-matrix in the  $j_B$ -th column of  $\mathbf{MB}$  is mapped to the  $k_M$ -th computation module within the  $j_B$ -th computation block. Each computation module receives optically encoded data from its corresponding modulation module and computes one GeMM operation of size  $[N_{FSR}, N_w] \times [N_w, N_h]$ . This mapping strategy enables the concurrent execution of  $N_B \times N_M$  independent GeMM operations per cycle, enhancing parallelism and resource utilization.

##### C. Memory Hierarchy

As illustrated in Fig. 4(a), the memory system uses SRAM banks as global memory interfaced with off-chip DRAM. Both input and computation blocks are equipped with local SRAMs for storing operands, and each computation block includes a dedicated output SRAM.

For a GeMM operation of size  $[M, K] \times [K, N]$ , sub-matrices of  $\mathbf{MB}$  along the  $N$  dimension showing in Fig. 4(b) are loaded into the SRAM of the  $N_B$  computation blocks. Meanwhile, segments of  $\mathbf{MA}$  along the  $K$  dimension are stored in the input block's local SRAM. After each computation cycle, the input block's data are refreshed along the  $M$  dimension, while  $\mathbf{MB}$  data remain unchanged within the computation blocks. This asymmetry in data reuse motivates a bandwidth-aware memory design: the input block is allocated higher memory bandwidth

to accommodate frequent updates, whereas computation blocks operate with lower bandwidth requirements.

This dataflow keeps **MB** stationary in computation blocks, substantially reducing data movement overhead. Data within computation modules are updated only after processing all sub-matrices of **MA** along the  $M$  dimension. Both **MA** and **MB** are partitioned into  $K/N_M$  segments along the  $K$  dimension. Each segment is processed exhaustively across  $M$  and  $N$  before loading the next, minimizing memory access and enhancing efficiency.

#### D. ADC sharing

The hierarchical architecture enables analog-domain accumulation of partial results, considerably reducing the number of analog-to-digital converters (ADCs) required. As described in Section IV-B, each computation block computes the product of a sub-matrix row from **MA** and a sub-matrix column from **MB**. Computation modules within a block generate partial sums for the same output sub-matrix. This allows analog outputs from TIAs across multiple modules to be aggregated using analog adders before being digitized by a single shared ADC per output channel. This ADC-sharing strategy significantly reduces both area and power overhead associated with data conversion, while maintaining high throughput. It leverages the analog nature of photonic computing to minimize electronic resource usage, enhancing the overall efficiency of the accelerator.

## V. EVALUATION

### A. Evaluation Setup

The FSR-GeMM's performance is evaluated by a python-based simulation framework. Utilizing PRACTI [17], the framework characterizes the memory subsystem to estimate area, leakage power, and access energy under a 14-nm process node. Though photonic devices can theoretically operate at tens to hundreds of GHz, their clock speed is conservatively set to 10 GHz with 4-bit fixed-point precision to accommodate practical electrical interface constraints. Laser power is calculated by compensating for optical losses while meeting photodetector sensitivity thresholds. Key parameters for photonic and electronic components are listed in Table IV.

### B. Scalability

Fig. 5 illustrates the scalability of FSR-GeMM in terms of computational density (operations per unit area,  $TOP/mm^2$ ) and energy efficiency (operations per watt,  $TOP/W$ ), with a fixed architecture size of  $N_M = N_B = 4$ . Computational density and energy efficiency are analyzed across increasing core sizes  $N$ , where  $N_{FSR} = N_w = N_h = N$ . As  $N$  grows from 4 to 24, computational density increases  $37.3\times$  ( $0.66$  to  $24.59$   $TOP/mm^2$ ) and energy efficiency rises  $5.7\times$  ( $5.94$  to  $33.59$   $TOP/W$ ), demonstrating the strong scalability.

### C. System Analysis

The stacked bars in Fig. 5 provide comprehensive breakdowns of (a) area and (b) power consumption for FSR-GeMM as core size  $N$  scales up. The Optical Source component

TABLE IV: Key Device Parameters

Device	Parameter	Value
MRR [18]	Energy Consumption	3.1 fJ/bit
	Insertion Loss	2 dB
	Area	$15 \times 15 \mu m^2$
MZM [19]	Energy Consumption	540 fJ/bit
	Insertion Loss	3.7 dB
	Area	$1200 \times 680 \mu m^2$
Y-branch [20]	Insertion Loss	0.3 dB
	Area	$1.8 \times 1.3 \mu m^2$
RAMZI	Insertion Loss	0.3 dB
	Area	$15 \times 25 \mu m^2$
Directional Coupler [21]	Insertion Loss	0.33 dB
	Area	$5.25 \times 2.4 \mu m^2$
Photodetector [22]	Power	1.1 mW
	Sensitivity	-25 dBm
	Area	$4 \times 10 \mu m^2$
Micro-comb [23]	Area	$550 \times 550 \mu m^2$
On-chip Laser	Wall-plug Efficiency	0.2 [24]
	Area	$400 \times 300 \mu m^2$
DAC [25]	Precision	8-bit
	Power	50 mW
	Area	$11,000 \mu m^2$
ADC [26]	Precision	8-bit
	Power	21.9 mW
	Area	$9,000 \mu m^2$
TIA [27]	Power	3 mW
	Area	$< 50 \mu m^2$

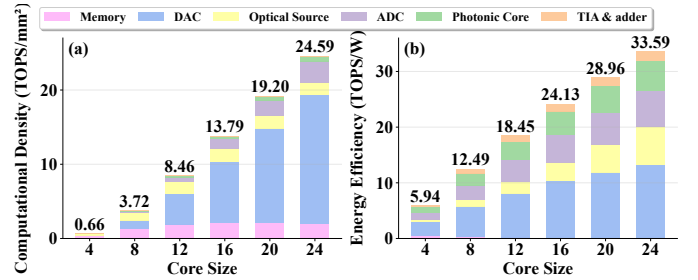


Fig. 5: (a) Computational density and (b) energy efficiency with area and power breakdowns of FSR-GeMM, scaled with increasing core size  $N$ .

includes both lasers and micro-combs, while Photonic Core comprises all photonic components plus photodetectors.

For both area and power breakdowns, DAC emerge as the dominant components, occupying 32.1% of area and 43.5% of power at size  $N = 8$ . This dominance arises significantly with size  $N$  due to the higher throughput at larger core sizes. In contrast, ADC observes a lower scaling effect with around 10% area and 20% power contributions at  $N = 24$ , proving the effectiveness of our ADC sharing strategy. The photonic core itself requires 0.7% to 2.9% of the area increasing with  $N$ , highlighting its exceptional efficiency achieved through wavelength-division multiplexing and compact design. The photonic core's power contribution remains consistent at approximately 7% of the total power across  $N$ . The optical source subsystem accounts for 6.5% to 20.3% increasing with  $N$ , as larger core size introduces higher optical loss. Auxiliary components like TIA & adder and memory have relatively small contributions, indicating their limited impact on overall efficiency.

### D. Computation Capacity

We configure two versions of FSR-GeMM by selecting core sizes  $N = 8$  and  $N = 16$ , termed as the base and large versions,

TABLE V: Base (B) and Large (L) configurations for FSR-GeMM.

Core Size	$N_w$	$N_{FSR}$	$N_h$	$N_B$	$N_M$
Base (B)	8	8	8	4	4
Large (L)	16	16	16	8	8

respectively. The architecture size is set at  $N_M = N_B = 4$  for base version, and  $N_M = N_B = 8$  for large version. Detailed parameters for these configurations are provided in Table V.

We evaluate the performance of the FSR-GeMM compared with three state-of-the-art photonic accelerator architectures: MZI mesh [6], MRR weight bank (WB) [8], Lightning-Transformer (LT) [12]. For a fair comparison, both the MZI mesh and the MRR weight bank are scaled to perform  $N_B \times N_M$  matrix-vector multiplications (MVM) of dimensions  $[1, N_w] \times [N_w, N_h]$ , using the same parameters as the base version of the FSR-GeMM. The LT is scaled to perform  $N_B \times N_M$  GeMMs with dimensions  $[N_{FSR}, N_w] \times [N_w, N_h]$ , maintaining consistent parameters for both its base and large versions as FSR-GeMM.

Fig. 6 illustrates the comparison of computation density and energy efficiency among the different photonic accelerators. As shown in Fig. 6(a), the proposed FSR-GeMM architecture exhibits a remarkable computation density, with its large version achieving  $18.45 \text{ TOPS}/\text{mm}^2$ , approximately  $31.7\times$  higher than MRR weight banks and  $54\times$  greater than LT’s large version. In terms of energy efficiency, the large version of FSR-GeMM reaches  $29.33 \text{ TOPS}/\text{W}$ , outperforming MRR-based systems, by  $13.8\times$ , and surpassing the LT-L’s energy efficiency of  $18.93 \text{ TOPS}/\text{W}$  by  $1.55\times$ .

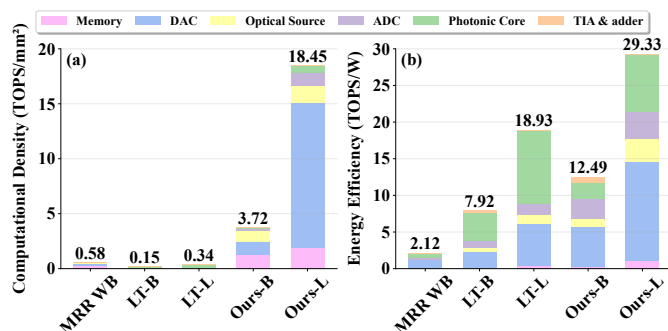


Fig. 6: (a) Computation density and (b) energy efficiency comparison between different photonic accelerators with area and power breakdown.

### E. Energy consumption

DeiT [28] and BERT [29] with 4-bit quantization in both weight and activation are used to evaluate the performance of different photonic accelerators. Both are well-recognized vision and NLP Transformer neural network models. Different versions: DeiT-T/S/B and BERT-B/L are used as workloads with  $1024 \times 1024$  input image and 1024 input sequence lengths, respectively.

Fig. 7(a) presents the energy consumptions across various photonic accelerators for these workloads. Conventional photonic solutions (MZI Mesh and MRR Weight Bank) exhibit significantly higher energy demands, approximately  $3.4\times$  and  $7.4\times$  higher than our FSR-GeMM base and large configurations. The performance advantage of our architecture is

still significant compared with the state-of-the-art photonic accelerator, LT. Specifically, FSR-GeMM achieves average energy reductions of  $27.3\%$  and  $24.9\%$  in the base and large configurations, respectively.

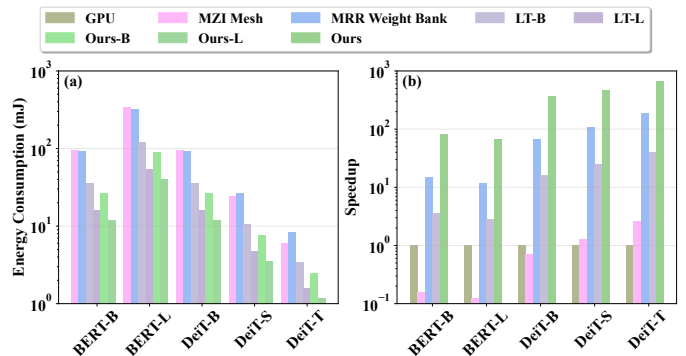


Fig. 7: (a) Energy consumption (mJ) comparison and (b) speedup comparison across different accelerators for various workloads.

### F. Speedup

For speedup comparison, we adjust the parameters  $N_B$  and  $N_M$  of all photonic accelerators to match the area of LT-B, which is around  $1000 \text{ mm}^2$ . A single Nvidia A30 GPU is used as the baseline for evaluating speedup. The results, depicted in Fig. 7(b), indicate that our FSR-GeMM consistently achieves the highest speedup, with over  $68\times$  to  $670\times$  speedup compared to the GPU. Compared with other photonic accelerators, the FSR-GeMM exhibits average speedups of approximately  $445\times$ ,  $5\times$ , and  $21\times$  relative to the MZI Mesh, MRR Weight Bank, and LT, respectively.

## VI. CONCLUSION

This work presents the FSR-parallel tensor core (FPTC) and FSR-GeMM accelerator, leveraging FSR multiplexing in dual-bus micro-ring resonators (MRRs) to overcome critical limitations in photonic GeMM computing. Our design features three key innovations: 1) A four-port MRR configuration that directly decomposes signed operands, eliminating preprocessing and non-negativity constraints; 2) Intra-MRR FSR-parallel processing maximizing optical bandwidth utilization; and 3) A hierarchical architecture integrating matrix tiling with analog-domain ADC sharing to minimize area and power overhead. These advances collectively enhance flexibility, computational density, and scalability for photonic accelerators.

Experimental evaluations demonstrate that FSR-GeMM achieves  $6\times$  to  $57\times$  higher area efficiency and  $1.55\times$  to  $13.8\times$  greater energy efficiency than state-of-the-art photonic accelerators. It also reduces energy consumption by  $27.3\%$  to  $70\%$  compared to leading solutions. Against MZI mesh, MRR weight bank, and Lightning-Transformer architectures, our design delivers average speedups of  $445\times$ ,  $5\times$ , and  $21\times$ , respectively. By eliminating preprocessing overhead, low bandwidth utilization, and operand constraints, FSR-GeMM offers a practical path toward energy-efficient, high-throughput optical computing for real-valued GeMM workloads.

## REFERENCES

- [1] Y. Liu, B. Hu, Z. Liu, P. Chen, L. Du, J. Liu, X. Li, W. Zhang, and J. Xu, "Fiona: Photonic-electronic cosimulation framework and transferable prototyping for photonic accelerator," in *2023 IEEE/ACM International Conference on Computer Aided Design (ICCAD)*, 2023, pp. 1–9.
- [2] E. Qin, A. Samajdar, H. Kwon, V. Nadella, S. Srinivasan, D. Das, B. Kaul, and T. Krishna, "Sigma: A sparse and irregular gemm accelerator with flexible interconnects for dnn training," in *2020 IEEE International Symposium on High Performance Computer Architecture (HPCA)*, 2020, pp. 58–70.
- [3] Y. Bai, X. Xu, M. Tan, Y. Sun, Y. Li, J. Wu, R. Morandotti, A. Mitchell, K. Xu, and D. J. Moss, "Photonic multiplexing techniques for neuromorphic computing," *Nanophotonics*, vol. 12, no. 5, pp. 795–817, 2023. [Online]. Available: <https://doi.org/10.1515/nanoph-2022-0485>
- [4] S. Afifi, F. Sunny, M. Nikdast, and S. Pasricha, "Accelerating neural networks for large language models and graph processing with silicon photonics," in *2024 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, 2024, pp. 1–6.
- [5] Y. Ning, H. Zhu, C. Feng, J. Gu, Z. Jiang, Z. Ying, J. Midkiff, S. Jain, M. H. Hlaing, D. Z. Pan, and R. T. Chen, "Photonic-electronic integrated circuits for high-performance computing and ai accelerators," *Journal of Lightwave Technology*, vol. 42, no. 22, pp. 7834–7859, 2024.
- [6] Y. Shen, N. C. Harris, S. Skirlo, M. Prabhu, M. 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, no. 7, pp. 441–446, 2017. [Online]. Available: <https://doi.org/10.1038/nphoton.2017.93>
- [7] H. Zhang, M. Gu, X. D. Jiang, J. Thompson, H. Cai, S. Paesani, R. Santagati, A. Laing, Y. Zhang, M. H. Yung, Y. Z. Shi, F. K. Muhammad, G. Q. Lo, X. S. Luo, B. Dong, D. L. Kwong, L. C. Kwek, and A. Q. Liu, "An optical neural chip for implementing complex-valued neural network," *Nature Communications*, vol. 12, no. 1, p. 457, 2021. [Online]. Available: <https://doi.org/10.1038/s41467-020-20719-7>
- [8] A. N. Tait, T. F. de Lima, E. Zhou, A. X. Wu, M. A. Nahmias, B. J. Shastri, and P. R. Prucnal, "Neuromorphic photonic networks using silicon photonic weight banks," *Scientific Reports*, vol. 7, no. 1, p. 7430, 2017. [Online]. Available: <https://doi.org/10.1038/s41598-017-07754-z>
- [9] W. Liu, W. Liu, Y. Ye, Q. Lou, Y. Xie, and L. Jiang, "Holylight: A nanophotonic accelerator for deep learning in data centers," in *2019 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, 2019, pp. 1483–1488.
- [10] F. Sunny, A. Mirza, M. Nikdast, and S. Pasricha, "Crosslight: A cross-layer optimized silicon photonic neural network accelerator," in *2021 58th ACM/IEEE Design Automation Conference (DAC)*, 2021, pp. 1069–1074.
- [11] S. R. 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, Apr 2024. [Online]. Available: <https://opg.optica.org/optica/abstract.cfm?URI=optica-11-4-542>
- [12] H. Zhu, J. Gu, H. Wang, Z. Jiang, Z. Zhang, R. Tang, C. Feng, S. Han, R. T. Chen, and D. Z. Pan, "Lightning-transformer: A dynamically-operated optically-interconnected photonic transformer accelerator," in *2024 IEEE International Symposium on High-Performance Computer Architecture (HPCA)*, 2024, pp. 686–703.
- [13] R. Cardoso, C. Zrouba, M. Abdalla, M. G. de Queiroz, P. Jimenez, I. O'Connor, and S. Le Beux, "Reconfigurable photonic gemm based on phase-change-materials and stochastic computing," *Journal of Lightwave Technology*, vol. 42, no. 22, pp. 8024–8031, 2024.
- [14] 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. [Online]. Available: <https://doi.org/10.1038/s41586-020-03070-1>
- [15] Y. Yuan, Y. Peng, W. V. Sorin, S. Cheung, Z. Huang, D. Liang, M. Fiorentino, and R. G. Beausoleil, "A 5 x 200 gbps microring modulator silicon chip empowered by two-segment z-shape junctions," *Nature Communications*, vol. 15, no. 1, p. 918, 2024. [Online]. Available: <https://doi.org/10.1038/s41467-024-45301-3>
- [16] L.-W. Luo, S. Ibrahim, A. Nitkowski, Z. Ding, C. B. Poitras, S. J. B. Yoo, and M. Lipson, "High bandwidth on-chip silicon photonic interleaver," *Opt. Express*, vol. 18, no. 22, pp. 23 079–23 087, Oct 2010. [Online]. Available: <https://opg.optica.org/oe/abstract.cfm?URI=oe-18-22-23079>
- [17] A. Shafaei, Y. Wang, X. Lin, and M. Pedram, "Fincacti: Architectural analysis and modeling of caches with deeply-scaled finfet devices," in *2014 IEEE Computer Society Annual Symposium on VLSI*, 2014, pp. 290–295.
- [18] D. W. U. Chan, X. Wu, C. Lu, A. P. T. Lau, and H. K. Tsang, "Efficient 330-gb/s pam-8 modulation using silicon microring modulators," *Opt. Lett.*, vol. 48, no. 4, pp. 1036–1039, Feb 2023. [Online]. Available: <https://opg.optica.org/ol/abstract.cfm?URI=ol-48-4-1036>
- [19] K. Li, D. J. Thomson, S. Liu, W. Zhang, W. Cao, C. G. Littlejohns, X. Yan, M. Ebert, M. Banakar, D. Tran, F. Meng, H. Du, and G. T. Reed, "An integrated cmos-silicon photonics transmitter with a 112 gigabaud transmission and picojoule per bit energy efficiency," *Nature Electronics*, vol. 6, no. 11, pp. 910–921, 2023. [Online]. Available: <https://doi.org/10.1038/s41928-023-01048-1>
- [20] D. P. Nair and M. Ménard, "A compact low-loss broadband polarization independent silicon 50/50 splitter," *IEEE Photonics Journal*, vol. 13, no. 4, pp. 1–7, 2021.
- [21] C. Ye and D. Dai, "Ultra-compact broadband  $2 \times 2$  3 db power splitter using a subwavelength-grating-assisted asymmetric directional coupler," *Journal of Lightwave Technology*, vol. 38, no. 8, pp. 2370–2375, 2020.
- [22] Z. Huang, C. Li, D. Liang, K. Yu, C. Santori, M. Fiorentino, W. Sorin, S. Palermo, and R. G. Beausoleil, "25&#x2009;&#x2009;gbps low-voltage waveguide si&#x2013;ge avalanche photodiode," *Optica*, vol. 3, no. 8, pp. 793–798, Aug 2016. [Online]. Available: <https://opg.optica.org/optica/abstract.cfm?URI=optica-3-8-793>
- [23] A. S. Raja, S. Lange, M. Karpov, K. Shi, X. Fu, R. Behrendt, D. Cletheroe, A. Lukashchuk, I. Haller, F. Karinou, B. Thomsen, K. Jozwik, J. Liu, P. Costa, T. J. Kippenberg, and H. Ballani, "Ultrafast optical circuit switching for data centers using integrated soliton microcombs," *Nature Communications*, vol. 12, no. 1, p. 5867, 2021. [Online]. Available: <https://doi.org/10.1038/s41467-021-25841-8>
- [24] H. Wang, R. Zhang, Q. Kan, D. Lu, W. Wang, and L. Zhao, "High-power wide-bandwidth  $1.55 - \mu\text{m}$  directly modulated dfb lasers for free space optical communications," in *2019 Conference on Lasers and Electro-Optics (CLEO)*, 2019, pp. 1–2.
- [25] P. Caragiulo, O. E. Mattia, A. Arbabian, and B. Murmann, "A compact 14 gs/s 8-bit switched-capacitor dac in 16 nm finfet cmos," in *2020 IEEE Symposium on VLSI Circuits*, 2020, pp. 1–2.
- [26] Y. Tao, M. Zhan, M. Gu, X. He, Y. He, Z. Zhang, Y. Zhong, L. Jie, and N. Sun, "An 8b 10gs/s 2-channel time-interleaved pipelined adc with concurrent residue transfer and quantization, and automatic buffer power gating," in *2025 IEEE International Solid-State Circuits Conference (ISSCC)*, vol. 68, 2025, pp. 440–442.
- [27] M. Rakowski, Y. Ban, P. De Heyn, N. Pantano, B. Snyder, S. Balakrishnan, S. Van Huylenbroeck, L. Bogaerts, C. Demeurisse, F. Inoue, K. J. Rebbis, P. Nolmans, X. Sun, P. Bex, A. Srinivasan, J. De Coster, S. Lardinois, A. Miller, P. Absil, P. Verheyen, D. Velenis, M. Pantouvaki, and J. Van Campenhout, "Hybrid 14nm finfet - silicon photonics technology for low-power tb/s/mm<sup>2</sup> optical i/o," in *2018 IEEE Symposium on VLSI Technology*, 2018, pp. 221–222.
- [28] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jegou, "Training data-efficient image transformers & distillation through attention," in *Proceedings of the 38th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, M. Meila and T. Zhang, Eds., vol. 139. PMLR, 18–24 Jul 2021, pp. 10 347–10 357. [Online]. Available: <https://proceedings.mlr.press/v139/touvron21a.html>
- [29] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, J. Burstein, C. Doran, and T. Solorio, Eds. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. [Online]. Available: <https://aclanthology.org/N19-1423/>